Inference of the past of random structures and other random problems

Simon Briend

université PARIS-SACLAY

Inference of the past of random structures and other random problems

Inférence du passé de structures aléatoires et autres problèmes aléatoires

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 574, école doctorale de Mathématiques Hadamard (EDMH)
Spécialité de doctorat: Mathématiques appliquées
Graduate School : Mathématiques. Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche **Laboratoire de Mathématique d'Orsay, (Université Paris-Saclay, CNRS)**, sous la direction de **Christophe GIRAUD**, professeur des universités, la co-direction de **Gábor LUGOSI**, professeur des universités

Thèse soutenue à Paris-Saclay, le 11 juin 2024, par

Simon BRIEND

Composition du jury

Membres du jury avec voix délibérative

Laurent MASSOULIÉ Professeur, INRIA Paris, DIENS PSL University	Président
Po Ling LOH Professeure, University of Cambridge	Rapporteuse & Examinatrice
Rui PIRES DA SILVA CASTRO Professeur associé, TU EIndhoven	Rapporteur & Examinateur
Nicolas CURIEN Professeur, Université Paris-Saclay	Examinateur
Vincent RIVOIRARD Professeur, Université Paris-Dauphine	Examinateur

THESE DE DOCTORAT
NNT : 2024UPASM013

université PARIS-SACLAY | ÉCOLE DOCTORALE de mathématiques Hadamard (EDMH)

Titre: Inférence du passé de structures aléatoires et autres problèmes aléatoires

Mots clés: Statistiques combinatoires, graphes aléatoires, archéologie des graphes, analyse de données, profondeur de Tukey, statistiques en grande dimension.

Résumé: Cette thèse est décomposée en trois parties disjointes. Les deux premières parties se concentrent sur des modèles de graphes aléatoires croissants de manière dynamique. Dans la première partie, nous inférons des informations sur le passé d'un graphe à partir d'une unique observation dudit graphe. Nous commençons par le problème de la recherche de racine, où l'objectif est de trouver un ensemble de confiance pour la racine. Nous proposons une méthode pour les ℓ-dags uniformes et analysons ses performances. À notre connaissance, il s'agit de la première méthode réalisant une archéologie du graphe dans des graphes généraux. Nous étendons ensuite naturellement la question de la recherche de racine à celle de la sériation. Étant donné un instantané d'un graphe, est-il possible de récupérer son ordre complet ? Nous présentons une méthode et une garantie statistique sur sa qualité dans le cas des arbres récursifs uniformes et des arbres d'attachement préférentiel linéaire. Pour conclure la section sur l'archéologie de graphe, nous étudions un problème de broadcasting, où l'on ne tente pas de retrouver la racine du graphe mais son état. Dans de tels problèmes, la racine se voit attribuer un bit, qui est ensuite propagé de manière bruité lors de la croissance du réseau. Dans les ℓ-dags, nous étudions un vote par majorité pour estimer le bit de la racine et identifions trois régimes, dépendants du niveau de bruit. Dans la deuxième partie, nous étudions l'arbre d'amitié aléatoire, qui est un modèle d'arbre récursif aléatoire avec redirection complète. Dans ce modèle apparaît un phénomène de rich-get-richer, mais à la différence du modèle d'attachement préférentiel celui ci découle d'un processus d'attachement local. Nous prouvons des conjectures sur la distribution des degrés, le diamètre et la structure locale. Enfin, nous plongeons dans le monde de l'apprentissage automatique théorique et de l'analyse de données. Nous étudions une approximation aléatoire de la profondeur de Tukey. La profondeur de Tukey est un outil puissant pour la visualisation des données et peut être considérée comme une extension des quantiles en dimension plus élevée (ils coïncident en dimension 1). Son calcul exact est NP-difficile, et nous étudions les performances d'une approximation aléatoire dans le cas de données échantillonnées à partir d'une distribution log-concave.

université PARIS-SACLAY | ÉCOLE DOCTORALE de mathématiques Hadamard (EDMH)

Title: Inference of the past of random structures and other random problems

Keywords: combinatorial statistics, random graphs, network archaeology, data analysis, Tukey depth, high dimensional statistics.

Abstract: This thesis is decomposed in three disjoint parts. The first two parts delve into dynamically growing networks. In the first part, we infer information about the past from a snapshot of the graph. We start by the problem of root finding, where the goal is to find confidence set for the root. We propose a method for uniform ℓ-dags and analyse its performance. It is, to the best of our knowledge, the first method achieving network archaeology in general graphs. Then, we naturally extend the question of root finding to the one of seriation. Given a snapshot of a graph, is it possible to retrieve its whole ordering? We present a method and statistical guarantee of its quality in the case of uniform random recursive trees and linear preferential attachment tree. To conclude the network archaeology section, we study the root bit finding problem, where one does not try to infer the position of the root but its state. In such problems, the root is assigned a bit and is then propagated through a noisy channel during network growth. In the ℓ-dag, we study majority voting to infer the bit of the root and we identify three different regimes depending on the noise level. In the second part of this thesis, we study the so called friendship tree, which is a random recursive tree model with complete redirection. This model display emerging properties, but unlike in the preferential attachment model they stem from a local attachment rule. We prove conjectures about degree distribution, diameter and local structure. Finally, we delve into the world of theoretical machine learning and data analysis. We study a random approximation of the Tukey depth. The Tukey depth is a powerful tool for data visualization and can be thought of as an extension of quantiles in higher dimension (they coincide in dimension 1). Its exact computation is NP-hard, and we study the performances of a classical random approximation in the case of data sets sampled from log-concave distribution.

0.1 Remerciements

First of all I want to thank my directeurs de Thèse, Christophe Giraud and Gábor Lugosi. Incredibly cool and patient, you guided me through those three years and taught me so much, about mathematics but also how to exist and thrive in the research community. Christophe, thank you for your time, your wise advices and your understanding of my somehow unique PhD organisation. Gábor, thanks to you I discovered research and so many cool problems, but maybe more importantly I met incredible people all over the world. You two convinced me I wanted to stay in academia without even trying, and I had the three best possible time working with you. You will always be examples for the rest of my career.

Thank you Pol-Ling Loh and Rui Castro for accepting to review my PhD. Thank you for your time and for carefully reading my manuscript, your comments were greatly appreciated. I also thank Nicolas Curien and Vincent Rivoirard who accepted to be part of my jury, and Laurent Massoulié who was the head of this Jury. I am glad I could share my work with you.

I also owe so much to the amazing co-authors I worked with during my PhD. First of all Louigi Addario-Berry, I admire your kindness and the unlimited attention you give to the people around you. I know how lucky I was to spend time and work with you, and you are really a model I want to follow later on in my career. Also at McGill, Luc Devroye, working with you is the most fun way to do mathematics. Infinite knowledge and good vibes go well together! Serte Donderwinkel, thank you so much for the help you gave me during this PhD, in our research projects but also when it came to sorting out my professional (and personal) future. Finally, Roberto Oliveira, my visits at IMPA are always so special and working there with you was one of the highlight of my PhD. I don't want to forget the other incredible people I worked with, Céline Kerriou, Anna Brandenberger, Rivka Mitchell, Francisco Calvillo, Déborah Sulem, who I hope will be my future colleagues. Meeting all of you around the world was by far my favourite part of this PhD.

I also want to thank the people at Université Paris-Saclay and Universitat Pompeu Fabra, who made this PhD possible and offered me so many incredible opportunities.

Pour finir je dois remercier mes proches, à qui je dois d'être arrivé là. Mes parents, pour m'avoir épaulé et guidés, à mes grands parents pour m'avoir appris beaucoup de la vraie vie (pas les maths quoi). Mes amis, qui permettent d'oublier le stress et le travail quand il le faut. Enfin, Rafaela, avec qui j'ai partagé ma vie durant toute ma thèse. Merci de m'avoir supporté, remis les pieds sur terre et surtout d'avoir rendu ces années si incroyables. Muito obrigado.

Contents

0.2 Thesis outline

Chapter I and II: Introduction

In the first chapter I introduce the problems studied in this PhD. I first introduce models of growing random graphs and their links to real life attachment and propagation phenomena. I start by presenting the general problem of inferring the past of growing random graphs and pinpoint what network archaeology problems we study in this PhD. Then, I present another growing random graph model, namely the friendship tree, that was introduced by physicists and that have the benefit of displaying a rich-get-richer phenomenon while having a local attachment rule. Finally, I delve into a theoretical machine learning problem and outline a few challenges faced by practitioners to visualize datasets in high dimension. Finally, I present a popular notion of depth used to order a high dimensional dataset, the Tukey depth. Unfortunately, this depth is hard to compute, which leads me to the introduction of a random approximation algorithm.

Chapter III to V: A contribution to network archaeology

Growing random networks are present in all aspects of our lives and deducing information about the history of the network from a snapshot of its current state is of great interest. It is useful to answer questions such as, which proteins interacted in long extinct species? Who was the first member of an online community? Where did a rumor originated online? Who was Covid's original patient? Until now, the problem that received the most attention is of *root finding* in trees, which consists in retrieving the first few vertices of a randomly growing tree (Brandenberger, Devroye, and Goh [21], Haigh [74]). In Chapter 3 and 4 we propose two extensions of the network archaeology toolbox. First, we extend results of root finding to more general graph models. Then we also present a method to not only infer the first vertex but the whole ordering of a randomly growing tree.

In fact, before this work, root finding for graphs was only explored in one very recent paper (Crane and Xu [43]). Extending network archaeology methods to graphs is of great interest. Indeed, most real life problems are best described by graphs, and even if the theoretical model is a tree, the final observation might have some errors resulting in the disappearance of the tree structure. We study the ℓ-dag model, a growing graph model that is an extension of the *uniform random recursive tree* (URRT). dag stands for *directed acyclic graph*, which can be confusing because in this thesis we will always observe the undirected version of the graphs. However, the term ℓ-dag was previously used for this model, so we decided to use it again even though we do not observe the directions of

the edges. In this model, at each step, the new vertex connects to ℓ ancestors chosen uniformly at random. We introduce a simple algorithm, running in cubic time, that retrieves a confidence set of size $K(\epsilon)$ that contains the oldest vertex with probability at least $1-\epsilon$, where

$$K(\epsilon) \leq \frac{c_0}{\epsilon} \log\left(\frac{1}{\epsilon}\right)^{\frac{c_1}{\ell}\log\frac{1}{\epsilon}},$$

for c_0 and c_1 two positive numerical constants.

In Chapter 4, we study the problem of ordering all vertices in a randomly growing tree, both in the URRT and the *linear preferential attachment* model (PA tree). Such problems of estimating latent variables (here, the arrival time of a node) are common in statistics. One particular instance of this problem in a setting close to ours is in seriation (Giraud, Issartel, and Verzelen [72]). Unlike our study, previous seriation works study graphs with no time structure (for example in random geometric graphs, one can infer the geometric position of vertices from an observation of the graph). They have applications as broad as phase synchronization or archaeology. Nonetheless, the addition of a time dependency, and thus of a model where vertices do not all have the same properties, calls for novel methods. This problem has been studied by Crane and Xu [42] for URRT, where they propose a method to infer an ordering but give no theoretical guarantees about its quality. First, we propose an error measure that takes into account the time dependency of the problem, assigning larger weights to the errors in older vertices. For an ordering procedure $\widehat{\sigma}$ on the tree $T=(E,V)$, we define, for $\alpha \in [1,2)$, the error

$$R_\alpha(\widehat{\sigma}) = \sum_{v\in V} \frac{|\widehat{\sigma}(v)-\sigma(v)|}{\sigma(v)^\alpha},$$

where $\sigma(v)$ is the true arrival time of vertex v. We prove a lower bound for this error, that is, for any ordering procedure $\widehat{\sigma}$ it stands that $R_\alpha(\widehat{\sigma}) \geq c_\alpha |V|^{2-\alpha}$. Then, we analyse the performances of ordering vertices by their Jordan centrality. In particular, for $\alpha \in [1,2)$ for the URRT model and $\alpha \in [1,5/4)$ in the PA tree model, we prove that

$$R_\alpha(\widehat{\sigma}_J) \leq C_\alpha |V|^{2-\alpha},$$

meaning that this method is optimal up to a constant factor. Then, we provide numerical illustrations of our results and numerical comparisons to other ordering methods, suggesting that the Jordan ordering is the best of the tested methods.

To conclude this work on inferring the past of growing network, in Chapter 5 we

study a problem where the goal is to infer the state of early vertices of a growing graph. The *broadcasting* problem consists in propagating information from a source vertex along edges of the graph, and at each propagation step there is a positive chance of altering the information. In the *root bit finding* problem, the information borne by a vertex takes values in $\{0,1\}$ and we want to infer the original state by observing the information received by (part of) the vertices of the graph. Like for root finding, this problem has been studied extensively for trees (Addario-Berry, Devroye, Lugosi, and Velona [2]), and is of great interest to extend to graphs. We study the root bit finding for the ℓ-dag model. We prove thresholds on the mutation's probability, corresponding to three distinct regimes of the information transmission in the ℓ-dag.

Chapter VI: The friendship tree

Chapter 6 focuses on discovering general properties of a novel model of growing random trees. Such properties are, for example, the depth, the degree sequence or the maximum degree. Those are well known for the most generic models, but to better capture the behaviour of real-world networks, theoreticians and practitioners alike introduce new models. Here, we study a twist of the URRT, that exhibits drastically different behaviours.

Random graphs with *redirection* are part of a family of growing networks where new vertices do not attach to a random ancestor but to someone "close" to a random ancestor. An example is the model in which new vertices connect to a random descendent of a random vertex. We study a model with *complete redirection* introduced by physicists (Saramäki and Kaski [125]). Here, the new vertex selects a random vertex and then connects to a random neighbour (or friend) of this vertex; because of this attachment rule we refer to it as the *random friend tree* (RFT). An intriguing feature of this model is that it has a very simple local attachment rule, yet it leads to interesting emergent properties such as a highly skewed degree sequence. This makes it an interesting toy model for growing real-life networks, since these also grow using only local information, while exhibiting striking global behaviour. We have done the first extensive rigorous study of the random friend tree and we prove conjectures formulated by Krapivsky and Redner [93]. We prove that, even though this model is a simple twist from the URRT, it has drastically different behaviour. For example, even if its diameter is of logarithmic order (like in a URRT), a vanishing fractions of vertices have degree at least two, when this fraction tends to $1/2$ in a URRT. Like in PA trees, a rich-get-richer phenomenon is at play, leading to the emergence

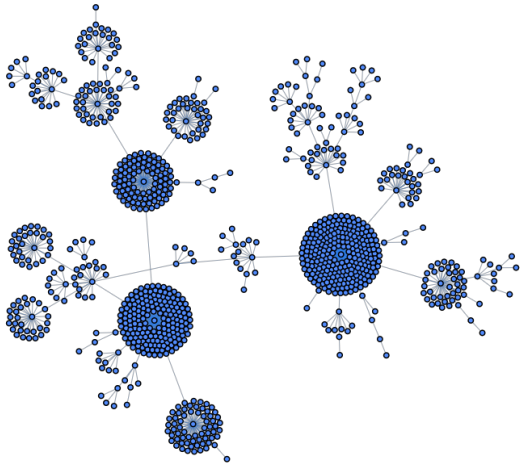

Figure 1: A random friendship tree realization with 1000 vertices.

of macro structures, such as linear degree vertices scattered everywhere in the graph.

Chapter VII: On the quality of randomized approximations of Tukey's depth

Finally, Chapter 7 delves into a theoretical machine learning problem. For some applications, it is important to order data by their *centrality* in the dataset. For example to visualize data, to detect outliers, or to train conformal predictors. In one dimension, a natural way to order data points is by their empirical quantiles. Of course, practitioners deal with data points of higher dimension, and the *Tukey depth* is a generalization of quantiles in higher dimension. Like quantiles in one dimension, it has nice properties of convergence, invariance under translation and linear rescaling. However, it is known that the Tukey depth is hard to compute (Bremner, Chen, Iacono, Langerman, and Morin [24], Chan [37]), and in particular has a complexity growing exponentially with the dimension. This is why a random approximation has been introduced. This approximation algorithm has a parameter k, that tunes the quality of the approximation and the time complexity. We study the quality of this approximation in the high dimensional regime, for points sampled from a log concave distribution. We prove that there exists three distinct behaviours for this random Tukey depth approximation. For points of low depth, that is, most of them in high dimension, k can be chosen independently from the dimension. For the most central points, k only has to grow polynomially with the dimension for the approximation to be accurate. But for all points lying in between, k has to be at least exponential in the dimension to

produce an accurate approximation. This is a problem for practitioners, especially since the most interesting depths to compute are the intermediate ones, to estimate the level sets of the Tukey depth.

Chapter 1

Introduction

Contents

1.1 Introduction to network archaeology

Numerous phenomena can be described by attachment or propagation processes. To name just a few, think of the spread of a disease, a computer virus or fake news, or the evolution of a social network. Graphs can be used to describe these phenomena, or more precisely, a sequence of graphs explaining the evolution of the said phenomenon. For example, at a given moment, a social network can be described by a graph, in which each individual corresponds to a vertex, and each friendship link corresponds to an edge. As this graph evolves, we can consider the sequence of graphs that describes the evolution of the social network. Induced by this sequence, a notion of *history* appears. Indeed, one vertex was added first and another one-thousand-two-hundred-and-forty-third. In this part of my thesis, we try to recover information about this history in the case where we are only observing the graph at a given moment (and not the sequence). To formulate this

problem mathematically, we need to define some random graph models. These models serve as toy models in which to develop techniques and algorithms in perfectly defined environments. Let us start with some commonly studied graph models, but with a fixed size.

Erdős–Rényi model

In this model, a parameter n (the size of the graph) is fixed. An Erdős-Rényi graph can refer to two similar models. The $\mathcal{G}(n,M)$ model, in which a graph is chosen uniformly at random from all graphs with n vertices and M edges, and the $\mathcal{G}(n,p)$ model, which is constructed by randomly connecting the n vertices. Each edge is included in the graph with probability p, independently of all other edges.

Random geometric graph

In this model, a metric space X with a measure is fixed. A parameter n (the size of the graph), a parameter ρ of connectivity and a distribution μ are also fixed. n points are randomly and independently drawn according to the distribution μ. Each of these points corresponds to a vertex of the graph, and each pair of vertices is connected if the corresponding points are less than ρ apart.

Stochastic block model (SBM)

In the simplest version of this model, vertices are split in two communities. The model is then similar to $\mathcal{G}(n,p)$, with the difference that the probability of an edge being added to the graph depends on the communities of the vertices. For a pair of vertices in the same community, the corresponding edge is added with probability p, and for a pair of vertices in different communities, the corresponding edge is added with probability q .

In this work, we study models that grow by recursively adding vertices and edges. Here are a few examples.

The uniform random recursive tree

Probably the simplest model for describing an attachment process. The first graph of the sequence consists of an isolated vertex. This graph is grown recursively by adding a new vertex, which connects to a vertex already present in the graph, chosen uniformly at random. It is easy to check that this process produces a tree, as no cycle is ever created.

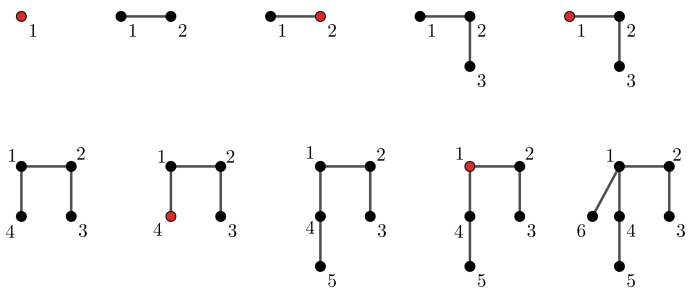

Figure 1.1: Illustration of the attachment process for a URRT. In red, the vertex chosen uniformly at random, where the new vertex attaches to.

The linear preferential attachment tree (PA tree)

This model, made famous by Barabási and Albert [14], is the simplest case of the broader class of preferential attachment. In the case studied in this thesis, the first graph of the sequence is an isolated vertex and the graph grows recursively by adding a new vertex, which connects to an already present vertex chosen at random with a probability proportional to its degree.

The ℓ-dag

Some of the real world objects mentioned above cannot be described by trees alone. This is why we introduce a recursive random graph model, based on the URRT. In this model, each new vertex connects not to a vertex chosen uniformly at random, but to ℓ vertices chosen uniformly at random (chosen with replacement).

Each of these models describes an attachment phenomenon. However, to describe propagation processes, it may be necessary to add a layer of complexity to these models. For example, to take into account the fact that as a social network grows, political ideas can propagate amongst its user. Broadcasting models do just that. Here, each new vertex not only connects to the past, but inherits a 0 or 1 bit (think voting Republican or Democrat). Here we present two broadcasting models.

Broadcasting in the URRT

Here, a bit is arbitrarily assigned to the first vertex. Each new vertex is then connected using the same process as in the URRT, but it inherits the bit of its ancestor with

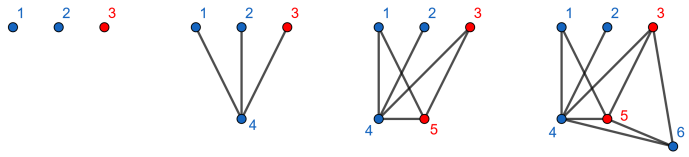

Figure 1.2: Illustration of the broadcasting process in an ℓ-dag.

probability $1-p$, and the opposite bit with probability p.

Broadcasting in an ℓ-dag

Here, the first graph in the sequence is a collection of ℓ isolated vertices, each arbitrarily assigned a bit. The connection process is the same as in the ℓ-dag. To assign the bit of the new vertex, a vector of ℓ bits is created with the bits of ℓ ancestors (some of which may be drawn several times and therefore appear several times in this vector). Then, each bit of this vector is independently flipped with probability p, called the mutation probability. Finally, a majority vote is taken to assign the bit of the new vertex.

These random graph models have been studied in numerous scientific work. Some of these can be grouped in the world of combinatorial statistics. It's impossible to give an exact definition of combinatorial statistics, but they encompass statistical problems where enumeration and combinatorial tools are used. Among these numerous problems, we can cite a few, taken in part from the notes of a course given at Saint-Flour, Lugosi [102].

The hidden clique

Define $\mathcal{G}(n,1/2,k)$ as an Erdős-Rényi model of parameter $1/2$ where the presence of a clique of size k is imposed. A natural question is to look for a statistical test to determine whether a graph has the law $\mathcal{G}(n,1/2)$ or $\mathcal{G}(n,1/2,k)$; or else to find the hidden clique (see for example the work of Alon, Krivelevich, and Sudakov [4]).

Dimension estimation in a random geometric graph

Given a geometric graph obtained with vertices uniformly distributed on the unit sphere of dimension d, and connecting points at distance less than $\sqrt{2}$. Is it possible to estimate the dimension d from observation of the graph alone (see, for example, the work of Atamanchuk, Devroye, and Lugosi [9])? Is it possible to create a test to differentiate this

graph from a $\mathcal{G}(n, 1/2)$?

Clustering

Given an SBM, for what values of p and q is it possible to retrieve (exactly or approximately) the two communities (see Lee and Wilkinson [97] for a review of the many results on this subject)? It's worth noting that the question of clustering can also apply to other types of data. For example, for points derived from a mixture of Gaussians (Even, Giraud, and Verzelen [68]). This remark allows us to bridge to the last chapter of this thesis, where questions close to those addressed in Chapters 3 and 4 are addressed, but when the data comes not from a graph but from the realization of a distribution on $\mathbb{R}^d$.

In the first three chapters following the introduction we study combinatorial statistics problems in recursive graphs. In particular, we ask how to infer information about the past of a random recursive graph.

1.1.1 Network archaeology in recursive random graphs

In recursive graphs, certain questions arise quite naturally, particularly concerning the inference of information about the process's past. For example, is it possible to find the first vertex of the graph? To estimate the order of arrival of all the vertices? To find out which bit had the first vertex in the broadcasting model? Let us start with the first question, that of *root finding*. Here, the aim is very simple: to find the first vertex. Several articles studying this problem have been published, with different formalizations of what *finding the first vertex* means. For example, Shah and Zaman [130] study a method that returns a single vertex, and prove that, in certain regimes, this vertex is indeed the vertex 1 with positive probability. Others have studied algorithms that only access a local observation of the graph to detect vertices of interest, for example Brautbar and Kearns [22]. Here, we formalize the root-finding task differently. We're looking for an algorithm which, given the graph $G = (V, E)$, returns a confidence set $S(\epsilon) \subset E$, containing vertex 1 with probability at least $1 - \epsilon$. This problem has been studied by Bubeck, Devroye, and Lugosi [33] in the case of URRTs and PA trees, Khim and Loh [86] have studied the case where the tree is obtained by diffusion on an infinite regular tree and Brandenberger et al. [21] the case of a size-conditioned Galton-Watson tree. A measure of the quality of these algorithms is the size of the confidence set, $K(\epsilon) = |S(\epsilon)|$. Note that here, we write $K(\epsilon)$ and not $K(|V|, \epsilon)$. This is because we want to find a confidence set whose size depends only on ϵ and not on the size of the graph. Looking at Figure 1.3, we realize that, in a URRT, identifying the first vertex is not straightforward. In fact, in what appears to be the "centre" of the graph, there

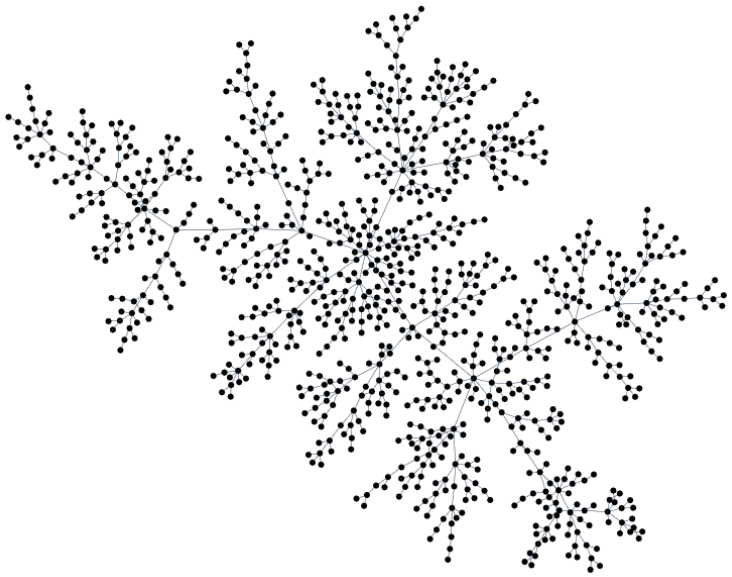

Figure 1.3: Realisation of a URRT of size 1000.

are both high-degree vertices and leaves. High-degree vertices are also found at peripheral positions in the graph. Thus, Bubeck et al. [33, Theorem 4] prove that the first vertex cannot be found exactly, and that, in a URRT, no matter which method is used,

$$K(\epsilon) \geq \exp\left(\sqrt{\frac{1}{30}\log\left(\frac{1}{2\epsilon}\right)}\right),$$

whether in a PA tree, one can not do better than

$$K(\epsilon) \geq \frac{c}{\epsilon},$$

for a positive constant c.

Recently, Contat, Curien, Lacroix, Lasalle, and Rivoirard [40] suggested an algorithm analyzing the degrees of pairs of vertices in the PA tree to locate vertex 1 in a confidence set of size $\epsilon^{-1+o(1)}$, which corresponds to the best possible performance. In the case of the URRT, there is still a gap between the lower bound and the performance of the best algorithm. To the best of our knowledge, the best algorithm is given by Bubeck et al. [33] and uses rumor centrality. It was later proved by Crane and Xu [42] that rumor centrality corresponds to ordering vertices by their likelihood of being vertex 1. This algorithm

locates vertex 1 in a confidence set of sub-polynomial size, of the order of $\exp\left(a\frac{\log 1/\epsilon}{\log 1/\epsilon}\right)$.

In Chapter 3, we develop a method for root-finding in graph models, as it is important to look at models that are more general than trees. Indeed, many problems are better described by graphs than by trees. For example, online communities or the world wide web are obviously not trees. Moreover, even if the theoretical model is a tree, in practice, during data acquisition, errors might be present and the tree structure destroyed. In the case of the URRT model, this is a real problem, as the methods analyzed so far depend entirely on the tree structure. In response to this problem, Crane and Xu [43] have studied the problem of root finding when noise is added to the tree. More precisely, they study the case where, in addition to the URRT or the PA tree, edges are added at random independently with the same probability for each pair of vertices (i.e. the edges of a graph $\mathcal{G}(n,p)$ are added to the edges of the tree). They introduce a Bayesian method and prove that it is possible to estimate the position of the root if the number of added edges is not too large. Here, we study two graph models and introduce a different method, based on the appearance of certain sub-graphs.

The two models we study are the ℓ-dag and a special case of the Cooper-Frieze model. The ℓ-dag model consists of a variant of the URRT model where, at each step, a new vertex connects not to a randomly chosen ancestor but to ℓ, chosen uniformly at random with replacement (multiple edges are then condensed into a single one). This model has been studied, for example, by Díaz Cort, Serna Iglesias, Spirakis, Torán Romero, and Tsukiji [54], Tsukiji and Mahmoud [135], Tsukiji and Xhafa [136] or Devroye and Janson [51]. This model is equivalent to considering the union of the edges of ℓ independent URRTs. Figure 1.4 illustrates this point of view for a 2-dag. The second model is a special case of the Cooper-frieze model, introduced by Cooper and Frieze [41]. Here, a parameter $\alpha \in (0,1)$ is fixed and the graph is grown from an isolated vertex. At each step, a Bernoulli random variable with parameter α is drawn, independently of past events. If the result is 0, a new vertex is added, which connects to an existing vertex chosen at random. Otherwise, a pair of vertices is drawn uniformly at random and an edge is added. If multiple edges appear, they are condensed into a single edge.

To select a confidence set $S(\epsilon)$, it is common practice to choose the most "central" vertices. Several notions of centrality exist. For example, Bubeck et al. [33] and Banerjee and Bhamidi [11] analyze Jordan centrality in URRTs. We can also think of rumor centrality, introduced by Shah and Zaman [130] and giving the best known algorithm for root finding in a URRT. It is also possible to use the likelihood of a vertex being vertex 1, as done by Crane and Xu [42] (as stated above, this turns out to coincide with rumor centrality). Our approach does not use these methods. Indeed, Jordan centrality, rumor centrality,

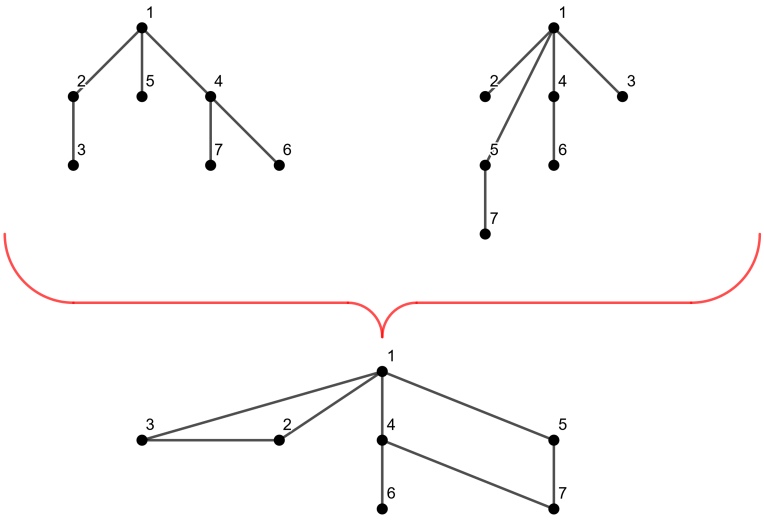

Figure 1.4: Illustration of a 2-dag.

or other popular notions of centrality, are only defined on trees. To analyze likelihood, Bubeck et al. [33] had already noticed that its analysis in the case of the URRT was too complex. They then proposed using a relaxed expression of the likelihood (corresponding to the rumor centrality), before Crane and Xu [42] realized that this relaxation did not change the order in which the vertices were ordered. This led to the study of the root finding algorithm, which involves selecting the vertices most likely to be vertex 1. In our case, likelihood has an even more complicated expression, and we were unable to simplify it, let alone show that its simplification orders the vertices in any meaningful way. Therefore, we decided to analyze another notion of centrality. We study the appearance of sub-graphs, and more precisely of double cycles. For the sake of clarity, the definition is deferred to Section 3.2. The confidence set is therefore the set of vertices present in small double cycles. We prove that this method locates vertex 1 in a set whose size does not depend on the size of the graph for the two models studied, and more precisely, Theorem 3.4 for the ℓ-dag model ensures that

$$K(\epsilon) \leq \frac{c_0}{\epsilon} \log\left(\frac{1}{\epsilon}\right)^{\frac{c_1}{\ell} \log \frac{1}{\epsilon}},$$

with probability at least $1 - \epsilon$. Theorem 3.5 ensures that, in the Cooper-frieze model,

$$K(\epsilon) \leq c_0 \log\left(\frac{1}{\epsilon}\right)^{c_1 \log \frac{1}{\epsilon}},$$

with probability at least $1 - \epsilon$. In contrast to the URRT model and the PA tree, we have not been able to find any lower bounds on the size of the confidence set. Let's take the case of the ℓ-dag model for example. Is it possible to do better than in the URRT model because the ℓ-dag is the superposition of ℓ independent URRTs? Or does destroying the tree structure make the problem strictly more difficult than in a URRT? These questions remain open today.

1.1.2 Arrival time estimation in random recursive trees

To extend our knowledge of network archaeology, we can think of two research directions. Choose a specific problem (e.g. root finding) and solve it in increasingly complex models or more efficiently. Among other things, this brings us closer to applications, for example by looking for robust algorithms, that can be applied to empirical data. Another possibility is to study more complex questions. The price to pay is to start working with simple models again, in the hope that, later on, some will be able to solve the same problem in more complex models. This is what we do in Chapter 4. A second problem of interest in the world of network archaeology is estimating the order of arrival of all vertices. Indeed, finding the first vertex provides only limited information on the history of the graph. But, retrieving the whole ordering of the graph can be very interesting, for example to track the history of the spread of fake news or rumors online. In this case, the problem is to estimate a latent random variable associated with each of the vertices: their arrival time. We study this problem in the case of the URRT and the PA tree. This type of problem has been widely studied, particularly in the field of seriation. In seriation problems, the aim is to estimate the order or relative positions of points by observing the affinity between them. This affinity is assumed to decrease with the distance between vertices in the latent space. This type of question appears in archaeology (Robinson [123]), bioinformatics (Recanati, Brüls, and d'Aspremont [121]) or matchmaking problems (Bradley and Terry [20]). A good example of data where affinities between pairs of vertices appear is the case of graphs, which are nothing other than the restriction to binary affinities (an edge is either present or not). In our case, we observe an adjacency matrix and try to estimate the position of vertices in the latent space of natural numbers. Gilbert [71], Giraud et al. [72], Janssen and Smith [82] have studied this problem in the case of random graphs. In particular, the example of the random geometric graph is emblematic of the seriation problem. The simplest example has as its latent space the unit circle in the plane. Figure 1.5 illustrates a realization of a random geometric graph.

Among seriation problems, Recanati et al. [121] study a problem close to ours, that

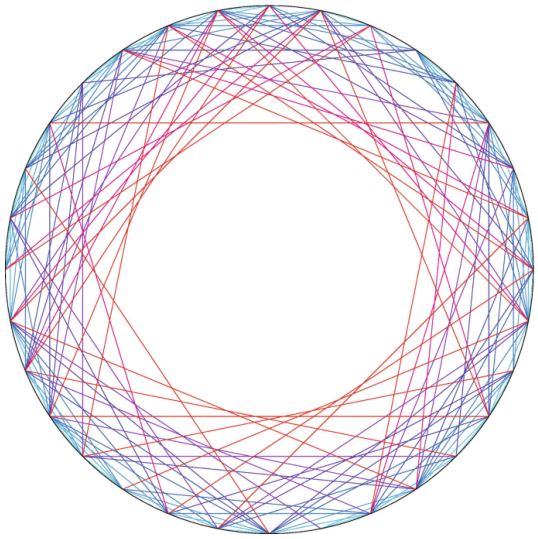

Figure 1.5: Illustration of a random geometric graph on the circle, taken from [7].

is, the observed data are perturbed Robinson matrices. A matrix is said to be Robinson when its entries are decreasing on the rows and columns away from the diagonal. In the case of a URRT or a PA tree, the expectation of the adjacency matrix is a Robinson matrix. Nevertheless, the results of Recanati et al. [121] do not apply and we show empirically in Section 4.4 that the method they propose performs poorly in our problem. To the best of our knowledge, the only theoretical result concerning the estimation of the order of arrival of vertices in a random graph comes from Crane and Xu [42]. In their paper, they present a general method for performing network archaeology tasks that can be applied to the case of arrival time estimation. This method consists in generating a tree ordering, with the same distribution as the true ordering of the model studied conditioned on the tree shape. Therefore, this method generates an order that has the same distribution as the true order, but they give no measure of the accuracy of this order. It was not obvious to us how to study its properties.

One of the reasons why the theoretical results of seriation do not apply to our problem, and why these same methods perform poorly empirically, is the temporal structure in our setting. In all the seriation problems we know, all the vertices have the same properties. For example, in the random geometric graph on the circle, the properties of all vertices are identical in law. This does not hold in our case, think for example of the degree of vertex 1 in the URRT, of the order of $\log(n)$ almost surely, whereas vertex n has a degree of 1 (where n is the size of the tree). This inhomogeneity has several consequences. First, the seriation methods introduced so far do not apply. Second, we need to define a new

way of measuring the quality of an estimator. Indeed, if we consider a uniform risk (e.g. the maximum distance between latent position and estimated position), then this error will be carried entirely by the leaves arriving at the end of the tree's growth. This is why we introduce a parametric family of risks as follows

$$R_\alpha(\widehat{\sigma}) \stackrel{\text{def}}{=} \mathbb{E}\left[\sum_{v\in V} \frac{|\widehat{\sigma}(v)-\sigma(v)|}{\sigma(v)^\alpha}\right],$$

for $\alpha > 0$. Here, $\sigma(v)$ denotes the arrival time of vertex v and $\widehat{\sigma}$ a method for estimating arrival times. This risk measure takes into account the inhomogeneity of the problem by putting more weight on vertices that arrive early in the graph. Indeed, we can verify that in a tree of size n, estimating the arrival time of a leaf makes an error of at least the order of n, whereas we can estimate the arrival time of older vertices much more precisely.

We study this risk in three stages, first by showing a lower bound on the best achievable performance, then by analyzing the performance of a vertex ordering procedure and finally by backing up our arguments with simulations. Our first results concern the best performance achievable by an ordering procedure. To do this, we need to limit ourselves to a restricted class of estimators. A common assumption, that makes practical sense, is that the ordering procedure is label invariant. In a few words, this means that the order returned depends only on the shape of the tree and not on the labels assigned to the vertices. An exact definition is given in Section 4.1. Under this assumption, it is possible to identify pairs of vertices in the URRT or PA tree that can not be ordered better than at random. Using these exchangeable pairs, we prove in Theorems 4.1 and 4.7 that the risk of a label ordering procedure is at least of the order of $n^{2-\alpha} \vee 1/2$, where n denotes the size of the tree. Quick calculations show that the maximum risk (obtained by ordering the vertices in the reverse order of the true order) results in an error of the order of $n^{2-\alpha}$ for $\alpha \in [0,1)$, of the order of $n\log(n)$ for $\alpha = 1$ and of the order of n for $\alpha > 1$. Thus, for α smaller than 1, the best and worst vertex orderings result in a risk of the same order of magnitude, suggesting that the renormalization induced in our risk is not interesting in the $\alpha \in [0,1)$ regime. Therefore, we only analyze the risk in the $\alpha \geq 1$ case.

Inspired by previous work on network archaeology in the URRT, we decided to analyze the method of ordering vertices by their Jordan centrality. To define it, we introduce, for a tree T and two disjoint vertices u and v, the subtree $(T,u)_v$. It corresponds to all vertices w for which v lies on the path between w and u on the tree T. The Jordan centrality of a vertex u is defined as

$$\psi(u,T) = \max_{v\in V(T),\ v\sim u} |(T,u)_v|,$$

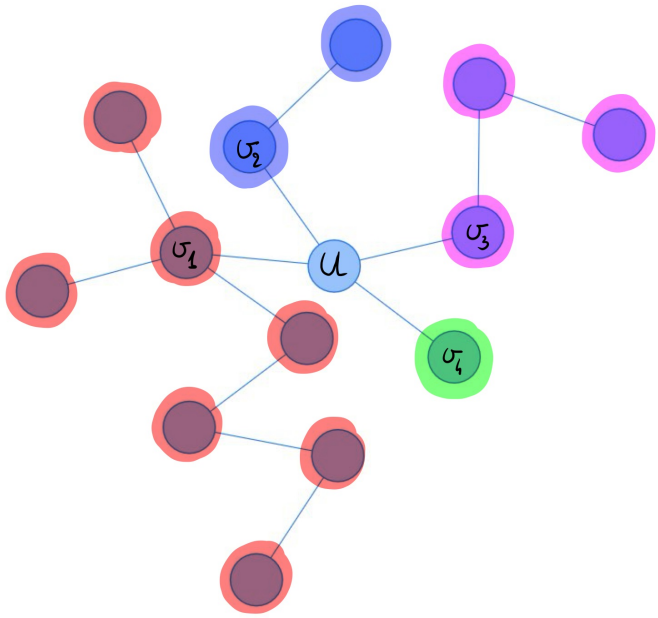

Figure 1.6: Illustration of the Jordan centrality. Here, vertex u has 4 neighbours, v_1, v_2, v_3 and v_4. Highlighted in red the subtree $(T,u)_{v_1}$, in blue the subtree $(T,u)_{v_2}$, in purple the subtree $(T,u)_{v_3}$ and in green the subtree $(T,u)_{v_4}$. Here, $\psi(u,T)=7$.

where $v \sim u$ indicates that vertices v and u are neighbors. See Figure 1.6 for an illustration of $(T,u)_v$ and the Jordan centrality. In particular, this ordering procedure is label invariant. It has the advantage of having been widely studied in the network archaeology problem, both in the URRT and PA tree models (Bubeck et al. [33], Moon [109], Wagner and Durant [139]). We use these results on the localization of vertex 1 as a first step to extend the analysis of the Jordan centrality to the ordering of all vertices. We prove bounds on the risk and in particular show in Theorems 4.4 and 4.8 that in the URRT model, for $\alpha \in [1,2)$ and in the PA tree, for $\alpha \in [1,5/4)$, the risk of this ordering procedure is of the order of $n^{2-\alpha}$, that is, of the order of the optimal risk. For larger α, we explain why this method cannot be optimal. In the URRT model, we propose a method using rumor centrality (Shah and Zaman [130]) and conjecture that its risk is optimal, up to constant factor, for all $\alpha \geq 1$.

Finally, we complete our discussion with simulations, to verify our results empirically, but more importantly to compare the performance of different ordering procedures.

In particular, we compare the performance of our estimator with a spectral method, studied by Recanati, Kerdreux, and d'Aspremont [122] for seriation problems where the affinities are Robinsonnian matrices (i.e. a framework close to ours). In the case of the PA tree, we also compare the performance of a pruning method, introduced by Navlakha and Kingsford [116]. It appears that Jordan centrality ordering is the only method we have tested whose risk grows at the optimal rate.

1.1.3 Broadcasting in random recursive trees

We can extend the horizon of problems in recursive graphs by introducing states on vertices. For example, we can define Covid-infected vertices and healthy vertices. Democrat-voting vertices and Republican-voting vertices. The list is infinite, and we formalize it mathematically in the simplest case of two states by assigning bits, 0 or 1, to each vertex. To describe practical problems, these states (or bits) are not assigned independently of the rest of the graph. In particular, in the broadcasting model, the implicit idea is that a vertex is more likely to inherit the state (bit) of its parents. This corresponds to different dependencies from the SBM case. In the SBM, the community is fixed a priori, and this has an impact on the attachment process. There, the bit has an impact on the structure of the graph. Indeed, new vertices are attached to the graph with a bit-dependent law. In the broadcasting models introduced earlier, the bit is assigned after the attachment process. Meaning that the way in which a vertex is attached is decided in a bit-independent way, and only then the vertex is assigned a bit. Note, therefore, that the heuristics differs between these two model classes. In the case of SBM, we are describing a process where similar vertices are more likely to connect (think social networks, for example). In broadcasting models, the idea is that connections take place independently of the state of the vertex, but that a vertex is more likely to inherit the state of the vertices to which it connects (think of Covid contamination or the inheritance of political convictions in a family, for example).

This new dimension in models opens the way to many new questions, in particular whether information is propagated throughout the entire graph. As the name suggests, this type of problem was motivated by the apparition of radio and television. Initially, deterministic graphs were studied, for example by Harutyunyan and Li [75] or Bhabak, Harutyunyan, and Tanna [17]. Like network archaeology problems, broadcasting problems are numerous. In the same way that the *root finding* problem appears naturally, its *rootbit finding* counterpart seems interesting. In this problem, an originator vertex (or vertices) is assigned a bit (or bits), which is (or are) then propagated from one vertex to the next in the graph. In the case of a tree, the problem is simplified because there is only one path between the originator vertex and any other vertex. This question was first formulated in

the case of general trees by Evans, Kenyon, Peres, and Schulman [67]. More recently, the case of random trees has been studied (Addario-Berry et al. [2], Desmarais, Holmgren, and Wagner [48]). Since then, a large number of variations of this problem have been studied, see Mossel [112] for a review of reconstruction problems on trees. In a spirit similar to that of the first project presented in this thesis, we have decided to study this problem in the case where bits propagate not on a tree but on a graph. We refer the reader to Section 1.1.1 for the motivations behind this generalisation to general graphs. In a similar problem, Antunović, Mossel, and Rácz [8] study the case of preferential attachment, where initial vertices are assigned a state and each new vertex has a colour assigned according to that of its neighbours. More recently, Makur, Mossel, and Polyanskiy [107] have studied another similar problem in a dag model different from the ℓ-dag, whose main parameters are the incoming degrees of the vertices and the number of vertices at distance k from vertex 1. It is also assumed that the position of vertex 1 is known. Two propagation processes are studied, a noisy majority voting process and a NAND-based decision process. They show that, if the number of vertices at depth k is of the order of $\Omega(\log(k))$, there is a threshold on the mutation probability below which it is possible to estimate the bit of the 1 vertex.

Here, we look at the broadcasting model on ℓ-dags, and in particular at the proportion of each bit. This is why we are making the link with Pólya urns. Indeed, if we are only interested in the number of vertices of a given bit, the broadcasting process on a ℓ-dag is such that the graph structure is no longer important. The model can thus be described as follows. An urn is filled with ℓ balls, blue (bit 0) or red (bit 1). When a new ball is added, its colour is decided by drawing ℓ balls successively with replacement. Their colours are observed, but with probability p a blue ball is observed as red (and vice versa). Finally, of the ℓ colours observed, the majority is passed on to the new ball. This connection to Pólya urns had already been made in the case of broadcasting in a URRT by Addario-Berry et al. [2]. In this case, the proportion of zero bits follows a Pólya urn with random replacement (the colour added is not a function of the colour of the ball drawn). These processes are called reinforcement processes, and we use results compiled by Pemantle [118] as well as non-convergence results studied by Pemantle [117]. We use as much as possible the description of the problem as a Pólya urn, partly because these properties have been widely studied in the literature (Janson [77], Knape and Neininger [91], Wei [140]). Many variations of this model have been studied, for example by increasing the number of colours (Bertoin [16]), by choosing several balls for each draw (Kuba and Mahmoud [94]), by making the choice of added colours non-deterministic (conditional on the draw) (Janson [80], Zhang [141]). The closest variation to our model that we are aware of comes from Crimaldi, Louis, and Minelli [44], with a multiple-draw model and linear random replacement. Our model also has a multiple draw but a non-linear random replacement law.

In the case where $\ell = 1$, the model is greatly simplified by the disappearance of majority voting and the fact that broadcasting then takes place in a tree. In this case, Addario-Berry et al. [2] propose two methods for rootbit finding. The first method consists of estimating the position of vertex 1, then using its state as an estimate of the state of vertex 1. This method therefore uses results from network archaeology and, in part, results based on the location of the vertex 1. In the case of ℓ-dags, these results are still unknown. We know of no natural method for defining a graph centre, let alone controlling the distance between this centre and vertex 1. We can estimate a confidence set for vertex 1, but we do not know how to control the typical distance of these vertices to vertex 1. However, since the size of this set is independent of the size of the graph, we can find exactly vertex 1 with positive probability (by choosing a vertex at random from this set). This is not the approach we have decided to follow, but we discuss it briefly. The second method proposed by Addario-Berry et al. [2] consists of estimating the bit of vertex 1 by the majority bit in the tree. This method has the advantage of being easily applicable to our model. It is also known that, in the case $\ell = 1$, this method is optimal for small probabilities of mutation (see Addario-Berry et al. [2]).

We define the majority bit at time n by b_n^{maj}, which is decided at random if both bits are present in equal numbers. We seek to determine for which regimes of (ℓ, p) this method makes it possible to find which bit was present in majority at the initialisation of the graph, that is, we seek for which values of the couple (ℓ, p) the probability of error $R^{maj}(n,p) = \mathbb{P}\left\{b_\ell^{maj} \neq b_n^{maj}\right\} < 1/2$. Note that this quantity depends on the initial conditions, but for the sake of clarity we do not take this into account in the notations. In the case $\ell = 1$, Addario-Berry et al. [2] prove that

(i) There exists a constant $c > 0$ such that

$$\limsup_{n\to\infty} R^{maj}(n,p) \leq cp\ .$$

(ii) Denoting R_n the proportion of bits 0 at time n,

$$\lim_{n\to\infty} R_n = \frac{1}{2} \text{ almost surely}\ .$$

(iii) For $p \in [0, 1/4)$

$$\limsup_{n\to\infty} R^{maj}(n,p) < \frac{1}{2}\ .$$

(iv) For $p \in [1/4, 1/2]$

$$\limsup_{n\to\infty} R^{maj}(n,p) = \frac{1}{2}\ .$$

Our work extends these results. We introduce for $\ell \geq 1$ odd

$$\alpha_\ell = \frac{1}{2^{\ell-2}} \sum_{i>\ell/2}^{\ell} \binom{\ell}{i} (i - \ell/2) \, .$$

Thus, $\alpha_1 = 1$, $\alpha_3 = 3/2$ and for ℓ going to infinity

$$\alpha_\ell \sim \sqrt{\frac{2\ell}{\pi}} \, .$$

We prove in Theorem 5.1 that

(i) If $p < \frac{1}{2} - \frac{1}{2\alpha_\ell}$, then there exists $\beta \in (0, 1/2)$ (of which the value only depends on ℓ and not on the initial conditions) such that

$$\mathbb{P}\{R_n \to \beta\} + \mathbb{P}\{R_n \to 1-\beta\} = 1 \quad \text{and} \quad \mathbb{P}\{R_n \to \beta\} < \mathbb{P}\{R_n \to 1-\beta\} \, .$$

In particular, independently of R_ℓ,

$$\limsup_{n\to\infty} R^{maj}(n,p) < \frac{1}{2} \, .$$

(ii) If $\frac{1}{2} - \frac{1}{2\alpha_\ell} \leq p < \frac{1}{2} - \frac{1}{4\alpha_\ell}$, then $R_n \to 1/2$ almost surely and

$$\limsup_{n\to\infty} R^{maj}(n,p) < \frac{1}{2} \, .$$

(iii) If $\frac{1}{2} - \frac{1}{4\alpha_\ell} \leq p \leq \frac{1}{2}$, then $R_n \to 1/2$ almost surely and

$$\lim_{n\to\infty} R^{maj}(n,p) = \frac{1}{2} \, .$$

Note that, for $\ell = 1$, the low mutation probability regime does not exist, which is in agreement with the results of Addario-Berry et al. [2]. For $\ell \geq 3$, three regimes exist. For small mutation probabilities the proportion converges to a value other than $1/2$. On the other hand, in the intermediate and high mutation probability regimes, the proportion tends towards $1/2$. However, as long as p does not get too close to $1/2$, even if the proportion of bits tends towards $1/2$, the majority bit is positively correlated with the majority bit when the graph is created. Note also that, as in the case of the preferential attachment studied by Antunović et al. [8], in all regimes no bit disappears (i.e. their proportion does not tend towards 0).

1.2 Introduction to the random friendship tree

The rest of this thesis deals with topics other than network archaeology. Initially, we remain in the world of random graphs and combinatorial statistics. All previously introduced models have been studied in detail. When a new model is introduced, for example to account for a different attachment process, it is interesting to study its most basic properties (diameter, maximum degrees, number of leaves, etc). This gives us a better understanding of the process and opens its comparison to reference models. Is this model simply a variant of the URRT or does it have completely different dynamics? Is there a rich-gets-richer phenomenon as in a PA tree, or is its attachment fairer? All these questions have been studied at length for more classical models such as the URRT, the ℓ-dag, the PA tree, the Erdős-Rényi, etc. In order to better describe certain phenomena, new models are regularly introduced. We produce the first rigorous analysis of one of these models, the random friendship tree (see Krapivsky and Redner [93]).

One characteristic that appears in many models is the presence of a 'rich gets richer' phenomenon. In other words, a vertex that is highly connected tends to reinforce its dominance over time. This behaviour is present in preferential attachment models and therefore in the PA tree. In the simplest definition of this model, the attachment rule is not local. In other words, in order to grow the tree according to the law of preferential attachment, we need access to the degrees of each vertex, or to all the edges present in the graph. Since it can be assumed that in practice attachment processes are local (for example in physical phenomena), many local attachment models exhibiting a rich-gets-richer phenomenon have been introduced. For example, Engländer, Iacobelli, Pete, and Ribeiro [64] have very recently introduced a random walk model constructing a tree. In their model, a tree is recursively built by a "walker" who moves randomly along the tree. At step n, with probability $n^{-\gamma}$, a new vertex is connected to the vertex where the walker is located. They show that this model corresponds to the PA tree, therefore proving that this model can be generated from a local attachment procedure (at each step the walker only needs to know its neighbourhood in order to progress).

Another way of creating a local attachment model where a rich-gets-richer phenomenon is at work is to introduce a redirection phenomenon. In other words, a model where a new vertex does not necessarily connect to a randomly chosen vertex but possibly to a neighbour (or close neighbour) of a randomly chosen vertex. Introduced by Kleinberg, Kumar, Raghavan, Rajagopalan, and Tomkins [90] in directed trees, the initial model consists of connecting each new vertex to a vertex chosen at random with probability $1-p$ or to its ancestor with probability p. This model gives rise to a preferential attachment process where each new vertex connects to a vertex chosen with probability proportional

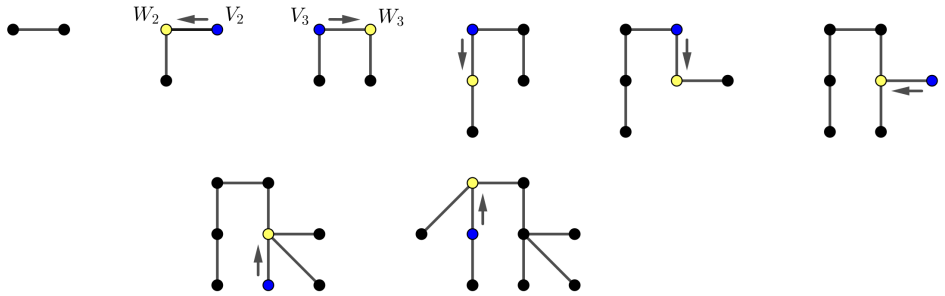

Figure 1.7: Illustration of the attachment process in a random friendship tree. In blue, vertex V_n, which is chosen uniformly at random among all vertices. In yellow, vertex W_n, which is chosen uniformly at random among all neighbours of V_n. Vertex $n+1$ attaches to W_n.

to $d-2+1/p$ (for d the degree of the vertex). Later, Saramäki and Kaski [125] introduced an undirected version of the model, later studied by Evans and Saramäki [66]. In their work and that of Evans and Saramäki [66] they claimed that this corresponds to a preferential attachment model, which was noted to be incorrect by Cannings and Jordan [35].

In the case of redirection in undirected trees, a tree is grown recursively by choosing a vertex uniformly at random, then starting at this vertex a random walk with k steps and finally attaching a new vertex to the last vertex reached by this random walk. Unlike the work of Engländer et al. [64], a new random walk is created at each step. In the case of $k=1$, we call this model the *random friendship tree* (RFT) because we can think of each new vertex becoming "friend" with a "friend" of a randomly chosen vertex. This family of processes, ranging from $k=0$ to $k=\infty$, has the property of containing the URRT model and the PA tree. Indeed, if $k=0$, then each new vertex connects to a vertex chosen uniformly at random and the model is therefore a URRT. If $k\to\infty$, then the resulting model is a PA tree. This is because a finite tree has a finite mixing time and the stationary distribution of a random walk on a tree is proportional to the degrees. Here, we will study the special case $k=1$, that is, each new vertex is attached as follows: at time $n+1$, a vertex is chosen uniformly at random, it is called V_n. Among the neighbours of V_n a vertex is chosen uniformly at random, it is called W_n. Vertex $n+1$ connects to W_n. Figure 1.7 illustrates this process.

Although the random friendship tree is part of a family of models containing the URRT and the PA tree, some of their properties are drastically different. Among these properties, the sequence of degrees or the modularity have been conjectured in an empirical work by Krapivsky and Redner [93]. The only theoretical result concerning this model of which I am aware concerns the number of leaves. Cannings and Jordan [35] proved that, almost surely, $n - o(n)$ vertices were leaves. We are extending the theoretical knowledge of RFT. Concerning small degree vertices, we show that at least $n - n^{0.9}$ vertices are leaves, whereas in the URRT or the PA tree only a fraction of the vertices are leaves. While for a fixed k there are on the order of $n/2^{k-1}$ vertices of degree at least k in the URRT (Janson [78]), this number is between $n^{0.1}$ and $n^{0.9}$ in the case of the RFT (see Theorem 6.11). Moreover, we show in Proposition 6.13 that most leaves will remain leaves forever, whereas in the URRT or PA tree the degree of each vertex tends towards infinity almost surely. Concerning high-degree vertices, we show in Theorem 6.2 that linear degree hubs appear. These hubs are clearly visible in Figure 1.8. In contrast, the maximum degree in a URRT is logarithmic (Devroye and Lu [52]) and in a PA tree of the order of $\sqrt{n}$ (Van Der Hofstad [138, Theorem 1.17]). We even show a stronger result: for each edge, at least one of its vertices will almost surely become a linear degree hub. This phenomenon is unprecedented and present in no other model of which we are aware. However, not everything is different from the URRT or PA models. In Theorem 6.6 we show that the diameter is logarithmic almost surely, as in a URRT (Addario-Berry and Ford [1, Corollary 1.3]), whereas it is at most logarithmic in a PA tree (Dommers, van der Hofstad, and Hooghiemstra [55]). Moreover, as we show in Theorems 6.7 and 6.8, in both a RFT and a URRT the farthest vertex from a leaf is at distance $\Theta(\log(n)/\log\log(n))$ from it.

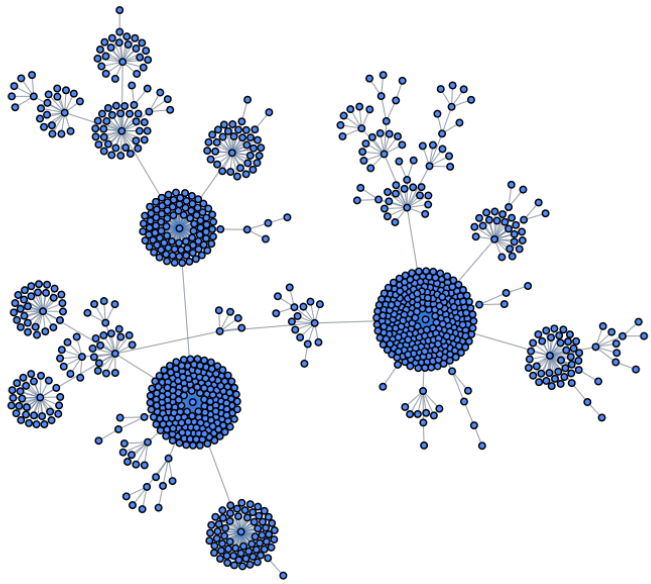

Figure 1.8: Illustration of a RFT of size 1000.

1.3 Introduction to the random Tukey depth

As was quickly mentioned when citing the clustering problem, it is possible to study similar questions in very different settings. In the last part of this thesis we study the problem of ordering high-dimensional data. The problem is therefore similar to the one studied in Chapter 4, and here we are also interested in a notion of centrality for ordering points in $\mathbb{R}^d$. However, the fact that the data being studied (a graph or a collection of points in high dimension) are completely different means that we have to use different statistical tools. Here, we enter deeper into the world of data analysis and machine learning. Because of their countless industrial applications, these are fields where research is fast and extensive. One of the major challenges facing practitioners today is the increasing size of the objects being studied. Many applications rely on high-dimensional data, and the growth of databases is only intensifying this problem. The fast progress of tools based on artificial intelligence and their media coverage are also a factor that pushes every statistician to confront himself to problems linked to high dimensionality. The problem in question here consists of ordering the points of a dataset by their centrality. It can be useful to define a partial order, to be able to visualise the data, from the most central to the extreme values. More than a visualisation tool, a centrality order is needed to train a conformal predictor

(in other words, a predictor that returns not a single point but a confidence interval). In dimension 1, an obvious way of ordering points in a set is by their empirical quantiles. Unfortunately, from dimension 2 onwards, the notion of quantile (and even median) is no longer so easily defined. Tukey [137] introduced a notion of depth, called the Tukey depth (or half-space depth), which is a popular tool for visualising the centrality of a point in a dataset. Many other depth measures have been introduced, such as simplex depth (Liu [99, 100]), projection depth (Liu [101], Zuo and Serfling [143]), zonoid depth (Koshevoy and Mosler [92, 92]) or a notion of outliers, (Donoho [56], Stahel [134]). Each of these notions has different properties of stability, invariance or computability, making them suitable for different applications. Tukey depth is defined as follows. Given a point x in R^d, for a direction $u \in S^{d-1}$ (where S^{d-1} is the Euclidean sphere of R^d) we define the closed half-space

$$H(x,u) = \{y \in \mathbb{R}^d \ : \ \langle y,u\rangle \leq \langle x,u\rangle\} \ .$$

Then, for a set of points $\{x_1,\cdots,x_n\}$, we define the depth in direction u by

$$r_n(x,u) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x_i \in H(x,u)} \ ,$$

which corresponds to the empirical quantile of $\langle x,u\rangle$ in the dataset projected on the direction u. We can then define the Tukey depth of x in the set $\{x_1,\cdots,x_n\}$ by

$$d_n(x) = \inf_{u\in S^{d-1}} r_n(x,u) \ .$$

Note that the principle is similar to that of Jordan centrality. We define a score, here the Tukey depth, and use this score to order points in $\mathbb{R}^d$. The Tukey depth has the expected attributes of a depth measure: it is invariant by affine transformation, tends to 0 as $\|x\| \to \infty$ and decreases on radii starting from the deepest point. Moreover, it is robust under conditions of symmetries (Donoho and Gasko [57]), and when data are drawn independently from the same distribution its contour lines converge rapidly (Brunel [31]). Nevertheless, one of the problems encountered with this depth measure is the difficulty of calculating it. This problem is highlighted by the need to process data of increasingly large dimensions. Thus, calculating even an approximate value is a NP hard problem, as shown by Amaldi and Kann [6], Bremner et al. [24], Johnson and Preparata [85]. Moreover, the calculation of the maximum depth is done in time $\mathcal{O}\left(n^{d-1}\right)$ (Chan [37]). Despite these difficulties, Tukey's depth has not been discarded by practitioners because other depth measures face the same challenges. This is why work is being done to develop and analyse approximation algorithms. Their importance is underlined by Nagy, Dyckerhoff, and Mozharovskyi [115]. Among the solutions proposed, Shao, Zuo, and Luo [131] study an algorithm based on

mcmc methods to calculate the projection depth in high dimension. For the Tukey depth, Zuo [142] proposed a new approach for calculating an approximation, but without giving any guarantees on its accuracy. Finally, Chen, Morin, and Wagner [38] studied the quality of various approximations to the Tukey depth.

Cuesta-Albertos and Nieto-Reyes [45] introduced a natural approximation algorithm, which does not take the minimum over all possible directions in $\mathbb{R}^d$ but limits itself to k directions chosen at random. Thus, for $U_1, \cdots, U_k$ independent uniform random variables on S^{d-1}, we can define the random Tukey depth of the point x within the set $\{x_1, \cdots, x_n\}$ by

$$D_{n,k}(x) = \min_{i=1,\cdots,k} r_n(x, U_i) .$$

It is easy to see that for each x, when k tends to infinity, the random Tukey depth converges with probability 1 to the Tukey depth. However, we need to know more to be able to use this approximation in practice. In particular, we need to know what size k must be for $D_{n,k}(x)$ to be a good approximation of $d_n(x)$. More precisely, for $\epsilon \in (0, 1/2)$, $\delta > 0$, how large must k be to ensure that $|D_{n,k} - d_n(x)| \leq \epsilon$ with probability at least $1 - \delta$? To investigate this question we need to restrict our study to "reasonable" datasets. Indeed, already in dimension 2, it is possible to construct an example of a dataset where k must be arbitrarily large to achieve a given precision. For example, for n even, let's define the points $x_i = \left(i/n, a(i/n)^2\right)$ where $a > 0$ is a parameter. Since it lies on the boundary of the convex envelope of $\{x_1, \cdots, x_n\}$, the depth of the point $x_{n/2}$ is 0. However, to "see" this depth we need to evaluate the depth in a direction u such that

$$\langle x_{n/2-1}, u \rangle > \langle x_{n/2}, u \rangle ,$$

and

$$\langle x_{n/2+1}, u \rangle > \langle x_{n/2}, u \rangle .$$

An illustration, presented in Figure 1.9, shows that by choosing a arbitrarily small, the area where u must be to detect the depth of $x_{n/2}$ becomes arbitrarily small, and consequently k must be chosen arbitrarily large to hope to estimate the Tukey depth.

This is why we have to assume a certain regularity in the dataset. We also assume that the points are independent and identically distributed, a natural assumption in machine learning. More precisely, we assume that the points $\{x_1, \cdots, x_n\}$ are independent realisations of an isotropic log-concave distribution. Note that the isotropic assumption

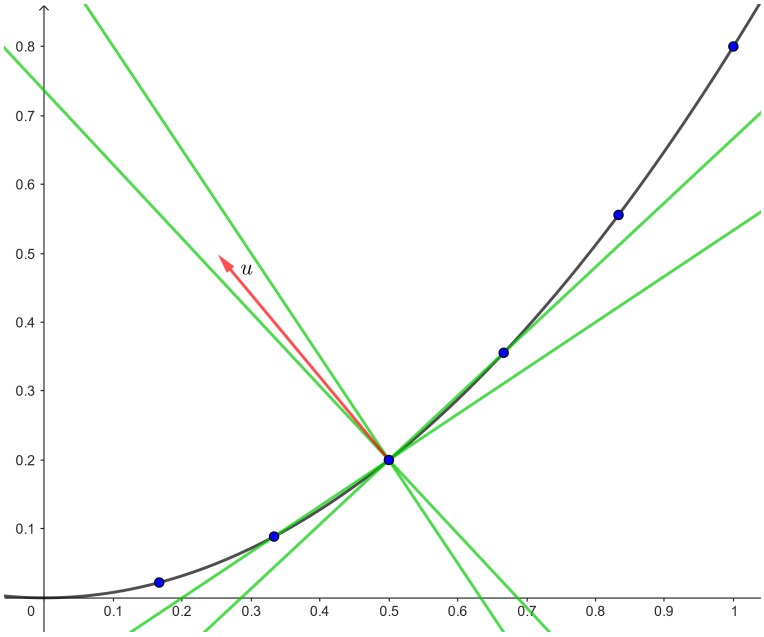

Figure 1.9: Illustration of the dataset $x_i = \left(i/6, 0.8(i/6^2\right)$ for $i \in [6]$. In green, the lines in between vector u must be to detect that the depth of x_3 is 0.

is not necessary, as it is possible to estimate precisely the covariance of a non-isotropic distribution, and therefore it is possible to put a dataset sampled from a log-concave distribution into a (quasi) isotropic position. For data sampled from a random variable, it is possible to introduce a version of Tukey's depth based solely on the distribution. Rather than counting the number of points in a half-space, we can simply measure this half-space for the distribution in question. We thus define a new version of d_n and $D_{n,k}$, respectively $\overline{d}$ and $\overline{D}_k$. A precise definition is given in Section 7.1.

We are not the first to be interested in the quality of this approximation algorithm. For example, Cuesta-Albertos and Nieto-Reyes [45] have produced experimental results suggesting that $\overline{D_k}$ is a good approximation of $\overline{d}$. From a theoretical point of view, Nagy et al. [115] have studied under what conditions $\sup_{x\in\mathbb{R}^d} |\overline{D}_k(x) - \overline{d}(x)| \to 0$ when $k \to \infty$. They also gave bounds on the speed of convergence. In contrast to this uniform approach, we study the quality of the approximation of $\overline{d}(x)$ by $\overline{D}_k(x)$ for a fixed point x. In particular, we show that the quality of this approximation depends strongly on the depth of x. Our results are presented in three parts. First, most of the points in the dataset have a shallow depth, which is easy to estimate. Second, estimating the Tukey depth of intermediate depth points is hard, as it requires k to grow exponentially with the dimension. Finally, if there is a point of depth $1/2$, it is easy to locate. More precisely, if the measure μ is log-concave isotropic on $\mathbb{R}^d$, we show in Corollary 7.2 that there exist universal constants $c, \kappa, C > 0$ such that for $\epsilon, \delta, \gamma > 0$, if

$$k = \left\lceil \max\left(C, \frac{4}{\epsilon} \log \frac{3}{\gamma}, \frac{2}{c} \log \frac{4}{\delta} \right) \right\rceil ,$$

and if the dimension d is greater than

$$d \geq \max\left(\left(\frac{3(k+1)}{\gamma} \right)^{1/\kappa}, \frac{64 \log(1/\epsilon) k}{\pi} \log \frac{3k}{\gamma}, \left(\frac{1}{c} \log \frac{6k}{\delta} \right)^2, \left(\frac{2}{\epsilon} \right)^{\kappa} \right) ,$$

then, with probability at least $1 - \delta$,

$$\mu\left(\left\{ x \in \mathbb{R}^d : \overline{D}_k(x) > \epsilon \right\} \right) < \gamma .$$

Since $\overline{D}_k(x) \geq \overline{d}(x)$, this corollary implies that, in the sense of the μ measure, most points have a small depth and that this is easily estimated. It is easy to understand why, in large dimensions, most points have a small depth. An isotropic log-concave distribution on $\mathbb{R}^d$ "looks like" the uniform distribution on the sphere of radius $\sqrt{d}$ in $\mathbb{R}^d$. We say that most of these points' depths are easy to estimate because, for a given precision, the parameter k

does not depend on the dimension. Although this Corollary proves that most points have a small depth, and that this depth is easy to estimate, it does not prove that for all points with a small depth the depth is easy to estimate. For example, consider the uniform distribution on $[-(3/2)^{1/3}, (3/2)^{1/3}]^d$, which is log-concave isotropic. The point $x = ((3/2)^{1/3}, 0, \cdots, 0)$ has depth 0, and yet we can show that $\overline{D}_k(x) \geq 1/4$ with high probability if k is not exponentially large in d.

We then show in Corollary 7.3 that intermediate depths are hard to estimate. Again, for μ log-concave isotropic on $\mathbb{R}^d$, we show that for $\delta \in (0,1)$, $\gamma \in (0,1/2)$, there exists a positive constant $c = c(\gamma)$ such that if $x \in \mathbb{R}^d$ is such that $\overline{d}(x) = \gamma$, then for $\epsilon < c$, if $k \leq \delta e^{d\epsilon^2 \log^2(1/\epsilon)/c}$, then with probability at least $1 - \delta$

$$|\overline{D}_k(x) - \overline{d}(x)| \geq \epsilon \ .$$

Finally, we show that the point of depth $1/2$, if it exists, is easy to locate. Note that there is no guarantee that such a point exists. The deepest point can have a depth smaller than $1/2$, but not arbitrarily small. Nagy, Schuett, and Werner [114, Theorem 3] show that $1/e \leq \sup_{x \in \mathbb{R}^d} \overline{d}(x) \leq 1/2$. In the case where $\sup_{x \in \mathbb{R}^d} \overline{d}(x) = 1/2$, the distribution is said to be half-space symmetric (Nagy et al. [114], Zuo and Serfling [144]). It is easy to see that, in this case, there is a single point of depth $1/2$, which is called the Tukey median. A symmetric distribution is half-space symmetric, but the converse is false. However, in the case of a uniform distribution on a convex K, then the distribution is half-space symmetric if and only if K is symmetric (Funk [70], Schneider [128]). Corollary 7.4 proves that, if μ is log-concave, isotropic and half-space symmetric, then if a point is such that $\overline{D}_k(x) \approx 1/2$ then x is close to the Tukey median. More precisely, consider $X_1, \cdots, X_n$ independent random variables distributed according to μ, m the Tukey median and $m_{n,k}$ the empirical Tukey median (i.e. $D_{n,k}(m_{n,k}) = \max_{x \in \mathbb{R}^d} D_{n,k}(x)$). Then there exist positive universal constants c and C such that for $\delta \in (0,1)$, $\gamma \in (0,c)$, if $n \geq Cd/\gamma^2$ and

$$k \geq c(d \log(d) + \log(1/\delta)) \ ,$$

then

$$\left\| m_{n,k} - m \right\| \leq C\gamma\sqrt{d} \ ,$$

with probability at least $1-\delta$. This means that by fixing γ of the order of $1/\sqrt{d}$ we can locate the Tukey median at constant distance with only $k \approx d \log(d)$. Note that since in high dimensions the mass of an isotropic log-concave distribution is concentrated on the sphere of diameter $\sqrt{d}$, locating the Tukey median at constant distance is not a trivial estimate.

Chapter 2

Introduction en Français

Contents

2.1 Introduction à l'archéologie dans les graphes

Autour de nous beaucoup de phénomènes peuvent être décrits par des processus d'attachement ou de propagation. Pour en citer quelques uns, pensez à la propagation d'une maladie, d'une fake news ou d'un virus informatique, ou à l'évolution d'un réseau social. Pour décrire ces phénomènes, il est possible d'utiliser des graphes, et plus précisément une séquence de graphes rendant compte du processus d'évolution de l'objet étudié. Par exemple, à un instant donné, un réseau social peut être décrit par un graphe où chaque individu correspond à un sommet et chaque lien d'amitié correspond à une arrête. Du fait que ce graphe évolue, on peut considérer la séquence de graphes qui décrit l'évolution du réseau social. Induite par cette séquence, une notion d'*histoire* apparaît. En effet, un sommet a été ajouté en premier et un autre mille-deux-cent-quarante-troisième. Dans cette

partie de ma Thèse, nous allons essayer de retrouver de l'information sur cette histoire dans le cas où nous n'observons le graphe qu'à un instant donné (et pas la séquence). Pour formuler ce problème mathématiquement, nous devons définir quelques modèles de graphes aléatoires. Ces modèles servent de bac à sable où développer des techniques et algorithmes dans des environnements parfaitement définis. Commençons par des modèles de graphe très couramment étudiés, mais dont la taille est fixée.

Modèle d'Erdős–Rényi

Dans ce modèle, un paramètre n (la taille du graphe est fixé). Lorsque l'on parle de graphe d'Erdős–Rényi on peut se référer à deux modèles similaires. Le modèle $\mathcal{G}(n, M)$, qui consiste à choisir uniformément au hasard un graphe parmi tous les graphes à n sommets et M arrêtes, et le modèle $\mathcal{G}(n, p)$, qui est construit en connectant aléatoirement les n sommets. Chaque arrête est incluse dans le graphe avec probabilité p, indépendamment de toutes les autres arrêtes.

Les graphes géométriques aléatoires

Dans ce modèle, un espace métrique X munie d'une mesure est fixé. Un paramètre n (la taille du graphe), un paramètre ρ de connectivité ainsi qu'un distribution μ sont aussi fixés. n points sont tirés au hasard de manière indépendantes selon la loi μ. Chacun de ses points correspond à un sommet du graphe, et chaque paire de sommets est connecté si les points correspondants sont à une distance inférieure à ρ.

Modèles à blocs stochastiques (SBM)

Dans la version la plus simple de ce modèle deux communautés sont créés. Le modèle est ensuite similaire au $\mathcal{G}(n, p)$, à la différence prêt que la probabilité qu'une arrête soit ajoutée au graphe dépend des communautés de ses extrémités. Si ces extrémités sont dans la même communauté, elle est ajoutée avec probabilité p, sinon avec probabilité q.

Dans ce travail nous étudierons des modèles *récursifs*, c'est à dire qui grandissent en ajoutant récursivement des sommets et des arrêtes. En voici quelques exemples.

L'arbre d'attachement uniforme (URRT)

Sans doute le modèle le plus simple pour décrire un processus d'attachement. Le premier graphe de la séquence consiste en un sommet isolé. Ce graphe est agrandi récursivement en ajoutant un nouveau sommet, connecté à un sommet déjà présent dans le graphe choisi uniformément au hasard. Il est facile de vérifier que ce processus produit

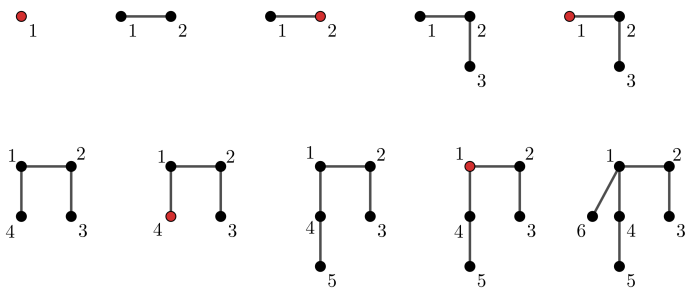

Figure 2.1: Illustration du processus d'attachement pour un URRT. En rouge le sommet choisi uniformément au hasard, où le nouveau sommet est connecté.

un arbre, car aucun cycle n'est jamais créé.

L'arbre d'attachement préférentiel linéaire (PA tree)

Ce modèle, rendu célèbre par Barabási and Albert [14], est le cas le plus simple de la classe plus large de l'attachement préférentiel. Dans le cas étudié dans cette thèse, le premier graphe de la séquence est un sommet isolé et le graphe grandit récursivement en ajoutant un nouveau sommet, qui se connecte à un sommet déjà présent choisi au hasard avec une probabilité proportionnel à son degré.

Le ℓ-dag

Certains des processus réels cités plus hauts ne peuvent pas être décrits seulement par des arbres. C'est pourquoi nous introduisons un modèle de graphe aléatoire récursif, sur le modèle de l'URRT. Cette fois, chaque nouveau sommet ne se connecte pas à un sommet choisi uniformément au hasard mais à ℓ sommets choisis uniformément au hasard (choisis avec remise).

Chacun de ces modèles décrivent un phénomène d'attachement. Cependant, pour décrire des processus de propagation il peut être nécessaire d'ajouter une couche de complexité à ces modèles. Par exemple, pour prendre en compte que lorsqu'un réseau social grandit, des idées politiques peuvent se propager au sein du graphe. Avec les modèles de broadcasting c'est chose faite. Ici, chaque nouveau sommet, non seulement se connecte au passé, mais hérite d'un bit 0 ou 1 (penser par exemple à voter Républicain ou

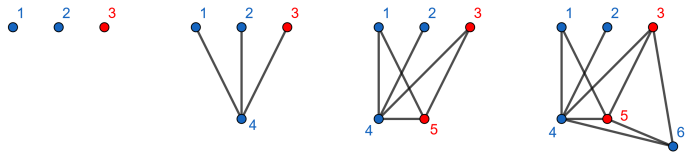

Figure 2.2: Illustration du processus de broadcasting dans un ℓ-dag.

Démocrate). Nous présentons ici deux modèles de broadcasting.

Broadcasting dans un URRT

Ici le premier sommet se voit arbitrairement attribué un bit. Chaque nouveau sommet est ensuite connecté selon le même processus que dans l'URRT mais il hérite du bit de sont ancêtre avec probabilité $1-p$, et du bit opposé avec probabilité p.

Broadcasting dans un ℓ-dag

Ici, le premier graphe de la séquence est une collection de ℓ sommets isolés qui se voient chacun arbitrairement attribués un bit. Le processus de connexion est le même que dans le ℓ-dag. Pour attribuer le bit du nouveau sommet, un vecteur de ℓ bits est créé avec les bits de ℓ ancêtres (certains peuvent être tiré plusieurs fois et apparaissent donc plusieurs fois dans ce vecteur). Puis, chaque bit de ce vecteur est indépendamment changé avec probabilité p, appelée probabilité de mutation. Enfin, un vote par majorité a lieu pour attribuer le bit du nouveau sommet.

Ces modèles de graphes aléatoires ont engendrés un grand nombre de travaux scientifiques. Une partie de ceux ci peuvent être regroupés dans le monde des statistiques combinatoires. Il est impossible de donner une définition exacte des statistiques combinatoires mais elles regroupent des problèmes statistiques où des outils de dénombrement sont utilisés. Parmi ces innombrables problèmes, nous pouvons en citer quelques uns, tirés en partie des notes d'un cours donné à Saint-Flour, Lugosi [102].

La clique cachée

Définissons $\mathcal{G}(n,1/2,k)$ comme un modèle d'Erdős–Rényi de paramètre $1/2$ où la présence d'une clique de taille k est imposée. Une question naturelle est de chercher un test statistique pour déterminer si un graphe a la loi $\mathcal{G}(n,1/2)$ ou $\mathcal{G}(n,1/2,k)$; ou alors de

retrouver la clique cachée (voir par exemple le travail de Alon et al. [4]).

Estimation de dimension dans un graphe géométrique aléatoire

Étant donné un graphe géométrique obtenu avec des sommets uniformément distribués sur la sphère unité de dimension d, pour la distance euclidienne et en connectant les points à distance inférieurs à $\sqrt{2}$. À partir de la seule observation du graphe, est il possible d'estimer la dimension d (voir par exemple le travail de Atamanchuk et al. [9])? Est il possible de créer un test pour différencier ce graphe d'un $\mathcal{G}(n, 1/2)$?

Clustering

Étant donné un SBM, pour quelles valeurs de p et q est il possible de retrouver (exactement ou de manière approchée) les deux communautés (voir Lee and Wilkinson [97] pour une review des nombreux résultats à ce sujet)? Remarquons que la question du clustering peut se poser sur d'autres types de données. Par exemple, pour des points issus d'un mélange de gaussiennes (Even et al. [68]). Cette remarque nous permet de faire un pont vers le dernier chapitre de cette thèse, où des questions proches de celles abordées aux Chapitres 3 et 4 sont traitées, mais lorsque les données ne viennent pas d'un graphe mais de la réalisation d'une distribution sur $\mathbb{R}^d$.

Dans les trois premiers chapitres suivant l'introduction nous étudions des problèmes de statistiques combinatoires dans des graphes récursifs. En particulier, nous nous demandons comment inférer de l'information sur le passé d'un graphe récursif aléatoire.

2.1.1 Retrouver la racine dans un graphe récursif aléatoire

Dans les graphes récursifs, certaines questions apparaissent très naturellement, en particulier concernant l'inférence d'information sur le passé du processus. Par exemple, est il possible de retrouver le premier sommet du graphe? D'estimer l'ordre d'arrivée de tous les sommets? De retrouver quel bit avait le premier sommet dans le modèle de broadcasting? Commençons par la première question, celle du *root finding*. Ici, le but est très simple, retrouver le premier sommet. Plusieurs articles étudiants ce problème ont été publiés, avec des formalisations différentes de ce que veut dire *retrouver le premier sommet*. Par exemple, Shah and Zaman [130] étudient une méthode qui renvoie un seul sommet, et prouvent que, dans certains régimes, ce sommet est bien le sommet 1 avec une probabilité positive. D'autres ont étudié des algorithmes n'ayant accès qu'à une observation locale du graphe pour détecter des sommets d'intérêt, par exemple Brautbar and

Kearns [22]. Ici, nous formalisons la tâche de root-finding différemment. Nous cherchons un algorithme qui, avec en entrée le graphe $G = (V, E)$, renvoie un ensemble de confiance, $S(\epsilon) \subset E$ contenant le sommet 1 avec probabilité au moins $1-\epsilon$. Cette question a été étudié par Bubeck et al. [33] dans le cas des URRT et d'arbres PA, Khim and Loh [86] ont étudié le cas où l'arbre est obtenu par diffusion sur un arbre régulier infini et Brandenberger et al. [21] le cas d'un arbre de Galton-Watson conditionné en taille. Une mesure de la qualité de ces algorithmes est la taille de l'ensemble de confiance, $K(\epsilon) := |S(\epsilon)|$. Remarquez qu'ici on écrit $K(\epsilon)$ et non pas $K(|V|, \epsilon)$. En effet, on souhaite trouver un ensemble de confiance dont la taille ne dépend que de ϵ et pas de $|V|$. En observant la Figure 2.3 on se rend compte que, dans un URRT, identifier le premier sommet n'est pas évident. En effet, dans ce qui apparaît comme "le centre" du graphe, on retrouve aussi bien des sommets de haut degrés que des feuilles. On retrouve aussi des sommets de haut degrés à des positions périphériques dans le graphe. Ainsi, Bubeck et al. [33, Theorem 4] prouvent que l'on ne peut pas retrouver exactement le premier sommet et que, dans un URRT, peu importe la méthode employée,

$$K(\epsilon) \geq \exp\left(\sqrt{\frac{1}{30}\log\left(\frac{1}{2\epsilon}\right)}\right),$$

alors que dans un arbre PA on ne peut pas faire mieux que

$$K(\epsilon) \geq \frac{c}{\epsilon},$$

pour une constante positive c.

Récemment Contat et al. [40] ont suggéré un algorithme analysant les degrés de paires de sommets dans l'arbre PA pour localiser le sommet 1 dans un ensemble de confiance de taille $\epsilon^{-1+o(1)}$, ce qui correspond à la meilleure performance possible. Dans le cas de L'URRT, il existe encore un gap entre la borne inf et les performances du meilleur algorithme. À ma connaissance, le meilleur algorithme est donné par Bubeck et al. [33] et utilise la centralité de rumeur. Il a été prouvé plus tard par Crane and Xu [42] que la centralité de rumeur correspond à ordonner les sommets par leur vraisemblance d'être le sommet 1. Cet algorithme permet de localiser le sommet 1 dans un ensemble de confiance de taille sous polynomial, de l'ordre de $\exp\left(a\frac{\log 1/\epsilon}{\log\log 1/\epsilon}\right)$.

Dans le Chapitre 3 nous développons une méthode pour le root-finding dans des modèles de graphe car il est important de s'intéresser à des modèles plus généraux que les arbres. En effet, beaucoup de problèmes sont mieux décrits par des graphes que par des arbres. Ainsi, les liens dans des communautés en ligne ou le world wide web

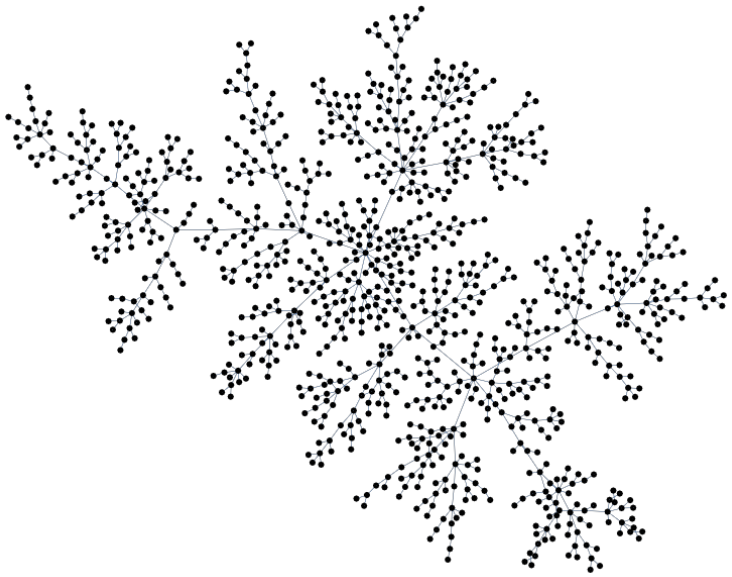

Figure 2.3: Réalisation d'un URRT de taille 1000.

ne sont évidemment pas des arbres. De plus, quand bien même le modèle théorique est un arbre, lors de l'acquisition de données, en pratique, des erreurs seront présentes et la structure d'arbre détruite. Dans le cas de l'URRT ceci est un réel problème car les méthodes analysés jusqu'alors dépendent entièrement de la structure d'arbre. En réponse à ce problème, Crane and Xu [43] ont étudié le problème du root finding lorsque du bruit est ajouté sur l'arbre. Plus précisément, ils étudient le cas où en sus de l'URRT ou de l'arbre PA des arrêtes sont ajoutés au hasard indépendamment avec la même probabilité pour chaque paire de sommets (c'est à dire que les arrêtes d'un graphe $\mathcal{G}(n,p)$ sont ajoutés aux arrêtes de l'arbre). Ils introduisent une méthode Bayésienne et prouvent qu'il est possible d'estimer la position de la racine si le nombre d'arêtes ajoutées n'est pas trop grand. Ici, nous étudions deux modèles de graphes et introduisons une méthode différente, reposant sur l'apparition de certains sous-graphes.

Les deux modèles que nous étudions sont le ℓ-dag et un cas particulier du modèle de Cooper-Frieze. Le modèle du ℓ-dag consiste en une variante de l'URRT où, à chaque étape, un nouveau sommet se connecte non pas à un ancêtre choisi au hasard mais à ℓ, choisis uniformément avec replacement (les arrêtes multiples sont alors condensés en une seule). Ce modèle a été étudié par exemple par Díaz Cort et al. [54], Tsukiji and Mahmoud [135], Tsukiji and Xhafa [136] ou Devroye and Janson [51]. Ce modèle est équivalent à considérer l'union des arrêtes de ℓ URRT indépendants. La Figure 2.4 illustre ce point de

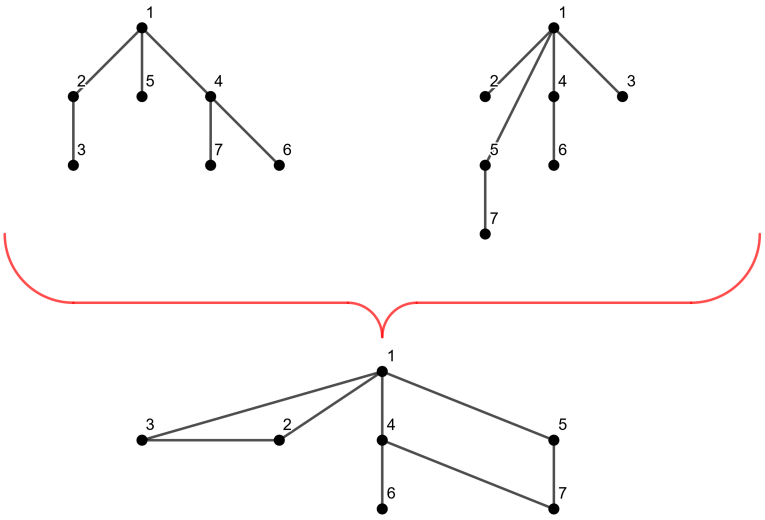

Figure 2.4: Illustration d'un 2-dag.

vue pour un 2-dag. Le second modèle est un cas particulier du modèle de Cooper-frieze, introduit par Cooper and Frieze [41]. Ici, un paramètre $\alpha \in (0,1)$ est fixé est le graphe est grandi à partir d'un sommet isolé. À chaque étape, une variable aléatoire de Bernoulli de paramètre α est réalisée, indépendamment des événements passés. Si le résultat est 0, un nouveau sommet est ajouté, qui se connecte à un sommet déjà présent choisi au hasard. Sinon, une paire de sommet est tirée uniformément au hasard et une arrête est ajoutée. Si des arrêtes multiples apparaissent elles sont condensées en une seule.

Pour choisir un ensemble de confiance $S(\epsilon)$, il est courant de choisir les sommets les plus "centraux". Plusieurs notions de centralité existent. Par exemple, Bubeck et al. [33] et Banerjee and Bhamidi [11] analysent la centralité de Jordan dans les URRT. On peut aussi penser à la centralité de rumeur, introduite par Shah and Zaman [130] et donnant le meilleur algorithme connu pour le root finding dans un URRT. Il est aussi possible d'utiliser la vraisemblance qu'un sommet soit le sommet 1, comme l'ont fait Crane and Xu [42] (comme dit plus haut, il s'avère que cela coïncide avec la centralité de rumeur). Notre approche n'utilise pas ces méthodes. En effet la centralité de Jordan, de rumeur, ou d'autres notions populaires de centralité, ne sont définies que sur des arbres. Pour analyser la vraisemblance, Bubeck et al. [33] avaient déjà remarqué que son analyse dans le cas de l'URRT était trop complexe. Ils ont alors proposé d'utiliser une expression relaxée de la vraisemblance (la centralité de rumeur), avant que Crane and Xu [42] ne réalisent que cette relaxation ne changeait pas l'ordre dans lequel les sommets étaient ordonnés. C'est

ainsi qu'a pu être étudié l'algorithme de root finding correspondant à choisir les sommets les plus probables d'être le sommet 1. Dans notre cas, la vraisemblance a une expression encore plus compliquée et nous n'avons pas été capables de la simplifier, encore moins de montrer que sa simplification ordonne les sommets de manière significative. Nous avons donc décidé d'analyser une autre notion de centralité. Nous étudions l'apparition de sous graphes, et plus précisément de double cycles. Par souci de clarté, la définition est reportée à la Section 3.2. L'ensemble de confiance est donc l'ensemble des sommets présents dans des doubles cycles de petite taille. Nous parvenons à prouver que cette méthode permet de localiser le sommet 1 dans un ensemble dont la ne dépend pas de la taille du graphe pour les deux modèles étudiés, et plus précisément, Théorème 3.4 pour le ℓ-dag assure que

$$K(\epsilon) \leq \frac{c_0}{\epsilon} \log\left(\frac{1}{\epsilon}\right)^{\frac{c_1}{\ell} \log \frac{1}{\epsilon}},$$

avec probabilité au moins $1 - \epsilon$. Le Théorème 3.5 assure que, pour le modèle de Cooper-Frieze,

$$K(\epsilon) \leq c_0 \log\left(\frac{1}{\epsilon}\right)^{c_1 \log \frac{1}{\epsilon}},$$

avec probabilité au moins $1 - \epsilon$. À contrario du modèle de l'URRT et de l'arbre PA, nous n'avons pas été capables de trouver des bornes infs sur la taille de l'ensemble de confiance. Prenons le cas du ℓ-dag par exemple. Est-il possible de faire mieux que dans le modèle de l'URRT car le ℓ-dag est la superposition de ℓ URRT indépendants? Ou bien le fait de détruire la structure d'arbre rend le problème strictement plus difficile que dans un URRT? Ces questions restent aujourd'hui ouvertes.

2.1.2 Estimer l'ordre d'arrivée dans un arbre récursif aléatoire

Pour étendre les connaissances sur l'archéologie des graphes, on peut penser à deux directions de recherche. Choisir un problème précis (par exemple le root finding) et le résoudre dans des modèles de plus en plus complexes ou de plus en plus efficacement. Ceci permet entre autre de nous rapprocher d'applications, par exemple en cherchant des algorithmes robustes, et donc applicables à des données empiriques. Une autre possibilité est d'étudier des questions plus complexes. Le prix à payer est alors de recommencer à travailler avec des modèles simples, en espérant que, plus tard, certains soient capables de résoudre ce même problème dans des modèles plus complexes. C'est ce que nous faisons dans le Chapitre 4. Une seconde question d'intérêt dans le monde de l'archéologie

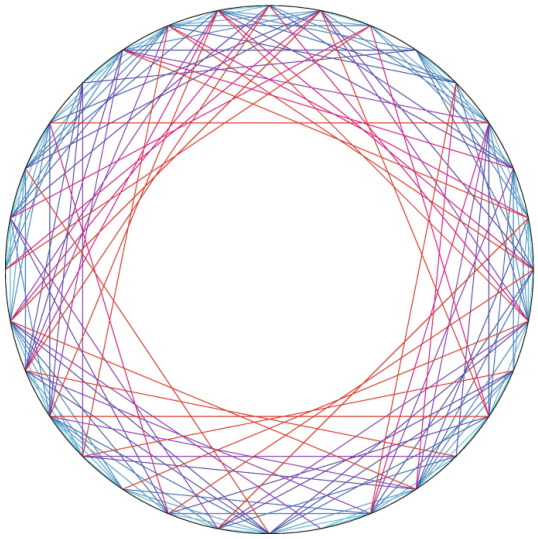

Figure 2.5: Illustration d'un graphe géométrique aléatoire sur le cercle, tiré de [7].

des graphes est d'estimer l'ordre d'arrivée de tous les sommets. En effet, retrouver le premier sommet n'apporte qu'une information limitée sur l'histoire du graphe. Mais il peut par exemple être très intéressant de retracer l'histoire de la propagation d'une fake news ou d'une rumeur en ligne. Dans ce cas, le problème n'est autre que d'estimer une variable aléatoire latente associée à chacun des sommets: leur temps d'arrivée. Nous étudions ce problème dans le cas de l'URRT et de l'arbre PA. Ce type de problème a été très largement étudié, en particulier dans le domaine de la sériation. Dans les problèmes de sériation, le but est d'estimer l'ordre ou les positions relatives de points grâce à l'observation d'affinité entre ces points. Cette similarité est supposée décroître avec la distance entre les sommets dans l'espace latent. Ce type de questions apparaît en archéologie (Robinson [123]), en bio-informatique (Recanati et al. [121]) ou encore dans des problèmes de matchmaking (Bradley and Terry [20]). Un bon exemple de données où des affinités entre paires de sommets apparaît est le cas des graphes, qui ne sont rien d'autres que la restriction a des affinités binaires (une arrête est présente ou non). Dans notre cas, nous observons une matrices d'adjacence et nous essayons d'estimer la position des sommets dans l'espace latent des entiers naturels. Gilbert [71], Giraud et al. [72], Janssen and Smith [82] ont étudié ce problème dans le cas de graphes aléatoires. En particulier, l'exemple du graphe géométrique aléatoire est emblématique du problème de sériation. L'exemple les plus simple a pour espace latent le cercle unité dans le plan muni de la distance Euclidienne. La Figure 2.5 illustre une réalisation d'un graphe géométrique aléatoire.

Parmi les problèmes de sériation, Recanati et al. [121] étudient un problème proche du nôtre, en ce sens que les données observées sont des matrices Robinsonniennes perturbées. Une matrice est dite de Robinson lorsque ses entrées sont décroissantes sur les lignes et colonnes s'éloignant de la diagonale. Dans le cas d'un URRT ou d'un arbre de PA, l'espérance de la matrice d'adjacence est une matrice de Robinson. Néanmoins, les résultats de Recanati et al. [121] ne s'appliquent pas et nous montrons empiriquement dans la Section 4.4 que la méthode qu'ils proposent a de mauvaises performances dans notre problème. À notre connaissance, le seul résultat théorique concernant l'estimation de l'ordre d'arrivée des sommets dans un graphe aléatoire vient de Crane and Xu [42]. Dans cet article, ils présentent une méthode générale pour réaliser des tâches d'archéologie des graphes qui peut être appliqué au cas de l'estimation des temps d'arrivée. Cette méthode consiste à générer un ordre de l'arbre avec la distribution attendue du modèle étudié conditionné sur la forme de l'arbre. Cette méthode permet donc de générer un ordre qui a la même loi que le vrai ordre, mais ils ne donnent aucune mesure de la qualité de cette méthode. Il ne nous ait pas apparu de manière évidente d'étudier sa qualité.

Une des raisons qui expliquent à la fois que les résultats théoriques de la sériation ne s'appliquent pas à notre problème et que ces mêmes méthodes ne sont pas bonnes empiriquement vient de la structure temporelle de notre problème. Dans tous les problèmes de sériation que nous connaissons, les sommets ont tous les même propriétés. Par exemple dans le graphe géométrique aléatoire sur le cercle, les propriétés de tous les sommets sont identiques en loi. C'est totalement faux dans notre cas, pensez par exemple au degrés du sommet 1 dans l'URRT, de l'ordre de $\log(n)$ presque sûrement, alors que le sommet n a un degrés de 1 (où n désigne la taille de l'arbre observé). Cette inhomogénéité a plusieurs conséquences. Premièrement, les méthodes de sériations introduites jusqu'à présent ne s'appliquent pas. Deuxièmement, il faut définir une nouvelle manière de mesurer la qualité d'un estimateur. En effet, si l'on considère un risque uniforme (par exemple la distance maximale entre position latente et position estimée), alors cette erreur sera intégralement porté par les feuilles arrivés à la fin de la croissance de l'arbre. C'est pourquoi nous introduisons une famille paramétrique de risques comme suit

$$R_\alpha(\widehat{\sigma}) \overset{\text{def}}{=} \mathbb{E}\left[\sum_{v\in V} \frac{|\widehat{\sigma}(v)-\sigma(v)|}{\sigma(v)^\alpha}\right],$$

pour $\alpha > 0$. Ici, $\sigma(v)$ désigne le temps d'arrivée du sommet v et $\widehat{\sigma}$ une méthode d'estimation des temps d'arrivées. Cette mesure d'erreur prend en compte l'inhomogénéité du problème en mettant plus de poids sur les sommets arrivés tôt dans le graphe. En effet, on peut

vérifier que dans un arbre de taille n, l'estimation du temps d'arrivée d'une feuille fait une erreur d'au moins l'ordre de n, alors que l'on peut estimer bien plus précisément le temps d'arrivée de sommets anciens.

Nous étudions ce risque en trois étapes, tout d'abord en montrant une borne inf sur la meilleure performance atteignable, puis en analysant les performances d'une méthode de classement des sommets et enfin en étayant nos propos par des simulations. Nos premiers résultats concernent la meilleure performance atteignable par une méthode de classement. Pour ce faire il faut se limiter à une classe restreinte d'estimateurs. Une hypothèse courante et faisant sens sur un plan pratique est de supposer que la méthode de classement est invariante par changement de labels. En quelques mots, cela signifie que l'ordre renvoyé ne dépend que de la forme de l'arbre et pas des labels assignés aux sommets. Une définition exacte est donnée dans la Section 4.1. Une fois cette hypothèse faite, il est possible d'identifier des paires de sommets de l'URRT ou de l'arbre PA qui ne peuvent pas être ordonnées mieux qu'au hasard. En utilisant ces paires échangeables, nous prouvons dans les Théorèmes 4.1 et 4.7 que le risque d'une méthode de classement invariante par changement de labels est au moins de l'ordre de $n^{2-\alpha}$, où n désigne la taille de l'arbre étudié. De rapides calculs permettent de vérifier que le risque maximal (obtenu en ordonnant les sommets dans l'ordre inverse du vraie ordre) résulte en une erreur de l'ordre de $n^{2-\alpha}$ pour $\alpha \in [0,1)$, de l'ordre de $n\log(n)$ pour $\alpha = 1$ et de l'ordre de n pour $\alpha > 1$. Ainsi, pour α plus petit que 1 le meilleur et le pire classement des sommets résultent en un risque du même ordre de grandeur, ce qui suggère que la renormalisation induite dans notre risque n'est pas intéressante dans le régime $\alpha \in [0,1)$. C'est pourquoi nous nous limitons ensuite à l'analyse du risque dans le cas $\alpha \geq 1$.

Inspirés par les travaux sur l'archéologie des graphes dans l'URRT nous avons décidé d'analyser la méthode consistant à ordonner les sommets par leur centralité de Jordan. Pour la définir, introduisons pour un arbre T et deux sommets disjoints u et v le sous arbre $(T,u)_v$, correspondant à tous les sommets w pour lesquels v se trouve sur le chemin entre w et u sur l'arbre T. La centralité de Jordan d'un sommet u est définie comme

$$\psi(u,T) = \max_{v \in V(T),\ v \sim u} |(T,u)_v|\,,$$

où $v \sim u$ indique que les sommets v et u sont voisins. Voir Figure 2.6 pour une illustration des sous arbres $(T,u)_v$ et de la centralité de Jordan. Cette méthode de classement est en particulier invariante par changement de labels. Elle a pour avantage d'avoir été largement étudié dans le problème de l'archéologie des graphes, aussi bien dans le modèle de l'URRT que de l'arbre PA (Bubeck et al. [33], Moon [109], Wagner and Durant [139]). Nous utilisons ces résultats de localisation du sommet 1 comme première étape pour étendre l'analyse

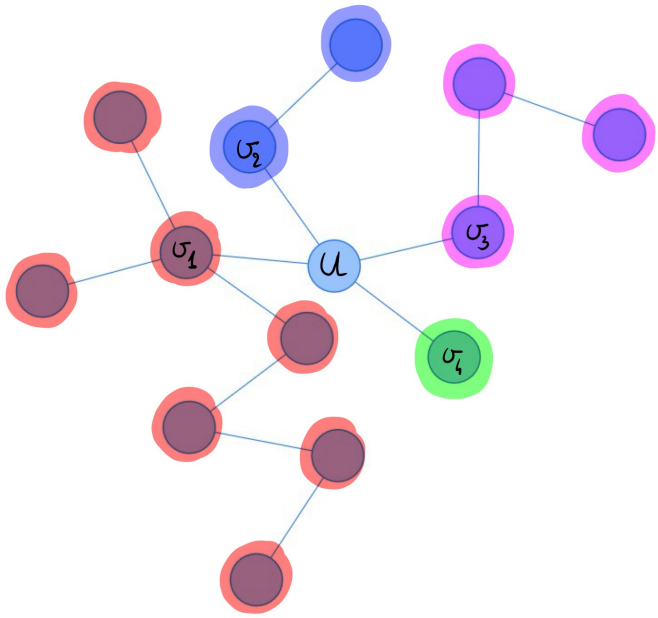

Figure 2.6: Illustration de la centralité de Jordan. Ici le sommet u a 4 voisins, v_1, v_2, v_3 et v_4. Surligné en rouge le sous arbre $(T,u)_{v_1}$, en bleu le sous arbre $(T,u)_{v_2}$, en violet le sous arbre $(T,u)_{v_3}$ et en vert le sous arbre $(T,u)_{v_4}$. Ici, $\psi(u,T)=7$.

de la centralité de Jordan au classement de tous les sommets. Nous prouvons des bornes sur le risque et en particulier nous montrons dans les Théorèmes 4.4 et 4.8 que dans le modèle de l'URRT, pour $\alpha \in [1,2)$ et dans le modèle de l'arbre PA, pour $\alpha \in [1,5/4)$, le risque de cette méthode de classement et de l'ordre de $n^{2-\alpha}$, c'est à dire de l'ordre du risque optimal. Pour α plus grand, nous expliquons pourquoi cette méthode ne peut pas être optimale. Dans le modèle URRT nous proposons une méthode utilisant la centralité de rumeur (Shah and Zaman [130]) et nous conjecturons que son risque est optimal à un facteur multiplicatif prêt, pour tout $\alpha \geq 1$.

Enfin, nous complétons notre discussion par des simulations, tout d'abord pour vérifier empiriquement nos résultats mais surtout pour comparer les performances de différentes méthodes de classement. En particulier, nous comparons les performances de notre estimateur à une méthode spectrale, étudiée par Recanati et al. [122] pour des problèmes de sériation où les affinités sont des matrices Robinsonniennes (i.e. un cadre

proche du notre). Dans le cas de l'arbre PA, nous comparons aussi les performances d'une méthode de pruning, introduite par Navlakha and Kingsford [116]. Il apparaît que le classement par la centralité de Jordan est la seule méthode que nous ayons testé dont le risque grandit au rythme optimal.

2.1.3 Broadcasting dans un graphe récursif aléatoire

Nous pouvons étendre l'horizon des problèmes dans les graphes récursifs en introduisant des états sur les sommets. Par exemple, il est possible de définir des sommets infectés par le Covid et des sommets sains. Des sommets votant Démocrate et des sommets votant Républicain. La liste est infinie, et nous la formalisons mathématiquement dans le cas le plus simple de deux états en assignant des bits, 0 ou 1, à chaque sommet. Pour décrire des problèmes pratiques, ces états (ou bits) ne sont pas attribués indépendamment du reste du graphe. En particulier, dans le modèle de broadcasting, l'idée implicite est qu'un sommet a plus de chance d'hériter de l'état (du bit) de ses parents. Cela correspond à des dépendances différentes du cas du SBM. Dans le SBM, la communauté est fixée à priori et celle ci impacte le processus d'attachement. Dans les modèles de broadcasting introduits précédemment, le bit est attribué après le processus d'attachement. La manière dont un sommet est attaché est décidé de manière indépendante des bits. Une fois connecté, le sommet se voit attribué un bit. On notera donc que l'heuristique diffère entre ces deux classes de modèle. Dans le cas du SBM, on décrit un processus ou des sommets similaires ont plus de chance de se connecter (penser par exemple aux réseaux sociaux). Dans les modèles de broadcasting, le parti pris est que les connections ont lieu indépendamment de l'état du sommet mais qu'un sommet a plus de chance d'hériter de l'état des sommets à qui il se connecte (penser par exemple aux contaminations par le Covid ou l'héritage de convictions politiques dans une famille).

Cette nouvelle dimension dans les modèles ouvrent la voie à de nombreuses nouvelles questions, en particulier de savoir si de l'information est propagée dans l'intégralité du graphe. Comme son nom l'indique, ce type de problèmes a été motivé par l'apparition de la radio et de la télévision. Dans un premier temps, des graphes déterministes ont été étudiés, par exemple par Harutyunyan and Li [75] ou Bhabak et al. [17]. Tout comme les problèmes d'archéologie des graphes, les problèmes de broadcasting sont multiples. De la même manière que le problème du *root finding* apparaît naturellement, son homologue du *rootbit finding* nous semble important à étudier. Dans ce problème, un (ou des) sommet d'origine se voit assigner un bit, qui se propage ensuite de proche en proche dans le graphe. Dans le cas d'un arbre le problème est simplifié car il existe un seul chemin entre le vertex d'origine et un sommet. Cette question a d'abord était formulée dans le cas

d'arbre généraux par Evans et al. [67]. Plus récemment, le cas d'arbre aléatoire a été étudié (Addario-Berry et al. [2], Desmarais et al. [48]). Depuis, un grand nombre de variations de ce problème ont été étudiées, se référer à Mossel [112] pour une review de problèmes de reconstruction sur des arbres. Dans un esprit similaire à celui du premier projet présenté dans cette thèse, nous avons décidé d'étudier ce problème dans le cas ou les bits ne se propagent pas sur un arbre mais sur un graphe. Nous référons le lecteur à la Section 2.1.1 pour les motivations de cette généralisation à des graphes généraux. Dans un problème similaire, Antunović et al. [8] étudient le cas de l'attachement préférentiel, où les sommets initiaux se voient attribués un état et chaque nouveau sommet a une couleur assignée en fonction de celle de ses voisins. Plus récemment, Makur et al. [107] ont étudié un autre problème similaire dans un modèle de dag différent du ℓ-dag, dont les paramètres principaux sont le degrés entrant des sommets et le nombre de sommets à distance k du sommet 1. Ils supposent aussi connaître la position du sommet 1. Deux processus de propagation sont étudiés, un processus de vote par majorité bruité et un processus de décision basé sur le NAND. Ils montrent que, si le nombre de sommets à profondeur k est de l'ordre de $\Omega(\log(k))$, il existe un seuil sur la probabilité de mutation en dessous duquel il est possible d'estimer le bit du sommet 1.

Ici, nous allons nous intéresser au le modèle de broadcasting sur les ℓ-dags, et en particulier à la proportion de chaque bits. C'est pourquoi nous faisons le lien avec les urnes de Pólya. En effet, si nous nous intéressons seulement au nombre de sommets d'un bit donné, le processus de broadcasting sur un ℓ-dag est tel que la structure de graphe n'a plus d'importance. Le modèle peut ainsi être décrit comme suit. Une urne est remplie de ℓ balles, bleues (bit 0) ou rouges (bit 1). Lorsqu'une nouvelle balle est ajoutée sa couleur est décidée en piochant ℓ balles successivement avec remise. Leurs couleurs sont observées, mais avec probabilité p une balle bleue est observée comme rouge (et vice versa). Enfin, parmi ces ℓ couleurs observées celle en majorité est transmise à la nouvelle balle. Ce lien avait déjà été fait dans le cas du broadcasting dans un URRT par Addario-Berry et al. [2]. Dans ce cas, la proportion des bits zéro suit une urne de Pólya avec replacement aléatoire (la couleur ajoutée n'est pas une fonction de la couleur de la balle tirée). Ces processus sont dits à renforcement, et nous utilisons des résultats compilés par Pemantle [118] ainsi que des résultats de non convergence étudiés par Pemantle [117]. Nous utilisons autant que possibles la description du problème comme une urne de Pólya, en partie car ces propriétés ont été très largement étudiés dans la littérature (Janson [77], Knape and Neininger [91], Wei [140]). De multiples variations de ce modèle ont été étudiés, par exemple en augmentant le nombre de couleurs (Bertoin [16]), en choisissant plusieurs balles à chaque tirage (Kuba and Mahmoud [94]), en rendant le choix des couleurs ajoutées non déterministes (conditionné sur le tirage) (Janson [80], Zhang [141]). La variation la plus proche de notre modèle dont nous ayons connaissance vient de Crimaldi et al. [44], avec

un modèle à tirage multiple et replacement aléatoire linéaire. Notre modèle a lui aussi un tirage multiple mais une loi de replacement aléatoire non linéaire.

Dans le cas où $\ell = 1$, le modèle est grandement simplifié par la disparition du vote par majorité et le fait que le broadcasting a alors lieu dans un arbre. Dans ce cas, Addario-Berry et al. [2] proposent deux méthodes pour le rootbit finding. Une première méthode consiste à estimer la position du sommet 1, puis à utiliser son état comme estimation de l'état du sommet 1. Cette méthode utilise donc des résultats d'archéologie des graphes, et en partie des résultats tenant de la localisation du sommet 1. Dans le cas du ℓ-dag ces résultats nous sont encore inconnus. Nous ne connaissons pas de méthode naturelle pour définir un centre du graphe, encore moins contrôler la distance entre ce centre et le sommet 1. Nous pouvons estimer un ensemble de confiance pour le sommet 1 mais nous ne savons pas contrôler la distance typique de ces sommets au sommet 1. Néanmoins, du fait que la taille de cet ensemble soit indépendante de la taille du graphe, nous pouvons retrouver exactement le sommet 1 avec probabilité positive (en choisissant un sommet au hasard dans cet ensemble). Ce n'est pas l'approche que nous avons décidé de suivre mais nous la discutons brièvement. La seconde méthode proposée par Addario-Berry et al. [2] consiste à estimer le bit du sommet 1 par le bit en majorité dans l'arbre. Cette méthode a l'avantage d'être facilement applicable dans notre modèle. Il est aussi connu que, dans le cas $\ell = 1$, cette méthode est optimale pour de petites probabilité de mutation(voir Addario-Berry et al. [2]).

Nous définissons le bit en majorité au temps n par b_n^{maj}, qui est décidé au hasard si les deux bits sont présents en nombre égaux. Nous cherchons à déterminer pour quels régimes de (ℓ, p) cette méthode permet de retrouver quel bit était présent en majorité à l'initialisation du graphe, c'est à dire, nous cherchons pour quelles valeurs du couple (ℓ, p) la probabilité d'erreur $R^{maj}(n,p) = \mathbb{P}\left\{b_\ell^{maj} \neq b_n^{maj}\right\} < 1/2$. Remarquons que cette grandeur dépend des conditions initiales, mais par souci de clarté nous ne le prenons pas en compte dans les notations. Dans le cas $\ell = 1$, Addario-Berry et al. [2] prouvent que

(i) Il existe une constant $c > 0$ telle que

$$\limsup_{n\to\infty} R^{maj}(n,p) \leq cp\ .$$

(ii) En notant R_n la proportion de bits 0 au temps n,

$$\lim_{n\to\infty} R_n = \frac{1}{2} \text{ presque sûrement}\ .$$

(iii) Pour $p \in [0, 1/4)$

$$\limsup_{n\to\infty} R^{maj}(n,p) < \frac{1}{2}\ .$$

(iv) Pour $p \in [1/4, 1/2]$

$$\limsup_{n\to\infty} R^{maj}(n,p) = \frac{1}{2}\,.$$

Notre travail étend ces résultats. Nous introduisons pour $\ell \geq 1$ impair

$$\alpha_\ell = \frac{1}{2^{\ell-2}} \sum_{i>\ell/2}^{\ell} \binom{\ell}{i} (i - \ell/2)\,.$$

Ainsi $\alpha_1 = 1$, $\alpha_3 = 3/2$ et pour ℓ tendant vers l'infini

$$\alpha_\ell \sim \sqrt{\frac{2\ell}{\pi}}\,.$$

Nous prouvons dans le Théorème 5.1 que

(i) Si $p < \frac{1}{2} - \frac{1}{2\alpha_\ell}$, alors il existe $\beta \in (0, 1/2)$ (dont la valeur ne dépend que de ℓ et pas des conditions initiales) telle que

$$\mathbb{P}\{R_n \to \beta\} + \mathbb{P}\{R_n \to 1-\beta\} = 1 \quad \text{et} \quad \mathbb{P}\{R_n \to \beta\} < \mathbb{P}\{R_n \to 1-\beta\}\,.$$

En particulier, indépendamment de R_ℓ,

$$\limsup_{n\to\infty} R^{maj}(n,p) < \frac{1}{2}\,.$$

(ii) Si $\frac{1}{2} - \frac{1}{2\alpha_\ell} \leq p < \frac{1}{2} - \frac{1}{4\alpha_\ell}$, alors $R_n \to 1/2$ presque sûrement et

$$\limsup_{n\to\infty} R^{maj}(n,p) < \frac{1}{2}\,.$$

(iii) Si $\frac{1}{2} - \frac{1}{4\alpha_\ell} \leq p \leq \frac{1}{2}$, alors $R_n \to 1/2$ presque sûrement et

$$\lim_{n\to\infty} R^{maj}(n,p) = \frac{1}{2}\,.$$

Remarquons que pour $\ell = 1$ le régime de faible probabilité de mutation n'existe pas, ce qui est en accord avec les résultats de Addario-Berry et al. [2]. Pour $\ell \geq 3$ trois régimes existent. Pour de petites probabilités de mutation la proportion converge vers une valeur différente de $1/2$. Par contre, dans les régimes de probabilité de mutation intermédiaire et élevé la proportion tend vers $1/2$. Cependant, tant que p ne s'approche pas trop de $1/2$, même si la proportion de bits tend vers $1/2$, le bit en majorité est positivement corrélé avec le bit en majorité à la création du graphe. Remarquons aussi que, comme dans le cas de l'attachement préférentiel étudié par Antunović et al. [8], il existe des régimes dans lesquels aucun bit ne disparaît (i.e. leur proportion ne tend pas vers 0).

2.2 Introduction à l'arbre d'amitié aléatoire

La suite de cette thèse s'intéresse à d'autres thèmes que l'archéologie des graphes. Dans un premier temps nous restons dans le monde des graphes aléatoires et de la statistique combinatoire. Tous les modèles introduits précédemment ont été étudiés en détail. Lorsqu'un nouveau modèle est introduit, par exemple pour rendre compte d'un processus d'attachement différent, il est intéressant d'en étudier les propriétés les plus basiques (diamètre, degrés maximal, nombre de feuilles, etc). Cela permet de mieux comprendre ce processus et de le comparer à des modèles de référence. Ce modèle est il une simple variante de l'URRT ou a-t-il une dynamique complètement différente? Observe-t-on un phénomène de rich-gets-richer comme dans un arbre PA ou l'attachement est plus équitable? Toutes ces questions ont été longuement étudiées pour des modèles plus classiques tels que l'URRT, le ℓ-dag, l'arbre PA, l'Erdős–Rényi, etc. Dans le but de mieux décrire certains phénomènes, de nouveaux modèles sont régulièrement introduits. Nous produisons la première analyse rigoureuse de l'un de ces modèle, l'arbre d'amitié aléatoire (voir Krapivsky and Redner [93]).

Une caractéristique qui apparaît dans de nombreux modèles est la présence d'un phénomène de "rich gets richer". Autrement dit, un sommet qui est beaucoup connecté à tendance à renforcer cette dominance avec le temps. Ce comportement est présent dans les modèles d'attachement préférentiel et donc dans l'arbre PA. Dans la définition la plus simple de ce modèle, la loi d'attachement n'est pas locale. C'est à dire que, pour faire grandir l'arbre selon la loi de l'attachement préférentiel, il faut avoir accès aux degrés de chaque sommet, ou à toutes les arrêtes présentes dans le graphe. Comme on peut supposer qu'en pratique les processus d'attachement sont locaux (par exemple dans des phénomènes physiques), de nombreux modèles d'attachement locaux exhibant un phénomène de rich-gets-richer ont été introduits. Par exemple, Engländer et al. [64] ont très récemment introduit un modèle de marche aléatoire construisant un arbre. Dans leur modèle un arbre est agrandi récursivement par un "marcheur" qui se déplace au hasard sur l'arbre. À l'étape n, avec probabilité $n^{-\gamma}$, un voisin est ajouté au sommet où se trouve le marcheur. Ils montrent que ce modèle correspond à l'arbre PA et constitue donc un processus d'attachement local (à chaque étape le marcheur n'a besoin que de connaître son voisinage pour progresser) exhibant un phénomène de rich-gets-richer.

Une autre manière de créer un modèle d'attachement local où un phénomène de rich-gets-richer est à l'œuvre est d'introduire un phénomène de redirection. C'est à dire, un modèle où un nouveau sommet ne se connecte pas forcément à un sommet choisi au hasard mais possiblement à un voisin (ou sommet proche) d'un sommet choisi au hasard. Introduits par Kleinberg et al. [90] dans des arbres dirigés, le modèle initial consistait à

connecter chaque nouveau sommet à un sommet choisi au hasard avec probabilité $1-p$ ou à son ancêtre avec probabilité p. Ce modèle donne lieu à un processus d'attachement préférentiel où chaque nouveau sommet se connecte à un sommet choisi avec probabilité proportionnelle à $d-2+1/p$ (pour d le degrés dudit sommet). Plus tard, Saramäki and Kaski [125] introduirent une version non dirigée du modèle, plus tard étudiée par Evans and Saramäki [66].

Dans le cas de la redirection dans l'arbre non dirigé, l'arbre est grandi de manière récursive en choisissant un sommet uniformément au hasard, puis en y démarrant une marche aléatoire avec k pas et enfin en attachant un nouveau sommet au dernier sommet atteint par cette marche aléatoire. Contrairement aux travaux de Engländer et al. [64], une nouvelle marche aléatoire est créé à chaque étape. Dans le cas $k=1$, nous appelons ce modèle l'*arbre d'amitié aléatoire* (RFT) car nous pouvons penser à chaque nouveau sommet devenant "ami" avec un "ami" d'un sommet choisi au hasard. Cette famille de processus, allant de $k=0$ à $k=\infty$ a la propriété particulière de contenir l'URRT et l'arbre PA. En effet, si $k=0$, alors chaque nouveau sommet se connecte à un sommet choisi uniformément au hasard et le modèle est donc un URRT. Si $k\to\infty$ alors on observe un arbre PA. En effet, un arbre fini a un temps de mélange fini et la distribution stationnaire d'une marche aléatoire sur un arbre est proportionnel aux degrés. Ici, nous étudierons le cas particulier $k=1$, c'est à dire que chaque nouveau sommet est attaché comme suit: au temps $n+1$, un sommet est choisi uniformément au hasard, il est appelé V_n. Parmi les voisins de V_n un sommet est choisi uniformément au hasard, il est appelé W_n. Le sommet $n+1$ se connecte à W_n. La Figure 2.7 illustre ce processus.

Bien que l'arbre d'amitié aléatoire fasse parti d'une famille de modèles contenant l'URRT et l'arbre PA, certaines de leurs propriétés sont drastiquement différentes. Parmi ces propriétés, la séquence des degrés ou la modularité ont été conjecturés dans un travail empirique par Krapivsky and Redner [93]. Le seul résultat théorique concernant ce modèle dont j'ai connaissance concerne le nombre de feuilles. Cannings and Jordan [35] ont prouvé que $n-o(n)$ sommets étaient des feuilles presque sûrement. Nous étendons grandement la connaissance théorique du RFT. Concernant les sommets de petit degrés, nous montrons qu'au moins $n-n^{0.9}$ sommets sont des feuilles, alors que dans l'URRT ou l'arbre PA il y a seulement une fraction des sommets qui sont des feuilles. Alors que pour un k fixé il y a dans l'URRT de l'ordre de $n/2^{k-1}$ sommets de degrés au moins k (Janson [78]), ce nombre se situe entre $n^{0.1}$ et $^{0.9}$ dans le cas du RFT (voir Théorème 6.11). De plus, nous montrons dans la Proposition 6.13 que la plupart des feuilles resteront des feuilles pour toujours, alors que dans l'URRT ou l'arbre PA le degrés de chaque sommet tend vers l'infini presque sûrement. Concernant les sommets de haut degrés, nous montrons dans le Théorème 6.2 que des "hubs" de degrés linéaires apparaissent. Ces hubs sont bien vis-

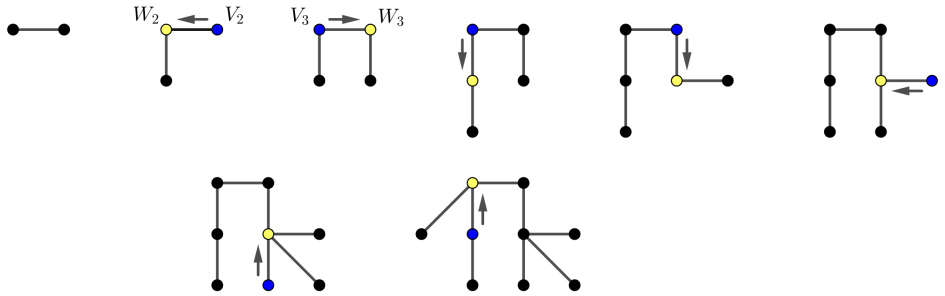

Figure 2.7: Illustration du processus d'attachement dans un arbre d'amitié aléatoire. En bleu le sommet V_n, choisi uniformément au hasard parmi tous les sommets. En jaune, le sommet W_n choisi uniformément au hasard parmi les voisins de V_n. Le sommet $n+1$ se connect au sommet W_n.

ibles dans la Figure 2.8. À contrario, le degrés maximal dans un URRT est logarithmique (Devroye and Lu [52]) et dans un arbre PA de l'ordre de $\sqrt{n}$ (Van Der Hofstad [138, Theorem 1.17]). Nous montrons même un résultat plus fort, pour chaque arrête, au moins un des sommets la constituant deviendra un hub de degrés linéaire presque sûrement. Ce phénomène est inédit et présent dans aucun autre modèle dont nous ayons connaissance. Cependant, tout n'est pas différent de l'URRT ou du PA. Dans le Théorème 6.6 nous montrons que le diamètre est logarithmique presque sûrement, comme dans un URRT (Addario-Berry and Ford [1, Corollary 1.3]), alors qu'il est au plus logarithmique dans un arbre PA (Dommers et al. [55]). De plus, comme nous le montrons dans les Théorèmes 6.7 et 6.8, à la fois dans un RFT et dans un URRT le sommet le plus éloigné d'une feuille est à distance $\Theta(\log(n)/\log\log(n))$ de celle ci.

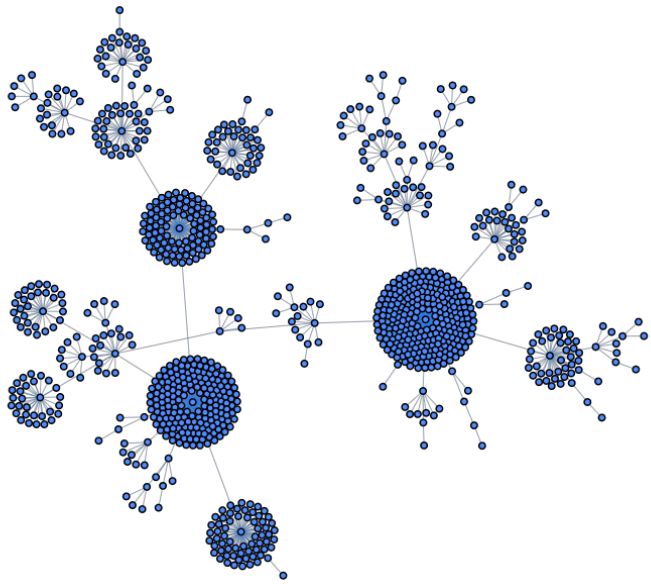

Figure 2.8: Illustration d'un RFT de taille 1000.

2.3 Introduction à la profondeur de Tukey aléatoire

Comme cela a été rapidement mentionné en citant le problème de clustering, il est possible d'étudier des questions similaires dans des settings très différents. Dans la dernière partie de cette thèse nous étudions le problème d'ordonner des données en grande dimension. La problème est donc similaire à celui étudié au Chapitre 4, et nous nous intéressons ici aussi à une notion de centralité pour ordonner des points dans $\mathbb{R}^d$. Cependant, le fait que les données étudiées (un graphe ou un nuage de points en grande dimension) soient complètement différentes nous fait découvrir des outils statistiques différents. Ici, nous entrons plus avant dans le monde de l'analyse de données et du machine learning. De par leurs innombrables applications industrielles ce sont des domaines où la recherche est rapide et fournie. Un des défis majeurs rencontré par les praticiens aujourd'hui vient de la dimension croissante des objets étudiés. Beaucoup d'applications reposent sur des données en grande dimension et la croissance des bases de donnée ne fait qu'intensifier cette problématique. Les progrès fulgurants des outils basés sur l'intelligence artificielle et leur médiatisation sont aussi un facteur qui pousse chaque statisticien à se confronter à des problèmes liés à la grande dimension. Le problème dont il est ici question consiste à ordonner les points d'un jeu de donnée par leur centralité. Il peut être utile de définir

un ordre partiel pour pouvoir visualiser les données, des plus centrales jusqu'aux valeurs extrêmes. Plus qu'un outil de visualisation, un ordre de centralité est nécessaire pour entraîner un prédicteur conforme (autrement dit, un prédicteur qui renvoie non pas un point unique mais un intervalle de confiance). En dimension 1, une manière évidente d'ordonner des points dans un ensemble est par leurs quantiles empiriques. Malheureusement, dès la dimension 2 la notion de quantile (et même de médiane) n'est plus définie aussi facilement. Tukey [137] introduit une notion de profondeur, appelée profondeur de Tukey (ou de demi espace), qui est un outil populaire pour visualiser la centralité d'un point dans un jeu de données. Bien d'autres mesures de profondeur ont été introduites, comme la profondeur de simplex (Liu [99, 100]), le profondeur de projection (Liu [101], Zuo and Serfling [143]), la profondeur de zonoïde (Koshevoy and Mosler [92, 92]) ou bien une notion de données aberrantes, (Donoho [56], Stahel [134]). Chacune de ses notions a des propriétés différentes de stabilité, d'invariance ou de computabilité les rendant adaptées à différentes applications. La profondeur de Tukey est définie comme suit. Étant donné un point x dans $\mathbb{R}^d$, on définit pour une direction $u \in S^{d-1}$ (où S^{d-1} est la sphère Euclidienne de $\mathbb{R}^d$) le demi espace fermé

$$H(x,u) = \{y \in \mathbb{R}^d \ : \ \langle y,u\rangle \leq \langle x,u\rangle\} \, .$$

Ensuite, pour un ensemble de points $\{x_1, \cdots, x_n\}$ nous pouvons définir la profondeur dans la direction u par

$$r_n(x,u) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x_i \in H(x,u)} \, ,$$

ce qui correspond au quantile empirique de $\langle x,u\rangle$ dans le jeu de donnée projeté sur la direction u. On peut ensuite définir la profondeur de Tukey de x dans le jeu de données $\{x_1, \cdots, x_n\}$ par

$$d_n(x) = \inf_{u \in S^{d-1}} r_n(x,u) \, .$$

Remarquons que le principe est similaire à celui de la centralité de Jordan. Nous définissons un score, ici la profondeur de Tukey, et utilisons ce score pour ordonner des points dans $\mathbb{R}^d$. La profondeur de Tukey a les attributs attendus d'une mesure de profondeur: elle est invariante par transformation affine, elle tend vers 0 à l'infini et décroît sur des rayons partant du point le plus profond. De plus, elle est robuste sous conditions de symétries (Donoho and Gasko [57]), et lorsque les données sont tirées indépendamment de la même distribution ses courbes de niveau convergent rapidement (Brunel [31]). Néanmoins, un des problèmes rencontrés avec cette mesure de profondeur est la difficulté à la calculer.

Ce problème est mis en exergue par la nécessité de traiter des données de dimension de plus en plus grandes. Ainsi, calculer même une valeur approchée est un problème NP hard, comme le montrent Amaldi and Kann [6], Bremner et al. [24], Johnson and Preparata [85]. De plus, le calcul de la profondeur maximale se fait en temps $\mathcal{O}\left(n^{d-1}\right)$ (Chan [37]). Malgré ces difficultés, la profondeur de Tukey n'a pas été écarté par les praticiens car les autres mesures de profondeur font face aux mêmes défis. C'est pourquoi un travail important est effectué pour développer et analyser des algorithmes d'approximation. Leur importance est soulignée par Nagy et al. [115]. Parmi les solutions proposées, Shao et al. [131] étudient un algorithme basé sur des mcmc pour calculer la profondeur de projection en grande dimension. Pour la profondeur de Tukey, Zuo [142] a proposé une nouvelle approche pour en calculer une approximation, mais sans donner de garanties sur sa précision. Enfin, Chen et al. [38] ont étudié la qualité de diverses approximations de la profondeur de Tukey.

Cuesta-Albertos and Nieto-Reyes [45] ont introduit un algorithme d'approximation naturel, qui consiste à ne pas prendre le minimum sur l'ensemble des directions possibles dans $\mathbb{R}^d$ mais de se limiter à k directions choisies au hasard. Ainsi, pour $U_1, \cdots, U_k$ des variables aléatoires indépendantes uniformes sur S^{d-1}, nous pouvons définir la profondeur de Tukey aléatoire du point x au sein du jeu de données $\{x_1, \cdots, x_n\}$ par

$$D_{n,k}(x) = \min_{i=1,\cdots,k} r_n(x, U_i) \, .$$

Il est facile de voir que pour chaque x, lorsque k tend vers l'infini, la profondeur de Tukey aléatoire converge avec probabilité 1 vers la profondeur de Tukey. Il faut toutefois en savoir plus pour pouvoir utiliser cette approximation dans la pratique. En particulier, on peut se demander quelle taille doit avoir k pour que $D_{n,k}(x)$ soit une bonne approximation de $d_n(x)$. Plus précisément, pour $\epsilon \in (0, 1/2)$, $\delta > 0$, à quel point k doit être grand pour assurer que $|D_{n,k} - d_n(x)| \leq \epsilon$ avec probabilité au moins $1 - \delta$? Pour étudier cette question nous devons limiter notre étude à des jeux de données "raisonnables". En effet, déjà en dimension 2, il est possible de construire un exemple de jeu de données ou k doit être arbitrairement grand pour atteindre un précision donnée. Ainsi, pour n paire, définissons les points $x_i = \left(i/n, a(i/n)^2\right)$ où $a > 0$ est un paramètre. Comme il est situé à la frontière de l'enveloppe convexe de $\{x_1, \cdots, x_n\}$, le point $x_{n/2}$ a comme profondeur 0. Cependant, pour "voir" cette profondeur de 0 il faut évaluer la profondeur dans une direction u telle que

$$\langle x_{n/2-1}, u \rangle > \langle x_{n/2}, u \rangle \, ,$$

et

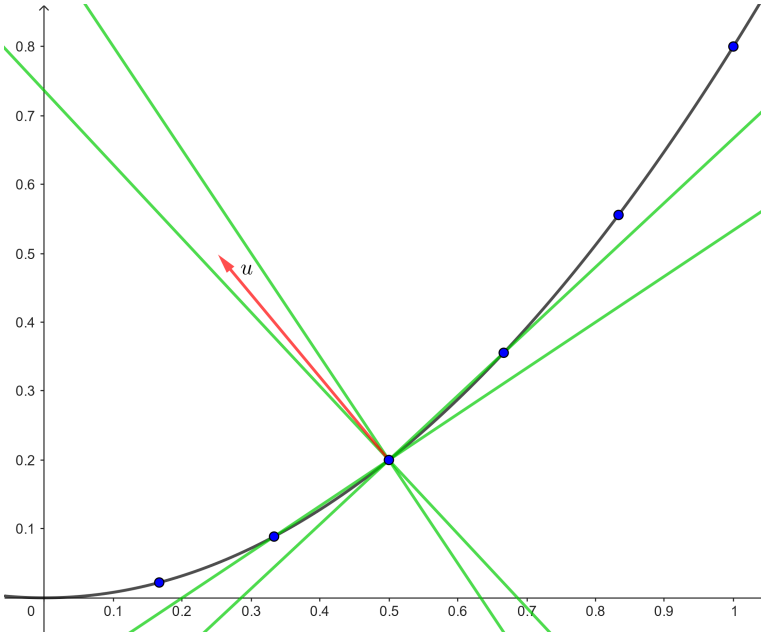

Figure 2.9: Illustration du jeu de données $x_i = \left(i/6, 0.8(i/6^2\right)$ pour $i \in [6]$. En vert les droites entre lesquels le vecteur u doit se trouver pour détecter la profondeur de 0 du point x_3.

$$\langle x_{n/2+1}, u\rangle > \langle x_{n/2}, u\rangle \,.$$

Une illustration, présentée dans la Figure 2.9 montre qu'en choisissant a arbitrairement petit, la zone où u doit se trouver pour détecter la profondeur de $x_{n/2}$ devient arbitrairement petite, et par conséquent k doit être choisi arbitrairement grand pour espérer estimer la profondeur de Tukey.

C'est pourquoi nous devons supposer une certaine régularité dans le jeu de données. Nous supposons aussi que les points sont indépendants et identiquement distribués, une hypothèse naturelle en machine learning. Plus précisément, nous supposons que les points $\{x_1, \cdots, x_n\}$ sont des réalisations indépendantes d'une loi log-concave isotrope. Remarquons que l'hypothèse isotrope n'est pas nécessaire, car il est possible d'estimer précisément la covariance d'une distribution isotrope et donc de ramener un jeu de données issu d'une loi log-concave en position (quasi) isotrope. Pour des données issues d'une variable aléatoire, il est possible d'introduire une version de la profondeur de Tukey ne reposant que sur la distribution. En effet, plutôt que de compter le nombre de points dans un demi espace,

on peut simplement mesurer ce demi espace pour la distribution en question. On définit ainsi une nouvelle version de d_n et $D_{n,k}$, respectivement $\overline{d}$ et $\overline{D}_k$. Une définition précise est donnée dans le Section 7.1.

Nous ne sommes pas les premiers à nous intéresser à la qualité de cet algorithme d'approximation. Ainsi, Cuesta-Albertos and Nieto-Reyes [45] ont produits des résultats expérimentaux suggérant que $\overline{D_k}$ est une bonne approximation de $\overline{d}$. D'un point de vue théorique, Nagy et al. [115] ont étudié dans quelles conditions $\sup_{x\in\mathbb{R}^d} |\overline{D}_k(x) - \overline{d}(x)| \to 0$ lorsque $k \to \infty$. Ils ont aussi donné des bornes sur la vitesse de convergence. À l'opposé de cette approche uniforme, nous étudions la qualité de l'approximation de $\overline{d}(x)$ par $\overline{D}_k(x)$ pour un point x fixé. Nous montrons en particulier que la qualité de cette approximation dépend fortement de la profondeur de x. Nos résultats sont ainsi présentés en trois parties. Premièrement, la plupart des points du jeu de données ont une faible profondeur et celle ci est facile à estimer. Deuxièmement, estimer la profondeur de Tukey de points de profondeur intermédiaire est dur, cela nécessite que k grandisse exponentiellement avec la dimension. Enfin, si il existe un point de profondeur $1/2$, celui ci est facile à localiser. Plus précisément, si la mesure μ est log-concave isotrope sur $\mathbb{R}^d$, nous montrons dans le Corollaire 7.2 qu'il existe des constantes universelles $c, \kappa, C > 0$ telles que pour $\epsilon, \delta, \gamma > 0$, si

$$k = \left\lceil \max\left(C, \frac{4}{\epsilon} \log \frac{3}{\gamma}, \frac{2}{c} \log \frac{4}{\delta} \right) \right\rceil ,$$

et si la dimension d est plus grande que

$$d \geq \max\left(\left(\frac{3(k+1)}{\gamma} \right)^{1/\kappa}, \frac{64 \log(1/\epsilon) k}{\pi} \log \frac{3k}{\gamma}, \left(\frac{1}{c} \log \frac{6k}{\delta} \right)^2, \left(\frac{2}{\epsilon} \right)^{\kappa} \right) ,$$

alors, avec probabilité au moins $1 - \delta$,

$$\mu\left(\left\{ x \in \mathbb{R}^d : \overline{D}_k(x) > \epsilon \right\} \right) < \gamma .$$

Comme $\overline{D}_k(x) \geq \overline{d}(x)$, ce corollaire implique, qu'au sens de la mesure μ, la plupart des points ont une faible profondeur et que celle ci est facilement estimée. On comprend intuitivement pourquoi, en grande dimension, la plupart des points ont une faible profondeur. En effet, une distribution isotrope log-concave sur $\mathbb{R}^d$ "ressemble" à la distribution uniforme sur la sphère de rayon $\sqrt{d}$ dans $\mathbb{R}^d$. On dit que la plupart de ces points sont faciles à estimer car, pour une précision donnée, le paramètre k ne dépend pas de la dimension! Même si ce Corollaire prouve que la plupart des points ont une faible profondeur,

et que celle ci est facile à estimer, il ne prouve pas que pour tous les points de faible profondeur la profondeur est facile à estimer. Ainsi, considérons la distribution uniforme sur $[-(3/2)^{1/3},(3/2)^{1/3}]^d$, qui est log-concave isotrope. Le point $x = (((3/2)^{1/3},0,\cdots,0)$ a pour profondeur 0, et pourtant on peut montrer que $\overline{D}_k(x) \geq 1/4$ avec grande probabilité si k n'est pas exponentiellement grand en d.

Nous montrons ensuite dans le Corollaire 7.3 que les profondeurs intermédiaires sont dures à estimer. Encore une fois, pour μ log-concave isotrope sur $\mathbb{R}^d$, nous montrons que pour $\delta \in (0,1)$, $\gamma \in (0,1/2)$, il existe une constante positive $c = c(\gamma)$ telle que si $x \in \mathbb{R}^d$ est tel que $\overline{d}(x) = \gamma$, alors pour $\epsilon < c$, si $k \leq \delta e^{d\epsilon^2 \log^2(1/\epsilon)/c}$, alors avec probabilité au moins $1-\delta$

$$|\overline{D}_k(x) - \overline{d}(x)| \geq \epsilon \ .$$

Enfin, nous montrons que le point de profondeur $1/2$, si il existe, est facile à localiser. Notons qu'il n'est pas garanti qu'un tel point existe. Le point le plus profond peut avoir une profondeur plus petite que $1/2$, mais pas arbitrairement petite. Nagy et al. [114, Theorem 3] montrent que $1/e \leq \sup_{x\in\mathbb{R}^d} \overline{d}(x) \leq 1/2$. Dans le cas où $\sup_{x\in\mathbb{R}^d} \overline{d}(x) = 1/2$, la distribution est dite demi espace symétrique (Nagy et al. [114], Zuo and Serfling [144]). Il est facile de voir que dans ce cas il existe un seul point de profondeur $1/2$, il est appelé médiane de Tukey. Une distribution symétrique est demi espace symétrique mais la contraposée est fausse. Cependant, dans le cas d'une distribution uniforme sur un convexe, alors un espace est demi espace symétrique si et seulement si il est symétrique (Funk [70], Schneider [128]). Le Corollaire 7.4 prouve que si μ est log-concave, isotrope et demi espace symétrique, alors si un point est tel que $\overline{D}_k(x) \approx 1/2$ alors x est proche de la médiane de Tukey. Plus précisément, considérons $X_1,\cdots,X_n$ des variables aléatoires indépendantes distribués selon μ, m la médiane de Tukey et $m_{n,k}$ la médiane de Tukey empirique (i.e. $D_{n,k}(m_{n,k}) = \max_{x\in\mathbb{R}^d} D_{n,k}(x)$). Alors, il existe des constantes universelles positives c et C tels que pour $\delta \in (0,1)$, $\gamma \in (0,c)$, si $n \geq Cd/\gamma^2$ et

$$k \geq c(d\log(d) + \log(1/\delta)) \ ,$$

alors

$$\left\| m_{n,k} - m \right\| \leq C\gamma\sqrt{d} \ ,$$

avec probabilité au moins $1-\delta$. Cela signifie qu'en fixant γ de l'ordre de $1/\sqrt{d}$ on peut localiser la médiane de Tukey à distance constante avec seulement $k \approx d\log(d)$. Remarquons que du fait qu'en grande dimension la masse d'une distribution log-concave isotrope

se concentre sur la sphère de diamètre $\sqrt{d}$, localiser la médiane de Tukey à distance constante n'est pas une estimation triviale.

Chapter 3

Archaeology of random recursive dags and Cooper-Frieze random networks

Contents

Abstract

We study the problem of finding the root vertex in large growing networks. We prove that it is possible to construct confidence sets of size independent of the number of vertices in the network that contain the root vertex with high probability in various models of random networks. The models include uniform random recursive dags and uniform Cooper-Frieze random graphs.

This Chapter is based on a joint work with Francisco Calvillo and Gábor Lugosi, published in Combinatorics, Probability and Computing (Briend, Calvillo, and Lugosi [26]).

3.1 Introduction

In order to develop a sound statistical theory for *network archaeology*, one usually models the growing network by simple stochastic growth dynamics. Perhaps the most prominent such growth model is the preferential attachment model, advocated by Albert and Barabási [3]. In these models, vertices of the network arrive one by one and a new vertex attaches to one or more existing vertices by an edge according to some simple probabilistic rule.

Arguably the simplest problem of network archaeology is that of *root finding*, when one aims at estimating the first vertex of a random network, based on observing the (unlabeled) network at a much later point of time.

The existing literature on the theory of network archaeology mostly focuses on the simplest possible kind of networks, namely trees, see Haigh [74], Shah and Zaman [129, 130], Bubeck, Mossel, and Rácz [32], Curien, Duquesne, Kortchemski, and Manolescu [47], Khim and Loh [86], Jog and Loh [83, 84], Bubeck, Eldan, Mossel, and Rácz [34], Bubeck, Devroye, and Lugosi [33], Lugosi and Pereira [103], Devroye and Reddad [53], Banerjee and Bhamidi [11], Crane and Xu [42], Addario-Berry, Devroye, Lugosi, and Velona [2], Brandenberger, Devroye, and Goh [21].

In various models of growing random trees, it is quite well understood up to what extent one may identify the origin of the tree (i.e., the root) by observing a large unlabeled tree. These models include uniform and linear preferential attachment trees and diffusion over regular trees. Remarkably, in all these models, the size of the tree does not play a role. In other words, there exist root-finding algorithms that are able to select a small number of nodes – independently of the size of the tree – such that the root vertex is among them with high probability.

Here we address the more difficult – and more realistic – problem of finding the origin of growing networks when the network is not necessarily a tree. The added difficulty stems from the fact that the centrality measures that proved to be successful in root estimation in trees crucially rely on properties of trees.

A notable exception in the literature is the recent paper of Crane and Xu [43] in which the authors allow for a "noisy" observation of the tree. In their model, the union of the tree of interest and an (homogeneous) Erdős-Rényi random graph is observed, and

the goal is to estimate the root of the tree.

In this chapter we study root estimation in two more complex network models. Both of these models may be viewed as natural extensions of the random recursive trees that were in the focus of most of the previous study of network archaeology. Recall that a uniform random recursive tree on the vertex set $[n]$ is defined recursively, such that each vertex $i \in \{2,3,\dots,n\}$ is attached by an edge to a vertex chosen uniformly at random among the vertices $\{1,\dots,i-1\}$, see, e.g., Drmota [58].

In particular, we study the problem of root finding in (1) uniform random recursive dags; and (2) uniform Cooper-Frieze random graphs.

Uniform random recursive dags

For a positive integer ℓ, a uniform random ℓ-dag is simply the union of ℓ independent uniform random recursive trees on the same vertex set $[n]$. Equivalently, a uniform random ℓ-dag may be generated recursively; each vertex $i \in \{2,3,\dots,n\}$ is attached by an edge to ℓ vertices chosen uniformly at random (with replacement) among the vertices $\{1,\dots,i-1\}$. Multiple edges are collapsed so that the resulting graph is simple. Random recursive dags have been studied by Broutin and Fawzi [30], Devroye and Janson [51], Díaz Cort, Serna Iglesias, Spirakis, Torán Romero, and Tsukiji [54], Mahmoud [106], Tsukiji and Mahmoud [135], Tsukiji and Xhafa [136], among others.

Definition 3.1. *Let $n,\ell \in \mathbb{N}$. For $i \in [\ell]$, let $G_i \in (V,E_i)$ be independent uniform random recursive trees on the vertex set $V = [n]$. A* uniform random recursive ℓ-dag *on n vertices is $G = (V, E_1 \cup \cdots \cup E_\ell)$.*

Uniform Cooper-Frieze random graphs

The other network model studied here was introduced by Cooper and Frieze [41] in an attempt to mathematically describe large web graphs, see also Frieze and Karoński [69]. In the Cooper-Frieze network model both vertices and edges are added sequentially to the network based on uniform or preferential attachment mechanisms. The model is quite general but here we focus on the simplest version when both vertices and edges are added by *uniform* attachment.

More precisely, the uniform Cooper-Frieze growth model is defined as follows. The procedure has a parameter $\alpha \in (0,1)$. The process is initialized by a graph containing a single vertex and no edges. At each time instance $t = 1,2,\dots$, an independent Bernoulli(α) random variable Z_t is drawn. If $Z_t = 0$, a new vertex is added to the vertex set along

with an edge that connects this vertex to one of the existing vertices, chosen uniformly at random. If $Z_t = 1$, then a new edge is added by choosing two existing vertices uniformly at random and connecting them. Note that the resulting graph may have multiple edges. In such cases, we may convert the graph into a simple graph by keeping only one of each multiplied edge.

If one runs the process for T steps for a large value of T, the graph has $n \sim 1 +$ Binomial$(T-1, 1-\alpha) \approx (1-\alpha)T$ vertices and $T \approx n/(1-\alpha)$ edges. If one removes the edges added at the times when $Z_t = 1$, the remaining graph is a tree, distributed as a uniform random recursive tree on n vertices. The remaining $T - n - 1$ edges are present approximately independently of each other and there is an edge between vertices i and j (where $1 \le i < j \le n$) if

$$\sum_{t=1}^{T} \sum_{\ell=j}^{n} \mathbb{1}_{t \in \{t_\ell + 1, t_{\ell+1} - 1\}} \mathbb{1}_{\text{the pair } (i,j) \text{ is selected at time } t} \ge 1 \ ,$$

where $1 = t_1 < t_2 < \cdots < t_n \le T$ are the times when new vertices are added, that is, when $Z_t = 1$. Since the probability that edge (i, j) is selected at time $t \in \{t_\ell + 1, t_{\ell+1} - 1\}$ is $1/\binom{\ell}{2}$, for large values of T, the probability that edge (i, j) is present in the graph after T steps is concentrated around

$$\frac{c_\alpha}{\max(i,j) - 1} \quad \text{where} \quad c_\alpha \stackrel{\text{def}}{=} \frac{2}{1-\alpha} \ ,$$

whenever $\max(i, j) - 1 \ge c_\alpha$. Hence, the uniform Cooper-Frieze model is essentially equivalent to the following random graph model. In order to avoid some tedious and uninteresting technicalities, we work with this modified model instead of the original recursive definition.

Definition 3.2. *Let $n \in \mathbb{N}$ and let c be a positive constant. Let $G_1 = (V, E_1)$ be a uniform random recursive tree on the vertex set $V = [n]$. Let $G_2 = (V, E_2)$ be a random graph on the same vertex set, independent of G_1, such that edges of G_2 are present independently of each other, such that for all $i \ne j$,*

$$\mathbb{P}\{(i,j) \in E_2\} = \min\left(\frac{c}{\max(i,j) - 1}, 1\right) .$$

Finally, the uniform Cooper-Frieze random graph *with parameters c and n is $G = (V, E_1 \cup E_2)$.*

Root estimation

The main result of this chapter is that finding the root is possible both in uniform random recursive dags and in uniform Cooper-Frieze random graphs. More precisely, one may find

confidence sets for the root vertex whose size does not depend on the number of vertices in the graph. To make such statements rigorous, consider the following definition.

Definition 3.3. *Let $\{G^{(n)}\}$ be a sequence of random graphs such that $G^{(n)}$ has vertex set $[n]$. We say that* root estimation *is possible if the following holds. For every $\epsilon > 0$, there exists a positive integer $K(\epsilon)$ such that, for every $n \in \mathbb{N}$, upon observing the graph $G^{(n)}$ without the vertex labels, one may find a set $S \subset [n]$ of vertices of size $|S| = K(\epsilon)$ such that*

$$\mathbb{P}\{1 \in S\} \geq 1 - \epsilon \, .$$

The set S in the above definiton is often called a confidence set for the root vertex.

As mentioned above, root estimation has mostly been studied for random recursive trees. Bubeck, Devroye, and Lugosi [33] show that root estimation is possible in the uniform random recursive tree and linear preferential attachment trees. They show that in the case of the uniform random recursive tree, one may take $K(\epsilon) \leq \exp(c \log(1/\epsilon)/\log\log(1/\epsilon))$ for some constant c. For linear preferential attachment trees one may take $K(\epsilon) = c\epsilon^{-2-o(1)}$, as shown by Banerjee and Bhamidi [11] who also show that root estimation is possible for a wide class of preferential attachment trees. Building on the papers of Shah and Zaman [129, 130], Khim and Loh [86] show that root estimation is possible in random trees obtained by diffusion on an infinite regular tree, and that one my take $K(\epsilon) = \exp(O(\log(1/\epsilon)/\log\log(1/\epsilon)))$. Brandenberger, Devroye, and Goh [21] study root estimation in size-conditioned Galton–Watson trees.

The sets S of constant size that establish the possibility of root estimation for various trees usually contain the set of most "central" vertices according to some notion of centrality such as *Jordan centrality* (as in [33], [11]) or *rumor centrality* introduced in [129, 130], see also [33], [86]. However, these notions are suited for trees only and when the observed network is more complex, new ideas need to be introduced. Crane and Xu [43] study a model in which the observed network consists of either a uniform attachment tree (i.e., uniform random recursive tree) or a preferential attachment tree, with random edges added (independently over all possible vertex pairs, with the same probability). They introduce a Bayesian method and prove that it is able to estimate the root as long as there are not too many edges, where the threshold value depends on the particular model. It is unclear if the method of [43] may be generalized to the random graph models studied here. Instead, we introduce an alternative root estimation method that is based on the appearance of certain subgraphs.

The main results of this chapter are summarized in the following two theorems.

Theorem 3.4. *Fix $\ell > 1$ and let $G = G^{(n)}$ be a uniform random ℓ-dag on n vertices. Root estimation is possible in G. In particular, there exist numerical constants $c_0, c_1, c_2 > 0$ such that, whenever $\epsilon \leq e^{-c_2 \ell}$, one may take*

$$K(\epsilon) \leq \frac{c_0}{\epsilon} \log(1/\epsilon)^{\frac{c_1}{\ell} \log(1/\epsilon)} .$$

Explicit values of the constants c_0, c_1, c_2 are given in the proof below. In the uniform Cooper-Frieze model we have a similar bound:

Theorem 3.5. *Let $G = G^{(n)}$ be a uniform Cooper-Frieze random graph on n vertices, with parameter c. Root estimation is possible in G. In particular, one may take*

$$K(\epsilon) \leq c_0 \log(1/\epsilon)^{c_1 \log(1/\epsilon)}$$

for some constants $c_0, c_1 > 0$ depending only on c.

The main results establish that, upon observing the graph after removing its vertex labels, one may find a set S of vertices of size independent of n such that S contains the root vertex (i.e., vertex 1) with probability at least $1 - \epsilon$. The size of the set is bounded by a function of ϵ only.

Observe that if ℓ is of the order of $\log(1/\epsilon)$, then the bound for $K(\epsilon)$ is $1/\epsilon$ times a poly-logarithmic term in $1/\epsilon$. On the other hand, when ℓ is a fixed constant, as $\epsilon \to 0$, the obtained bounds are super-polynomial in $1/\epsilon$, significantly larger than the analogous bounds obtained for uniform and preferential attachment trees. In all ranges of ℓ, these bounds are inferior to the best upper bounds available for the case $\ell = 1$ (i.e., uniform random recursive trees). We do not claim optimality of this bound. It is an interesting open question whether much smaller vertex sets may be found with the required guarantees. We conjecture that for any $\ell > 1$, root finding is easier in a uniform random ℓ-dag than in a uniform random recursive tree. If that is the case, one should be able to take $K(\epsilon)$ as $\exp(O(\log(1/\epsilon)/\log\log(1/\epsilon)))$. Similar remarks hold for the bound of Theorem 3.5.

In order to prove Theorems 3.4 and 3.5, we propose a root estimation procedure and prove that the same procedure works in both models. The procedure looks for certain carefully selected subgraphs that we call *double cycles*. The set S of candidate vertices are certain special vertices of such double cycles.

The rest of the chapter is organized as follows. In Section 3.2 we introduce the proposed root estimation procedure. The proof of Theorem 3.4 is given in Section 3.3 while Theorem 3.5 is proved in Section 3.4.

3.2 Double cycles

In this section we define the root estimation method that we use to prove the main results. In order to determine the set S of vertices that are candidates for being the root vertex, we define "double cycles".

Let s, t be positive integers. We say that a vertex $v \in [n]$ is an *anchor of a double cycle* of size (s,t) if there exists an integer $0 < p \leq \min(s,t)/2$ and $s+t-1-p$ different vertexes $i_1, i_2, \dots i_{s+t-2-p} \in [n]$, such that

- vertices $v, i_1, \dots, i_{s-1}$ form a cycle of length s in G (in this order);
- vertices $v, i_{s+1-p}, \dots, i_{s+t-1-p}$ form a cycle of length t in G (in this order).

Note that the two cycles are disjoint, except for the common path $v \sim \cdots \sim i_{p-1}$ (so p is the number of common vertices in both cycles). Also note that i_{p-1} is another anchor of the same double cycle. If $p = 1$, we declare $i_0 = v$. In that case the two cycles intersect in the single vertex v and the double cycle has a unique anchor v, see Figure 3.1.

In other words, if two vertices $v, u \in [n]$ are connected by three disjoint paths such that the sum of the lengths of the first and second paths is s and the sum of the lengths of the second and third paths is t, then v and u are anchors of a double cycle of size s and t. Also, v is the anchor of a double cycle of size (s,t) if vertex v is the unique common vertex of two cycles of lengths s and t.

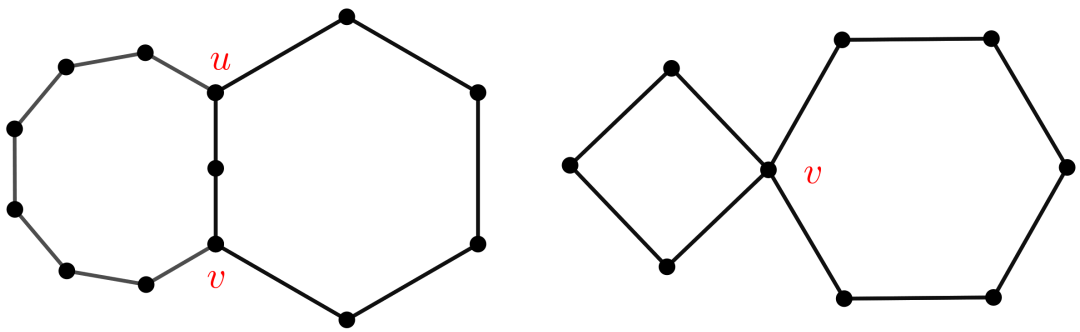

Figure 3.1: Examples of double cycles

For a positive integer m, let $S_m \subset [n]$ be the set of vertices i such that i is an anchor of a double cycle of size (s,t) for some $s \leq m$ and $t \leq m$.

In order to prove Theorem 3.4, it suffices to show that for any given $\epsilon \in (0, 1/100)$,

one may take $m = m_\epsilon = \left\lceil \frac{30}{\ell} \log(1/\epsilon) \right\rceil$ such that

$$\mathbb{P}\left\{1 \in S_m \text{ and } |S_m| \le \frac{4}{\epsilon} \ell^{2m} (2m)!\right\} \ge 1 - \epsilon \,.$$

This follows if we prove that we have both

$$\mathbb{P}\{1 \in S_m\} \ge 1 - \frac{\epsilon}{2} \tag{3.2.1}$$

and

$$\mathbb{P}\left\{|S_m| \le \frac{4}{\epsilon} \ell^{2m} (2m)!\right\} \ge 1 - \frac{\epsilon}{2} \,. \tag{3.2.2}$$

We prove (3.2.1) in Section 3.3.1 and (3.2.2) in Section 3.3.2.

Remark. The reader may wonder why the proposed method looks for double cycles as opposed to simpler small subgraphs such as triangles or a clique of size 4 with an edge removed, etc. The reason is that such simpler subgraphs are either too abundant in the sense that vertices with high index may be contained in (many of) them or the root vertex is not contained in any of them with some probability that is bounded away from zero. This may happen in spite of the fact that the expected number of such small subgraphs containing the root vertex goes to infinity as $n \to \infty$. Double cycles guarantee the appropriate concentration expressed in (3.2.1).

3.3 Proof of Theorem 3.4

As it is explained in the previous section, in order to prove Theorem 3.4, it is enough to prove the inequalities (3.2.1) and (3.2.2), where S_m is the set of those vertices that are anchors of a double cycle of size (s, t) for some $s, t \le m$.

3.3.1 The root vertex is the anchor of a small double cycle

First we consider the case when $\ell = 2$. Then the observed graph G is the union of two independent random recursive trees T_1 and T_2. To prove (3.2.1) we need to ensure that vertex 1 is the anchor of a double cycle of small size, with probability at least $1 - \epsilon/2$. To do so, it suffices to show that there are two edges $(1, i)$ and $(1, j)$ that are present in T_2 but not in T_1 where i and j are "small"– whose meaning is specified below. Indeed, in this case there are two cycles containing vertex 1 formed as follows:

- the unique path from vertex 1 to i in T_1 loops back to 1 thanks to edge $(1,i)$, present in T_2;
- the unique path from vertex 1 to j in T_1 loops back to 1 thanks to edge $(1,j)$, present in T_2.

The only intersection of those two cycles is the intersection of the paths in T_1 from vertex 1 to i and from vertex 1 to j. In a tree, the intersection of two paths is either empty or a path itself. Here the intersection is not empty since both paths contain vertex 1. Thus, vertex 1 is in two cycles which only intersect in a path having vertex 1 as an extremity, meaning that vertex 1 is the anchor of a double cycle. Next we show that two such edges indeed exist, with high probability.

For a vertex $i \in [2,n]$, the probability that the edge $(1,i)$ is present in T_2 is $1/(i-1)$. The probability that it is absent in T_1 is $1-1/(i-1)$. By independence of T_1 and T_2, the probability that the edge $(1,i)$ is present in T_2 and absent in T_1 is $(1-1/(i-1))/(i-1)$. Let X_k denote the number of edges of form $(1,i)$ for some $i \in [k]$, that are not edges in T_1. Then X_k may be written as a sum of independent random variables,

$$X_k \;=\; \sum_{i=2}^{k} B_i$$

where B_i is a Bernoulli random variable with parameter $\frac{1}{i-1}\left(1-\frac{1}{i-1}\right)$.

If $X_k \geq 2$, there exist two edges of form $(1,i)$ with $i \leq k$ that are present in T_2 but not in T_1. By a standard bound for the lower tail for for sums of nonnegative independent random variables, see [19, Exercise 2.9], we have

$$\mathbb{P}\{X_k \geq 2\} \;\geq\; 1-\exp\left(-\frac{(\mathbb{E}[X_k]-1)^2}{2\mathbb{E}[X_k]}\right).$$

Since $\mathbb{E}[X_k]$ is easily seen to fall between $\log(k)-2$ and $\log(k)-1$, we have

$$\mathbb{P}\{X_k \geq 2\} \;\geq\; 1-\exp\left(-\frac{1}{2}\log(k)+\frac{5}{2}-\frac{1}{\log(k)-1}\right).$$

Hence, for $k_\epsilon = \left\lceil 16e^5/\epsilon^2 \right\rceil$, we have $\mathbb{P}\{X_{k_\epsilon} \geq 2\} \geq 1-\epsilon/4$. This implies that, with probability at least $1-\epsilon/4$, vertex 1 is the anchor of a double cycle such that all vertices in the double cycle are in $[k_\epsilon]$. To conclude the proof of (3.2.1) we need to check that indeed the size of the double cycle containing vertex 1 is at most m. Such double cycles are formed by a path in T_1, closed by an additional edge coming from T_2. Therefore, both cycles contained in

the double cycle of interest have a size bounded by the height of the subtree of T_1 induced by the vertex set $[k_\epsilon]$, plus 1. By well-known bounds for the height of a uniform random recursive tree (see, e.g., Drmota [58], Devroye [49], Pittel [119]) we have that the depth of a uniform random recursive tree on k vertices is bounded by $e\log(k) + e\log(4e/\epsilon)$ with probability at least $1 - \epsilon/4$, see Drmota [58, p. 284].

Plugging in the value of k_ϵ, we get that for any $\epsilon \leq 10^{-2}$, the diameter of a uniform recursive random tree of size k_ϵ is at most $15\log(1/\epsilon)$, with probability at least $1 - \epsilon/4$.

Putting these bounds together, we have that, in the case $\ell = 2$, with probability at least $1 - \epsilon/2$, vertex 1 is an anchor of a double cycle of size (s,t) with $s,t \leq 15\log(1/\epsilon)$, implying (3.2.1) for $\ell = 2$.

It remains to extend the above to the general case of $\ell \geq 2$. Since G is the union of ℓ independent uniform random recursive trees, it contains the union of $\lfloor \ell/2 \rfloor$ independent, identically distributed random uniform 2-dags. Using the result proved for random uniform 2-dags above, the probability than in G, vertex 1 is not the anchor of a double cycle of size at most $15\log(\epsilon^{2/(\ell-1)})$ is at most ϵ. This concludes the proof of (3.2.1) in the general case.

3.3.2 High-index vertices are not anchors of double cycles

In order to prove (3.2.2) we need to show that no vertex with high index is the anchor of a double cycle of size smaller than m. We bound the probability that there exists $v > K$ such that $v \in S_m$, where recall that $K = K(\epsilon)$. To this end, we count $C_{s,t}(v)$, the number of double cycles of size (s,t) having vertex v as an anchor. Then, by the union bound,

$$\mathbb{P}\{\exists v > K : v \in S_m\} \leq \sum_{v \geq K} \sum_{s,t \leq m_\epsilon} \mathbb{P}\{C_{s,t}(v) \geq 1\} \leq \sum_{v \geq K} \sum_{s,t \leq m_\epsilon} \mathbb{E}C_{s,t}(v)\,. \tag{3.3.1}$$

In order to bound $\mathbb{E}C_{s,t}(v)$, we may assume, without loss of generality, that $s \leq t$.

For a permutation $\sigma \in \Pi_{s+t-2-p}$ the set of permutation of $[s+t-2-p]$, we denote by $C(s,t,p,v,\sigma,i_1,\ldots,i_{s+t-p-2})$ the following event:

- if $p = 1$,

$$\begin{aligned} &C(s,t,1,v,\sigma,i_1,\ldots,i_{s+t-2}) \\ &\quad = \left\{v \sim i_{\sigma(1)} \sim \cdots \sim i_{\sigma(s-1)} \sim v \sim i_{\sigma(s)} \sim \cdots \sim i_{\sigma(s+t-2)} \sim v\right\}, \end{aligned}$$

- and if $p > 1$

$$C(s,t,p,v,\sigma,i_1,\ldots,i_{s+t-p-2})$$

$$= \left\{ v \sim i_{\sigma(1)} \sim \cdots \sim i_{\sigma(s-1)} \sim v \sim i_{\sigma(s)} \sim \cdots \sim i_{\sigma(s+t-2-p)} \sim i_{\sigma(s-p)} \right\} .$$

where $i \sim j$ denotes that vertices i and j are joined by an edge. Thus, $C(s,t,p,v,\sigma,i_1,\ldots,i_{s+t-p-2})$ is the event that the double cycle of size s,t ($s \le t$) having p vertices in the intersection, with v as an anchor and on the set of vertices $\{i_1,\ldots,i_{s+t-p-2}\}$ ordered by σ as illustrated in Figure 3.2 is present.

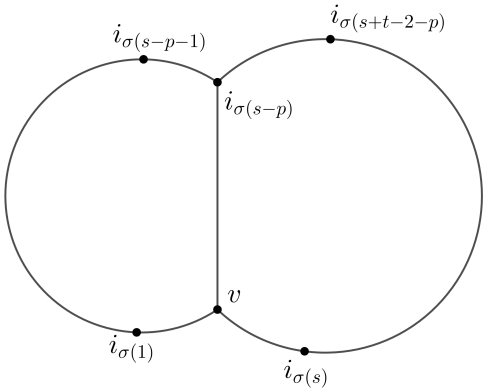

Figure 3.2: Index ordering in a double cycle

With this notation, we may write $C_{s,t}(v)$ as follows:

$$C_{s,t}(v) = \sum_{p=1}^{\lfloor s/2 \rfloor} \sum_{i_1 < \cdots < i_{s+t-2-p}} \sum_{\sigma \in \Pi_{s+t-2-p}} \mathbb{1}_{C(s,t,p,v,\sigma,i_1,\ldots,i_{s+t-p-2})} , \tag{3.3.2}$$

in order to bound the expected number $\mathbb{E}C_{s,t}(v)$ of double cycles of size (s,t) anchored at v, we need to estimate $\mathbb{P}\left\{C(s,t,p,v,\sigma,i_1,\ldots,i_{s+t-p-2})\right\}$.

This exact probability is difficult to compute. Instead, we make use of the following proposition that establishes that a uniform random ℓ-dag is dominated by an appropriately defined inhomogeneous Erdős-Rényi random graph. This random graph is defined as a graph on the vertex set $[n]$ such that each edge is present independently of the others and the probability that vertex i and vertex j are connected by an edge equals

$$\pi(i,j) \stackrel{\text{def}}{=} \min\left(1, \frac{\ell}{\max(i,j)-1}\right) .$$

The next proposition shows that every fixed subgraph is at most as likely to appear in a uniform random ℓ-dag as in the inhomogeneous Erdős-Rényi random graph.

Proposition 3.6. *Let $G = (V, E)$ be a uniform random ℓ-dag on the vertex set $V = [n]$. For some $k \le \binom{n}{2}$, let $(a_1, b_1), \ldots, (a_k, b_k)$ be distinct pairs of vertices such that $a_i \neq b_i$ for all $i \le k$. Then*

$$\mathbb{P}\{(a_1, b_1), \ldots, (a_k, b_k) \in E\} \le \prod_{i=1}^{k} \pi(a_i, b_i) \, .$$

Proof. Recall that the edge set of G may be written as $E = \cup_{j=1}^{\ell} E_j$, where $(V, E_1), \ldots, (V, E_\ell)$ are independent uniform random recursive trees. We may assume, without loss of generality, that $b_i > a_i$ for all $i \in [k]$.

We prove the proposition by induction on k. For $k = 1$, the inequality follows from the union bound:

$$\mathbb{P}\{(a_1, b_1) \in E\} \le \sum_{j=1}^{\ell} \mathbb{P}\Big\{(a_1, b_1) \in E_j\Big\} = \frac{\ell}{\max(a_1, b_1) - 1} \, . \tag{3.3.3}$$

For the induction step, suppose the claim of the proposition holds for up to k edges and consider $k+1$ distinct pairs $(a_1, b_1), \ldots, (a_{k+1}, b_{k+1})$. Then, by the induction hypothesis,

$$\begin{aligned}
&\mathbb{P}\{(a_1, b_1), \ldots, (a_{k+1}, b_{k+1}) \in E\} \\
&\quad = \quad \mathbb{P}\{(a_1, b_1), \ldots, (a_k, b_k) \in E\} \; \mathbb{P}\{(a_{k+1}, b_{k+1}) \in E \mid (a_1, b_1), \ldots, (a_k, b_k) \in E\} \\
&\quad \le \quad \mathbb{P}\{(a_{k+1}, b_{k+1}) \in E \mid (a_1, b_1), \ldots, (a_k, b_k) \in E\} \prod_{i=1}^{k} \pi(a_i, b_i) \, .
\end{aligned}$$

Thus, it suffices to show that for all distinct pairs $(a_1, b_1), \ldots, (a_{k+1}, b_{k+1})$,

$$\mathbb{P}\{(a_{k+1}, b_{k+1}) \in E \mid (a_1, b_1), \ldots, (a_k, b_k) \in E\} \le \pi(a_{k+1}, b_{k+1}) \, .$$

First, consider the simpler case when for all $i \in [k]$, $b_i \neq b_{k+1}$. Then, for every fixed $j \in [\ell]$, the events $\{(a_1, b_1) \in E_j, \ldots, (a_k, b_k) \in E_j\}$ and $\{(a_{k+1}, b_{k+1}) \in E_j\}$ are independent. Moreover since the ℓ uniform random recursive trees are independent, the events $\{(a_1, b_1) \in E, \ldots, (a_k, b_k) \in E\}$ and $\{(a_{k+1}, b_{k+1})\} \in E$ are also independent, and therefore

$$\mathbb{P}\{(a_{k+1}, b_{k+1}) \in E \mid (a_1, b_1), \ldots, (a_k, b_k) \in E\} = \mathbb{P}\{(a_{k+1}, b_{k+1}) \in E\} \le \pi(a_{k+1}, b_{k+1}) \, ,$$

by (3.3.3).

Now, suppose that there exist some $i \in [k]$ such that $b_i = b_{k+1}$. We may assume that there exists a $w \in [k]$ such that $b_1, \ldots, b_w = b_{k+1}$ and for all $i \in [w+1, k]$, $b_i \neq b_{k+1}$. Since each (V, E_j) is a recursive tree, for $i \in [w]$, $(a_i, b_{k+1}) \in E_j$ and $(a_{k+1}, b_{k+1}) \in E_j$ cannot happen at

the same time. Thus, edge (a_{k+1}, b_{k+1}) can only be present in the sets E_j that do not contain any of the edges (a_i, b_{k+1}). Hence, introducing $A = \#\left\{j \in [\ell] : \ E_j \cap \{(a_1, b_{k+1}), \ldots, (a_w, b_{k+1})\} \neq \emptyset\right\}$, we have, for all $a \in [\ell]$,

$$\mathbb{P}\{(a_{k+1}, b_{k+1}) \in E \mid (a_1, b_1), \ldots, (a_k, b_k) \in E \text{ and } A = a\} \ = \ \mathbb{P}\left\{(a_{k+1}, b_{k+1}) \in \cup_{j=1}^{\ell-a} E_j\right\} .$$

Using the union bound again,

$$\mathbb{P}\left\{(a_{k+1}, b_{k+1}) \in \cup_{j=1}^{\ell-a} E_j\right\} \leq \frac{\ell - a}{b_{k+1} - 1} \leq \frac{\ell}{b_{k+1} - 1} .$$

Since this holds for all a, we have

$$\mathbb{P}\{(a_{k+1}, b_{k+1}) \in E \mid (a_1, b_1), \ldots, (a_k, b_k) \in E\} \leq \frac{\ell}{b_{k+1} - 1} ,$$

as desired.

■

To count $C_{s,t}(v)$ we split the sum in (3.3.2) by adding a parameter r in order to separate the vertices $i_1, \ldots, i_{s+t-2-p}$ according to whether they are smaller or larger than v, obtaining

$$C_{s,t}(v) = \sum_{p=1}^{\lfloor s/2 \rfloor} \ \sum_{r=0}^{s+t-p-2} \ \sum_{\sigma \in \Pi_{s+t-2-p}} \ \sum_{i_1 < \cdots < i_r < v} \ \sum_{v < i_{r+1} < \cdots < i_{s+t-2-p}} \mathbb{1}_{C(s,t,p,v,\sigma,i_1,\ldots,i_{s+t-p-2})} .$$

From Proposition 3.6 we know that the probability of each given double cycle is upper bounded by the product of $\pi(i,j) = \ell/(max(i,j)-1)$. Thus we introduce $E_\sigma(j) \in \{0,1,2,3,4\}$ counting the number of vertices neighboring vertex i_j in the double cycle, that have indices smaller than i_j. By convention we write $E_\sigma(0)$ for the analogous quantity for vertex v. Doing so, we may write

$$\begin{aligned} \mathbb{E}C_{s,t}(v) \leq \sum_{p=1}^{\lfloor s/2 \rfloor} \ell^{s+t-p} \ \sum_{r=0}^{s+t-p-2} \ \sum_{\sigma \in \Pi_{s+t-2-p}} \ & (v-1)^{-E_\sigma(0)} \\ & \times \left(\sum_{i_1 < \cdots < i_r < v} \ \prod_{j=1}^{r} (i_j - 1)^{-E_\sigma(j)} \right) \times \left(\sum_{v < i_{r+1} < \cdots < i_{s+t-2-p}} \ \prod_{j=r+1}^{s+t-2-p} (i_j - 1)^{-E_\sigma(j)} \right) . \end{aligned} \tag{3.3.4}$$

This allows us to decompose the sum in two parts; the sum involving the r vertices with index smaller than v and the $s+t-2-p-r$ vertices with index larger than v. If we fix p, m and σ, we need to upper bound both

$$A(\sigma, p, r) := A = \sum_{i_1 < \cdots i_r < v} \ \prod_{j=1}^{r} (i_j - 1)^{-E_\sigma(j)}$$

and

$$B(\sigma,p,r) := B = \sum_{v<i_{r+1}<\cdots<i_{s+t-2-p}} \prod_{j=r+1}^{s+t-2-p} (i_j-1)^{-E_\sigma(j)} .$$

This may be done with the help of the next two lemmas.

Lemma 3.7. *Fix a vertex v, vertices $i_1<\cdots<i_r<v<i_{r+1}<\cdots<i_{s+t-p-2}$ and an ordering σ of a double cycle on this set of vertices with v as an anchor. Then, for every $k\in[r]$ we have*

$$k-1 \geq \sum_{i=1}^{k} E_\sigma(i) .$$

Proof. For $k\in[r]$, we define $G(k)$ as the subgraph of the double cycle in which we only keep the k vertices of smallest index, so that $\sum_{i=1}^{k} E_\sigma(i)$ is the number of edges in $G(k)$.

Since $G(k)$ does not contain v, there are no cycles in $G(k)$, and therefore it is a forest. Since $|G(k)|=k$, it follows that $G(k)$ has at most $k-1$ edges. ∎

Lemma 3.8. *Fix a vertex v, vertices $i_1<\cdots<i_r<v<i_{r+1}<\cdots<i_{s+t-p-2}$ and an ordering σ of a double cycle on this set of vertices with v as an anchor. Then, $\forall k\in[s+t-2-p-r]$ we have*

$$k+1 \leq \sum_{i=1}^{k} E_\sigma(s+t-1-p-i) .$$

Proof. For $k\in[s+t-2-p-r]$, we define $G'(k)$ as the subgraph of the double cycle in which we only keep the k vertices of largest index. Vertex $i_{s+t-2-p-k}$ has at least two neighbors in the double cycle. From the definition of $E_\sigma(s+t-p-1-k)$, $E_\sigma(s+t-p-1-k)$ is then at least 2 minus the number of neighbors of $i_{s+t-2-p-k}$ in the double cycle with larger index. The number of such neighbors of $i_{s+t-2-p-k}$ is exactly the number of edges in $G'(k)$ minus the number of edges in $G'(k-1)$. Denoting $G'(k)=(V'(k),E'(k))$, it leads to

$$E_\sigma(s+t-p-1-k) \geq 2-(\#E'(k)-\#E'(k-1)) ,$$

implying

$$\sum_{i=1}^{k} E_\sigma(s+t-1-p-i) \geq 2k-\#E'(k) .$$

Since $G'(k)$ does not contain v, it is a forest. Moreover $|G'(k)|=k$ so $G'(k)$ has at most $k-1$ edges, which concludes the proof. ∎

We may decompose A as follows:

$$A = \sum_{i_r:\ i_r<v} (i_r-1)^{-E_\sigma(r)} \cdots \sum_{i_1:\ i_1<i_2} (i_1-1)^{-E_\sigma(1)} .$$

From Lemma 3.7 with $k+1$, we know that $-E_\sigma(1) \geq 0$, leading to

$$\sum_{i_1:\ i_1<i_2} (i_1-1)^{-E_\sigma(1)} \leq (i_2-1)^{1-E_\sigma(1)} ,$$

which in turn leads to

$$A \leq \sum_{i_r:\ i_r<v} (i_r-1)^{-E_\sigma(r)} \cdots \sum_{i_2:\ i_2<i_3} (i_2-1)^{1-E_\sigma(1)-E_\sigma(2)} .$$

Once again, by Lemma 3.7 with $k=2$, we have $1-E_\sigma(1)-E_\sigma(2) \geq 0$, leading to

$$\sum_{i_2:\ i_2<i_3} (i_2-1)^{1-E_\sigma(1)-E_\sigma(2)} \leq (i_3-1)^{2-E_\sigma(1)-E_\sigma(2)} .$$

Iterating this scheme r times, using Lemma 3.7 at each step leads to

$$A \leq (v-1)^{r-\sum_{i=1}^{r} E_\sigma(i)} . \tag{3.3.5}$$

Similarly, we decompose B as

$$B = \sum_{i_{r+1}:\ i_{r+1}>v} (i_{r+1}-1)^{-E_\sigma(r+1)} \cdots \sum_{i_{s+t-p-2}:\ i_{s+t-p-2}>i_{s+t-p-3}} (i_{s+t-p-2}-1)^{-E_\sigma(s+t-p-2)} .$$

It follows from Lemma 3.8 that $E_\sigma(s+t-p-2) \geq 2$, and therefore

$$\sum_{i_{s+t-p-2}:\ i_{s+t-p-2}>i_{s+t-p-3}} (i_{s+t-p-2}-1)^{-E_\sigma(s+t-p-2)} \leq (i_{s+t-p-3}-1)^{1-E_\sigma(s+t-p-2)} .$$

Following an analogous reasoning to the upper bound of A, iterating this scheme $s+t-2-p-r$ times, using Lemma 3.8 at each step leads to

$$B \leq (v-1)^{s+t-2-p-r-\sum_{j=r+1}^{s+t-2-p} E_\sigma(j)} . \tag{3.3.6}$$

Substituting (3.3.5) and (3.3.6) into (3.3.4), we obtain

$$\mathbb{E}XPC_{s,t}(v) \leq \sum_{p=1}^{\lfloor s/2 \rfloor} \sum_{r=0}^{s+t-p-2} \sum_{\sigma \in \Pi_{s+t-2-p}} \ell^{s+t-p} (v-1)^{-E_\sigma(0)} \times (v-1)^{s+t-2-p-r-\sum_{j=r+1}^{s+t-2-p} E_\sigma(j)} \times (v-1)^{r-\sum_{i=1}^{r} E_\sigma(i)} .$$

Since

$$\sum_{j=0}^{s+t-2-p} E_\sigma(j) = s+t-p\ ,$$

we have

$$\mathbb{E}XPC_{s,t}(v) \le \frac{1}{(v-1)^2} \sum_{p=1}^{\lfloor s/2 \rfloor} \sum_{r=0}^{s+t-p-2} \sum_{\sigma \in \Pi_{s+t-2-p}} \ell^{s+t-p}\ ,$$

leading to

$$\begin{aligned}\mathbb{E}\left[C_{s,t}(v)\right] &\le \sum_{p=1}^{\lfloor s/2 \rfloor} \ell^{s+t-p}(s+t-p-2)!(s+t-p-2)\frac{1}{(v-1)^2} \\ &\le 2\ell^{s+t}\frac{(s+t)!}{(v-1)^2}\ .\end{aligned}$$

Finally, we plug this bound in (3.3.1):

$$\mathbb{P}(\exists v \ge K\ :\ v \in S_m) \le \sum_{v \ge K} \sum_{s,t \le m_\epsilon} 2\ell^{s+t}\frac{(s+t)!}{(v-1)^2} \tag{3.3.7}$$

$$\le 4\ell^{2m_\epsilon}(2m_\epsilon)!\frac{1}{K}\ . \tag{3.3.8}$$

Choosing $K = 8\frac{1}{\epsilon}\ell^{2m_\epsilon}(2m_\epsilon)!$ concludes the proof of (3.2.2) and therefore Theorem 3.4 follows.

3.4 Proof of Theorem 3.5

The proof of Theorem 3.5 is analogous to that of Theorem 3.4. In order to avoid repeating essentially the same argument, we only highlight the differences in the proofs.

It is enough to prove that, choosing $m = m_\epsilon = \lceil (9+12/c)\log(1/\epsilon) \rceil$ one has

$$\mathbb{P}\left\{1 \in S_m \text{ and } |S_m| \le \frac{4}{\epsilon}(c+1)^{2m}(2m)!\right\} \ge 1-\epsilon\ .$$

This follows if we prove that

$$\mathbb{P}\{1 \in S_m\} \ge 1 - \frac{\epsilon}{2} \tag{3.4.1}$$

and

$$\mathbb{P}\left\{|S_m| \leq \frac{4}{\epsilon}(c+1)^{2m}(2m)!\right\} \geq 1 - \frac{\epsilon}{2} \tag{3.4.2}$$

both hold.

Recall that the uniform Cooper-Frieze model is the union of a uniform random recursive tree G_1 and an inhomogeneous Erdős-Rényi random graph G_2 (with edges probabilities $\min(c/\max(i,j)-1,1)$).

Proving (3.4.1) and (3.2.1) shares the same basic argument. In order to show that the root vertex is an anchor of a double cycle of size (s,t) for some $s,t \leq m$, one may show that, with the desired probability, there exist at least two vertices i,j with sufficiently small index such that the edges $(1,i)$ and $(1,j)$ are not present in the uniform random recursive tree but they are present in the inhomogeneous Erdős-Rényi random graph G_2. This follows by similar concentration arguments (for sums of independent Bernoulli random variables and for the height of a uniform random recursive tree) as in the proof of Theorem 3.4.

The proof of (3.4.2) is once again analogous to the proof of (3.2.2). We remind the reader than the main step of the proof of Theorem 3.4 relies on the fact that a uniform random ℓ-dag is dominated by an inhomogeneous Erdős–Rényi random graph with of edges probabilities $\ell/(\max(i,j)-1)$, as shown in Proposition 3.6. Using a similar reasoning as in Proposition 3.6, one may prove that a uniform Cooper-Frieze random graph is dominated by an an inhomogeneous Erdős–Rényi random graph with edge probabilities $(c+1)/(\max(i,j)-1)$. The remainder of the proof is exactly the same as that of the proof of (3.2.2) and concludes the proof of Theorem 3.5.

3.5 Concluding remarks

In this chapter we addressed the problem of finding the first vertex in dynamically growing networks, based on observing a present-day snapshot of the unlabeled network. This problem has mainly been studied for trees and the main purpose of the chapter is to study root finding in more complex networks. The main results show that in certain natural models it is possible to construct confidence sets for the root vertex whose size does not depend on the observed network. These confidence sets contain the root vertex with high probability, and their size only depends on the required probability of error. We prove this property in two models of random networks, namely uniform ℓ-dags and a simplified model inspired by a general random network model of Cooper and Frieze. In both models, the constructed confidence set contains all vertices that are anchors of certain small subgraphs that we call "double cycles."

We leave a number of questions open. We conjecture that the upper bounds obtained for the size of the confidence set are suboptimal (as a function of the probability of error ϵ). To substantially improve on these bounds one may need to consider "global" measures, reminiscent to the centrality measures employed in the case of root finding in recursive trees, as opposed to the "local" method proposed here. However, their use and analysis appears substantially more challenging.

Deriving lower bounds for the size of the confidence set is another interesting open question.

Another path for further research is to extend the network models beyond the uniform ones considered here. The most natural extensions are preferential attachment versions of the models. Such a graph is grown recursively by connecting each new vertex to ℓ vertices chosen with probability being a function of their degree. In those models, the graph is not the union of independent trees. Proofs presented in this chapter are not yet adapted to deal with those dependencies. However, in preferential attachment graphs, degree centrality is known to achieve graph archaeology, see Banerjee and Huang [12].

We end by noting that the methodology based on double cycles also works in a variant of the uniform Cooper-Frieze model in which the uniform random recursive tree is removed. More precisely, one may consider an inhomogeneous Erdős-Rényi random graph on the vertex set $[n]$ with edge probabilities $\min(c/(\max(i,j)-1),1)$, where $c>1$ is a constant. In this case one may prove the following.

Theorem 3.9. *Let $c>1$ and let $G=G^{(n)}$ be an inhomogeneous Erdős-Rényi random graph on n vertices, with edge probabilities $p_{i,j}=\min(c/(\max(i,j)-1),1)$. Root estimation is possible in G. In particular, there exist constants $c_0,c_1>0$, depending on c only, such that one may take*

$$K(\epsilon)\le\left(\frac{c_0}{\epsilon^{c_1}}\right)^{\frac{c_0}{\epsilon^{c_1}}}.$$

The outline of the proof is similar to that of Theorems 3.4 and 3.5. The only difference is in the proof that the root vertex is an anchor of a sufficiently small double cycle. To prove this, we may write G as the union of two independent inhomogeneous Erdős-Rényi random graphs as follows. Let k be a sufficiently large integer (only depending on ϵ). Then we may define $G_1=([n],E_1)$ and $G_2=(n,[E_2])$ as independent inhomogeneous Erdős-Rényi random graphs such that for all $1\le i<j\le n$,

$$\mathbb{P}\{(i,j)\in E_1\}=\begin{cases}\frac{c}{k} & \text{if } j\le k\\ 0 & \text{otherwise}\end{cases}$$

and

$$\mathbb{P}\{(i,j) \in E_2\} = \begin{cases} \frac{p_{i,j} - \frac{c}{k}}{1 - \frac{c}{k}} & \text{if } j \leq k \\ p_{i,j} & \text{otherwise} \end{cases}$$

Clearly, $G = ([n], E_1 \cup E_2)$. The subgraph of G_1 induced by the vertex set $[k]$ is a supercritical Erdős-Rényi random graph and therefore, with high probability, it has a connected "giant" component of size that is linear in k. Then one may easily show that, with high probability, there are three edges in G_2 of the form $(1, i)$, where i belongs to the giant component. This is enough for vertex 1 to be an anchor of a double cycle.

The rest of the proof is identical to that of Theorem 3.4.

Finally, the double cycle method can be implemented in polynomial time. Indeed, checking if the pair of vertices u, v are the anchors of a double cycle of size at most m can be achieved by finding the three shortest disjoint path between them. This can be achieved in $O(n)$ time by running three times a modified Dijkstra algorithm, see Bhandari [18]. To find all double cycle anchor, a naive method is just to run this on every possible pair of vertices. Thus, it is possible to compute the confidence set of the double cycle method in $O(n^3)$ time.

Acknowledgement

We would like to thank Umberto De Ambroggio for his help correcting and improving this chapter.

Chapter 4

Estimating the history of a random recursive tree

Contents

Abstract

This chapter studies the problem of estimating the order of arrival of the vertices in a random recursive tree. Specifically, we study two fundamental models: the uniform attachment model and the linear preferential attachment model. We propose an order estimator based on the Jordan centrality measure and define a family of risk measures to quantify the quality of the ordering procedure. Moreover, we establish a minimax lower bound for this problem, and prove that the proposed estimator is nearly optimal. Finally, we numerically demonstrate that the proposed estimator outperforms degree-based and spectral ordering procedures.

This Chapter is based on a joint work with Christophe Giraud, Gábor Lugosi and Déborah Sulem (Briend, Giraud, Lugosi, and Sulem [29]).

4.1 Introduction

In this chapter, we consider the problem of estimating the entire history of the network, that is, the arrival times of all the vertices in a random recursive tree. One may consider this as a question of latent variable estimation. A related statistical problem is the so-called *seriation*. Seriation is the problem of inferring an ordering of points, based on pairwise similarity or on the adjacency information between two points. This similarity measure is assumed to statistically decrease with the distance in a latent space and informs on the latent global order of the points. The seriation problem has been studied in various fields, such as in archaeology (Robinson [123]), bioinformatics (Recanati et al. [121]), and matchmaking (Bradley and Terry [20]). It has been theoretically analyzed in random graph models such as geometric graphs and graphons (Giraud et al. [72], Janssen and Smith [82]). In recursive trees, the pairwise affinity between nodes is encoded in the adjacency matrix, and the latent space and latent positions are respectively the temporal line and the arrival times of the vertices. Estimating the temporal order of the vertices in a recursive tree can therefore be interpreted as an instance of the seriation problem.

To estimate the vertices' order, we propose a procedure based on a centrality measure, specifically on the *Jordan centrality*. We prove that this procedure is nearly optimal in two random recursive tree models, namely, the uniform random recursive tree (urrt) and the preferential attachment (pa) model. In these models, a tree of $n \geq 1$ vertices is grown by adding and connecting one vertex at each time step. To describe the growing process, we assume that the vertices have intrinsic *labels* from 1 to n. At each step $t = 1, \dots, n$ of the growth, a new vertex, say of label j_t, is picked arbitrarily among the set of nodes not yet in the tree, and added to the tree with the *rank* t. At $t = 1$, the first sampled vertex is the root of the tree. We denote by $\sigma : \{1, \dots, n\} \to \{1, \dots, n\}$ the ordering (or, ranking) map of vertices such that $\sigma(j_t) = t$. In other words, σ is a permutation.

The urrt and pa models differ by the attachment rule used to connect a new vertex at each step $t = 2, \dots, n$ of the growth process. In the urrt model, the vertex j_t is connected by an undirected edge to a vertex sampled uniformly among the vertices of the current tree. In the pa model, the vertex of the tree is sampled with a probability proportional to its degree. We denote by $T = T_n$ the obtained tree *structure*, that is, the set of nodes with labels in $\{1, \dots, n\}$ and the undirected edges between them. In the statistical problem considered in this chapter, after the tree is grown, the rank or arrival time $\sigma(i)$ of each node

i is not observed on T. The random growing process defines a probability distribution on trees. We denote by $\mathbb{P}$ the corresponding probability distribution and $\mathbb{E}$ the associated expectation.

We note that in these random models, the sampling process is independent of the labels chosen to identify the vertices, here $\{1,\dots,n\}$. Therefore, any coherent ordering procedure should be *label invariant*, that is, independent of these labels. Saying that $\widehat{\sigma}$ is label invariant means that for any fixed T and ranking σ,

$$\widehat{\sigma}(T,\sigma) \overset{\mathcal{L}}{=} \widehat{\sigma}(T^{\sigma'},\sigma\circ\sigma')\,, \tag{4.1.1}$$

for a permutation σ', where $T^{\sigma'}$ denotes the tree with label i replaced by $\sigma'(i)$. Note that the equality in distribution is a simple equality if the ordering procedure $\widehat{\sigma}$ is deterministic. Let us also remark that an easy way to transform any ordering procedure into a label invariant ordering procedure is by applying a random permutation to the labels of the tree before feeding it to the ordering procedure.

In order to measure the quality of an estimator of the history, we introduce a family of risk measures that takes into account the error in the estimated arrival time of each vertex, weighted by a function of the arrival time. We define the following family of risk measures

$$R_\alpha(\widehat{\sigma}) \overset{\text{def}}{=} \mathbb{E}\left[\sum_{i=1}^{n}\frac{|\widehat{\sigma}(i)-\sigma(i)|}{\sigma(i)^\alpha}\right], \tag{4.1.2}$$

where $\alpha>0$. The parameter α tunes the importance given to vertices with small true rank $\sigma(i)$: the higher α, the more weight is given to vertices with low rank. Perhaps the most natural choice is $\alpha=1$. In that case the risk corresponds to normalizing the error on the estimation of the arrival time of a vertex by its true arrival time. We note that it is often the early stages of a propagation phenomenon that are more relevant, for example, for designing prevention strategies. Additionally, in random growing trees, it is harder to accurately order the high-rank vertices, due to the inherent model symmetries (Sreedharan, Magner, Grama, and Szpankowski [133]).

One way to construct an estimator $\widehat{\sigma}$ of the ranking map is to choose a score function on the set of vertices, and order vertices by increasing (or decreasing) values. Such a score function could be based on the likelihood under the tree model. However, the latter is generally difficult to compute, see Bubeck et al. [33]. Instead, score functions based on the degree (Navlakha and Kingsford [116]) or the so-called rumor centrality (Cantwell, St-Onge, and Young [36]) can be computed in polynomial time.

Another approach are iterative algorithms that recursively infer previous states of the tree such as the history sampling algorithms (Cantwell et al. [36], Crane and Xu [42])

and the Peeling procedure (Sreedharan et al. [133]), which is related to the depth centrality score. These methods are guaranteed to recover a recursive ordering of the vertices. Moreover, Crane and Xu [42] show that the history sampling algorithm outputs confidence sets for the arrival time of a single vertex with valid frequentist coverage. Besides, Sreedharan et al. [133] demonstrate that the partial ordering retrieved by the Peeling procedure has good properties in settings where the root of the tree can be unambiguously identified. Nonetheless, there are not yet guarantees on the quality of the global ordering provided by these methods.

The ordering procedure we propose is based on the Jordan centrality, defined, for a vertex $u \in T$ belonging to a tree T, as

$$\psi_T(u) = \max_{v \in V(T),\, v \sim u} |(T, u)_v| \,. \tag{4.1.3}$$

where (T, u) denotes the tree T rooted at u, where $u \sim v$ means that u and v are neighbors in T, and where $(T, u)_v$ denotes the subtree of T containing all vertices w such that v lies on the path connecting w to u (see Figure 4.1). Somewhat informally, we call $(T, u)_v$ the subtree hanging from v in the rooted tree (T, u). The maximum in (4.1.3) is taken over vertices v of the tree that are connected to vertex u by an edge. Intuitively, if a vertex is *central*, then none of the subtrees hanging from it can be too large. Therefore, the lower $\psi_T(i)$, the more central is vertex i. It is straightforward to see that $(\psi_T(u))_{u \in T}$ only depends on the structure of the tree and not on the labels of its vertices. We then define $\widehat{\sigma_J} : \{1, \dots, n\} \to \{1, \dots, n\}$ the ordering obtained by ranking the vertices by increasing value of Jordan centrality–breaking ties at random. This estimator is label invariant. An equivalent formulation of this algorithm is to estimate the position of vertex 1 by the Jordan centroid, rooting the tree at this vertex and then ordering vertices by the size of their hanging subtree in the rooted tree. Thus, if the exact position of vertex 1 was known, we would be ordering vertices by the number of their descendants.

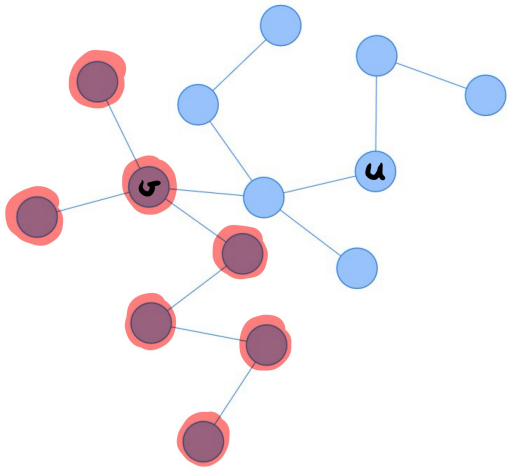

Figure 4.1: An illustration of the subtree $(T, u)_v$, corresponding to nodes highlighted in red.

While the risk defined in (4.1.2) can be computed for any value $\alpha > 0$, we restrict α to a range of values which are relevant for our ordering problem. Specifically, we only consider $\alpha \geq 1$, since for $\alpha < 1$ the problem becomes trivial, since even a random permutation has a risk which is minimax optimal up to constant factor (see Appendix 4.5.1). In Theorems 4.1 and 4.7, we provide minimax lower bounds for the risk $R_\alpha(\widehat{\sigma})$ in the urrt and pa model, for any label-invariant estimator $\widehat{\sigma}$. Then, in Theorems 4.4 and 4.8 upper bounds for the risk of the Jordan ordering are obtained. Finally, in Corollaries 4.5 and 4.9 we prove that our proposed estimator is minimax optimal up to constant factors, in a non-trivial range of parameters α. In the following table, we summarise our findings. For $\alpha \geq 1$, we denote by R^*_α the optimal risk, and $R_\alpha(\widehat{\sigma_J})$ the risk of the Jordan ordering.

	urrt	pa
R^*_α	$\geq n^{2-\alpha}/65 \vee 1/2$	$\geq n^{2-\alpha}/70 \vee 1/2$
$R_\alpha(\widehat{\sigma_J})$	$= \mathcal{O}\left(n^{2-\alpha} + \log^4(n)\right)$	$= \mathcal{O}\left(n^{2-\alpha} + n^{3/4}\right)$

We also compare numerically the performance of the Jordan estimator with other ordering procedures in a simulation study.

In the rest of this section, we review previous works and introduce some notation. Then, in Section 4.2, we analyze the Jordan ordering in the urrt model. Next, we consider

the pa model and prove analogous results in Section 4.3. Finally, in Section 4.4, we report the results of our simulation study and compare the empirical performance of the Jordan estimator to alternative methods based on the degree centrality, a peeling method (Navlakha and Kingsford [116, Section 2.3]) and a spectral method commonly used in seriation problems (Recanati et al. [122]).

4.1.1 Related work

Most methods for ranking the vertices of a random recursive tree have been introduced for the root-finding problem, that is, recovering a vertex (or a set of vertices) that is (contains) the root. For this problem, maximum likelihood estimators (Brandenberger et al. [21], Bubeck et al. [33], Haigh [74]) and estimators based on rumor centrality (Shah and Zaman [129, 130]) have been proposed and analyzed. Jordan centrality is another measure of centrality used by Bubeck et al. [33, 34] to construct confidence sets. Banerjee and Bhamidi [11], Jog and Loh [83, 84] study the persistence of the most central nodes in random recursive trees. Furthermore, while the vertex with maximum degree is generally not a good estimator of the root in the urrt model, in the pa model pairs degree centrality is useful for retrieving the first vertex (Banerjee and Bhamidi [10], Contat et al. [40]). Some recent work studies root-finding in Galton-Walton trees (Brandenberger et al. [21]) and more general graphs (Briend et al. [26], Crane and Xu [43]).

Crane and Xu [42] propose a general history-sampling procedure for network archaeology, which can be applied to the problem of estimating arrival times. The history sampling algorithm outputs a confidence set of rankings that contains the true one with high probability. However, there is no known bound of the size nor the average global error of an ordering in this confidence set.

The vertex arrival-time estimation problem bears some similarity to the seriation problem, though in the former, the dependence is intrinsically related to the tree structure. For example, in a random geometric graph (Gilbert [71]), the seriation problem is to estimate the position of the random points. Since there is no time structure in seriation, different metrics for the error are used, such as the maximum distance between the true and estimated latent position. Examples of methods are provided by Giraud et al. [72]. Another widely studied seriation method consists in ordering latent points by a spectral method on the graph Laplacian (see Section 4.4 and Recanati et al. [122] for details). They give guarantees for the quality of their method when the observed adjacency matrix is a perturbed Robinson matrix. The expected adjacency matrices of both urrt and pa trees are Robinson, and therefore the models studied here can be viewed as perturbed Robinson matrices. Nonetheless, none of the above-mentioned papers gives any insight about

seriation in urrt and pa trees.

4.1.2 Notation

Let π_n be the set of permutations of $[n] := \{1, 2, \dots, n\}$, and let $\mathbb{T}_n$ be the set of un-labelled trees of size n. We denote by urrt(n) the distribution of a tree $\mathcal{T}_n$ of size $n \geq 1$, generated from the uniform attachment model. Similarly, we denote by pa(n) the distribution of a tree sampled from the preferential attachment model. Moreover, we decompose the tree as $\mathcal{T}_n = (T_n, \sigma_n)$, where T_n is the shape of the tree and σ_n is the recursive ordering of the vertices in $\mathcal{T}_n$. For simplicity, we drop the subscript n when the size of the tree is fixed and clear from the context. We denote by $\mathbb{P}$ the probability distribution under the tree growing process and $\mathbb{E}$ the corresponding expectation.

Recall that for a tree T and a vertex $u \in T$, we denote by (T, u) the tree rooted at u. For a rooted tree (T, u) and a vertex v we denote by $(T, u)_v$ the subtree of T consisting of all vertices w such that v lies on the path connecting w to u (see Figure 4.1). For two vertices $u, v \in T$, $u \sim v$ means that u is a neighbor of v in T (and reciprocally). In a rooted tree (T, u), we say that w is a child of v if w is in $(T, u)_v$. We denote by $\mathrm{de}_n(u) = |(\mathcal{T}_n, 1)_u|$ the number of descendants of u in $\mathcal{T}_n$. For simplicity, we drop the subscript n and use $\mathrm{de}(u)$ when the size of the tree is fixed and clear from the context.

Recall the definition of the Jordan centrality; for a tree T and vertex $u \in T$,

$$\psi_T(u) = \max_{v \in T, v \sim u} |(T, u)_v| \,. \tag{4.1.4}$$

We denote by c a centroid of T, defined as $c = \arg\min_{u \in T} \psi_T(u)$. It is well-known that any tree has at least one and at most two centroids. Moreover, for a vertex $u \in T$ that is not a centroid, the subtree $(T, u)_v$, $v \sim u$ with maximum size, contains all centroids.

The Jordan ordering procedure consists in ordering points by increasing values of ψ (ties being broken randomly). Equivalently, it consists in rooting the tree at c and order vertices by $|(T, c)|_u$. We use $\widehat{\sigma_J}$ to refer to the Jordan ordering of T_n. As noted in the introduction, the Jordan centrality does not depend on the labelling of the tree (only on its shape), and so is a label invariant ordering.

4.2 The uniform attachment model

In this section, we focus on the uniform attachment model as the random growing process of the tree. We first present a lower bound for the risk $R_\alpha(\widehat{\sigma})$ of any label-invariant

estimator $\widehat{\sigma}$ of the vertices order.

4.2.1 A lower bound

In the next proposition, we provide a lower bound for the risk $R_\alpha(\widehat{\sigma})$ for any label-invariant estimator of the recursive ordering in the urrt model. Define, for any $n \geq 1$, the optimal risk by

$$R_\alpha^* := \min_{\widehat{\sigma} \in \Pi_n} R_\alpha(\widehat{\sigma}) \,,$$

where Π_n is the set of label-invariant recursive orderings.

Theorem 4.1. *In the* urrt *model, we have, for all $\alpha > 0$ and $n \geq 200$,*

$$R_\alpha^* \geq \frac{n^{2-\alpha}}{65} \,.$$

Proof. For a tree T and an ordering of its vertices σ, let $\tau = \sigma^{-1}$ (i.e, $\tau(i)$ is the label of the vertex that arrives at time i). We start by recalling that

$$R_\alpha(\widehat{\sigma}) = \mathbb{E}\left[\sum_{j=1}^{n} \frac{|\widehat{\sigma}(j) - \sigma(j)|}{\sigma(j)^\alpha}\right] = \mathbb{E}\left[\sum_{j=1}^{n} \frac{|\widehat{\sigma} \circ \tau(j) - j)|}{j^\alpha}\right] ,$$

which is lower bounded as follows

$$R_\alpha(\widehat{\sigma}) \geq \sum_{j=\lfloor n/2 \rfloor + 1}^{\lfloor 3n/4 \rfloor} \frac{|\widehat{\sigma} \circ \tau(j) - j)|}{j^\alpha} + \sum_{j=\lfloor 3n/4 \rfloor + 1}^{n} \frac{|\widehat{\sigma} \circ \tau(j) - j)|}{j^\alpha} \tag{4.2.1}$$

$$\geq \frac{1}{n^\alpha} \sum_{j=\lfloor n/2 \rfloor + 1}^{\lfloor 3n/4 \rfloor} \mathbb{E}\left[\left|\widehat{\sigma} \circ \tau(j) - j)\right| + \left|\widehat{\sigma} \circ \tau\left(\left\lfloor \frac{n}{4} \right\rfloor + j\right) - \left\lfloor \frac{n}{4} \right\rfloor - j\right|\right] . \tag{4.2.2}$$

We associate summands of the risk by pairs to later use the fact that some pairs of vertices are indistinguishable. It is convenient to create pairs of nodes as above to exploit the way we will identify indistinguishable pairs of vertices. The problem is reduced to a control of each term of the summand. For a labelled tree T and a permutation γ, we denote by T^γ the tree with γ applied to its labels. For $j \geq \lceil n/2 \rceil$, fix $\gamma = (\tau(j), \tau(\lfloor n/4 \rfloor + j))$, that is, the permutation sending j to $\lfloor n/4 \rfloor + j$ and vice versa, while keeping all other elements of $[n]$ in place. Introduce the event

$$\Omega_j := \{\tau(j) \text{ and } \tau(\lfloor n/4 \rfloor + j) \text{ are leaves, connected to vertices of rank } \leq n/2\} \,.$$

First, we check that Ω_j is an event whose probability is bounded away from 0. We note that for Ω_j to occur, it suffices that

- vertex j connects to a vertex of rank at most $n/2$. This happens with probability $\lfloor n/2 \rfloor/(j-1)$.
- For times ranging from $j+1$ to $\lfloor n/4 \rfloor + j - 1$ new vertices connect to vertices different from j. This happens with probability

$$\prod_{k=j+1}^{\lfloor n/4 \rfloor + j - 1} \frac{k-2}{k-1}.$$

- Vertex $\lfloor n/4 \rfloor + j$ connects to a vertex of rank at most $n/2$. This happens with probability $\lfloor n/2 \rfloor/(\lfloor n/4 \rfloor + j - 1)$.
- For times ranging from $\lfloor n/4 \rfloor + j + 1$ to n new vertices connect to vertices different from j and $\lfloor n/4 \rfloor + j$. This happens with probability

$$\prod_{k=\lfloor n/4 \rfloor + j + 1}^{n} \frac{k-3}{k-1}.$$

Finally, note that from the definition of the urrt model, the four events corresponding to the four items above are independent. Thus

$$\begin{aligned}
\mathbb{P}\left\{\Omega_j\right\} &= \frac{\lfloor n/2 \rfloor}{j-1} \cdot \frac{\lfloor n/2 \rfloor}{\lfloor n/4 \rfloor + j} \cdot \prod_{k=j+1}^{\lfloor n/4 \rfloor + j - 1} \frac{k-2}{k-1} \cdot \prod_{k=\lfloor n/4 \rfloor + j + 1}^{n} \frac{k-3}{k-1} \\
&= \frac{\lfloor n/2 \rfloor}{j-1} \cdot \frac{\lfloor n/2 \rfloor}{\lfloor n/4 \rfloor + j} \cdot \frac{j-2}{\lfloor n/4 \rfloor + j - 2} \cdot \frac{(\lfloor n/4 \rfloor + j - 2)(\lfloor n/4 \rfloor + j - 1)}{(n-1)n},
\end{aligned}$$

which simplifies to

$$\mathbb{P}\left\{\Omega_j\right\} \geq \frac{1}{4}\left(1 - \frac{1}{n-2}\right)^3. \tag{4.2.3}$$

The first step is to use (4.2.3) to control one of the summands in (4.2.1) by conditioning on Ω_j.

$$\begin{aligned}
&\mathbb{E}\left[|\widehat{\sigma} \circ \tau(j) - j| + |\widehat{\sigma} \circ \tau(\lfloor n/4 \rfloor + j) - (\lfloor n/4 \rfloor + j)|\right] \\
&\quad \geq \quad \frac{1}{4}\left(1 - \frac{1}{n-2}\right)^3 \mathbb{E}\left[\left|\widehat{\sigma} \circ \tau(j) - j\right| + \left|\widehat{\sigma} \circ \tau\left(\left\lfloor \frac{n}{4} \right\rfloor + j\right) - \left(\left\lfloor \frac{n}{4} \right\rfloor + j\right)\right| \,\middle|\, \Omega_j\right].
\end{aligned}$$

We then decompose on each possible realization of a recursive tree

$$\begin{aligned}
&\mathbb{E}\left[\left|\widehat{\sigma}\circ\tau(j)-j\right|+\left|\widehat{\sigma}\circ\tau\left(\left\lfloor\frac{n}{4}\right\rfloor+j\right)-\left(\left\lfloor\frac{n}{4}\right\rfloor+j\right)\right|\,\middle|\,\Omega_j\right]\\
&=\sum_{t\in\mathbb{T}}\mathbb{P}\left\{T=t\mid\Omega_j\right\}\mathbb{E}\left[\left|\widehat{\sigma}\circ\tau(j)-j\right|\mid\Omega_j,\,T=t\right]\\
&\quad+\sum_{t\in\mathbb{T}}\mathbb{P}\left\{T=t^{\gamma}\mid\Omega_j\right\}\mathbb{E}\left[\left|\widehat{\sigma}\circ\tau\left(\left\lfloor\frac{n}{4}\right\rfloor+j\right)-\left(\left\lfloor\frac{n}{4}\right\rfloor+j\right)\right|\,\middle|\,\Omega_j,\,T=t^{\gamma}\right],
\end{aligned}$$

which is a valid decomposition since $t\mapsto t^{\gamma}$ is a bijection from $\mathbb{T}$ to itself. Theorem 4 of Crane and Xu [42] states that, in the urrt model, two trees having the same shape but different recursive orders have the same probability. Since on the event Ω_j, t is recursive if and only if t^{γ} is recursive, then

$$\mathbb{P}\left\{T=t\mid\Omega_j\right\}=\mathbb{P}\left\{T=t^{\gamma}\mid\Omega_j\right\}.$$

As a consequence, the above expression factorizes to

$$\begin{aligned}
&\mathbb{E}\left[\left|\widehat{\sigma}\circ\tau(j)-j\right|+\left|\widehat{\sigma}\circ\tau\left(\left\lfloor\frac{n}{4}\right\rfloor+j\right)-\left\lfloor\frac{n}{4}\right\rfloor-j\right|\,\middle|\,\Omega_j\right]=\sum_{t\in\mathbb{T}}\mathbb{P}\left\{T=t\mid\Omega_j\right\}\times\\
&\left(\mathbb{E}\left[\left|\widehat{\sigma}\circ\tau(j)-j\right|\mid\Omega_j,\,T=t\right]+\mathbb{E}\left[\left|\widehat{\sigma}\circ\tau\left(\left\lfloor\frac{n}{4}\right\rfloor+j\right)-\left(\left\lfloor\frac{n}{4}\right\rfloor+j\right)\right|\,\middle|\,\Omega_j,\,T=t^{\gamma}\right]\right).
\end{aligned}\tag{4.2.4}$$

The label invariant condition implies that

$$\widehat{\sigma}[T^{\gamma}]\circ\gamma\overset{\mathcal{L}}{=}\widehat{\sigma}[T],$$

and in particular,

$$\left(\widehat{\sigma}(j)\mid\Omega_j,T=t\right)\overset{\mathcal{L}}{=}\left(\widehat{\sigma}(\lfloor n/4\rfloor+j)\mid\Omega_j,T=t^{\gamma}\right),$$

which directly implies that

$$\mathbb{E}\left[\left|\widehat{\sigma}\circ\tau(j)-j\right|\mid\Omega_j,\,T=t\right]\;+\;\mathbb{E}\left[\left|\widehat{\sigma}\circ\tau\left(\left\lfloor\frac{n}{4}\right\rfloor+j\right)-\left\lfloor\frac{n}{4}\right\rfloor-j\right|\,\middle|\,\Omega_j,\,T=t^{\gamma}\right]\quad\geq\quad\frac{n}{4}\quad.$$

By plugging the above inequality in (4.2.4)

$$\mathbb{E}\left[\left|\widehat{\sigma}\circ\tau(j)-j\right|+\left|\widehat{\sigma}\circ\tau\left(\left\lfloor\frac{n}{4}\right\rfloor+j\right)-\left\lfloor\frac{n}{4}\right\rfloor-j\right|\right]\geq\frac{n}{16}\left(1-\frac{1}{n-2}\right)^{3}.$$

Now, plugging the above inequality in (4.2.1) yields

$$R_{\alpha}(\widehat{\sigma})\geq\frac{n^{2-\alpha}}{65},$$

for all $n\geq 200$. ∎

Remark. In the urrt, vertices 1 and 2 are indistinguishable. Indeed, when the tree has size 2, vertices 1 and 2 have exactly the same properties. Thus, no label invariant ordering procedure can assign order 1 to vertex 1 with probability higher than $1/2$. As a result, we obtain, for any α, the trivial lower bound

$$R^*_\alpha \geq \frac{1}{2}\,,$$

which improves the bound of Theorem 4.1 for $\alpha \geq 2$, and therefore Theorem 4.1 is non-trivial when $\alpha < 2$.

4.2.2 An auxiliary "descendant-ordering" procedure

In the sequel, we estabish upper bounds for the risk of Jordan ordering. Since this is a label-invariant procedure, we may assume, without loss of generality, that $\sigma = \mathrm{Id}$ is the identity permutation. In other words, the arrival time of a vertex and its label are the same. When the context is clear, vertex labels and arrival times are used interchangeably.

In order to analyze the Jordan ordering, we introduce an auxiliary centrality measure and the corresponding estimator of vertex arrival times. As observed in the introduction, the Jordan ordering procedure consists in estimating the position of vertex 1 by the Jordan centroid c and ordering vertices according to the values of $|(T,c)_u|$. If c was replaced by vertex 1, this measure would correspond to the number of descendants of u. Thus, a natural ordering is to order vertices by the number of their descendants, that is, the ordering according to the values of $|(T,1)_u|$. We call this *descendant ordering*, noting that, as before, ties are broken at random. Note that descendant ordering is not a valid procedure, since the location of the root vertex is not known. On the other hand, the number of descendants is easily analyzed by Pólya urns, and our approach is based on comparing Jordan ordering to this auxiliary procedure. In this section we prove an upper bound for difference of the risk of both procedures. For a tree T and for each $u \in T$, we define the descendant centrality

$$\psi'_T(u) = n - \mathrm{de}(u)\,,$$

where $\mathrm{de}(u) = |(T,1)_u|$ is the number of descendants of u, as defined in Section 4.1.2. We denote by $\widehat{\sigma}'$ the ordering of the vertices induced by sorting the values of ψ'_T in increasing order.

In the following lemma, we first prove that for the urrt model, the Jordan centrality ψ_T, defined in (4.1.3), and the descendant centrality ψ'_T coincide for most vertices. Furthermore, we prove bounds on both the number of nodes for which ψ_T may differ from

ψ'_T and the estimated rank of vertex 1. We recall that 1 and c denote respectively the root and the rank of a centroid of the tree.

Lemma 4.2. *Let $T \sim$ urrt, let $c \in [n]$ be a centroid of T and let $\{1 \to c\}$ be the set of vertices on the path connecting 1 to c in T. Then*

- *for any $v \in [n] \backslash \{1 \to c\}$, we have*

$$\psi_T(v) = \psi'_T(v) \; ;$$

- *there exists a universal constant K such that c is stochastically dominated by an exponential random variable with mean K;*
- *for $\epsilon \leq 0.2$, with probability at least $1 - 5\epsilon$*

$$\widehat{\sigma_J}(1) \leq 2.5 \frac{\log(1/\epsilon)}{\epsilon} \; .$$

Proof. Let $T \sim$ urrt. First, we decompose the vertices of T in four sets as shown in Figure 4.2: case 1 corresponds to the set $\{1 \to c\}$ of nodes connecting the root to the centroid, case 2 to the vertices of $(T,1)_c \setminus \{c\}$, case 3 to the vertices of $(T,c)_1 \setminus \{1\}$ and finally case 4 to the vertices of $(T,1)_i \setminus \{i\}$ for $i \in \{1 \to c\} \setminus \{1,c\}$. As we mentioned before, it is well known that for a non-centroid vertex u, its neighbor maximizing $|(T,u)_v|$ is such that $|(T,u)_v|$ contains any centroid. Note that for each vertex u in cases 2, 3 and 4, for $v \sim u$ such that the subtree $(T,u)_v$ contains c, $(T,u)_v$ also contains vertex 1. As a consequence, $\psi_T(u) = |(T,u)_{\mathrm{pa}(u)}|$, where $\mathrm{pa}(u)$ is the "parent" of u. But by definition, $|(T,u)_{\mathrm{pa}(u)}| = n - \mathrm{de}(u) = \psi'_T(u)$, concluding the proof of the first part of the lemma.

Moon [109] showed that the rank of the centroid c is dominated by an exponential random variable of mean K, for a universal constant. The third statement follows from Theorem 3 of Bubeck et al. [33]. ∎

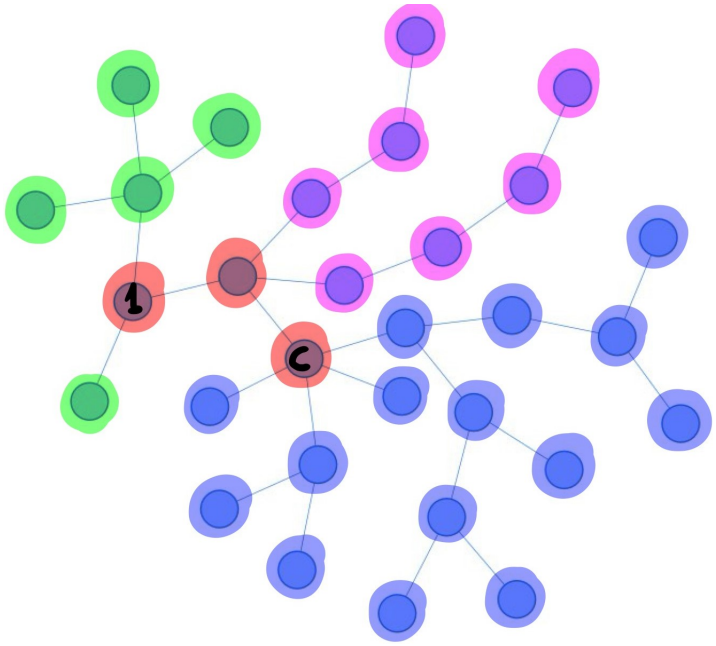

Figure 4.2: Sketch of a tree and its centroid. Circled in red are the vertices of the path $\{1 \to c\}$ (case 1). Blue vertices correspond to case 2, green to case 3 and purple vertices to case 4.

In the following lemma, we bound the risk of $\widehat{\sigma_J}$ by that of the descendant ordering $\widehat{\sigma'}$.

Lemma 4.3. *Let* $T \sim$ urrt. *For* $\alpha > 0$

$$R_\alpha(\widehat{\sigma_J}) \le R_\alpha(\widehat{\sigma'}) + K \sum_{i=1}^{n} \frac{1}{i^\alpha} + C \log^4(n)\,,$$

where $C > 0$ *is a constant (not depending on* α*).*

Proof. Recall that $\sigma = \mathrm{Id}$, that is, we use the same integer to denote the label of a vertex and its arrival time. We first decompose the global risk $R_\alpha(\widehat{\sigma_J})$ into

$$R_\alpha(\widehat{\sigma_J}) = \mathbb{E}\left[\sum_{i \in \{1 \to c\}} \frac{\left|\widehat{\sigma_J}(i) - i\right|}{i^\alpha}\right] + \mathbb{E}\left[\sum_{i \notin \{1 \to c\}} \frac{\left|\widehat{\sigma_J}(i) - i\right|}{i^\alpha}\right].$$

Since by Lemma 4.2 vertices outside of the path $\{1 \to c\}$ are put in the same order by the Jordan ordering and the descendant ordering, for $i \notin \{1 \to c\}$, $|\widehat{\sigma_J}(i) - \widehat{\sigma}(i)| \leq D+1$, where D is the distance between 1 and c. Thus, we can control the second term of the right-hand side as follows:

$$\mathbb{E}\left[\sum_{i \notin \{1\to c\}} \frac{\left|\widehat{\sigma_J}(i) - i\right|}{i^\alpha}\right] = \mathbb{E}\left[\sum_{i \notin \{1\to c\}} \frac{\left|\widehat{\sigma'}(i) - i + \widehat{\sigma_J}(i) - \widehat{\sigma'}(i)\right|}{i^\alpha}\right]$$
$$\leq \mathbb{E}[D] \sum_{i=1}^{n} \frac{1}{i^\alpha} + \mathbb{E}\left[\sum_{i \notin \{1\to c\}} \frac{|\widehat{\sigma'}(i) - i|}{i^\alpha}\right].$$

Since D is at most the arrival time of the centroid, Lemma 4.2 implies that $\mathbb{E}[D] \leq \mathbb{E}[c] \leq K$. On the other hand,

$$\mathbb{E}\left[\sum_{i \in \{1\to c\}} \frac{\left|\widehat{\sigma_J}(i) - i\right|}{i^\alpha}\right] \leq \mathbb{E}\left[\sum_{i \in \{1\to c\}} i\right] + \mathbb{E}\left[\sum_{i \in \{1\to c\}} \widehat{\sigma_J}(i)\right].$$

Clearly,

$$\mathbb{E}\left[\sum_{i \in \{1\to c\}} i\right] \leq \mathbb{E}[cD].$$

Since $D \leq c$ and since c is dominated by an exponential random variable of mean K, by Lemma 4.2,

$$\mathbb{E}[cD] \leq \mathbb{E}\left[c^2\right] \leq 2K^2.$$

In addition, since on the path $\{1 \to c\}$, $\widehat{\sigma_J}$ is decreasing, we have

$$\mathbb{E}\left[\sum_{i \in \{1\to c\}} \widehat{\sigma_J}(i)\right] \leq \mathbb{E}\left[D\widehat{\sigma_J}(1)\right].$$

Since D and $\widehat{\sigma_J}(1)$ are bounded by n, they have finite moments. Using Hölder's inequality, for any $\gamma > 0$,

$$\mathbb{E}\left[D\widehat{\sigma_J}(1)\right] \leq \left(\mathbb{E}\left[D^{\frac{1+\gamma}{\gamma}}\right]\right)^{\frac{\gamma}{1+\gamma}} \left(\mathbb{E}\left[\widehat{\sigma_J}(1)^{1+\gamma}\right]\right)^{\frac{1}{1+\gamma}}. \tag{4.2.5}$$

Since $D \leq c$ which is dominated by an exponential random variable,

$$\left(\mathbb{E}\left[D^{\frac{1+\gamma}{\gamma}}\right]\right)^{\frac{\gamma}{1+\gamma}} \leq C\frac{1+\gamma}{\gamma}, \tag{4.2.6}$$

for some positive constant C. Next, using Lemma 4.2,

$$\mathbb{P}\left\{\widehat{\sigma_J}(1) \geq f(\epsilon)\right\} \leq 5\epsilon,$$

where $f(\epsilon) = 2.5\frac{\log(1/\epsilon)}{\epsilon}$. f is a non-increasing function and therefore $f\left(5\log^2(k)/k\right) \leq k$ for all $k \geq 1$. Therefore,

$$\mathbb{P}\left\{\widehat{\sigma_J}(1) \geq k\right\} \leq 25\frac{\log^2(k)}{k}\,, \quad \text{for all } k \geq 1,$$

so for any $\gamma > 0$,

$$\mathbb{P}\left\{\widehat{\sigma_J}(1)^{1+\gamma} \geq k\right\} = \mathbb{P}\left\{\widehat{\sigma_J}(1) \geq k^{\frac{1}{1+\gamma}}\right\} \leq 25\frac{\frac{1}{(1+\gamma)^2}\log^2(k)}{k^{\frac{1}{1+\gamma}}}\,.$$

It follows that

$$\begin{aligned}
\mathbb{E}\left[\widehat{\sigma_J}(1)^{1+\gamma}\right] &= 1 + \int_{k=1}^{n^{1+\gamma}} \mathbb{P}\left\{\widehat{\sigma_J}(1)^{1+\gamma} \geq k\right\}\, dk \leq 1 + \int_{k=1}^{n^{1+\gamma}} \frac{\frac{25}{(1+\gamma)^2}\log^2(k)}{k^{\frac{1}{1+\gamma}}}\, dk \\
&\leq 1 + \frac{25}{(1+\gamma)^2}(1+\gamma)^2\log^2(n)\frac{\gamma+1}{\gamma}\left(n^{1+\gamma}\right)^{\frac{\gamma}{1+\gamma}} \\
&= 1 + 25\frac{(1+\gamma)\log^2(n)}{\gamma}n^{\gamma}\,.
\end{aligned}$$

Plugging the obtained inequality in (4.2.5) and recalling (4.2.6), we obtain

$$\mathbb{E}\left[D\widehat{\sigma_J}(1)\right] \leq C\frac{1+\gamma}{\gamma}\left(1 + 25\frac{1+\gamma}{\gamma}\log^2(n)n^{\gamma}\right).$$

Choosing $\gamma = 1/\log(n)$ we get

$$\mathbb{E}\left[D\widehat{\sigma_J}(1)\right] \leq C'\log^4(n)\,.$$

This concludes the proof of the lemma. ■

4.2.3 Performance of Jordan ordering in the URRT model

In this section, we prove upper bounds for the risk $R_\alpha(\widehat{\sigma_J})$. In particular, we prove that for $\alpha \in [1, 2)$, the risk $R_\alpha(\widehat{\sigma_J})$ has the same order as the optimal risk R^*_α, defined in Section 4.2.1.

Theorem 4.4. *Let $T \sim$ urrt. Then there exist positive constants C, K such that for $1 \leq \alpha < 2$*

$$R_\alpha(\widehat{\sigma_J}) \leq K(\alpha) n^{2-\alpha} + K \sum_{i=1}^{n} \frac{1}{i^\alpha} + C \log^4(n) \,,$$

where $K(\alpha) = \left(\frac{2}{2-\alpha} + \frac{2e^2}{(2-\alpha)^2} + \frac{2}{(2-\alpha)^3} \right)$. Moreover, for $\alpha \geq 2$

$$R_\alpha(\widehat{\sigma_J}) \leq C \log^4(n) \,.$$

Before proving Theorem 4.4, we state a corollary that is a direct consequence of Theorems 4.1 and 4.4. This corollary notably implies that the Jordan ordering has a risk of optimal order for $\alpha \in [1, 2)$. Note that one cannot hope to match the established lower bounds for the optimal risk in a broader range of α for this method. Indeed, in a uniform random recursive tree of size n, the probability that vertex 1 is a leaf is $1/n$. Since leaves are ordered last by $\widehat{\sigma_J}$, and that there are roughly $n/2$ leaves, $\mathbb{P}\left\{\widehat{\sigma_J}(1) \geq n/2\right\} \approx 1/n$. This implies that $\mathbb{E}[\widehat{\sigma_J}(1)] \gtrsim \log(n)/2$, so $\widehat{\sigma_J}$ has a risk of order at least $\log(n)$, while the lower bound is of constant order for $\alpha \geq 2$. We discuss in Appendix 4.5.2 the possibility of estimating the position of vertex 1 better, namely using the rumor centroid. We conjecture this alternative method has a risk of optimal order for any $\alpha \geq 1$.

Corollary 4.5. *Let $T \sim$ urrt. For $\alpha = 1$*

$$R_\alpha(\widehat{\sigma_J}) \leq (1 + o(1)) 1170 R_1^* \,,$$

and for $\alpha \in (1, 2)$,

$$R_\alpha(\widehat{\sigma_J}) \leq (1 + o(1)) \left(\frac{1}{2-\alpha} + \frac{3}{(2-\alpha)^2} + \frac{1}{(2-\alpha)^3} \right) 65 R_\alpha^* \,.$$

Proof of Theorem 4.4. By the triangle inequality,

$$R_\alpha(\widehat{\sigma'}) \leq \sum_{i=1}^{n} \mathbb{E}\left[\frac{\widehat{\sigma'}(i)}{i^\alpha} \right] + \sum_{i=1}^{n} \frac{i}{i^\alpha} \,,$$

Let $i \in [n]$ be a vertex of T. We first note that

$$\mathbb{E}\left[\widehat{\sigma'}(i)\right] \leq \mathbb{E}\left[\sum_{j:\ j \neq i} \mathbb{1}_{\mathrm{de}(j) \geq \mathrm{de}(i)} \right] . \tag{4.2.7}$$

Moreover, for any $\tau_{i,j} \in \mathbb{R}$,

$$\mathbb{P}\{\mathrm{de}(j) \geq \mathrm{de}(i)\} \leq \mathbb{P}\left\{\frac{\mathrm{de}(j)}{n} \geq \tau_{i,j}\right\} + \mathbb{P}\left\{\frac{\mathrm{de}(i)}{n} \leq \tau_{i,j}\right\}.$$

Therefore, we may upper bound (4.2.7) by

$$\mathbb{E}[\widetilde{\sigma}'(i)] \leq i + 1 + \sum_{j>i+1} \mathbb{P}\left\{\frac{\mathrm{de}(j)}{n} \geq \tau_{i,j}\right\} + \mathbb{P}\left\{\frac{\mathrm{de}(i)}{n} \leq \tau_{i,j}\right\}. \tag{4.2.8}$$

Let $j > i+1$ be a vertex of T. Note that the distributions of $\mathrm{de}(i)$ and $\mathrm{de}(j)$ in a urrt model follow a Pólya urn model. In particular, for any vertex $k \in [n]$,

$$\mathbb{P}\{k \in (T,1)_v\} = \frac{de_{k-1}(j)+1}{k-1},$$

and each connection of a new vertex to a descendant of j is independent of the previous ones, conditionally on $\mathrm{de}(j)$. Let $N_n := n - j$ and let $\widetilde{W}_N = \mathrm{de}(j)$ be the number of descendants of j at time n. We thus have that $\widetilde{W}_N = \mathrm{de}(j)$ follows a Pólya urn distribution, where the Pólya urn process has balls of two colours, it is started when j is added to T, and it is run for a maximum number of steps N_n. From Mahmoud [105, Section 3.2], we have that

$$\mathbb{E}[\mathrm{de}(j)+1] \;=\; \frac{n}{j},$$

and also that

$$\mathbb{P}\{\mathrm{de}(j) = k\} \;=\; \frac{k!(j-1)(j)\cdots(n-k-2)}{j(j+1)\cdots(n-1)}\binom{n-j}{k}. \tag{4.2.9}$$

In Appendix 4.5.4, we derive from these formulas the following upper-bounds.

Lemma 4.6. *$i \geq 2$ and $j > i+1$, by choosing $\tau_{i,j} = \frac{1}{j}\log\frac{j}{i}$,*

$$\mathbb{P}\left\{\frac{\mathrm{de}(j)}{n} \geq \tau_{i,j}\right\} + \mathbb{P}\left\{\frac{\mathrm{de}(i)}{n} \leq \tau_{i,j}\right\} \leq 2e^2 e^{-\log\frac{j}{i}} + \frac{i}{j}\log\frac{j}{i} \leq \frac{i}{j}\left(2e^2 + \log\frac{j}{i}\right),$$

and that for $i = 1$, $j > i+1$, choosing $\tau_{1,j} = \frac{1}{j}\log(j)$,

$$\mathbb{P}\left\{\frac{\mathrm{de}(j)}{n} \geq \tau_{1,j}\right\} + \mathbb{P}\left\{\frac{\mathrm{de}(1)}{n} \leq \tau_{1,j}\right\} \leq \frac{1}{j}\left(2e^2 + \log(j)\right) + \frac{1}{n-1}.$$

Once plugged into the expression of R_α, for $n \geq 60$, this leads to (details in Appendix 4.5.4)

$$\sum_{i=1}^{n} \mathbb{E}\left[\frac{|\widehat{\sigma}'(i) - i|}{i}\right] \leq 18n\,.$$

For $1 < \alpha < 2$, a similar computation yields (details in Appendix 4.5.4)

$$\sum_{i=1}^{n} \mathbb{E}\left[\frac{|\widehat{\sigma}'(i) - i|}{i^\alpha}\right] \leq \left(\frac{2}{2-\alpha} + \frac{2e^2}{(2-\alpha)^2} + \frac{2}{(2-\alpha)^3}\right) n^{2-\alpha}\,.$$

Lemma 4.3 concludes the proof.

4.3 Preferential attachment tree

In this section, we consider the preferential attachment model and investigate the performance of the Jordan ordering procedure. Since the arguments have a similar structure to the urrt model analyzed in Section 4.2, we omit some details of the proofs and report them to the Appendices. Similarly to the previous section, we first prove a minimax lower bound for the risk of any label-invariant estimator.

4.3.1 A lower bound

Theorem 4.7. *In the* pa *model, we have, for* $\alpha = 1$ *and* $n \geq 300$

$$R^*_\alpha \geq \frac{n^{2-\alpha}}{70}\,.$$

The proof is deferred to Appendix 4.5.5.

Remark. In the same way as in the case of the urrt model, we have

$$R^*_\alpha \geq \frac{1}{2}\,,$$

which is better than the result of Theorem 4.7 for $\alpha > 2$.

4.3.2 Performance of the Jordan ordering in the PA model

Similarly to Section 4.2.3, we establish upper bounds for $R_\alpha(\widehat{\sigma_J})$. In a subsequent corollary, we bound the risk $R_\alpha(\widehat{\sigma_J})$ in terms of the optimal risk R^*_α.

Theorem 4.8. *Let $T \sim$ PA. Then, there exist positive constants C, K, such that for $\alpha \in [1, 5/4)$*

$$R_\alpha(\widehat{\sigma_J}) \leq \left(\frac{2}{2-\alpha} + \frac{1}{(\alpha - 5/4)(\alpha - 2)}\right) n^{2-\alpha} + K \sum_{i=1}^{n} \frac{1}{i^\alpha} + C \log^2(n)\sqrt{n}\,.$$

For $\alpha \geq 5/4$,

$$R_\alpha(\widehat{\sigma_J}) \leq \frac{2}{2-\alpha} n^{2-\alpha} + \frac{32}{3}\zeta\left(\alpha - \frac{1}{4}\right) n^{3/4}\,,$$

where ζ denotes the Riemann zeta function.

Corollary 4.9 is a direct consequence of Theorems 4.7 and 4.8. It states that the Jordan ordering has a risk of optimal order for $\alpha \in [1, 5/4)$. Let us remark that, here, the boundary value $5/4$ does not appear for the same reason as in the urrt case. In the urrt, the optimality result is limited to $\alpha < 2$ because of the error originating from the estimation of vertex 1. Here, the limitation to $\alpha < 5/4$ has a different origin than in the URRT model. Indeed, our analysis of the descendant ordering only proves that the risk is optimal up to canstant factor for $\alpha < 5/4$. It means that even if the position of vertex 1 was known, ordering vertices by the number of their descendants would not result in a risk bound that matches the lower bound for $\alpha \geq 5/4$.

Corollary 4.9. *Let $T \sim$ PA. For $\alpha \in [1, 5/4)$*

$$R_\alpha(\widehat{\sigma_J}) \leq (1 + o(1))\, 70 \left(\frac{2}{2-\alpha} + \frac{1}{(\alpha - 5/4)(\alpha - 2)}\right) R^*_\alpha\,.$$

The proof of Theorem 4.8 is reported to Appendix 4.5.7.

4.4 Simulations

In this section, we first report a numerical illustration of our theoretical results on trees generated from the urrt and pa models. Then, we compare the performance of the Jordan

ordering to other ordering procedures. For computational reasons, we display results for the descendant ordering procedure. The descendant ordering can be computed in time $\mathcal{O}(n\log(n))$. Also, one can find the Jordan centroid in linear time.

Note that the bounds of Lemmas 4.3 and 4.11 show that the risk of the descendant ordering is a good approximation of the risk of Jordan ordering.

In the first experiment, we compute the risk $R_\alpha(\widehat{\sigma'})$ (see (4.1.2)) of the descendant ordering and display the theoretical upper bound and minimax lower bound from Theorems 4.1 and 4.4 (Theorems 4.7 and 4.8 in the pa model).

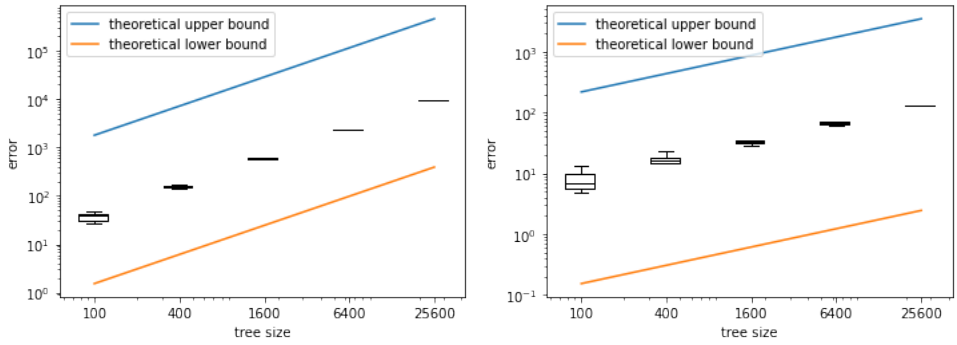

Figure 4.3: Risk R_α of the descendant ordering versus the tree size n in logarithmic scales, for $\alpha = 1$ (left panel) and for $\alpha = 1.5$ (right panel), and for trees simulated from the urrt model. Here, we sample 10 trees for each size, and report a boxplot with the median, first, and last quartiles, for each tree size.

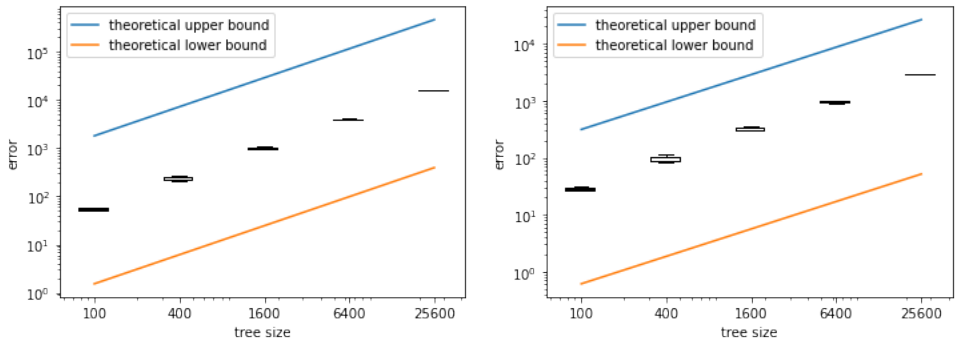

Figure 4.4: Risk R_α of the descendant ordering versus the tree size n in logarithmic scales, for $\alpha = 1$ (left panel) and for $\alpha = 1.2$ (right panel), and for trees simulated from the pa model. Here, we sample 10 trees for each size, and report a boxplot with the median, first, and last quartiles, for each tree size.

In the second experiment, we perform an empirical comparison of Jordan ordering with the three following ordering methods:

- **Degree** ordering, which orders the vertices by decreasing degree. Again, we break ties at random. Degree ordering is justified by the fact that the lower the rank of a vertex, the higher its expected degree is. Note, however, that ordering vertices by degree does not necessarily produce a recursive ordering.
- **Spectral** method by Recanati et al. [122]. This method is widely used in seriation problems and consists of finding the eigenvector associated to the second smallest eigenvalue of the Laplacian of the graph. Then, considering the entries of this eigenvector as a score function, the estimated ordering is derived by sorting these entries by increasing values.
- **Reverse DMC** algorithm, proposed by Navlakha and Kingsford [116]. This algorithm is analogous to a pruning method, which consists of ordering the vertices by sequentially removing all leaves from the tree and ordering the leaves removed at each step. In Reverse DMC, a score is computed for each leaf and the algorithm sequentially removes the leaf with the highest score. This score function corresponds to the likelihood of the leaf being the last vertex in the current tree, therefore, at each step, the leaf which is the most likely to be the last vertex arrived in the tree is removed.

Remark. We note that the spectral method is a reasonable method to compare with in our setting of recursive trees since (i) spectral methods recovers the order of a Robinson matrix Recanati et al. [122], and (ii) in the urrt and pa models, the expected value of the adjacency matrix is a Robinson matrix.

Similarly to the previous experiment, we compute the risk R_α for the four methods, on trees simulated from the urrt or pa models, in multiple settings. From Figures 4.5 and 4.6 we see that the Jordan estimator has the lowest risk, for all values of the trees sizes, and that the degree method is the second best one. In fact, it is not surprising that the degree method performs well for pa trees, since, in this model, the degree has a power law distribution and the order by degree correlates well with the arrival times of the vertices. However, this result is more surprising for the urrt model, where degree-centrality is known to be sub-optimal in the root-finding problem, see Bubeck et al. [33]. This is discussed further in Appendix 4.5.3. Moreover, the spectral method has the poorest performance in both models. A possible explanation for this is that the random fluctuations of the adjacency matrix in the considered recursive tree models are large, leading to a large difference between the expected and empirical adjacency matrices in spectral norm. This absence of concentration, which is generally required in spectral ordering methods, could explain why this method poorly performs in our setting. Finally, the Reverse DMC algorithm performs similarly poorly as the spectral method in the pa model.

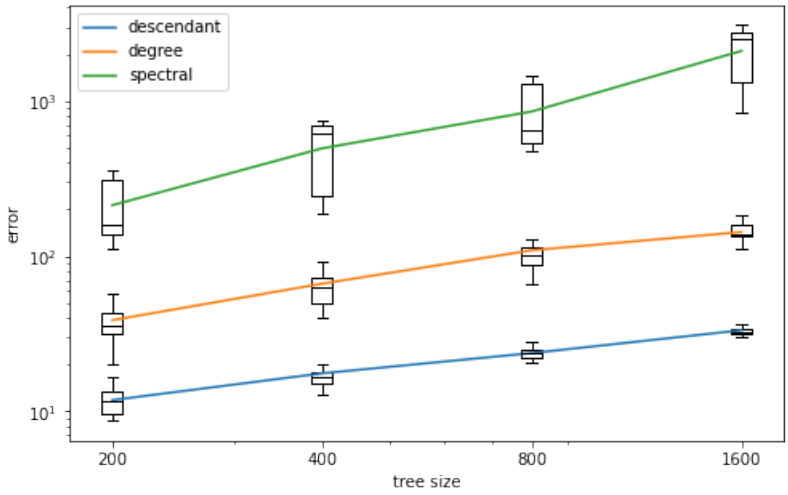

Figure 4.5: Risk R_α versus the tree size n in logarithmic scales, for $\alpha = 1.5$, and for trees simulated from the urrt model. Here, we sample 10 trees for each size. We compare the risk of descendant (blue), degree (orange), and spectral methods (green), and report a boxplot with the median, first, and last quartiles, for each tree size. In all settings, the descendant ordering largely outperforms the other methods.

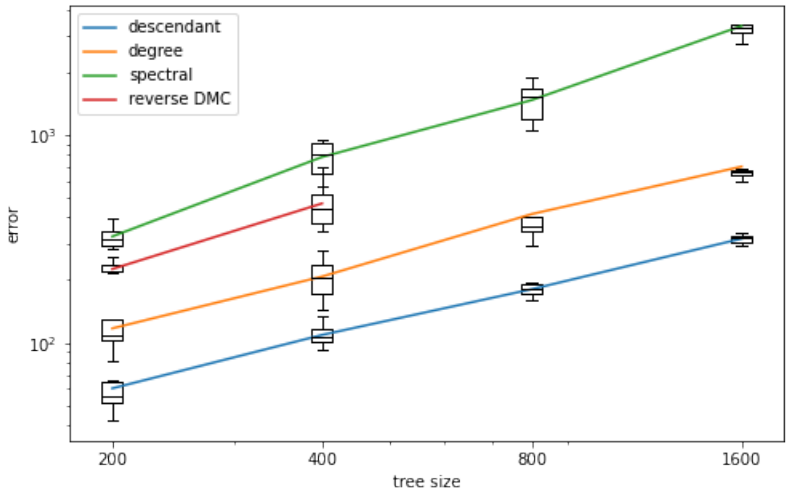

Figure 4.6: Risk R_α versus the tree size n in logarithmic scales, for $\alpha = 1.2$, and for trees simulated from the pa model. Here, we sample 10 trees for each size, only considering small trees for the reverse DMC method due to its high computational cost. We compare the risk of descendant (blue), degree (orange), spectral (green), and reverse DMC (red) methods, and report a boxplot with the median, first, and last quartiles, for each tree size. Just like in the urrt model, the descendant ordering outperforms the other methods.

4.5 Appendix

In this appendix we discuss some issues concerning the choice of the parameter α in the definition of the risk and ordering according to degrees. Some elements of the proofs for the urrt model are also reported here. The final sections contain the proofs of all results in the pa model.

4.5.1 A remark on the choice selection of α

The risk R_α defined in (4.1.2) leads to a meaningful performance measure in the urrt and pa models only for some values of α. In particular, for $\alpha < 1$, it is easy to see that the risk of a random permutation is of the same order as the established lower bound, both in the urrt and pa models. More precisely, let Σ be a permutation chosen uniformly at random. Simple computation for $\alpha < 1$ leads to

$$R_\alpha(\Sigma) \leq c_\alpha n^{2-\alpha} ,$$

for some positive constant c_α. On the other hand, Theorems 4.1 and 4.7 imply that for $\alpha < 1$, $R^*_\alpha \geq c'_\alpha n^{2-\alpha}$. Therefore, with $\alpha < 1$, the a random ordering has a risk of the same order as that of the optimal one. This is why we restrict our analysis of the risk to $\alpha \geq 1$. Our analysis of the Jordan ordering proves that this method has a risk of optimal order for $\alpha \in [1, 2)$ in the urrt case and $\alpha \in [1, 5/4)$ in the pa tree.

4.5.2 A remark on rumor centrality

We conjecture that in the urrt model, there exists an ordering procedure whose risk is of the order of $n^{2-\alpha}$ for any $\alpha \geq 1$, matching that of the minimax lower bound. Indeed, in our analysis, the risk is decomposed in two parts. First, a part coming from the difference between the Jordan and the descendant ordering (i.e, the error made by estimating the position of vertex 1 by the Jordan centre), second the risk of the descendant ordering. A possible way to improve our bound on the risk is to estimate the position of vertex 1 more precisely. To do so, using the rumor centrality appears to be a promising option. Indeed, due to recent results from Crane and Xu [42], in the urrt model, the rumor centrality orders vertices by their likelihood of being vertex 1. In particular, using the rumor centrality is optimal for minimizing the size of a conficence set containing the root, outperforming Jordan centrality (Bubeck et al. [33]). However, one step in the analysis is missing. Copying the proof of Lemma 4.3, with the rumor center instead of the Jordan center, one needs to

bound the moment of order $1+\gamma$ of the arrival time of the rumor centre. Bounding it by a constant (for any value of γ) would be sufficient to prove that this new ordering procedure has a risk of optimal order for any $\alpha \geq 1$.

4.5.3 A remark on ordering by degree

As discussed in Section 4.4, a simple ordering procedure is by the degrees. Simulations suggest that it does not perform as well as Jordan ordering, and it may produces non-recursive ordering. Nonetheless, it is a simple procedure worth mentioning. Since in a pa tree the degree of a given node follows a Pólya urn distribution, analysing the performance of the degree ordering is similar to the analysis carried out for Jordan centrality. However, the simulations results displayed in Figure 4.7 suggest that for any $\alpha \in [1, 5/4)$ the risk of the degree ordering grows at a faster rate than $n^{2-\alpha}$. Both in the URRT and PA model, for sizes of trees $\{1000, 2000, 4000, 8000\}$, we sample 10 trees at each size and compute the risk of the descendant and degree ordering for different values of α. Then, for each value of α, we perform a linear regression on the log-plot of the risk, to estimate the exponent of the polynomial.

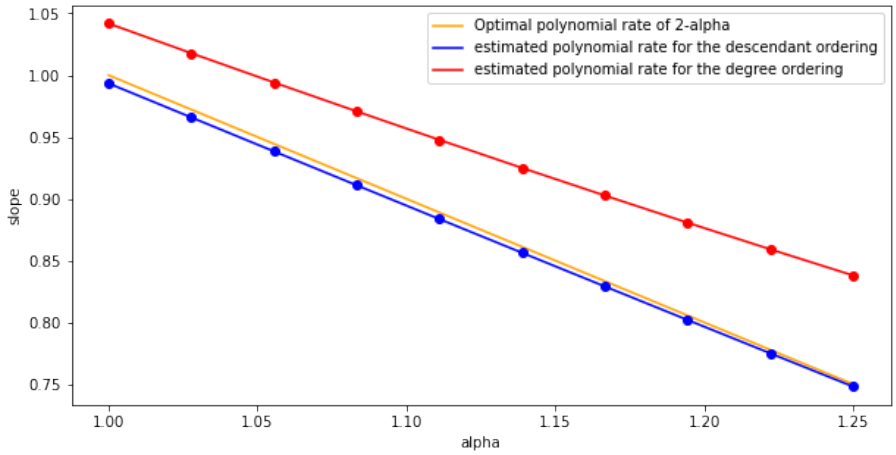

Figure 4.7: An estimation of the rate at which the risk increases with the size of the tree for different values of α in the pa tree. Here, we compare the case of the descendant and degree ordering. Proposition 4.7 shows that the optimal risk grows as $n^{2-\alpha}$, and that the the risk of descendant ordering grows at the same rate (see Proposition 4.8). This experiment confirms these results. For the degree ordering, this plot suggests that the risk grows faster than $n^{2-\alpha}$.

On the other hand, ordering vertices by their degree in the urrt is known to be sub-optimal for finding the root, as there are many vertices with much higher degree (Eslava [65]). Simulation results displayed in Figure 4.8 suggest that, for most values of α, the risk of ordering by degree in the urrt model grows at a faster rate than $n^{2-\alpha}$ for any $\alpha \in (1,2)$. On the other hand, observing Figure 4.8, it seems like, for $\alpha = 1$, the degree ordering may have a risk growing at the optimal rate of n.

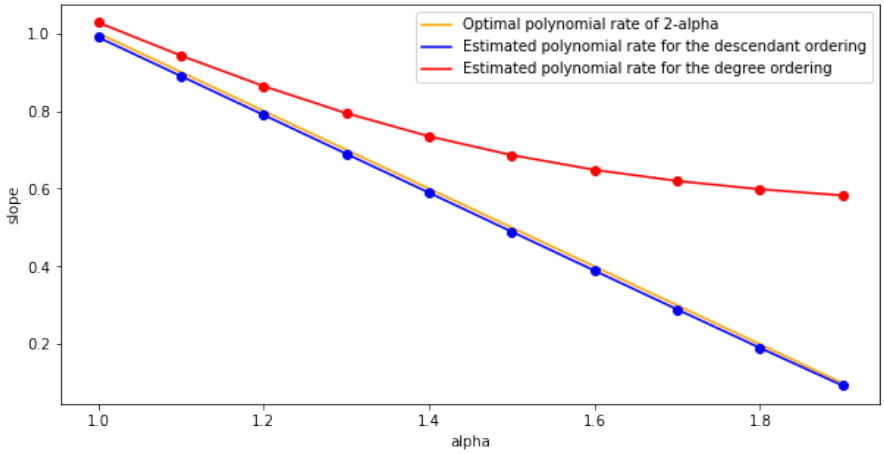

Figure 4.8: Estimation of the rate at which the risk increases with the size of the tree for different values of α in the urrt. Here we compare the case of descendant and degree ordering. Proposition 4.1 shows that the optimal risk grows as $n^{2-\alpha}$, and that the Jordan and descendant ordering's risks grow at the same rate (see Proposition 4.4). This experiment is in accordance with these results. For the degree ordering, this plot suggests that the risk grows at a rate faster than $n^{2-\alpha}$.

In Section 4.4, where the empirical performance of different ordering procedures are compared, the degree ordering is the method with second best performance. The above simulations suggest that all the other tested methods have risks growing at a faster rate than $n^{2-\alpha}$.

4.5.4 Proof of Theorem 4.4

Here, we present the arguments to complete the proof of Theorem 4.4. We recall that we need to upper bound $\sum_{i=1}^{n} \mathbb{E}\left[\frac{|\widehat{\sigma}'(i)-i|}{i^{\alpha}}\right]$, and that we reduced in (4.2.8) the problem to upper

bounding $\mathbb{P}\left\{\frac{\mathrm{de}(j)}{n} \geq \tau_{i,j}\right\} + \mathbb{P}\left\{\frac{\mathrm{de}(i)}{n} \leq \tau_{i,j}\right\}$, which is done in Lemma 4.6.

Proof. [Proof of Lemma 4.6] From (4.2.9), we have

$$\mathbb{P}\{\mathrm{de}(j) = k\} \;=\; \frac{k!(j-1)(j)\cdots(n-k-2)}{j(j+1)\cdots(n-1)}\binom{n-j}{k}.$$

Re-arranging the factors,

$$\begin{aligned}\mathbb{P}\{\mathrm{de}(j) = k\} \;&=\; k!\frac{(n-k-2)!}{(j-2)!}\frac{(j-1)!}{(n-1)!}\frac{(n-j)!}{k!(n-j-k)!}\\ &= (j-1)\frac{(n-k-2)!}{(n-j-k)!}\frac{(n-j)!}{(n-1)!}\end{aligned}$$

Since $j \geq 3$,

$$\begin{aligned}\mathbb{P}\{\mathrm{de}(j) = k\} \;&= (j-1)\frac{(n-j-k+1)\cdots(n-k-2)}{(n-j+1)\cdots(n-1)}\\ &= \frac{j-1}{n-1}\frac{n-j-k+1}{n-j+1}\cdots\frac{n-k-2}{n-2}\\ &\leq\; \frac{j-1}{n-1}\left(1-\frac{k}{n-2}\right)^{j-2}.\end{aligned}$$

Therefore, for $n, j \geq 3$, we can upper bound the second term of (4.2.8) by

$$\begin{aligned}\mathbb{P}\left\{\frac{\mathrm{de}(j)}{n} \geq \tau_{i,j}\right\} \;&\leq \sum_{(\tau_{i,j}n)\leq k\leq n}\frac{j-1}{n-1}\left(1-\frac{k}{n}\right)^{j-2}\\ &\leq \frac{(j-1)n}{n-1}\int_{\tau-\frac{1}{n}}^{1}(1-t)^{j-2}dt = \frac{n}{n-1}\left(1-\tau_{i,j}+\frac{1}{n}\right)^{j-1} \leq 2e^2e^{-j\tau_{i,j}},\end{aligned}$$

using $\log(1+x) \leq x$. Moreover, for $i \geq 2$, we bound the third term of (4.2.8) by

$$\begin{aligned}\mathbb{P}\left\{\frac{\mathrm{de}(i)}{n} \leq \tau_{i,j}\right\} &\leq \sum_{k\in[1,\tau_{i,j}n]}\frac{i-1}{n-1}\left(1-\frac{k}{n-2}\right)^{i-2}\\ &\leq \tau_{i,j}i\left(1-\frac{1}{n-2}\right)^{i-2} \leq \tau_{i,j}i\,.\end{aligned}$$

Since for $i = 1$, $\mathbb{P}\{\mathrm{de}(1) = k\} = 1/(n-1)$,

$$\mathbb{P}\left\{\frac{\mathrm{de}(1)}{n} \leq \tau_{i,j}\right\} \leq \sum_{k\in[1,\tau_{i,j}n]}\frac{1}{n-1} \leq \tau_{i,j} + \frac{1}{n-1}\,.$$

Therefore, for $i \geq 2$ and $j > i+1$, by choosing $\tau_{i,j} = \frac{1}{j}\log\frac{j}{i}$, we obtain that

$$\mathbb{P}\left\{\frac{\mathrm{de}(j)}{n} \geq \tau_{i,j}\right\} + \mathbb{P}\left\{\frac{\mathrm{de}(i)}{n} \leq \tau_{i,j}\right\} \leq 2e^2 e^{-\log\frac{j}{i}} + \frac{i}{j}\log\frac{j}{i} \leq \frac{i}{j}\left(2e^2 + \log\frac{j}{i}\right),$$

and for $i = 1$, $j > i+1$, choosing $\tau_{1,j} = \frac{1}{j}\log(j)$ we obtain that

$$\mathbb{P}\left\{\frac{\mathrm{de}(j)}{n} \geq \tau_{1,j}\right\} + \mathbb{P}\left\{\frac{\mathrm{de}(1)}{n} \leq \tau_{1,j}\right\} \leq \frac{1}{j}\left(2e^2 + \log(j)\right) + \frac{1}{n-1}.$$

■

Plugging the upper-bounds of Lemma 4.6 in (4.2.8), for $\alpha = 1$, we get

$$\begin{aligned}
&\sum_{i=1}^{n}\mathbb{E}\left[\frac{|\widehat{\sigma}'(i)-i|}{i}\right] \leq n + \sum_{i=1}^{n}\mathbb{E}\left[\frac{\widehat{\sigma}'(i)}{i}\right] \\
&\leq n + \sum_{i=2}^{n}\frac{1}{i}\left(i+1+\sum_{j=i+2}^{n}\frac{i}{j}\left(2e^2+\log\frac{j}{i}\right)\right) + 2 + \sum_{j=3}^{n}\left(\frac{1}{j}\left(2e^2+\log(j)\right)+\frac{1}{n-1}\right) \\
&\leq 2n + 3 + \log(n) + \sum_{j=3}^{n}\frac{1}{j}\sum_{i=1}^{j-2}\left(2e^2+\log\frac{j}{i}\right). \qquad (4.5.1)
\end{aligned}$$

Since

$$\sum_{i=1}^{j-2}\log\frac{j}{i} \leq \log\left(\frac{j^j}{j!}\right),$$

and that the Stirling formula implies that $j! \geq (j/3)^j$,

$$\sum_{i=1}^{j-2}\log\frac{j}{i} \leq j\log(3).$$

Plugging this in (4.5.1) yields

$$\sum_{i=1}^{n}\mathbb{E}\left[\frac{|\widehat{\sigma}'(i)-i|}{i}\right] \leq 3 + \log(n) + (2 + 2e^2 + \log 3)n,$$

which in turn proves that for $n \geq 60$,

$$\sum_{i=1}^{n}\mathbb{E}\left[\frac{|\widehat{\sigma}'(i)-i|}{i}\right] \leq 18n.$$

For $1 < \alpha < 2$, a similar calculation yields

$$\begin{aligned}
\sum_{i=1}^{n} \mathbb{E}\left[\frac{|\widehat{\sigma}'(i) - i|}{i^{\alpha}}\right] &\leq \frac{1}{2-\alpha} n^{2-\alpha} + \sum_{i=1}^{n} \mathbb{E}\left[\frac{\widehat{\sigma}'(i)}{i^{\alpha}}\right] \\
&\leq \frac{1}{2-\alpha} n^{2-\alpha} + 1 + \sum_{i=1}^{n} \frac{1}{i^{\alpha}} \left(i + 1 \sum_{j=i+2}^{n} \frac{i}{j} \left(2e^2 + \log \frac{j}{i} \right) \right) \\
&\leq \frac{2}{2-\alpha} n^{2-\alpha} + 1 + \zeta(\alpha) + \sum_{j=3}^{n} \frac{1}{j} \sum_{i=1}^{j-2} \frac{2e^2}{i^{\alpha-1}} + \frac{1}{i^{\alpha-1}} \log \frac{j}{i} \\
&\leq \frac{2}{2-\alpha} n^{2-\alpha} + 1 + \zeta(\alpha) + \frac{2e^2}{(2-\alpha)^2} n^{2-\alpha} + \sum_{j=3}^{n} \frac{1}{j} \sum_{i=1}^{j-2} \frac{1}{i^{\alpha-1}} \log \frac{j}{i} \,.
\end{aligned}$$

Recall that ζ denotes the Riemann zeta function. We may upper bound

$$\sum_{i=1}^{j-2} \frac{1}{i^{\alpha-1}} \log \frac{j}{i} \leq 2 \int_{1}^{j} \frac{1}{t^{\alpha-1}} \log\left(\frac{j}{t}\right) ,$$

which in turn can be evaluated by integration by parts, leading to

$$\sum_{i=1}^{j-2} \frac{1}{i^{\alpha-1}} \log \frac{j}{i} \leq \frac{2}{(2-\alpha)^2} j^{2-\alpha} \,.$$

Finally

$$\sum_{i=1}^{n} \mathbb{E}\left[\frac{|\widehat{\sigma}'(i) - i|}{i^{\alpha}}\right] \leq \left(\frac{2}{2-\alpha} + \frac{2e^2}{(2-\alpha)^2} + \frac{2}{(2-\alpha)^3} \right) n^{2-\alpha} \,.$$

For $\alpha \geq 2$, we similarly get

$$\sum_{i=1}^{n} \mathbb{E}\left[\frac{|\widehat{\sigma}'(i) - i|}{i^{\alpha}}\right] \leq C \,,$$

for some positive constant C.

4.5.5 Proof of the minimax lower bound in the PA model

Here we prove Theorem 4.7

Proof. The proof follows the same argument as that of Theorem 4.1. It suffices to check that the event

$$\Omega_j := \{\tau(j) \text{ and } \tau(\lfloor n/4 \rfloor + j) \text{ are leaves, connected to vertices of rank in } [n/2]\}$$

has a probability bounded away from 0. Proceeding as in the proof of Theorem 4.1, we get

$$\begin{aligned}
\mathbb{P}\left\{\Omega_j\right\} &= \frac{2\left(\lfloor n/2\rfloor - 1\right)}{2(j-2)} \prod_{k=j+1}^{\lfloor n/4\rfloor + j - 1} \frac{2k-3}{2(k-1)} \frac{2\left(\lfloor n/2\rfloor - 1\right)}{2\left(\lfloor n/4\rfloor + j - 2\right)} \prod_{k=\lfloor n/4\rfloor + j + 1}^{n} \frac{2k-4}{2(k-1)} \\
&= \frac{\left(\lfloor n/2\rfloor - 1\right)}{(j-2)} \frac{2j-1}{2\left(\lfloor n/4\rfloor + j - 2\right)} \frac{\left(\lfloor n/2\rfloor - 1\right)}{\left(\lfloor n/4\rfloor + j - 2\right)} \frac{\left(\lfloor n/4\rfloor + j - 1\right)\left(\lfloor n/4\rfloor + j\right)}{(n-1)(n-2)} \\
&= \frac{\left(\lfloor n/2\rfloor - 1\right)^2}{(n-1)(n-2)} \frac{2j-1}{2j-4} \frac{\left(\lfloor n/4\rfloor + j - 1\right)\left(\lfloor n/4\rfloor + j\right)}{\left(\lfloor n/4\rfloor + j - 2\right)^2} \\
&\geq \frac{1}{4}\left(1 - \frac{5}{n}\right)^5 .
\end{aligned}$$

■

4.5.6 Descendant ordering in the PA model

Here, we analyze the descendant ordering in the pa model. Recall the notation introduced in Section 4.2.2: the centrality measure $\psi'(u) = n - \mathrm{de}(u)$, and the corresponding ordering $\widehat{\sigma'}$. In the next lemma we prove that ψ' and ψ coincide for most vertices and provide a control both on the number of vertices for which they differ and the estimated arrival time of vertex 1.

Lemma 4.10. *Let c be the rank of a Jordan's centroid, and let $\{1 \to c\}$ be the set of vertices on the path from the root to the centroid. Then*

1. $\forall v \in [n] \backslash \{1 \to c\},\ \psi_T(v) = \psi'_T(v)$;
2. *there exists an universal constant K such that c is stochastically dominated by an exponential random variable with parameter K;*
3. *for any $\epsilon > 0$, with probability at least $1 - \epsilon$*

$$\widehat{\sigma_J}(1) \leq \frac{C}{\epsilon^2} \exp\left(\sqrt{C \log\left(\frac{1}{\epsilon}\right)}\right).$$

Proof. The first part of the proof is identical to the proof of Lemma 4.2. First, we use Theorem 6 of Wagner and Durant [139], which extend the result of Moon [109] from uniform random recursive trees to preferential attachment trees. Using their result, we obtain that

$$\mathbb{P}\{c \geq k\} \leq \sum_{j=k}^{\infty} \frac{(-\log(2)/2)^j}{j!},$$

so there exists an exponential random variable of parameter K such that $c \leq \mathcal{E}(K)$. Using Corollary 3.3.b of Banerjee and Bhamidi [11], we have that the event

$$\widehat{\sigma_J}(1) \leq \frac{C}{\epsilon^2} \exp\left(\sqrt{C \log\left(\frac{1}{\epsilon}\right)}\right),$$

holds with probability at least $1 - \epsilon$. This concludes the proof of the lemma. ∎

The next lemma allows us to compare the risk of Jordan and descendant ordering.

Lemma 4.11. *Let $T \sim PA$. Then, there exist positive constants C, K, such that, for $\alpha > 0$*

$$R_\alpha(\widehat{\sigma_J}) \leq R_\alpha(\widehat{\sigma'}) + K \sum_{i=1}^{n} \frac{1}{i^\alpha} + C \log^2(n)\sqrt{n}\,.$$

Proof. The proof is similar to the one of Lemma 4.3. Recalling that D is the distance between vertices 1 and c, we have

$$\mathbb{E}\left[\sum_{i \in \{1 \to c\}} \frac{\left|\widehat{\sigma_J}(i) - i\right|}{i^\alpha}\right] \leq \mathbb{E}\left[\sum_{i \in \{1 \to c\}} i\right] + \mathbb{E}\left[\sum_{i \in \{1 \to c\}} \widehat{\sigma_J}(i)\right]$$

$$\leq \frac{1}{2}\mathbb{E}\left[D^2\right] + \mathbb{E}\left[D\widehat{\sigma_J}(1)\right].$$

As in Lemma 4.3, we use the fact that $D \leq c$ and the domination of c by an exponential random variable (see Lemma 4.10) to get that

$$\frac{1}{2}\mathbb{E}\left[D^2\right] \leq K^2\,.$$

Then, it follows from Hölder's inequality that

$$\mathbb{E}\left[D\widehat{\sigma_J}(1)\right] \leq \left(\mathbb{E}\left[D^{\frac{1+\gamma}{\gamma}}\right]\right)^{\frac{\gamma}{1+\gamma}} \left(\mathbb{E}\left[\widehat{\sigma_J}(1)^{1+\gamma}\right]\right)^{\frac{1}{1+\gamma}}\,. \tag{4.5.2}$$

Using once again the domination of D by an exponential random variable,

$$\left(\mathbb{E}\left[D^{\frac{1+\gamma}{\gamma}}\right]\right)^{\frac{\gamma}{1+\gamma}} \leq C\frac{1}{1+\gamma}\gamma\,,$$

for some positive constant C. Next, using Lemma 4.10,

$$\mathbb{P}\left\{\widehat{\sigma_J}(1) \geq f(\epsilon)\right\} \leq \epsilon\,,$$

where $f(\epsilon) = \frac{C}{\epsilon^2}\exp\left(\sqrt{C\log\left(\frac{1}{\epsilon}\right)}\right)$. The function f is a non-increasing, therefore $f\left(\frac{C}{\sqrt{k}}\exp\left(\sqrt{C\log(k)}\right)\right) \leq k$. So

$$\mathbb{P}\left\{\widehat{\sigma_J}(1) \geq k\right\} \leq \frac{C}{\sqrt{k}}\exp\left(\sqrt{C\log(k)}\right).$$

Following the same steps as in Lemma 4.3, and choosing $\gamma = 1/\log(n)$, yields

$$\mathbb{E}\left[D\widehat{\sigma_J}(1)\right] \leq C\log^2(n)\sqrt{n},$$

which concludes the proof of the lemma. ∎

4.5.7 Performance of Jordan ordering in the PA model

In this section we prove Theorem 4.8.

Proof. Similarly to the urrt case, in the pa model, the number of descendants of a vertex is distributed as a Pólya urn. This well-know fact is easily seen since in the pa model, sampling a vertex with a probability proportional to its degree is the same as sampling an edge uniformly at random and picking one of its endpoints at random. In turn, it is the same as picking a half edge uniformly at random. Therefore, the resulting Pólya urn has slightly different initial conditions than in the urrt. Such Pólya urns are well understood. In particular, by Mahmoud [105, Section 3.2], for a vertex $i \in [n]$, the distribution of $\mathrm{de}(i)$ is given by

$$\mathbb{P}\{\mathrm{de}(i) = k\} = \frac{(1\,3\cdots(2k-1))\,((2i-3)(2i-1)\cdots(2n-2k-5))}{(2i-2)2i\cdots(2n-4)}\binom{n-i}{k}.$$

Re-arranging the terms in the above expression,

$$\begin{aligned}\mathbb{P}\{\mathrm{de}(i) = k\} \quad = \quad & \underbrace{\frac{1\cdot 3\cdots(2k-1)}{k!}}_{\stackrel{\text{def}}{=}A}\cdot\underbrace{\frac{(2i-3)(2i-1)\cdots(2n-5)}{(2i-2)2i\cdots(2n-4)}}_{\stackrel{\text{def}}{=}B} \\ & \cdot\underbrace{\frac{(n-i)!/(n-i-k)!}{(2n-2k-3)(2n-2k-1)\cdots(2n-5)}}_{\stackrel{\text{def}}{=}C}.\end{aligned} \tag{4.5.3}$$

We bound each term on the right-hand side of (4.5.3). First, for $k \geq 1$,

$$A = \frac{1}{k}\prod_{j=1}^{k-1}\frac{2j+1}{j} = \frac{2^{k-1}}{k}\prod_{j=1}^{k-1}\left(1+\frac{1}{2j}\right).$$

Since

$$\prod_{j=1}^{k-1}\left(1+\frac{1}{2j}\right)=\exp\left(\sum_{j=1}^{k-1}\log\left(1+\frac{1}{2j}\right)\right)\le\exp\left(\sum_{j=1}^{k-1}\frac{1}{2j}\right)\le\sqrt{k}\,,$$

then,

$$A\le\frac{2^{k-1}}{\sqrt{k}}\,.$$

Second, we have

$$\begin{aligned}B&=\prod_{j=i}^{n-1}\frac{2j-3}{2j-2}=\prod_{j=i}^{n-1}\left(1-\frac{1}{2j-2}\right)\\&=\exp\left(\sum_{j=i}^{n-1}\log\left(1-\frac{1}{2j-2}\right)\right)\le\exp\left(-\sum_{j=i}^{n-1}\frac{1}{2j-2}\right)\le 2\sqrt{\frac{i}{n}}\end{aligned}$$

Finally, we have that

$$C=\prod_{j=n-k-2}^{n-3}\frac{j-i-3}{2j+1}=\frac{1}{2^{k-1}}\prod_{j=n-k-2}^{n-3}\left(1-\frac{i+2.5}{j+0.5}\right)\le\frac{1}{2^{k-2}}\left(1-\frac{k}{n}\right)^{i}\,.$$

Plugging these bounds into (4.5.3), we get

$$\mathbb{P}\{\mathrm{de}(i)=k\}\le 4\sqrt{\frac{i}{kn}}\left(1-\frac{k}{n}\right)^{i}\,.$$

Thus, for any $\tau>0$,

$$\mathbb{P}\left\{\frac{\mathrm{de}(i)}{n}\le\tau\right\}\le\sum_{k=1}^{n\tau}4\sqrt{\frac{i}{kn}}\le 4\sqrt{i\tau}\,. \tag{4.5.4}$$

Now, for $j\in[n]$, we have

$$\begin{aligned}\mathbb{P}\left\{\frac{\mathrm{de}(j)}{n}\ge\tau\right\}&\le\sum_{k=n\tau}^{n}4\sqrt{\frac{j}{kn}}\left(1-\frac{k}{n}\right)^{j}\le 4\sqrt{\frac{j}{\tau}}\sum_{k=n\tau}^{n}\frac{1}{n}\left(1-\frac{k}{n}\right)^{j}\\&\le\frac{4}{\sqrt{j\tau}}\,.\end{aligned} \tag{4.5.5}$$

Combining (4.5.4) and (4.5.5) with $\tau=1/\sqrt{ij}$,

$$\mathbb{P}\{\mathrm{de}(i)\le\mathrm{de}(j)\}\le 8\left(\frac{i}{j}\right)^{1/4}\,.$$

Following similar calculations as in Section 4.2.3,

$$\sum_{i=1}^{n} \mathbb{E}\left[\frac{|\widehat{\sigma}'(i)-i|}{i^{\alpha}}\right] \leq \frac{1}{2-\alpha} n^{2-\alpha} \sum_{i=1}^{n} \frac{1}{i^{\alpha}} \left(i + \sum_{j=i+1}^{n} 8 \left(\frac{i}{j}\right)^{1/4} \right)$$
$$\leq \frac{2}{2-\alpha} n^{2-\alpha} + 8 \sum_{j=1}^{n} \left(\frac{1}{j^{1/4}} \sum_{i=1}^{j-1} \frac{1}{i^{\alpha-1/4}} \right).$$

For $\alpha \in [1, 5/4)$ we obtain

$$\sum_{i=1}^{n} \mathbb{E}\left[\frac{|\widehat{\sigma}'(i)-i|}{i^{\alpha}}\right] \leq \frac{2}{2-\alpha} n^{2-\alpha} + \frac{8}{\alpha - 5/4} \sum_{j=1}^{n} \frac{1}{j^{\alpha-1}}$$
$$\leq \left(\frac{2}{2-\alpha} + \frac{1}{(\alpha-5/4)(\alpha-2)} \right) n^{2-\alpha},$$

while for $\alpha \geq 5/4$

$$\sum_{i=1}^{n} \mathbb{E}\left[\frac{|\widehat{\sigma}'(i)-i|}{i^{\alpha}}\right] \leq \frac{2}{2-\alpha} n^{2-\alpha} + 8\, \zeta\left(\alpha - \frac{1}{4}\right) \sum_{j=1}^{n} \frac{1}{j^{1/4}}$$
$$\leq \frac{2}{2-\alpha} n^{2-\alpha} + \frac{32}{3} \zeta\left(\alpha - \frac{1}{4}\right) n^{3/4}.$$

which concludes the proof of Theorem 4.8. ■

Chapter 5

Broadcasting in random recursive dags

Contents

Abstract

A uniform ℓ-dag generalizes the uniform random recursive tree by picking ℓ parents uniformly at random from the existing nodes. IN tihs version of the model, it starts with ℓ "roots". Each of the ℓ roots is assigned a bit. These bits are propagated by a noisy channel. The parents' bits are flipped with probability p, and a majority vote is taken. When all nodes have received their bits, the ℓ-dag is shown without identifying the roots. The goal is to estimate the majority bit among the roots. We identify the threshold for p as a function of ℓ below which the majority rule among all nodes yields an error $c + o(1)$ with $c < 1/2$. Above the threshold the majority rule errs with probability $1/2 + o(1)$.

This Chapter is based on a joint work with Luc Devroye and Gábor Lugosi (Briend, Devroye, and Lugosi [27]).

5.1 Introduction

The problem we study in this chapter is the one of broadcasting on random graphs. We study the setting where a bit propagates with noise and we want to infer the value of the original bit. The question is not if and how the information propagates, but if there is a signal propagating on the graph, or only noise. Variations of this binary classification problem have been studied. For example, in the root-bit estimation problem, the root of a tree has a bit 0 or 1. The value of this bit propagates from the root to the leafs, and at each propagation from a vertex to the next it mutates (flips the bit) with probability p. One can try to infer the root's bit value from observing all the bits of the graph or only the leaf bits. This question was first formulated in Evans et al. [67] on general trees, where it was shown that root bit reconstruction is possible depending upon a condition on the branching number. More recently, the case of random recursive trees (Addario-Berry et al. [2], Desmarais et al. [48]) has been studied. Other variations of these problems on trees include looking at asymmetric flip probabilities (Sly [132]), non-binary vertex values (Mossel [112]) and robustness to perturbation (Janson and Mossel [81]). We refer the reader to Mossel [113] for a survey of reconstruction problems on trees. Many problems are described by more general graphs rather than trees. The original broadcasting question has been studied on deterministic graphs (Harutyunyan and Li [75]) and Harary graphs (Bhabak et al. [17], for example). We are interested in the problem of noisy propagation in the spirit of the root-bit reconstruction (Evans et al. [67]), but on a class of random graphs that we call ℓ-dag (for directed acyclic graph). A similar problem – for a different class of random dags – has been studied in Makur et al. [107]. In a related probelm, Antunović et al. [8] studied the case of the preferential attachment model, where initial nodes have a color and the color of the new nodes is a function of the colors of their neighbors.

Since we track the proportion of zero bits in our graph, we cast the process as an urn model. A similar reformulation was already done in Addario-Berry et al. [2] to study majority voting properties of broadcasting on random recursive trees. The proportion of zero bits and the bit assignment procedure can be viewed as random processes with reinforcement. A review of results can be found in Pemantle [118] and is extensively used, alongside results of non-convergence found in Pemantle [117]. As in Addario-Berry et al. [2], we make ample use of the properties of Pólya urns (Janson [77], Knape and Neininger

[91], Wei [140]). Variations of the Pólya urn model that are useful for our analysis include an increase of the number of colors over time (Bertoin [16]), the selection of multiple balls in each draw (Kuba and Mahmoud [94]), and randomization in the color of the new ball (Janson [80], Zhang [141]). We note, in particular, the multi-ball draw with a linear randomized replacement rule of Crimaldi et al. [44]. In the present chapter, we consider multi-ball draws, but with non-linear randomized replacement.

The chapter is organized as follows. After introducing the mathematical model in Section 5.1.1, in Section 5.1.2 we present the main result of the chapter (Theorem 5.1) that shows that there are three different regimes of the value of the mutation probability that characterize the asymptotic behavior of the majority rule. In Section 5.2 we discuss the three regimes of p. In Section 5.3 we establish convergence properties of the global proportion of both bit values assigned to vertices and in Section 5.4 we finish the proof of Theorem 5.1 by studying the probability of error in all three regimes. Finally, in Section 5.5 we establish a lower bound for the probability of error that holds uniformly for all mutation probabilities. We conclude the chapter by discussing avenues for further research.

5.1.1 The model

We start by describing the evolution of the uniform random recursive ℓ-dag and the assigned bit values that we represent by two colors; red and blue (let us remark that the model is slightly different than the one defined in Chapter 3).

Let us fix an odd integer $\ell > 0$. The growth process is initiated at time ℓ. At time ℓ, the graph consists of ℓ isolated vertices. A fraction R_ℓ are red and a fraction $B_\ell = 1 - R_\ell$ are blue. We set $R_1 = \cdots = R_\ell$ and $B_1 = \cdots = B_\ell$. The network is grown recursively by adding a new colored vertex and at most ℓ edges at each time step. At time n, a new vertex n connects to a sample of ℓ vertices chosen uniformly at random with replacement among the $n-1$ previous vertices. (Possible multiple edges are collapsed into one so that the graph remains simple.) The color of vertex n is determined by the following randomized rule:

- the colors of the ℓ selected parents are observed;
- each of these is independently flipped with probability p (if a parent is selected more than once, its color is flipped independently for each selection);
- the color of vertex n is chosen according to the majority vote of the flipped parent colors.

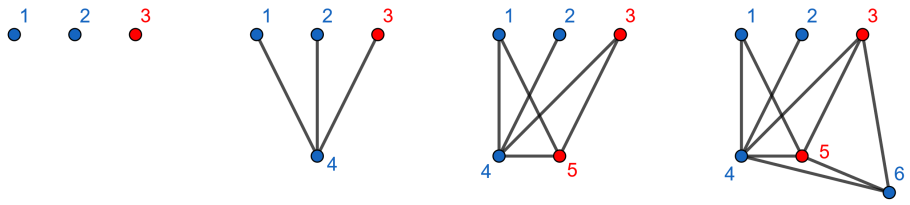

Figure 5.1: A realisation of the process up to time 6, for $\ell = 3$, starting with $R_3 = 1/3$.

If one is only interested in the evolution of the proportion of red and blue vertices (but not the structure of the graph), one may equivalently describe it by an urn model with multiple draws and random (nonlinear) replacement. The urn process is defined as follows. The urn is initialized with an odd number of ℓ balls, a fraction R_ℓ being red and $B_\ell = 1 - R_\ell$ blue.

- ℓ balls are drawn from the urn, uniformly at random with replacement, and returned to the urn;
- the color of each drawn ball is flipped with probability p (i.e., a drawn ball that is red is observed as blue with probability p);
- a new ball is added to the urn, whose color is chosen as the majority of the ℓ observed colors.

In the root-bit estimation problem considered here, the statistician has access to an unlabelled and undirected version of the graph at time n, along with the vertex colors. The goal of the statistician is to estimate the colors assigned to the ℓ roots. More precisely, based on the observed graph, one would like to guess the majority color at time ℓ.

This problem has been studied in depth by [2] in the case when $\ell = 1$, that is, when the produced graph is a uniform random recursive tree. Two types of methods for root-bit estimation were studied in [2]. One is based on first trying to localize the root of the tree– disregarding the vertex colors. If one finds a vertex that is close to the root, one may use the color of that vertex as a guess for the root color. Such a vertex is the centroid of the tree. Indeed, it is shown in [2] that the color of the centroid is a nearly optimal estimate of the root color. In the same paper, the majority rule is also studied. This method disregards

the structure of the tree and guesses the root color by taking a majority vote among all vertices. It is shown that for small mutation probabilities the majority rule is also nearly optimal.

In the more general problem considered in this chapter, one may also try to estimate the colors of the ℓ roots by finding nearby vertices. However, this problem becomes significantly more challenging as the ℓ-dag does not have a natural centroid. The interested reader is referred to Chapter 3 on root finding in random ℓ-dags. Instead of pursuing this direction, we focus on the majority vote. More precisely, we are interested in characterizing the values of the mutation probability p such that the asymptotic probability of error is strictly better than random guessing.

At time n, the majority vote, denoted by b_n^{maj}, is defined as follows:

$$b_n^{maj} = \begin{cases} \text{“R” (red) if } R_n > 1/2 \\ \text{“B”(blue) if } R_n < 1/2 \\ \text{Ber}(1/2) \text{ if } R_n = 1/2 \text{ (a random coin flip)} . \end{cases}$$

We define the probability of error by

$$R^{maj}(n,p) = \mathbb{P}\left\{ b_n^{maj} \neq b_\ell^{maj} \right\} .$$

Note that b_ℓ^{maj} depends on the initial vertex colors that are assumed to be chosen arbitrarily. Hence, $R^{maj}(n,p)$ is a function of the initial proportion R_ℓ but to avoid heavy notation, we supress this dependence.

5.1.2 Related results and our contribution

Our broadcasting model is an extension of the broadcasting on uniform random recursive trees that was extensively studied in Addario-Berry et al. [2]. In this problem, $\ell = 1$ and the only parameter is p, the mutation probability. For the majority voting rule, they prove the following:

(i) There exists a constant $c > 0$ such that

$$\limsup_{n\to\infty} R^{maj}(n,p) \leq cp .$$

(ii) For all $p \in (0, 1/2]$,

$$\lim_{n\to\infty} R_n = \frac{1}{2} \quad \text{with probability one .}$$

(iii) For $p \in [0, 1/4)$

$$\limsup_{n\to\infty} R^{maj}(n,p) < \frac{1}{2} .$$

(iv) For $p \in [1/4, 1/2]$

$$\limsup_{n\to\infty} R^{maj}(n,p) = \frac{1}{2} .$$

In other words, even though the proportion of vertices that have the same color as the root converges to $1/2$, for mutation probabilities smaller than $1/4$, sufficient information is preserved about the root color for the majority vote to work with a nontrivial probability.

We generalize these results to ℓ-dags and characterize the values of p for which majority voting outperforms random guessing. In order to state the main result of the chapter, we introduce some notation.

For any odd positive integer ℓ, let

$$\alpha_\ell := \frac{1}{2^{\ell-2}} \sum_{i>\ell/2}^{\ell} \binom{\ell}{i} (i - \ell/2) = 4\mathbb{E}\left[\left(\mathrm{Bin}(\ell, 1/2) - \frac{\ell}{2}\right)_+\right] . \tag{5.1.1}$$

For example, $\alpha_1 = 1$, $\alpha_3 = 3/2$, and by a simple application of the central limit theorem, for large ℓ,

$$\alpha_\ell \sim \sqrt{\frac{2\ell}{\pi}} . \tag{5.1.2}$$

In the statement of our main theorem, we assume, without loss of generality, that initially red vertices are in majority, that is, $R_\ell > 1/2$.

Theorem 5.1. *Let ℓ be an odd positive integer and consider the broadcasting process on a random ℓ-*dag *described above. Assume that initially $R_\ell > 1/2$.*

(i) If $p < \frac{1}{2} - \frac{1}{2\alpha_\ell}$, then there exist $\beta_1 \in (0, 1/2)$ and $\beta_2 = 1 - \beta_1$ (whose value only depends on ℓ but not on the initial color configuration) such that

$$\mathbb{P}\{R_n \to \beta_1\} + \mathbb{P}\{R_n \to \beta_2\} = 1 \quad \textit{and} \quad \mathbb{P}\{R_n \to \beta_1\} < \mathbb{P}\{R_n \to \beta_2\} \ .$$

In particular, regardless of the initial value of R_ℓ,

$$\limsup_{n\to\infty} R^{maj}(n,p) < \frac{1}{2} \ .$$

(ii) If $\frac{1}{2} - \frac{1}{2\alpha_\ell} \le p < \frac{1}{2} - \frac{1}{4\alpha_\ell}$, then $R_n \to 1/2$ a.s. and

$$\limsup_{n\to\infty} R^{maj}(n,p) < \frac{1}{2} \ .$$

(iii) If $\frac{1}{2} - \frac{1}{4\alpha_\ell} \le p \le \frac{1}{2}$ then $R_n \to 1/2$ a.s. and

$$\lim_{n\to\infty} R^{maj}(n,p) = \frac{1}{2} \ .$$

Theorem 5.1 shows that for all $\ell \ge 3$, there are three regimes of the value of the mutation probability. In the low-rate-of-mutation regime the proportion of red balls almost surely converges to one of two numbers, both different from $1/2$. Moreover, the limiting proportion is positively correlated with the initial value. In the intermediate phase, the vertex colors are asymptotically balanced, but there is enough signal for the majority vote to perform strictly better than random guessing. Finally, in the high-rate-of-mutation regime, the majority vote is equivalent to a coin toss, at least asymptotically.

Note that for $\ell = 1$, $\alpha_1 = 1$, so $1/2 - 1/(2\alpha_1) = 0$, and therefore the low-rate-of-mutation regime does not exist. Of course, this is in accordance with the results of [2] cited above.

On the other hand, for $\ell = 3$ the two thresholds are $1/2 - 1/(2\alpha_3) = 1/6$ and $1/2 - 1/(4\alpha_1) = 1/3$, meaning that from $\ell = 3$ onward the three different regimes can be observed. For large ℓ, both threshold values are of the order $1/2 - \Theta(1/\sqrt{\ell})$.

A closely related model has been studied by Makur et al. [107]. They study different random dags, where important parameters are the number of vertices at distance ℓ from the root and the indegree of vertices. They also suppose that the position of the root vertex

is known. Two rules of root bit estimation are studied: a noisy majority rule and the NAND rule. Makur et al. [107] show that if the number of vertices of depth k is $\Omega(\log(k))$ then there is a threshold on the mutation probability for which root bit estimation is possible.

As a first step, we study the convergence of the proportion of red balls. To this end, it suffices to study the generalized urn process defined above. We mention here that Crimaldi et al. [44] study a somewhat related urn process, though with linear replacement rules.

5.2 Different regimes

We start by studying the evolution of R_n. Let us denote by c_n the color of the n-th vertex appearing in the graph. After possible mutation, each edge connecting vertex $n+1$ to an older vertex carries a signal. This signal is red with probability

$$f(R_n) := (1-p)R_n + p(1-R_n) = (1-2p)R_n + p \ .$$

Because the ℓ parents are chosen independently and that the color is chosen by the majority,

$$\mathbb{P}\{c_{n+1} = R\} \ = \ \mathbb{P}\{\mathsf{Bin}\,(\ell, f(R_n)) \geq \ell/2\} \ , \tag{5.2.1}$$

where, conditionally on R_n, $\mathsf{Bin}\,(\ell, f(R_n))$ is a binomial random variable. Moreover, we know that the number of red vertices evolves as $(n+1)R_{n+1} = nR_n + \mathbb{1}(c_{n+1} = R)$, where $\mathbb{1}$ is the indicator function. We rewrite this as

$$R_{n+1} = R_n + \frac{\mathbb{1}(c_{n+1} = R) - R_n}{n+1} \ . \tag{5.2.2}$$

A key to understanding R_n is then to study the random variable $\mathbb{1}(c_{n+1} = R) - R_n$. We define, for $t \in [0,1]$,

$$g(t) \ := \ \mathbb{E}\left[\mathbb{1}(c_{n+1} = R) - R_n | R_n = t\right] \ = \mathbb{P}\{\mathsf{Bin}(\ell, f(t)) > \ell/2\} - t \ . \tag{5.2.3}$$

The evolution of R_n is entirely determined by the function g. Observe first that for any $t \in [0,1]$, $f(1-t) = 1 - f(t)$. Also, since ℓ is odd,

$$\mathbb{P}\{\mathsf{Bin}(\ell, 1-f(t)) > \ell/2\} = 1 - \mathbb{P}\{\mathsf{Bin}(\ell, f(t)) > \ell/2\} \ ,$$

which implies that

$$g(1-t) = -g(t)\,.$$

The extremal values of g are

$$g(0) = \mathbb{P}\{\mathsf{Bin}(\ell, p) > \ell/2\} > 0\,,$$

and

$$g(1) = \mathbb{P}\{\mathsf{Bin}(\ell, 1-p) > \ell/2\} - 1 < 0\,.$$

Since g is continuous, the polynomial g has at least one root. From the symmetry property we have $g(1/2) = -g(1-1/2) = -g(1/2)$, so $g(1/2) = 0$. Moreover we obtain

$$g'(1/2) = \frac{1-2p}{2^{\ell-2}} \sum_{i>\ell/2}^{\ell} \binom{\ell}{i} (i - \ell/2) - 1\,.$$

Recalling the definition of α_ℓ from (5.1.1), we have $g'(1/2) = (1-2p)\alpha_\ell - 1$. Since $\alpha_\ell \geq 1$, we conclude:

$$g'\left(\frac{1}{2}\right) \begin{cases} < 0 & \text{if } p > \frac{1}{2} - \frac{1}{2\alpha_\ell}\,, \\ \\ > 0 & \text{if } p < \frac{1}{2} - \frac{1}{2\alpha_\ell}\,. \end{cases}$$

To understand the other potential zeros of g, let us study its convexity.

Lemma 5.2. *The function g is strictly convex on $(0, 1/2)$ and strictly concave on $(1/2, 1)$.*

Proof. We may use the elementary identities

$$\mathbb{P}\left\{\mathsf{Bin}(\ell, x) \geq \frac{\ell+1}{2}\right\} = \mathbb{P}\left\{\mathsf{Beta}\left(\frac{\ell+1}{2}, \frac{\ell+1}{2}\right) < x\right\}\,, \tag{5.2.4}$$

where $\mathsf{Beta}(a, b)$ is a beta(a, b) random variable. Hence,

$$g(t) = \int_0^{f(t)} (x(1-x))^{\frac{\ell-1}{2}} \frac{\Gamma(\ell+1)}{\Gamma^2\left(\frac{\ell+1}{2}\right)} dx \quad - t\,,$$

and therefore

$$g'(t) = (1-2p)\,(f(t)(1-f(t)))^{\frac{\ell-1}{2}} \frac{\Gamma(\ell+1)}{\Gamma^2\left(\frac{\ell+1}{2}\right)} - 1\,. \tag{5.2.5}$$

Since $f(t)(1-f(t)) = -(1-2p)t(t-1) + p(1-p)$ is increasing for $t \in (0,1/2)$ and decreasing for $t \in (1/2,1)$, g is strictly convex on $(0,1/2)$ and strictly concave on $(1/2,1)$.

■

In summary, if $p > \frac{1}{2} - \frac{1}{2\alpha_\ell}$, then $g'(1/2) < 0$, and thus g is monotonically decreasing on $[0,1]$ and has only one zero in $[0,1]$. If $g'(1/2) = 0$, then there is only one zero (at $1/2$) and g exhibits an inflection point at $1/2$. If $p < \frac{1}{2} - \frac{1}{2\alpha_\ell}$, then $g'(1/2) > 0$ and thus g has exactly one zero in $(0,1/2)$ and by symmetry, it also has one zero on $(1/2,1)$. We denote these zeros by β_1 and β_2, respectively.

Figure 5.2 shows two examples of the graph of the function g.

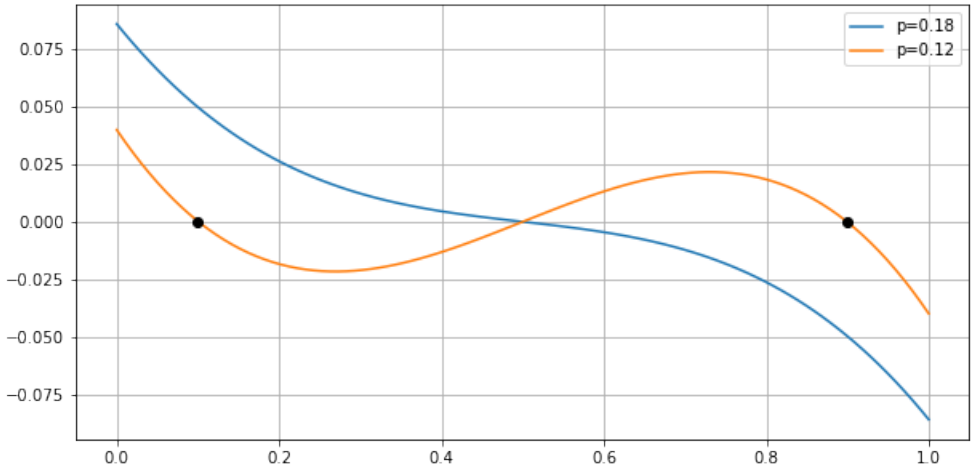

Figure 5.2: g as a function of $t \in [0,1]$, for $\ell = 3$, with the choices $p = 0.18 > 1/6$ and $p = 0.12 < 1/6$.

It is also interesting to know the position of β_1 (recall that $\beta_2 = 1 - \beta_1$). First, we note that for fixed ℓ, if p tends to the threshold $1 - 1/(2\alpha_\ell)$, then β_1 tends to $1/2$. In the following lemma we study the case when p is far enough from the threshold, that is, when $p \le \frac{1}{2} - \frac{C}{2\alpha_\ell}$, for a sufficiently large constant C.

Lemma 5.3. *Let $p \le \frac{1}{2} - \frac{C}{2\alpha_\ell}$ for $C \ge \sqrt{\frac{8\log(2)}{\pi}}$. Then*

$$\beta_1 \le \exp\left(-\frac{\ell(1-2p)^2}{8}\right).$$

Proof. β_1 is the smallest root of $g(t)$ and since $g(0) > 0$, its smallest root is smaller than the smallest root of any upper bound of g. On the other hand,

$$g(t) = \mathbb{P}\left\{\text{Bin}(\ell, f(t)) \geq \frac{\ell}{2}\right\} - t \leq \exp\left(-2\ell\left(\frac{1}{2} - f(t)\right)^2\right) = \exp\left(-2\ell(1-2p)^2\left(\frac{1}{2} - t\right)^2\right) - t\,.$$

Thus, β_1 is at most the first zero of $b(t) := \exp\left(c_1\left(\frac{1}{2} - t\right)^2\right) - t$, for $c_1 = 2\ell(1-2p)^2$. Since $b(0) > 0$, if for some t^*, $b(t^*) \leq 0$ then the first zero of b and therefore β_1 is at most t^*. Taking $t^* = e^{-c_1/16}$, we have

$$b(t^*) \leq 0 \iff \left(\frac{1}{2} - e^{-c_1/16}\right)^2 \geq 1/16 \iff c_1 \geq 32\log(2)\,.$$

From (5.1.2) and the expression of c_1, we have that by taking $C \geq \sqrt{\frac{8\log(2)}{\pi}}$,

$$2\ell(1-2p)^2 \geq 32\log(2)\,.$$

This shows that for $p \leq \frac{1}{2} - \frac{C}{2\alpha_\ell}$, we have

$$\beta_1 \leq \exp\left(-\frac{\ell(1-2p)^2}{8}\right).$$

■

5.3 Convergence of the proportion of red balls

In order to analyze the probability of error of the majority vote, first we establish convergence properties of R_n. The two possible regimes of g suggest that there are two distinct regimes of the evolution of R_n. From (5.2.2) we note that R_n has a positive drift if $g(R_n)$ is positive, and a negative drift otherwise. This suggests that in the high-rate-of-mutation regime, R_n converges to $1/2$ and in the low-rate-of-mutation regime it converges to either β_1 or β_2. The following section investigates this intuition, using Lemma 2.6 and Corollary 2.7 from Pemantle [118] about the convergence of reinforced random processes. We state them here.

Lemma 5.4 (Pemantle [118])**.** *Let $\{X_n;\ n \geq 0\}$ be a stochastic process in $\mathbb{R}$ adapted to a filtration $\{\mathcal{F}_n\}$. Suppose that X_n satisfies*

$$X_{n+1} - X_n = \frac{1}{n}\left(F(X_n) + \xi_{n+1} + E_n\right),$$

where F is a function on $\mathbb{R}$, $\mathbb{E}[\xi_{n+1} \mid \mathcal{F}_n] = 0$ and the remainder term E_n goes to 0 and satisfies $\sum_{n=1}^{\infty} n^{-1}|E_n| < \infty$ almost surely. Suppose that F is bounded and that $\mathbb{E}\left[\xi_{n+1}^2 \mid \mathcal{F}_n\right] < K$ for some finite constant ℓ. If for $a_0 < x < b_0$, $F(x) \geq \delta$ for some $\delta > 0$, then for any $[a, b] \subset (a_0, b_0)$ the process $\{X_n\}$ visits $[a, b]$ finitely many times almost surely. The same result holds if $F(x) \leq -\delta$.

Corollary 5.5 (Pemantle [118])**.** *If F is continuous on $\mathbb{R}$, then X_n converges almost surely to the zero set of F.*

5.3.1 The high-rate-of-mutation regime $\left(\frac{1}{2} - \frac{1}{2\alpha_\ell} \leq p \leq \frac{1}{2}\right)$

Rewrite (5.2.2) as

$$R_{n+1} - R_n = \frac{1}{n+1}\left(\mathbb{P}\left\{\text{Bin}(\ell, f(R_n)) \geq \frac{\ell}{2}\right\} - R_n\right) + \frac{1}{n+1}\left(\mathbb{1}(c_{n+1} = R) - \mathbb{P}\left\{\text{Bin}(\ell, f(R_n)) \geq \frac{\ell}{2}\right\}\right).$$

Since $g(R_n) = \mathbb{P}\{\text{Bin}(\ell, f(R_n)) \geq \ell/2\} - R_n$, we see that

$$R_{n+1} - R_n = \frac{g(R_n) + \xi_{n+1}}{n+1}, \tag{5.3.1}$$

where $\xi_{n+1} = \mathbb{1}(c_{n+1} = R) - \mathbb{P}\{\text{Bin}(\ell, f(R_n)) \geq \ell/2\}$. Because g is continuous and $\mathbb{E}[\xi_{n+1}|R_n] = 0$, our process satisfies all the requirements for Corollary 5.5. It states that R_n converges almost surely to the set of zeros of g. In this regime, this implies that R_n converges to $1/2$ almost surely.

5.3.2 The low-rate-of-mutation regime $\left(0 < p < \frac{1}{2} - \frac{1}{2\alpha_\ell}\right)$

In this regime, the requirements of Corollary 5.5 are still met. So R_n converges almost surely to the set of zeros of g, which is $\{\beta_1, 1/2, \beta_2\}$. We first show that R_n does not converge to $1/2$: $1/2$ seems to be an unstable equilibrium point, since the drift in the process has a tendency to pull R_n away from $1/2$. We state Theorem 2.9 from Pemantle [118] here:

Theorem 5.6 (Pemantle [118])**.** *Suppose $\{X_n\}$ satisfies the conditions of Lemma 5.4 and that for some $w \in (0, 1)$ and $\epsilon > 0$, sign$F(x) =$ sign$(x - w)$ for all $x \in (w - \epsilon, w + \epsilon)$. For $\xi_{n+1}^+ = \max(\xi_{n+1}, 0)$ and $\xi_{n+1}^- = \max(-\xi_{n+1}, 0)$, suppose that $\mathbb{E}[\xi_{n+1}^+ \mid \mathcal{F}_n]$ and $\mathbb{E}[\xi_{n+1}^- \mid \mathcal{F}_n]$ are bounded above and below by positive numbers when $X_n \in (w - \epsilon, w + \epsilon)$. Then $\mathbb{P}\{X_n \to w\} = 0$.*

Corollary 5.7. *In the low-rate-of-mutation regime, almost surely the process R_n does not converge to $\frac{1}{2}$.*

Proof. Since the conditional distribution of ξ_{n+1}, given $R_n = 1/2$ does not depend on n, it is immediate that

$$c < \mathbb{E}[\xi^+_{n+1} | R_n = 1/2] < 1 \ ,$$

and

$$c < \mathbb{E}[\xi^-_{n+1} | R_n = 1/2] < 1 \ ,$$

for some $c > 0$ that does not depend on n. Since $t \mapsto \mathbb{E}[\xi^\pm_{n+1} | R_n = t]$ is continuous and does not depend on n, there exists $\epsilon > 0$ such that for all $t \in (1/2 - \epsilon, 1/2 + \epsilon)$,

$$\frac{c}{2} < \mathbb{E}[\xi^\pm_{n+1} | R_n = t] < 2 \ .$$

Moreover, g is negative on $(1/2 - \epsilon, 1/2)$ and positive on $(1/2, 1/2 + \epsilon)$. So, by Theorem 5.6,

$$\mathbb{P}\left\{R_n \mapsto \frac{1}{2}\right\} = 0 \ .$$

■

Corollary 5.8. *In the low-rate-of-mutation regime, the process R_n converges almost surely, either to β_1 or to β_2, that is,*

$$\mathbb{P}\{R_n \to \beta_1\} + \mathbb{P}\{R_n \to \beta_2\} = 1 \ .$$

Proof. It suffices to check that R_n converges to β_1 or β_2 and does not oscillate between them. Between $1/2$ and β_2 the function g is positive, so there exists $1/2 < a_0 < a_1 < \beta_2$ and $\delta > 0$ such that for all $t \in (a_0, a_1)$, $g(t) > \delta$.

Lemma 5.4 shows that R_n visits any set $[a, b] \subset (a_0, a_1)$ finitely often almost surely. Because the step sizes of R_n are of order $1/n$, if R_n visits $[a, b]$ finitely many times, it crosses it finitely many times. Indeed, for n large enough it cannot cross $[a, b]$ without visiting $[a, b]$. Since R_n converges almost surely to the set $\{\beta_1, \beta_2\}$, but R_n crosses the set (a_0, a_1) finitely many times, we see that R_n converges almost surely either to β_1 or β_2, as claimed. ■

5.4 Is majority voting better than random guessing?

As a first step of understanding if majority voting is better than random guessing, we prove the following lemma. It gives an equivalent condition to the success of majority voting in terms of the first time the majority flips.

Lemma 5.9. *Let T denote the random time at which the majority flips for the first time, that is,*

$$T = \min\left\{n \in \mathbb{N} : \ b_n^{maj} \neq B_\ell^{maj}\right\} .$$

Then $\limsup_{n\to\infty} R^{maj}(n,p) < 1/2$ *if and only if* $\mathbb{P}\{T = +\infty\} > 0$.

Lemma 5.9 states that, once the proportion of red balls reaches $1/2$, since the broadcasting process is symmetric, inference of the original configuration is impossible if one only counts vertices.

Proof. From the definition of $R^{maj}(n,p)$,

$$\limsup_{n\to\infty} R^{maj}(n,p) = 1 - \liminf_{n\to\infty} \mathbb{P}\left\{b_n^{maj} = B_\ell^{maj}\right\} .$$

Fix a positive ϵ. Since the sequence of events $\{\forall i \in [n] : \ b_i^{maj} = B_\ell^{maj}\}$ is decreasing, and $\{T = +\infty\} = \{\forall i > \ell;\ b_i^{maj} = B_\ell^{maj}\}$, by the monotonicity of measure we can choose n such that

$$\mathbb{P}\left\{\forall i \in [n] : \ b_i^{maj} = B_\ell^{maj}\right\} \leq \mathbb{P}\{T = +\infty\} + \epsilon .$$

For $N \geq n+1$, we have

$$\begin{aligned}\mathbb{P}\left\{b_N^{maj} = B_\ell^{maj}\right\} =& \mathbb{P}\left\{b_N^{maj} = B_\ell^{maj} \text{ and } \forall i \in [n] : \ b_i^{maj} = B_\ell^{maj}\right\} \\ &+ \mathbb{P}\left\{b_N^{maj} = B_\ell^{maj} \text{ and } \exists i \in [n] : \ b_i^{maj} \neq B_\ell^{maj}\right\} .\end{aligned} \tag{5.4.1}$$

The second term on the right-hand side decomposes as

$$\begin{aligned}&\mathbb{P}\left\{b_N^{maj} = B_\ell^{maj} \text{ and } \exists i \in [n] : \ b_i^{maj} \neq B_\ell^{maj}\right\} \\ &= \left(1 - \mathbb{P}\left\{\forall i \in [n] : \ b_i^{maj} = B_\ell^{maj}\right\}\right)\mathbb{P}\left\{b_N^{maj} = B_\ell^{maj} \,\middle|\, \exists i \in [n] : \ b_i^{maj} \neq B_\ell^{maj}\right\} .\end{aligned}$$

From the definition of our process, if $R_i = 1/2$, then, conditionally on this event, the distribution of R_N for $N > i$ is symmetric. Therefore

$$\mathbb{P}\left\{b_N^{maj} = B_\ell^{maj} \,\middle|\, \exists i \in [n] : \ b_i^{maj} \neq B_\ell^{maj}\right\} = \frac{1}{2} . \tag{5.4.2}$$

Plugging this into (5.4.1) yields

$$\begin{aligned}\mathbb{P}\left\{b_N^{maj} = B_\ell^{maj}\right\} =& \mathbb{P}\left\{b_N^{maj} = B_\ell^{maj} \ \cap \ \forall i \in [n] : \ b_i^{maj} = B_\ell^{maj}\right\} \\ &+ \frac{1}{2}\left(1 - \mathbb{P}\left\{\forall i \in [n] : \ b_i^{maj} = B_\ell^{maj}\right\}\right) ,\end{aligned} \tag{5.4.3}$$

The first term of the right-hand side is bounded from below by $\mathbb{P}\{T = +\infty\}$, which transforms (5.4.3) into

$$\mathbb{P}\left\{b_N^{maj} = B_\ell^{maj}\right\} \geq \frac{1}{2} + \mathbb{P}\{T = +\infty\} - \frac{1}{2}\mathbb{P}\left\{\forall i \in [n] :\ b_i^{maj} = B_\ell^{maj}\right\} .$$

Taking the limit on N and recalling the choice of n gives

$$\liminf_{N\to\infty} \mathbb{P}\left\{R_N > \frac{1}{2}\right\} \geq \frac{1}{2} + \frac{1}{2}\mathbb{P}\{T = +\infty\} - \frac{\epsilon}{2} .$$

Since the above holds for any ϵ, if $\mathbb{P}\{T = +\infty\} > 0$ then $\liminf_{N\to\infty} \mathbb{P}\left\{b_N^{maj} = B_\ell^{maj}\right\} > 1/2$. This proves the "if" direction of the statement.

On the other hand, from (5.4.3),

$$\mathbb{P}\left\{b_N^{maj} = B_\ell^{maj}\right\} \leq \mathbb{P}\left\{\forall i \in [n] :\ b_i^{maj} = B_\ell^{maj}\right\} + \frac{1}{2}\left(1 - \mathbb{P}\left\{\forall i \in [n] :\ b_i^{maj} = B_\ell^{maj}\right\}\right) .$$

Taking the limit on N and recalling the choice of n yields

$$\begin{aligned}\liminf_{N\to\infty}\ \mathbb{P}\left\{b_N^{maj} = B_\ell^{maj}\right\} &\leq \frac{1}{2} + \frac{1}{2}\mathbb{P}\left\{\forall i \in [n] :\ b_i^{maj} = B_\ell^{maj}\right\} \\ &\leq \frac{1}{2} + \frac{1}{2}\mathbb{P}\{T = +\infty\} + \frac{\epsilon}{2} .\end{aligned}$$

As this holds for any positive ϵ, if $\liminf_{N\to\infty} \mathbb{P}\left\{b_N^{maj} = B_\ell^{maj}\right\} > 1/2$, then $\mathbb{P}\{T = +\infty\} > 0$. This concludes the proof. ∎

Lemma 5.10. *If*

$$\limsup_{n\to\infty} R^{maj}(n,p) \geq \frac{1}{2} ,$$

then

$$\lim_{n\to\infty} R^{maj}(n,p) = \frac{1}{2} .$$

Proof. If $\limsup_{n\to\infty} R^{maj}(n,p) \geq \frac{1}{2}$ then Lemma 5.9 shows that T is almost surely finite. But since

$$\mathbb{P}\left\{b_n^{maj} \neq B_\ell^{maj} \mid T \leq n\right\} = \frac{1}{2} ,$$

this implies

$$\mathbb{P}\left\{b_n^{maj} \neq B_\ell^{maj},\ T \leq n\right\} = \frac{1}{2}\mathbb{P}\{T \leq n\} .$$

Moreover, since T is finite almost surely, $\lim_{n\to\infty} \mathbb{P}\{T \leq n\} = 1$ and by the monotonicity of measure,

$$\lim_n \mathbb{P}\left\{b_n^{maj} \neq B_\ell^{maj},\ T \leq n\right\} = \mathbb{P}\left\{b_n^{maj} \neq B_\ell^{maj}\right\} .$$

This concludes the proof of the the lemma. ∎

5.4.1 The low-rate-of-mutation regime $\left(0 < p < \frac{1}{2} - \frac{1}{2\alpha_\ell}\right)$

As explained in Section 5.3.2, if $p < \frac{1}{2} - \frac{1}{2\alpha_\ell}$, then R_n converges to either β_1 or β_2. Next we show that if $R_1 > 1/2$, then R_n is more likely to converge to β_2 than to β_1. To do so, recall (5.2.2) and write it as

$$R_{n+1} = \frac{n}{n+1} R_n + \frac{1}{n+1} B_n(g(R_n) + R_n) \,,$$

where the B_n are independent Bernoulli random variables. We fix $\tau \in (1/2, \beta_2)$. From the analysis of g we know that $g(\tau) > 0$. Since $g(t) + t = \mathbb{P}\{\mathsf{Bin}(\ell, f(t)) \geq \ell/2\}$ and f is increasing, for all $t \geq \tau$,

$$g(t) + t \geq g(\tau) + \tau \,.$$

Fix a positive integer N and introduce the mapping

$$t \mapsto h(t) : \begin{cases} h(t) = 1/2 \text{ if } t < \tau \\ \\ h(t) = g(\tau) + \tau \text{ otherwise} \,. \end{cases}$$

Then define $D_\ell = 1$. For $n \geq \ell$, let

$$D_{n+1} = \frac{n}{n+1} D_n + \frac{1}{n+1} B'_n(h(D_n)) \,,$$

where B'_n are independent Bernoulli random variables. From the definition of the process (D_n), on the event $\{D_n \geq \tau,\ \forall n \geq 1\}$

$$nD_n \geq D_\ell + \mathsf{Bin}(n - \ell, g(\tau) + \tau) \,.$$

Hence, by the union bound and Hoeffding's inequality,

$$\mathbb{P}\{\exists i \geq N :\ D_i \leq \tau \mid \forall n \in [\ell, N] :\ D_n \geq \tau\} \leq \sum_{i \geq N} \mathbb{P}\{\mathsf{Bin}(i - \ell, g(\tau) + \tau) \leq i\tau\} \leq \frac{2e^{-(N-\ell)g(\tau)^2}}{1 - e^{-2g(\tau)^2}} \,.$$

Choosing N such that the last term above is less than one yields

$$\mathbb{P}\{\forall i \geq N :\ D_i \geq \tau \mid \forall n \in [\ell, N] :\ D_n \geq \tau\} > 0 \,.$$

Since

$$\mathbb{P}\{\forall i \geq \ell :\ D_i \geq \tau\} = \mathbb{P}\{\forall i \in [\ell, N] :\ D_i \geq \tau\} \times \mathbb{P}\{\forall i \geq N :\ D_i \geq \tau \mid \forall n \in [\ell, N];\ D_n \geq \tau\} \,,$$

we just proved that

$$\mathbb{P}\{\forall i \geq \ell : \ D_i \geq \tau\} > 0 \, . \tag{5.4.4}$$

Define the stopping time $T' = \min\{n \geq \ell ; \ D_n \leq \tau\}$. Since for all $t \geq \tau$, $g(t) + t \geq g(\tau) + \tau$, on the event $\{R_\ell \geq D_\ell \geq \tau\}$, there exists a coupling of the Bernoulli random variables B and B' such that

$$\forall n \in [\ell, T'] : \ B_n \geq B'_n \, ,$$

and thus a coupling of the random variables R_n and D_n such that

$$\forall n \in [\ell, T'] : \ R_n \geq D_n \, .$$

From this coupling and (5.4.4) we have

$$\mathbb{P}\left\{\forall n \geq \ell : \ R_n > \frac{1}{2}\right\} > 0 \, ,$$

which, thanks to Lemma 5.9, proves that in the regime $p < 1/2 - 1/(2\alpha_\ell)$,

$$\limsup_{n\to\infty} R^{maj}(n,p) < \frac{1}{2} \, ,$$

proving the first statement of Theorem 5.1.

5.4.2 The high-rate-of-mutation regime $\left(\frac{1}{2} - \frac{1}{2\alpha_\ell} \leq p \leq \frac{1}{2}\right)$

In the range $p > 1/2 - 1/(2\alpha_\ell)$ the proportion of red balls converges to $1/2$. It does not mean that majority voting can not be better than random guessing. Indeed, the proportion can converge to $1/2$ from above. This is this possibility that will now be investigated.

Extreme rate

First, we examine the "extreme" case when the rate of mutation is near $1/2$, more precisely when $p > 1/2 - 1/(4\alpha_\ell)$. Define the linear function h by $h(t) := g'(1/2)\,(t - 1/2)$. Then

$$g(t) \begin{cases} \geq h(t), \text{ if } t \in [0, 1/2], \\ \leq h(t), \text{ if } t \in [1/2, 1] \, . \end{cases}$$

In Figure 5.3 we plot h and g.

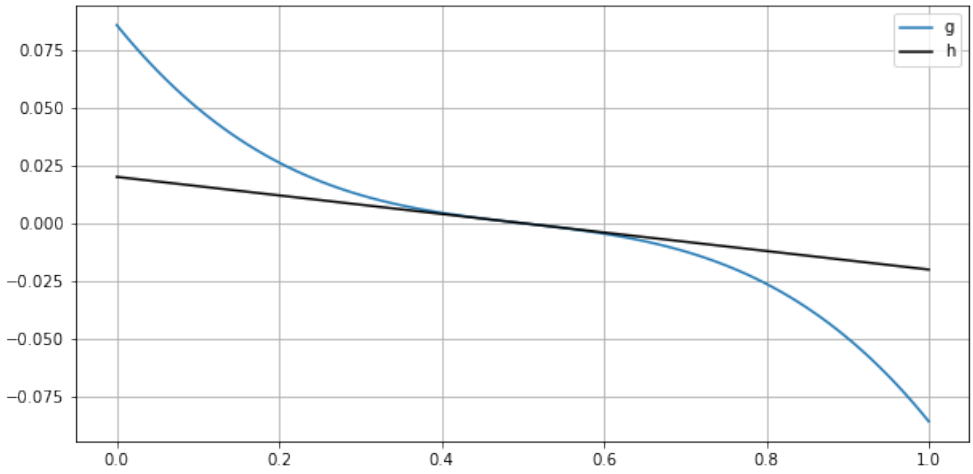

Figure 5.3: A linear lower bound for $|g|$, $\ell = 3$ and $p = 0.18$.

Let us define an auxiliary process R^*_n by the stochastic recursion $R^*_\ell = 1$ and for $n \geq \ell$

$$R^*_{n+1} = R^*_n + \frac{B_n\left(h(R^*_n) + R^*_n\right) - R^*_n}{n+1} ,$$

where $B_n\left(h(R^*_n) + R^*_n\right)$ is a Bernoulli random variable with parameter $h(R^*_n) + R^*_n$, conditionally independent of R^*_n. In particular,

$$\mathbb{E}\left[B_n\left(h(R^*_n) + R^*_n\right) - R^*_n | R^*_n = t\right] = h(t) .$$

Since the value of g (for (R_n)) and h (for (R^*_n)) represents a drift in the processes R_n and R^*_n we expect that the process (R^*_n) is further away from $1/2$. Indeed, we may introduce a coupling as follows. Define the stopping time T^* as the first time R^* reaches $1/2$:

$$T^* := \min\left\{n \geq \ell : \ R^*_n \leq \frac{1}{2}\right\} .$$

Since for the times $n \in [\ell, T^*]$, $h(R^*_n) \geq g(R_n)$, we may use a similar coupling argument as in Section 5.4.1. Thus, there is a coupling of R^* and R such that

$$\forall n \in [\ell, T^*]; \ R_n \leq R^*_n .$$

From this coupling, for T defined in Lemma 5.9 we have

$$\mathbb{P}\{T = +\infty\} \leq \mathbb{P}\{T^* = +\infty\} \ . \tag{5.4.5}$$

Observe that in the case of $\ell = 1$, g is linear and the two processes R_n and R_n^* coincide. The linear case was analyzed in Addario-Berry et al. [2] and we may use their results to understand the behavior of R_n^*. Indeed, the process defined in Addario-Berry et al. [2] is the same as R^* if one sets the flip probability of Addario-Berry et al. [2] equal to $-g'(1/2)/2$ and starts at time ℓ. They prove that if $p \geq 1/4$, then, for the process starting at time 1, majority voting has an error probability of $1/2+o(1/2)$. Lemma 5.9 implies that this process reaches $1/2$ in finite time almost surely. So even conditioned on its value being 1 at time ℓ it will reach $1/2$ in finite time almost surely. This proves that even for R_n^* starting at time ℓ its error probability is $1/2 + o(1)$. According to Lemma 5.9 this implies that for this range of p, $\mathbb{P}\{T^* = +\infty\} = 0$. Hence, using Lemma 5.9 and (5.4.5), shows that if $g'(1/2) \leq -\frac{1}{2}$, then

$$\limsup_{n\to\infty} R^{maj}(n,p) = \frac{1}{2} \ .$$

Lemma 5.10 shows that $\lim_{n\to\infty} R^{maj}(n,p) = 1/2$. Because $g'(1/2) = (1-2p)\alpha_\ell - 1$, we just proved that if $p \geq 1/2 - 1/4\alpha_\ell$, then

$$\lim_{n\to\infty} R^{maj}(n,p) = \frac{1}{2} \ ,$$

completing the proof of the third statement of Theorem 5.1.

Intermediate rate

It remains to study the "intermediate" case $p \in [1/2-1/(2\alpha_\ell), 1/2-1/(4\alpha_\ell))$. To this end, we may couple R_n to a process for which majority voting outperforms random guessing. Let us fix $p \in [1/2 - 1/(2\alpha_\ell), 1/2 - 1/(4\alpha_\ell))$, which implies that $g'(1/2)/2 > -1/4$. Then choose $q = -g'(1/2)/2 + \epsilon$ with $\epsilon > 0$ small enough so that $q < 1/4$ and $g(0) > h(0)$. We define the linear function $h(t) := -2q(t - 1/2)$, and as illustrated in Figure 5.4, we denote by a and b the intersection points between h and g (apart from $1/2$). More precisely a and b are defined as the the roots of $g-h$ distinct from 0. Since $g-h$ is strictly convex on $(0,1/2)$ and $(g-h)(0) > 0$, $(g-h)'(1/2) < 0$, a and b are well defined and sit respectively in $(0,1/2)$ and $(1/2,1)$.

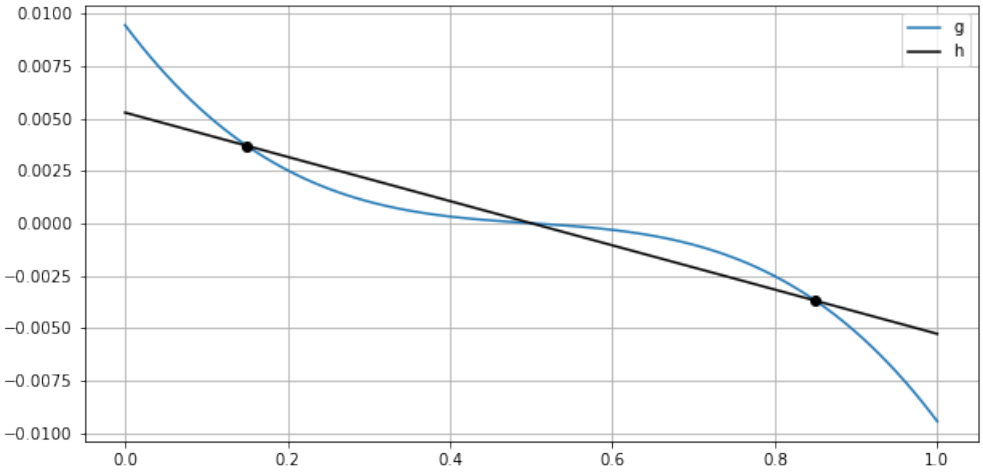

Figure 5.4: Comparison of h and g, for $\ell = 3$ and $p = 0.34$ (rescaled for clarity).

We define R^*_n similarly as in the previous section but now with $h(t) = -2q(t - 1/2)$, that is $R^*_\ell = 1$ and

$$R^*_{n+1} = R^*_n + \frac{B_n\,(h(R^*_n) + R^*_n) - R^*_n}{n+1}\,,$$

where the B_n are conditionally independent Bernoulli random variables. In particular,

$$\mathbb{E}\,[B_n\,(h(R^*_n) + R^*_n) - R^*_n | R^*_n = t] = -2q\left(t - \frac{1}{2}\right).$$

Just as in the previous section, we may use the analysis of Addario-Berry et al. [2] for the case $\ell = 1$ with mutation probability of q. Addario-Berry et al. [2] state that for the process starting at time 1 and for $q < 1/4$ majority voting is better than random guessing. A simple coupling from the process starting at time 1 and R^*_n proves that this statement holds for R^*_n. Thus, from Lemma 5.9 it follows that

$$\mathbb{P}\left\{\forall n \geq \ell :\ R^*_n > \frac{1}{2}\right\} > 0\,.$$

Now, from Lemma 5.4 we deduce that both processes R_n and R^*_n converge almost surely to $1/2$ and exceed b only finitely many times. Thus, there exists an almost surely finite random time T' such that and $\forall n \geq T'$; $R_n \leq b$ and $R^*_n \leq b$. We use similar coupling arguments as in Section 5.4.1. So, on the event that R^* does not reach $1/2$ we can couple R_n and R^*_n from T' onwards such that $R_n \geq R^*_n$. This proves that

$$\mathbb{P}\left\{\forall n \geq T' : \ R_n > \frac{1}{2} \mid T'\right\} > 0 \, .$$

Using that T' is finite almost surely and Lemma 5.9 we conclude that majority voting is better than random guessing in this regime. More precisely, if $1/2 - 1/2\alpha_\ell \leq p < 1/2 - 1/4\alpha_\ell$, then

$$\limsup_{n\to\infty} R^{maj}(n,p) < \frac{1}{2} \, .$$

This completes the proof of Theorem 5.1.

5.5 A general lower bound

In this final section we derive a lower bound for the probability of error that holds for all mutation probabilities. In particular we show the following.

Proposition 5.11. *Let m be a positive odd integer and let $\ell/2 < m < \ell$. Assume that initially there are m red vertices, that is $R_\ell = m/\ell$. Letting*

$$h_\ell := \mathbb{P}\left\{Beta\left(\frac{\ell+1}{2}, \frac{\ell+1}{2}\right) \geq 1 - \frac{1}{\ell}\right\} \, ,$$

the probability of error of the majority rule satisfies

$$\inf_{\substack{0\leq p\leq 1 \\ n\geq 2\ell}} \mathbb{P}\left\{b_n^{maj} \neq B_\ell^{maj}\right\} \geq \frac{1}{2} h_\ell^{2m-\ell} \, .$$

Proof. The proposition follows by simply considering the event that the first $2m - \ell$ new vertices are all blue. In that case, at time $2m$ the number of red and blue vertices are equal. We may write, for any $n \geq 2m$,

$$\mathbb{P}\left\{b_n^{maj} \neq B_\ell^{maj}\right\} \geq \mathbb{P}\left\{b_n^{maj} \neq B_\ell^{maj} \mid c_{\ell+1} = \cdots = c_{2m} = B\right\} \times \mathbb{P}\{c_{\ell+1} = \cdots = c_{2m} = B\} \, .$$

From the symmetry of our model, $\mathbb{P}\left\{b_n^{maj} \neq B_\ell^{maj} \mid c_{\ell+1} = \cdots = c_{2m} = B\right\} = 1/2$. Thus

$$\mathbb{P}\left\{b_n^{maj} \neq B_\ell^{maj}\right\} \geq \frac{\mathbb{P}\{c_{\ell+1} = \cdots c_{2m} = B\}}{2} \, .$$

To estimate the probability on the right-hand side, we use (5.2.4), which implies

$$\mathbb{P}\{c_i = B\} = \int_{f(R_i)}^{1} (x(1-x))^{\frac{\ell-1}{2}} \frac{\Gamma(\ell+1)}{\Gamma^2\left(\frac{\ell+1}{2}\right)} dx \,.$$

If $R_m = m/\ell$ and $c_{\ell+1} = \cdots = c_{i-1} = B$, where $\ell < i \le 2\ell$, then $R_{i-1} = m/i$. Since $0 \le p \le 1/2$,

$$f(R_{i-1}) = (1-2p)\frac{m}{i} + p \le \max\left(\frac{1}{2}, \frac{m}{i}\right) \le \frac{\ell-1}{\ell} = 1 - \frac{1}{\ell} \,.$$

Therefore,

$$\min_{\ell < i \le 2\ell} \mathbb{P}\{c_i = B \mid c_{\ell+1} = \cdots = c_{i-1} = B\} \ge h_\ell \,,$$

as claimed. ■

5.6 Concluding remarks

In this chapter we study the majority rule for guessing the initial bit values at the roots of a random recursive ℓ-dag in a broadcasting model. The main result of the chapter characterizes the values of the mutation probability for which the majority rule performs strictly better than random guessing. Even in this exact model, many interesting questions remain open. For example, we do not have sharp bounds for the probability of error. It would also be interesting to study other, more sophisticated, classification rules that take the structure of the observed ℓ-dag into account. In particular, the optimal probability of error (as a function of ℓ and the mutation probability p) is far from being well understood. For an initial study of localizing the root vertices, we refer the interested reader to Chapter 3.

Chapter 6

The random friendship tree

Contents

Abstract

We study a random recursive tree model featuring complete redirection called the random friend tree and introduced by Saramäki and Kaski [125]. Vertices are attached in a sequential manner one by one by selecting an existing target vertex and connecting to one of its neighbours (or friends), chosen uniformly at random. This model has interesting emergent properties, such as a highly skewed degree sequence. In contrast to the preferential attachment model, these emergent phenomena stem from a local rather than a global attachment mechanism. The structure of the resulting tree is also strikingly different from both the preferential attachment tree and the uniform random recursive tree: every edge is incident to a macro-hub of asymptotically linear degree, and with high probability all but at most $n^{9/10}$ vertices in a tree of size n are leaves. We prove various results on the neighbourhood of fixed vertices and edges, and we study macroscopic properties such as the diameter and the degree distribution, providing insights into the overall structure of the tree. We also present a number of open questions on this model and related models.

This Chapter is based on a joint work with Louigi Addario-Berry, Luc Devroye, Serte Donderwinkel, Céline Kerriou and Gábor Lugosi (Berry, Briend, Devroye, Donderwinkel, Kerriou, and Lugosi [15]).

6.1 Introduction

Growing networks. Various real-life phenomena, including contagion, social networks, rumour spreading, and the internet, have been described by models of growing networks (see e.g. Kumar, Bhat, and Panda [95]). Among these models, preferential attachment, introduced by Barabási and Albert [14] is arguably the most well-studied. In this model, vertices arrive one by one, and at each time a new vertex connects to one or more existing vertices with probability proportional to their degree. The degree sequence of this model satisfies the so-called scale-free property, which is often also observed in real-world models. In contrast to models such as the configuration model or the inhomogeneous random graph model, this property of the degree sequence is an intrinsic result of the dynamics rather than a pre-imposed characteristic. This makes the preferential attachment and its related models popular tools for understanding why real-world models may develop in this way. These models however require the knowledge of the full degree sequence in order to attach a new vertex. This requirement is unnatural for real-world networks and impractical in implementation.

The model. The friend tree is a randomly growing network of which the dynamic also autonomously produces highly skewed degree sequences, but whose attachment rule is based on redirection and requires only local information. Models involving redirection were introduced by Kleinberg et al. [90] in directed, rooted graphs. In these models, a new vertex connects to a uniformly random vertex, or, with probability p, it connects to the ancestor of a randomly selected vertex. This mechanism, called *directed redirection*, yields a shifted linear preferential attachment rule, where the new vertex connects to a vertex with degree d with probability proportional to $d-2+1/p$. A variant was studied by Banerjee, Bhamidi, and Huang [13], where a new vertex attaches to the graph by randomly sampling a vertex and attaching to the endpoint of a path of random length directed towards the root. The undirected version of the model that we study was introduced by Saramäki and Kaski [125] and yields strikingly different graphs. In the works of Saramäki and Kaski [125] and Evans and Saramäki [66], the authors make the claim that the tree has the same law as a preferential attachment tree. This turns out to be inaccurate, as was noted by Cannings and Jordan [35]. In the undirected version, newly added vertices connect to a neighbour of a randomly selected vertex. More precisely, the starting tree T_2 consists of a single

edge joining vertices labelled 1 and 2. Inductively, for $n \geq 2$, let $V_n \in \{1,\dots,n\}$ be chosen uniformly at random and let W_n be a uniformly random neighbour of V_n in T_n. Then build T_{n+1} from T_n by attaching a new vertex labelled $n+1$ to the vertex W_n, see Figure 6.1. Note that, for all $n \geq 2$, the vertex set of T_n is $\{1,\dots,n\}$. Moreover, setting $W_1 = 1$, then the edge set of T_n is $\{\{m+1, W_m\}, 1 \leq m \leq n-1\}$. We call T_n the 1-*step friend tree*, inspired by the following picture. In a 1-step friend tree a person selects a random stranger and befriends a uniformly random friend of theirs. A 2-step friend tree would correspond to, instead, befriending a uniformly random friend of the stranger's random friend. In most of this work, we refer to the 1-*step friend tree* simply as the *random friend tree (RFT)*.

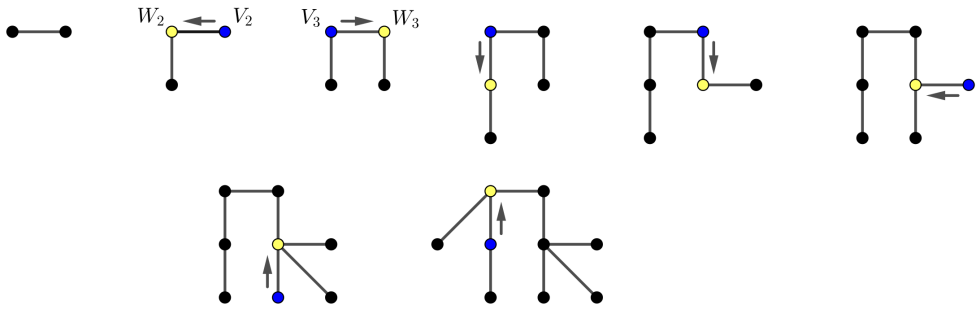

Figure 6.1: A realisation up to $n = 9$, with V_n in blue and W_n in yellow.

Motivation and challenges. This model gives rise to numerous interesting emergent phenomena that make it worth studying. A first feature of the model is a rich-get-richer mechanism. In the preferential attachment model, it is embedded into the dynamics that vertices with many neighbours accumulate more neighbours. In the friend tree, vertices that have many neighbours *with low degree* accumulate more neighbours. It turns out that, the highest degree vertices mostly have leaves as neighbours, so the growth of high-degree vertices is mostly determined by their degree. In fact, the reinforcement is strengthened by a second effect: the larger the degree of a vertex, the less likely it is for its neighbours to increase their degree, so the more likely it is for the high degree vertex itself to increase its degree. Unlike in the preferential attachment model, this is a result of the dynamics rather than a built-in feature.

Furthermore, the model is part of a whole family of models that in some sense interpolate between the uniform random recursive tree (URRT) and the linear preferential

attachment tree (PA tree). Indeed, in the k-step friend tree, new vertices attach to the endpoint of a random walk of length k that starts at a random vertex. If $k = 0$, the resulting model is the URRT. If k was chosen so large that the random walk is perfectly mixed, the resulting model would be a PA tree. In other words, for k large, the starting point of the random walk has a negligible effect and the end point is distributed proportionally to the degree of each vertex. Our work demonstrates several features of the 1-step friend tree which are remarkably different from both the URRT and the PA tree. It is therefore worth investigating what range of behaviour can be observed in the entire family. In a related model, Engländer et al. [64] study a "random walk tree builder" where a tree is grown by a random walk. More precisely, a walker is moving at random on the tree and at each time step n, with probability $n^{-\gamma}$, a neighbour is added to the vertex where the walker is. They prove that this model is actually a PA tree for an appropriate choice of γ. Unlike the friend tree model, where a new random walk is started at every time step, a single random walk is able to produce a tree displaying a rich gets richer phenomenon. This further motivates the study of the friend tree model in its whole range, and suggests the investigation of its possible links to the "random walk tree builder" model.

The challenges of studying 1-step friend trees are numerous. For example, even if the process grows locally (one needs to know the neighbours of the randomly picked vertex to understand the connection probabilities), tracking only local information does not suffice to study how degrees evolve over time. Indeed, to understand how the vertex degrees change within two time steps, one must keep track of the degrees of second neighbours, and in general, for t time steps one must track the t'th neighbourhood of every vertex. Note that this challenge does not arise in either directed redirection or the preferential attachment model; in those models the growth of a vertex degree only depends on the vertex degree itself, so degrees can be tracked on their own without considering the global structure of the tree. For k-step friend trees, the dependencies grow stronger as k increases, bringing in new challenges that we do not attempt to tackle in the current work.

Results and comparison. Some of the structural properties of random friend trees are comparable to those of the URRT. In Theorem 6.6, we show that the diameter is of logarithmic order almost surely (like for the URRT and the PA tree ([**?**])). Moreover, as we show in Theorems 6.7 and 6.8 respectively, both in the random friend tree and in the URRT, the largest distance to the nearest leaf in the n-vertex tree is $\Theta(\log(n)/\log\log(n))$ in probability.

However, the interaction between neighbouring vertices in the attachment procedure yields significant structural differences between the random friend tree and both the URRT and the PA tree. The degree sequence might be the most illustrative of this difference. Regarding the high-degree vertices, we prove in Theorem 6.2 that "hubs" of linear

degree appear almost surely, whereas in a URRT the maximum degree is logarithmic (Devroye and Lu [52]) and in linear preferential attachment tree the largest degree is of order $\sqrt{n}$ (Van Der Hofstad [138, Theorem 1.17]). In fact, the dynamics of low-degree vertices 'feeding their neighbours' implies that for every edge, at least one of the endpoints has asymptotically linear degree almost surely, so that a highly modular network emerges. This phenomenon has not been observed in any other random tree model, as far as the authors of this chapter are aware. The existence of linear degree hubs also implies that two uniformly random vertices in T_n are at distance two from each other with probability bounded away from zero, while in both the URRT and the PA tree typical distances grow logarithmically [50],[**?** , Theorem 8.1]. As for low-degree vertices, both in PA trees and URRT an asymptotically positive fraction of vertices have degree at least two, but we shall show that a random friend tree of size n has $n - o(n^{0.9})$ leaves. While a URRT of size n has on average $n/2^{k-1}$ vertices of degree at least k for fixed k (Janson [78]), for friend trees, asymptotically, this number sits between $n^{0.1}$ and $n^{0.9}$ (see Theorem 6.11). The proliferation of hubs and the interaction between neighbours block most leaves from ever growing their degree. Proposition 6.13 shows that most leaves remain leaves forever[1], whereas for URRT and PA trees, the degree of every vertex is a.s. unbounded.

Earlier work. The only previous rigorous result on random friend trees the authors are aware of was obtained by Cannings and Jordan [35], who show that in the 1-step friend tree, $n - o(n)$ of the vertices are leaves almost surely. Random friend trees were also studied in the physics literature by Krapivsky and Redner [93]. In that work, the authors use simulations and non-rigorous arguments to study the distribution of size of the largest degrees and the order of growth of the number of non-leaves, and estimate the degree distribution restricted to the bounded-degree vertices. They conjecture that, for any fixed k, the number of degree k vertices is of the order of n^{μ} for $\mu \approx 0.566$. Moreover, they conjecture that among the non-leaf vertices, the proportion of degree k vertices is of the order of $k^{-(1+\mu)}$. We discuss their estimates in Section 6.7.

Outline. First, in Section 6.2 we introduce notation that we use throughout the chapter. We then present our main results for friend trees. Our findings can naturally be divided into local and global properties, which we present in Section 6.3 and 6.4 respectively. In Sections 6.5 and 6.6 we prove our main results. Finally, Section 6.7 contains some open questions about random friend trees.

[1] So one might argue that 'random loneliness tree' is in fact a more appropriate name for our model.

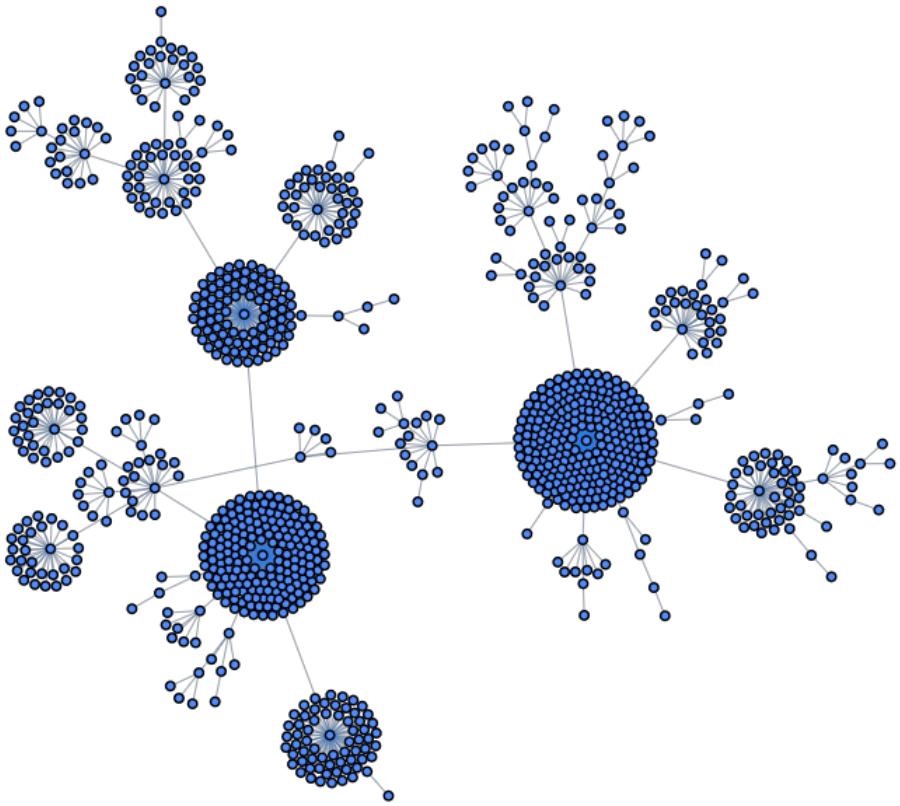

Figure 6.2: A realisation of T_n with $n = 1000$.

6.2 Notation

For a graph G and a vertex v of G, write $\mathcal{N}(v;G)$ for the neighbourhood of v in G and $\mathcal{L}(v;G)$ for the set of leaf neighbours of v in G (i.e. vertices of degree one in $\mathcal{N}(v;G)$). In the rest of the chapter, T_n denotes a tree of size n obtained from the random friend tree model. The set of vertices of T_n is $[n] := \{1,\dots,n\}$, where the label of the vertex is its time of arrival in the tree. Since every integer $k \in \mathbb{N}$ is a vertex of T_n for all $n \geq k$, we take the liberty of referring to integers of $\mathbb{N}$ as vertices. For $v \in T_n$, let $D_n(v) = |\mathcal{N}(v;T_n)|$ be the degree and let $L_n(v) = |\mathcal{L}(v;T_n)|$ be the number of leaf neighbours of v in T_n. Define the random variable

$$Z_v := \liminf_{n\to\infty} \frac{D_n(v)}{n} .$$

A vertex $v \in \mathbb{N}$ is called a *hub* if $Z_v > 0$, that is, if the degree of v is of linear order asymptotically. A vertex w is said to be a *child* of vertex v in T_n if $w \in \mathcal{N}(v;T_n)$ and $v < w$. If w is

a child of v, then v is the *parent* of w. For $i, j \in [n]$, denote by $d_n(i,j)$ the graph distance between vertices i and j in T_n. We also introduce $\mathrm{Diam}_n := \max_{i,j\in[n]} d_n(i,j)$, the diameter of T_n, and let $M_n := \max_{i\le n} \min_{\{\ell:\, D_n(\ell)=1\}} d_n(i,\ell)$ be the maximal distance of any vertex of T_n to its nearest leaf. For integers $n, k \ge 1$ we let $X_n^k = \{v \in [n] : D_n(v) = k\}$ be the number of vertices of degree k in T_n, and let $X_n^{\ge k} = \sum_{j\ge k} X_n^j$.

For any sequence $(x_n)_{n\ge1}$, we define, for all $n \ge 1$, $\Delta x_n := x_{n+1} - x_n$. For non-negative $(x_n)_{n\ge1}$ and positive $(y_n)_{n\ge1}$ we write $x_n = O(y_n)$ and $y_n = \Omega(x_n)$ if $\limsup_{n\to\infty} \frac{x_n}{y_n} < \infty$ and we write $x_n = o(y_n)$ and $y_n = \omega(x_n)$ if $\lim_{n\to\infty} \frac{x_n}{y_n} = 0$. We say that $x_n = \Theta(y_n)$ if $x_n = O(y_n)$ and $x_n = \Omega(y_n)$ both hold. We also use this notation with a p subscript meaning that the property holds in probability. So, for sequences (X_n) and (Y_n) of non-negative random variables, we write $X_n = O_p(Y_n)$ and $Y_n = \Omega_p(X_n)$ if for all $\varepsilon > 0$, there exist $M > 0$ and $K > 0$ such that for all $n \ge M$, $\mathbf{P}\{X_n > KY_n\} < \varepsilon$. We write $X_n = o_p(Y_n)$ and $Y_n = \omega_p(X_n)$ if for all $\varepsilon > 0$, $\delta > 0$, there exists $M > 0$ such that for all $n \ge M$, $\mathbf{P}\{X_n > \delta Y_n\} < \varepsilon$. We write $X_n = \Theta_p(Y_n)$ if both $X_n = O_p(Y_n)$ and $X_n = \Omega_p(Y_n)$.

6.3 Local results

In this section we state our results regarding the properties of individual vertices and their close neighbourhoods. First, we state a convergence result for the normalised degree.

Theorem 6.1 (Convergence of normalised degree)**.** *For vertex $u \in \mathbb{N}$, the random variables $D_n(u)$, $L_n(u)$ and Z_u, defined in Section 6.2, are such that*

$$\frac{D_n(u)}{n} \to Z_u \text{ and } \frac{L_n(u)}{n} \to Z_u$$

almost surely as $n \to \infty$.

We observe that $\sum_{i\ge1} Z_i \le 1$ almost surely. Indeed, on the event that this sum exceeds 1, then there must be $k \in \mathbb{N}$ and $\delta > 0$ such that $Z_1 + \cdots + Z_k = 1 + \delta$. But this would imply that there is a finite n such that the number of leaves neighbouring vertices $1, \dots, k$ in T_n satisfies $L_n(1) + \cdots + L_n(k) \ge (1 + \delta/2)n$. However, the number of leaves in T_n is at most $n - 1$ deterministically, so this gives a contradiction.

We conjecture that $\sum_{i\ge1} Z_1 = 1$ almost surely, implying that, asymptotically, all but a negligible proportion of the vertices are a leaf next to a hub. We discuss this and other open questions on the law of $(Z_i)_{i\ge1}$ in Section 6.7.

A striking property of the friend tree concerns the degree of an edge.

Theorem 6.2 (Abundance of hubs)**.** *The degree of every edge is almost surely asymptotically linear. That is, for any $m \geq 1$, if W_m is the parent of $m+1$, then $Z_{m+1} + Z_{W_m} > 0$ almost surely.*

We also study the limit of the proportion of vertices that are adjacent to W_n. This theorem shows that the mean of the empirical law of $D_n(W_n)/n$ given T_n converges almost surely, implying that this global property of T_n "stabilizes" as n grows large. We conjecture that, in fact, the empirical law itself converges almost surely to $\sum_{i\geq 0} Z_i \delta_{Z_i}$ with respect to the Prokhorov topology. This is in fact equivalent to the conjecture that $\sum_{i\geq 1} Z_i = 1$ almost surely.

Theorem 6.3 (Expected degree of W_n)**.** *As $n \to \infty$, $\frac{1}{n}\mathbf{E}[D_n(W_n)|T_n]$ has a positive almost sure limit. Moreover, $\frac{1}{n}\mathbf{E}[D_n(W_n)]$ converges to some positive number as $n \to \infty$.*

Another notable property of random friend trees concerns the probability of a vertex having bounded degree.

Theorem 6.4 (Bounded degree)**.** *Let k be a positive integer. Fix $v \in \mathbb{N}$. Then, for all $n \geq v$,*

$$\mathbf{P}\left\{D_{n+j}(v) = k \ \forall j \geq 0 \mid T_n\right\} > c_k \mathbb{I} D_n(v) = k,$$

where the constant $c_k > 0$ only depends on k.

In particular, Theorem 6.4 implies that it is impossible to 'diagnose' which vertices are the hubs, even at a very large time. Indeed, every vertex, no matter how large its degree is, has probability bounded away from zero to never acquire any new neighbours. We prove Theorems 6.1 and 6.2 in Section 6.5.1, the proof of Theorem 6.3 can be found in Section 6.5.2 and Theorem 6.4 is proven in Section 6.5.3.

6.4 Global results

We now present our results on the global properties of random friend trees.

Theorem 6.5 (Typical distances)**.** *For U_n and V_n two uniformly random vertices in T_n, it holds that the distance between U_n and V_n is equal to 2 with probability bounded away from zero. Moreover, the distance between vertex 1 and U_n is at most 2 with probability bounded away from 0.*

The previous result sets the random friend tree apart from the "universality class" of logarithmic trees (and in particular from the uniform random recursive tree and preferential attachment trees), in which typical distances are logarithmic.

The next result shows that, while distances between typical vertices can be very small, the diameter of random friend trees is indeed logarithmic.

Theorem 6.6 (Diameter of random friend trees)**.** *Almost surely*

$$1 \leq \liminf_n \frac{\mathrm{Diam}_n}{\log(n)} \leq \limsup_n \frac{\mathrm{Diam}_n}{\log(n)} \leq 4e\,.$$

In particular, this means that, asymptotically, a path of arbitrary length is present in the tree. Since for every edge at least one of its endpoints is a hub, this implies that an asymptotically unbounded number of hubs are present. We strengthen this statement in Theorem 6.9, where we prove an almost sure polynomial lower bound for the number of hubs in T_n.

We also study the leaf-depth in T_n. The next result implies that, although each vertex is at distance at most 1 from a hub, and will therefore eventually be at distance at most two from the nearest leaf, at fixed times there are still exceptional locations in the graph where the nearest leaf is much further away.

Theorem 6.7 (Leaf depth)**.** *Let M_n be the maximal distance of any vertex to its nearest leaf in T_n. Then*

$$M_n = \Theta\left(\frac{\log(n)}{\log\log(n)}\right) \text{ in probability.}$$

An input to the proof of Theorem 6.7 is the corresponding result for the URRT, which we state as a separate theorem.

Theorem 6.8 (Leaf depth in URRT)**.** *Let M_n' be the maximal distance of any vertex to its nearest leaf in a URRT. Then,*

$$M_n' = \Theta\left(\frac{\log(n)}{\log\log(n)}\right) \text{ in probability.}$$

We prove Theorem 6.7 using Theorem 6.8 and a coupling between URRTs and random friend trees under which distances are at most a factor of two larger in the URRT than in the random friend tree to which it is coupled. This coupling is presented in Lemma 6.17, below.

Another global property of interest is the degree distribution of the tree. We study degree statistics at both ends of the spectrum, for both sub-linear-degree vertices and bounded-degree vertices. We get the following lower bound on the number of hubs in T_n.

Theorem 6.9 (Number of hubs)**.** *There exists a constant $\delta > 0.1$ such that*

$$\frac{\#\{u \in [n] : Z_u > 0\}}{n^\delta} \to \infty \text{ a.s.}$$

The following theorem concerns the number of high degree vertices.

Theorem 6.10 (Abundance of high degree vertices)**.** *For any sequence $(m_n)_{n\geq 1}$, satisfying $m_n = o(n)$, almost surely*

$$\lim_{n\to\infty} X_n^{\geq m_n} = \infty.$$

The next theorem gives polynomial upper and lower bounds for the number of bounded degree vertices.

Theorem 6.11 (Polynomial bounds on low-degree vertices)**.** *There exist constants $0.1 < \delta \leq \lambda < 0.9$ such that, for any $k \geq 2$, almost surely*

$$\lim_{n\to\infty} \frac{X_n^{\geq k}}{n^\delta} = \infty\,,$$

$$\lim_{n\to\infty} \frac{X_n^{\geq k}}{n^\lambda} = 0\,.$$

It has been conjectured by Krapivsky and Redner [93] that a stronger statement is

true for the RFT. They conjecture that there exists a constant $\mu \approx 0.566$ such that, $n^{-\mu} X_n^{\geq k} \to X_k$, where X_k is a non-degenerate random variable. Moreover, they conjecture that $X_n^k / X_n^{\geq 2}$ has an almost sure limit which is $\Theta(k^{-(1+\mu)})$.

Finally, we show that $X_n^{\geq k} = \Theta(X_n^{\geq 2})$ almost surely, for any fixed k.

Theorem 6.12 (Comparing low-degree vertices)**.** *There exists a sequence $(c_k)_{k\geq 2}$ of positive real numbers, such that for any $k \geq 2$,*

$$c_k < \liminf_{n \geq \infty} \frac{X_n^{\geq k+1}}{X_n^{\geq k}} \leq 1$$

almost surely.

We prove Theorem 6.5 in Section 6.6.1 and we prove Theorem 6.6 in Section 6.6.2. The proofs of Theorems 6.7 and 6.8 are in Section 6.6.3. Finally, Theorems 6.9 and 6.10 are proven in Section 6.6.4 and Theorems 6.11 and 6.12 are proven in Section 6.6.5. Although Theorem 6.11 is invoked in the proofs of several of the earlier results, we postpone its proof to later in the chapter, as it is quite technical.

6.5 Proofs of local properties

6.5.1 Hubs

[Proof of Theorems 6.1 and 6.2] We prove Theorem 6.1 with a submartingale argument, deferring a crucial step of the proof to Section 6.6.

Proof. [Proof of Theorem 6.1] Fix $v \in T_n$. Note that $L_{n+1}(v) = L_n(v) + 1$ if $W_n = v$. For $V_n = u$ a neighbour of v, $\mathbf{P}\{W_n = v | V_n = u\} = 1/D_n(u)$. Since V_n is a uniform sample from $[n]$, it follows that

$$\mathbf{P}\{L_{n+1}(n) = L_n(v) + 1 \mid T_n\} = \sum_{u \in \mathcal{N}(v;T_n)} \frac{1}{n} \frac{1}{D_n(u)} .$$

Next, $L_{n+1}(v) = L_n(v) - 1$ if $V_n = v$ and $W_n \in \mathcal{L}(v;T_n)$. Thus,

$$\mathbb{E}\{L_{n+1}(v) \mid T_n\} = L_n(v) - \frac{1}{n}\frac{L_n(v)}{D_n(v)} + \sum_{u \in \mathcal{N}(v;T_n)} \frac{1}{n}\frac{1}{D_n(u)} .$$

Using that $L_n(v) \leq D_n(v)$ and, for a leaf v, $D_n(v) = 1$, we can lower bound the above by $L_n(v) - \frac{1}{n} + \frac{L_n(v)}{n}$. By rearrangement it follows that

$$\mathbb{E}\left\{ \frac{L_{n+1}(v)-1}{n+1} \;\middle|\; T_n \right\} \geq \frac{L_n(v)-1}{n}.$$

Thus, for any $m \in \mathbb{N}$, the process $((L_n(v)-1)/n, n \geq m-1)$ is a submartingale relative to the filtration generated by the random friend tree process. It is bounded, so it converges almost surely. Furthermore, by the trivial inequalities

$$\frac{L_n(u)}{n} \leq \frac{D_n(u)}{n} \leq \frac{L_n(u)+X_n^{\geq 2}}{n},$$

the joint convergence follows from Theorem 6.11 below, stating that $n^{-1}X_n^{\geq 2} \to 0$ almost surely. ∎

Recall that a vertex u is a hub if $Z_u > 0$. Theorem 6.2 implies that each edge has at least one endpoint that is a hub, which, in particular, shows that hubs exist.

Proof. [Proof of Theorem 6.2] Fix $m \in \mathbb{N}$. For $n \geq m+1$, write $D_n := D_n(m+1) + D_n(W_m)$ for the total number of neighbours of vertex $m+1$ and vertex W_m at time n. Write $L_n := L_n(m+1) + L_n(W_m)$ for the number of those neighbours that are leaves. We have $L_{m+1} \geq 1$ and $D_{m+1} \geq 3$. Note that D_n is non-decreasing and that if $D_{n+1} = D_n + 1$ then also $L_{n+1} = L_n + 1$, so $\Delta(L_n, D_n) \in \{(1,1),(0,0),(-1,0)\}$.

Moreover,

$$\mathbf{P}\{\Delta(L_n, D_n) = (1,1) \mid (L_i, D_i), m+1 \leq i \leq n\} \geq \frac{1}{n}\left(L_n + \frac{1}{D_n}\right), \tag{6.5.1}$$

since, to have $\Delta(L_n, D_n) = (1,1)$, it suffices that either $V_n \in \mathcal{L}(m+1; T_n) \cup \mathcal{L}(W_m; T_n)$ or else that $\{V_n, W_n\} = \{m+1, W_m\}$. We also have

$$\mathbf{P}\{\Delta(L_n, D_n) = (-1,0) \mid (L_i, D_i), m+1 \leq i \leq n\} \leq \frac{\min(2, L_n)}{n}, \tag{6.5.2}$$

since if $\Delta(L_n, D_n) = (-1,0)$, then $V_n \in \{m, W_m\}$ and $W_n \in \mathcal{L}(m; T_n) \cup \mathcal{L}(W_m; T_n)$.

We claim that $D_n \to \infty$ almost surely. We fix $k \in \mathbb{N}$ and show that the stopping time $\tau_k = \min\{n \geq m+1 : D_n \geq k\}$ is finite almost surely. On the event that $\tau_k < n$, (6.5.1) implies that $\Delta D_n \mid T_n$ stochastically dominates a $\text{Bernoulli}(\frac{1}{nk})$ random variable. If $\mathcal{B}_i$ are independent $\text{Bernoulli}(\frac{1}{ik})$ random variables, $\sum_{i=m+1}^{n} \mathcal{B}_i \to \infty$ almost surely, so $\tau_k < \infty$ almost surely. It then follows that $\#\{n : L_n > 0\} = \infty$ almost surely. Indeed, $D_n \to \infty$ implies that $\Delta(D_n, L_n) = (1,1)$ infinitely many times, so if L_n ever hits zero it almost surely becomes positive again.

Let $(J_k, k \geq 0)$ be the sequence of jump times of the process $(\Delta(L_n, D_n))_n$, that is $J_0 = m+1$ and $J_k = \min\{\ell > J_{k-1} : \Delta(L_{\ell-1}, D_{\ell-1}) \neq (0,0)\}$ for $k \geq 1$. This sequence has infinite length because $D_n \to \infty$ almost surely.

Our goal is to show that L_n grows linearly. To do so, we couple L_n to an urn process. Inequalities (6.5.1) and (6.5.2) suggest that the growth of L_n is similar to the growth of the number of black balls in a standard Pólya urn with black and white balls. One difference being that, at time n, with probability at most $2/n$, a black ball is replaced by a white ball. Nonetheless, we exhibit a coupling between L_n and the number of black balls in a standard Pólya urn of black and white balls of size n. More precisely, a coupling where L_n is greater than the number of black balls in the Pólya urn of size n. This coupling can fail, meaning that it is valid until a random time S that is finite with positive probability. We say that the coupling succeeds if $S = \infty$. We show that this coupling succeeds with probability greater than 0, and that, if it fails, we may try again by starting a new coupling at a subsequent time. This guarantees that one of the coupling attempts is successful, proving that L_n grows linearly, because the number of black balls in a standard Pólya urn grows linearly almost surely.

The coupling is started at a time where L_n is at least 10. The bounds (6.5.1) and (6.5.2) on the transition probabilities show that the process $(L_{J_k}, k \geq 0)$ stochastically dominates a simple symmetric random walk reflected at 0. This, in particular, implies that there are infinitely many n such that $L_n \geq 10$, because $J_k \to \infty$ almost surely.

Let ρ_1 be the first time for which $L_{\rho_1} \geq 10$ (and so $L_{\rho_1} = 10$). A first coupling is started from time ρ_1. For any k, if the kth coupling fails, let ρ_{k+1} be the first time after the failure at which $L_{\rho_{k+1}} \geq 10$ and start the $(k+1)$st coupling from that time. We show that there is a $c > 0$ so that for each k, given that couplings $1, \ldots, k-1$ all failed, the kth coupling succeeds with probability at least c. This implies that there is an almost surely finite M so that the Mth coupling succeeds.

So let us fix some $N > 0$ and condition on $\rho_k = N$. Now define B_n the number of black balls in a Pólya urn starting at time N with 5 black balls and $N-5$ white balls (with the standard replacement rule that a drawn ball is replaced along with one extra ball of the same colour). Note that $L_N = L_{\rho_k} \geq 10$ so $L_N - B_N \geq 5$. We couple L_n and B_n from time N onwards and we say the coupling fails at time S if S is the first time $S > N$ at which $L_S - B_S \leq 4$. If the coupling never fails we set $S = \infty$. Then, for $n \in [N, S]$, we can couple L_n and B_n such that if $B_{n+1} = B_n + 1$ then $L_{n+1} = L_n + 1$. From this coupling, for $n \geq N$,

$$\mathbf{P}\{\Delta(L_n - B_n) = 1 \mid S > n\} \geq \frac{L_n - B_n}{n}$$
$$\mathbf{P}\{\Delta(L_n - B_n) = -1 \mid S > n\} \leq \frac{2}{n}.$$

This means that until the coupling fails, $L_n - B_n$ can be coupled to a symmetric random walk for which an increment with value 1 is twice as likely as an increment with value -1. We introduce R_k, a random walk with $R_0 = 5$ and

$$R_{k+1} - R_k = \begin{cases} +1 \text{ with probability } \frac{2}{3}, \\ -1 \text{ with probability } \frac{1}{3}\,. \end{cases}$$

Set $I_0 = N$ and let $I_{k+1} = \min\{j \geq I_k : \Delta(L_n - B_n) \neq 0\}$ be the kth jump time of $L_n - B_n$. Then $(R_k, k \geq 0)$ and $(L_n - B_n, n \geq N)$ can be coupled such that if $I_k \leq S$

$$L_{I_k} - B_{I_k} \geq R_k.$$

With positive probability $R_k > 4$ for all k, so with positive probability, not depending on N, $S = \infty$. This shows that, almost surely, one of the coupling attempt succeeds. Suppose that the kth coupling succeeds and that $\rho_k = N$. Then, for B_n as above, $L_n \geq B_n$ for $n \geq N$, so since

$$\lim_{n\to\infty} \frac{B_n}{n} > 0 \text{ almost surely,}$$

by a standard result on Pólya urns (see Mahmoud [105, Section 3.2]), we also get that

$$\liminf_{n\to\infty} \frac{L_n}{n} > 0 \text{ almost surely,}$$

which implies the statement. ∎

6.5.2 Expected degree of W_n

[Proof of Theorem 6.3]

We show the statement using the almost sure martingale convergence theorem by identifying a supermartingale. Set $Y_n := \mathbf{E}[D_n(W_n)|T_n]$ so that Y_n is adapted to $\sigma(T_n)$ and note that

$$Y_n = \frac{1}{n}\sum_{i\in[n]} \mathbb{E}\{D_n(W_n) \mid T_n, V_n = i\} = \frac{1}{n}\sum_{i\in[n]} \left(\frac{1}{D_n(i)} \sum_{j\sim_{T_n} i} D_n(j) \right).$$

To identify the supermartingale, we study

$$\mathbb{E}\{(n+1)Y_{n+1} \mid T_n\} = \mathbb{E}\left\{ \sum_{i\in[n+1]} \sum_{j\sim_{T_{n+1}} i} \frac{D_{n+1}(j)}{D_{n+1}(i)} \,\middle|\, T_n \right\}.$$

Observe that the only randomness in Y_{n+1}, conditional on T_n, comes from the choice of W_n. It holds that $D_{n+1}(W_n) = D_n(W_n) + 1$, and the new neighbour of W_n (vertex $n+1$) has degree 1. Moreover, $D_{n+1}(n+1) = 1$, because every vertex starts as a leaf, and its neighbour is W_n. Finally, $D_{n+1}(i) = D_n(i)$ for all other i. Therefore, we get the following equalities for the different terms in $\sum_{i\in[n+1]} \sum_{j\sim_{T_{n+1}} i} \frac{D_{n+1}(j)}{D_{n+1}(i)}$:

$$\sum_{j\sim_{T_{n+1}} i} \frac{D_{n+1}(j)}{D_{n+1}(i)} = \begin{cases} \frac{1}{D_n(W_n)+1} + \sum_{j\sim_{T_n} W_n} \frac{D_n(j)}{D_n(W_n)+1} & \text{for } i = W_n \\ D_n(W_n)+1 & \text{for } i = n+1 \\ \frac{\mathbb{I}i\sim_{T_n} W_n}{D_n(i)} + \sum_{j\sim_{T_n} i} \frac{D_n(j)}{D_n(i)} & \text{otherwise.} \end{cases}$$

Combining these cases, we see that

$$\begin{aligned} &\mathbb{E}\{(n+1)Y_{n+1} \mid T_n\} \\ &= \sum_{i\in[n]} \sum_{j\sim_{T_n} i} \frac{D_n(j)}{D_n(i)} \\ &\quad + \mathbb{E}\left\{ \frac{1}{D_n(W_n)+1} - \left(\frac{1}{D_n(W_n)} - \frac{1}{D_n(W_n)+1} \right) \sum_{j\sim_{T_n} W_n} D_n(j) \,\middle|\, T_n \right\} \\ &\quad + \mathbb{E}\left\{ D_n(W_n) + 1 + \sum_{j\sim_{T_n} W_n} \frac{1}{D_n(j)} \,\middle|\, T_n \right\}. \end{aligned}$$

Then, using that $D_n(j) \geq 1$ for all $j \sim_{T_n} i$ we get that the second term on the right hand side is positive. To get an upper bound for the third term, we again use that $D_n(j) \geq 1$ to get that

$$\mathbb{E}\{(n+1)Y_{n+1} \mid T_n\} \leq \sum_{i\in[n]} \sum_{j\sim_{T_n} i} \frac{D_n(j)}{D_n(i)} + \mathbb{E}\{2D_n(W_n) + 1 \mid T_n\} \leq (n+2)Y_n + 1,$$

so that $\mathbb{E}\{Y_{n+1}/(n+2) \mid Y_n\} \leq Y_n/(n+1) + 1/(n+1)^2$, and therefore $Y_n/(n+1) - \sum_{i=1}^{n} 1/i^2$ is a supermartingale in the filtration generated by T_n, and therefore has an almost sure limit. Since $1/i^2$ is summable, it follows that $\frac{1}{n}\mathbb{E}\{D_n(W_n) \mid T_n\}$ has an almost sure limit.

To see that the limit is positive, note that Theorem 6.2 implies that for any $\varepsilon > 0$, there is a $\delta > 0$ so that $\mathbf{P}\{Z_1 + Z_2 > \delta\} > 1 - \varepsilon$.

Then, observe that

$$\mathbb{E}\left\{ \tfrac{1}{n} D_n(W_n) \,\middle|\, T_n \right\} > \tfrac{1}{n} D_n(1)\mathbf{P}\{W_n = 1 \mid T_n\} + \tfrac{1}{n} D_n(2)\mathbf{P}\{W_n = 2 \mid T_n\}$$

$$\geq \frac{L_n(1)^2 + L_n(2)^2}{n^2},$$

since $D_n(i) \geq L_n(i)$ and $\mathbf{P}\{W_n = i \mid T_n\} \geq L_n(i)/n$ for $i = 1, 2$. Then, note that, if $L_n(1)+L_n(2) > \delta n/2$ then

$$\frac{L_n(1)^2 + L_n(2)^2}{n^2} \geq (\max\{L_n(1)/n, L_n(2)/n\})^2 \geq (\delta/4)^2,$$

so

$$\mathbf{P}\left\{\tfrac{1}{n}\mathbb{E}\{D_n(W_n) \mid T_n\} > \delta^2/16\right\} \geq \mathbf{P}\{L_n(1) + L_n(2) > \delta n/2\}.$$

But, $\frac{1}{n}(L_n(1) + L_n(2)) \to Z_1 + Z_2$ almost surely, so

$$\liminf_{n\to\infty} \mathbf{P}\{L_n(1) + L_n(2) > \delta n/2\} > 1 - \varepsilon$$

so also

$$\liminf_{n\to\infty} \mathbf{P}\left\{\tfrac{1}{n}\mathbb{E}\{D_n(W_n) \mid T_n\} > \delta^2/16\right\} > 1 - \varepsilon,$$

which implies the statement.

The convergence in expectation follows from the bounded convergence theorem, since $\frac{1}{n}D_n(W_n) \leq 1$ deterministically.

6.5.3 Eternal leaves and eternal degree k vertices

[Proof of Theorem 6.4] Note that, for any integer i, $D_n(i)$ is increasing in n and therefore it has an almost sure limit (that might be infinite). For a vertex ℓ that is a leaf at time m, we say it is *temporary* if $\lim_{n\to\infty} D_n(\ell) > D_m(\ell) = 1$. Otherwise we call it *eternal*. Similarly, we call a vertex v that has degree k at time m *temporary* if $\lim_{n\to\infty} D_n(v) > D_m(v) = k$ and otherwise we call it *eternal*. Informally, our next proposition says that, only a bounded number of leaves next to a given hub ever stop being a leaf.

Proposition 6.13. *For $n \geq v$, let $S_n(v)$ be the number of temporary leaves attached to v at time n. If $\mathbf{P}\{Z_v > 0\} > 0$, then conditional on $Z_v > 0$, $S_n(v)$ is tight, that is, for all $\varepsilon > 0$ there exists a constant $M > 0$ such that $\mathbf{P}\{S_n(v) > M \mid Z_v > 0\} < \varepsilon$ for all $n \geq v$.*

Proof.

Fix $\varepsilon > 0$. Suppose v is a hub, that is $Z_v = \lim_{n\to\infty} \frac{D_n(v)}{n} > 0$. This implies that there is a δ and a $N > v$ such that $\mathbf{P}\{\forall n \geq N\ L_n(v) > \delta n \mid Z_v > 0\} > 1 - \varepsilon/2$. We show that there exists a constant K such that $\mathbf{E}[S_k(v)\mathbb{I}\forall n \geq N\ L_n(v) > \delta n] < K$ for all $k \geq N$. We first show that this

implies the statement. Write $\mathbf{P}\{Z_v > 0\} = \rho$, so that $\mathbf{E}[S_k(v)\mathbb{I}\forall n \geq N\ L_n(v) > \delta n \mid Z_v > 0] < K/\rho$. Observe that $L_n(v) > \delta n\ \forall n \geq N$ implies $Z_v > 0$. Then, we see that for $k \geq N$ and $\varepsilon > 0$,

$$\begin{aligned}
&\mathbf{P}\left\{S_k(v) > \tfrac{2K}{\rho\varepsilon} \mid Z_v > 0\right\} \\
&\leq \mathbf{P}\{\exists n \geq N : L_n(v) \leq \delta n \mid Z_v > 0\} + \mathbf{P}\left\{S_k(v) > \tfrac{2K}{\rho\varepsilon}, \forall n \geq N\ L_n(v) > \delta n \mid Z_v > 0\right\} \\
&\leq \varepsilon/2 + \mathbf{P}\left\{S_k(v) > \tfrac{2K}{\rho\varepsilon} \mid \forall n \geq N\ L_n(v) > \delta n\right\}\mathbf{P}\{L_n(v) > \delta n\ \forall n \geq N \mid Z_v > 0\} \\
&\leq \varepsilon/2 + \tfrac{\varepsilon\rho}{2K}\mathbb{E}\{S_k(v) \mid \forall n \geq N\ L_n(v) > \delta n\}\mathbf{P}\{L_n(v) > \delta n\ \forall n \geq N \mid Z_v > 0\} \\
&= \varepsilon/2 + \tfrac{\varepsilon\rho}{2K}\mathbb{E}\{S_k(v)\mathbb{I}\forall n \geq N\ L_n(v) > \delta n \mid Z_v > 0\} < \varepsilon,
\end{aligned}$$

where we use Markov's inequality in the penultimate line.

We now show that there exists a constant K such that $\mathbf{E}[S_k(v)\mathbb{I}\forall n \geq N\ L_n(v) > \delta n] < K$ for all $k \geq N$. Note that, at any time $M \geq k$, if w is a leaf neighbouring v, then w stops being a leaf if $V_M = v$ and $W_M = w$. Conditionally on T_M, this occurs with probability $1/(MD_M(v))$, so the probability that this happens for some vertex in $\mathcal{L}(v;T_k)$ is at most $k/(MD_M(v))$, because $|\mathcal{L}(v;T_k)| < k$. Therefore, the probability that the number of leaves in $\mathcal{L}(v;T_k)$ that are no longer leaves increases at time $M \geq k$ satisfies

$$\begin{aligned}
&\mathbf{P}\{\Delta|\mathcal{L}(v;T_k) \cap \mathcal{L}(v;T_M)^c| = 1, L_n(v) > \delta n\ \forall n \geq N\} \\
&\leq \mathbf{P}\{\Delta|\mathcal{L}(v;T_k) \cap \mathcal{L}(v;T_M)^c| = 1, L_M(v) > \delta M\} \leq \frac{k}{\delta M^2}.
\end{aligned}$$

Therefore

$$\mathbf{E}[S_k(v)\mathbb{I}L_n(v) > \delta n\ \forall n \geq N] \leq \sum_{M \geq k} \frac{k}{\delta M^2} \leq K$$

for some constant K not depending on k, which proves the claim. ∎

The next corollary, stating that the number of eternal leaves attached to any edge grows linearly with high probability, is a consequence of the proposition above and Theorem 6.2. Indeed, since every edge has linear degree almost surely, given the presence of edge (u,v), either u or v has probability at least $1/2$ of being a hub.

Corollary 6.14. *Fix an edge (u,v) and for $n \geq \max(u,v)$ define $E_n(u,v) := |\{w \in \mathcal{N}(u;T_n) \cup \mathcal{N}(v;T_n) : w$ is an eternal leaf$\}|$. Then, $E_n(u,v)$ grows linearly with high probability, that is, there exists $\delta > 0$ such that for every $\varepsilon > 0$ there is a $N = \mathcal{N}(u,v)$ such that*

$$\mathbf{P}\{\exists n > N : E_n(u,v) < \delta n\} < \varepsilon\ .$$

We now prove Theorem 6.4.

Proof. [Proof of Theorem 6.4] Fix $v \in \mathbb{N}$. We first prove the statement for $k = 1$, and then discuss how to adapt the proof to general k. If $D_n(v) = 1$ then let w denote the unique neighbour of v in T_n. We show that v is an eternal leaf with positive probability by showing that, with positive probability, the vertex w acquires a large number of leaf neighbours, ensuring that both the degree of w grows and that, when a new vertex is attached to a uniform neighbour of w, it is unlikely that v is chosen. Fix N and let

$$\tau = \min\{t \in \mathbb{N} : \#\{n < i \le t : V_i \in \{v, w\}\} = N\}$$

be the random time at which a new vertex is attached to a random neighbour of either v or w exactly N times since time n. Define A_N as the event $\{\#\{n < i \le \tau : V_i = v\} = N\}$. Since V_i is chosen uniformly at random, with probability 2^{-N}, $V_i = v$ exactly N times between times n and τ. So, $\mathbf{P}\{A_N\} = 2^{-N}$. Recall that $D_n(v) = 1$, so conditionally on A_N, $D_\tau(v) = 1$ because for all $i \in [n, \tau]$, $V_i \neq u$, and thus $W_i \neq v$. Moreover, $L_\tau(w) \ge N$, because, conditioned on A_N, when $V_i = v$, $W_i = w$, so that a new leaf is attached to w. Since for all $i \in [n, \tau]$ $V_i \neq u$, these leaves stay leaves until time τ. We show that, on the event A_N, v is an eternal leaf with positive probability, because it is likely that w continues to acquire many leaf neighbours beyond time τ, making it unlikely for the degree of v to grow.

For $j \ge 0$, set $a_j = \tau \cdot 2^j$ and $\ell_j = N(5/4)^j$ so that $a_{j+1} - a_j = a_j$ and $\ell_{j+1} - \ell_j = \ell_j/4$. Define the following events for $j \ge 0$:

$$\begin{aligned}
E_j &:= \{D_{a_{j+1}}(v) > 1\},\\
F_j &:= \{\#\{i \in (a_j, a_{j+1}] : V_i = w\} > m(j+1)\},\\
G_j &:= \{L_{a_{j+1}}(w) < \ell_{j+1}\},\\
B_j &:= E_j \cup F_j \cup G_j,
\end{aligned}$$

where $m > 0$ is such that $m < N/20$. In words, E_j is the event that vertex v is not a leaf in $T_{a_{j+1}}$. The event F_j corresponds to the event that between steps $(a_j, a_{j+1}]$, new vertices attach to a random neighbour of w more than $m(j+1)$ times and G_j corresponds to the event that w has less than ℓ_{j+1} leaf neighbours in $T_{a_{j+1}}$. It suffices to prove that

$$\mathbf{P}\left\{\bigcup_{j\ge 0} B_j \mid A_N\right\} < 3/4\,. \tag{6.5.3}$$

Indeed,

$$\mathbf{P}\{v \text{ is an eternal leaf}\} = \mathbf{P}\left\{\cap_{j\ge 0} E_j^c\right\} \ge \mathbf{P}\left\{\cap_{j\ge 0} B_j^c\right\}\,,$$

since $B_j^c \subset E_j^c$. By (6.5.3),

$$\mathbf{P}\left\{\cap_{j\geq 0} B_j^c\right\} \geq \mathbf{P}\left\{\cap_{j\geq 0} B_j^c,\ A_N\right\} = \mathbf{P}\{A_N\}\,\mathbf{P}\left\{\cap_{j\geq 0} B_j^c \mid\ A_N\right\} \geq \frac{1}{4}2^{-N}\ .$$

This implies

$$\mathbf{P}\{v \text{ is an eternal leaf}\} \geq 2^{-(N+2)}\ ,$$

proving Theorem 6.4 for $k = 1$. To show (6.5.3) we begin by noting that

$$\mathbf{P}\left\{ E_j \,\middle|\, \cap_{i=0}^{j-1} B_i^c, A_N \right\} \leq \mathbf{E}\left[\mathbb{E}\left\{ \sum_{i=a_j}^{a_{j+1}} \frac{1}{i\ell_j} \,\middle|\, \tau \right\} \right] \leq \frac{1}{\ell_j}(1+\log(2))\ ,$$

where the first inequality holds since the conditioning implies that vertex w has degree at least ℓ_j from time a_j onwards. Next, since the probability of $V_i = w$ equals to $1/i \leq 1/a_j$ for all $i \geq a_j$, we deduce that, for $j \geq 1$,

$$\mathbf{P}\left\{ F_j \,\middle|\, \cap_{i=0}^{j-1} B_i^c, A_N \right\} \leq \mathbf{P}\left\{\mathsf{Bin}(a_{j+1}-a_j, 1/a_j) > m(j+1)\right\} \leq e^{-m(j+1)/3}\ ,$$

by a Chernoff bound ([76, Theorem 2.1.]). Lastly, under the previous conditioning and given F_j^c, the probability of G_j is less than the probability of creating fewer than $\ell_{j+1}+m(j+1)-\ell_j$ new leaf neighbours for w between steps $(a_j, a_{j+1}]$. Conditionally on F_j^c, the probability of attaching vertex $i+1$ to w at time $i \in (a_j, a_{j+1}]$ is at least $(\ell_j - m(j+1))/a_{j+1}$. Thus

$$\begin{aligned}\mathbf{P}\left\{ G_j \,\middle|\, \cap_{i=0}^{j-1} B_i^c, F_j^c, A_N \right\} &\leq \mathbf{P}\left\{ \mathsf{Bin}\left(a_{j+1}-a_j, \frac{\ell_j - m(j+1)}{a_{j+1}}\right) \leq \ell_{j+1} + m(j+1) - \ell_j \right\} \\ &\leq e^{-\ell_j/90},\end{aligned}$$

where we use that $m < N/20$, so that $m(j+1) \leq \frac{1}{10}\ell_j$ and $\ell_j - m(j+1) > 0$ for all $j \geq 0$, and the bound then follows from the Chernoff bound. Putting everything together gives us that,

$$\mathbf{P}\left\{ B_j \,\middle|\, \cap_{i=0}^{j-1} B_i^c, A_N \right\} \leq \ell_j^{-1}(1+\log(2)) + e^{-m(j+1)/3} + e^{-\ell_j/90},$$

and

$$\begin{aligned}\mathbf{P}\left\{ \bigcup_{j\geq 0} B_j \,\middle|\, A_N \right\} &\leq \sum_{j\geq 0} \mathbf{P}\left\{ B_j \,\middle|\, \cap_{i=0}^{j-1} B_i^c, A_N \right\} \\ &\leq \frac{1}{N}\sum_{j\geq 0}(4/5)^j(1+\log(2)) + \sum_{j\geq 1} e^{-m(j+1)/3} + \sum_{j\geq 0} e^{-\mathcal{N}(5/4)^j/90}.\end{aligned}$$

Now, choose m sufficiently large so that the second sum is smaller than $1/4$, then choose N large enough so that $m < N/20$ and the first and the third sum are both smaller than $1/4$; this proves the statement for $k = 1$. Next, we adapt the statement for general k. Conditionally on T_n, if $D_n(v) = k$ then we define τ as the random time at which a new vertex has been attached to a random friend of vertex v or one of its k neighbours exactly kN times. That is, $\tau = \min\{t : \#\{n < i \le t : V_i \in \{v\} \cup \mathcal{N}(v; T_n)\} = kN\}$. Then, with probability at least $(k+1)^{-kN}$ exactly N of these kN new leaves are attached to each of the k neighbours of v. We call this event $A_N = \{\forall w \in \mathcal{N}(v; T_n) : \#\{n < i \le \tau : V_i = w\} = N\}$. Conditionally on A_N, we show that v is an eternal degree k vertex with positive probability, because it is likely that all of the neighbours of v continue to acquire many leaf neighbours beyond time τ, making it unlikely for the degree of v to grow.

Let $w_1, \dots, w_k$ denote the neighbours of v in T_n. Define the events

$$\begin{aligned}
E_j^k &:= \{D_{a_{j+1}}(v) > k\},\\
F_j^k &:= \bigcup_{l=1}^{k} \{\#\{i \in (a_j, a_{j+1}] : V_i = w_l\} > m(j+1)\},\\
G_j^k &:= \bigcup_{l=1}^{k} \{L_{a_{j+1}}(w_l) < \ell_{j+1}\},\\
B_j^k &:= E_j \cup F_j \cup G_j\,.
\end{aligned}$$

Following the same arguments used in the case of $k = 1$, it follows that, choosing m and N sufficiently large, we have

$$\mathbf{P}\left\{\bigcup_{j\ge 0} B_j^k \mid A_N\right\} < 3/4,$$

which implies the statement for general k. ∎

Note that it follows from Theorem 6.4 that, for E_n^k the number of eternal degree k vertices in T_n,

$$\mathbf{E}\left[E_n^k\right] > c_k\, \mathbf{E}\left[X_n^k\right]. \tag{6.5.4}$$

6.6 Proofs of global properties

6.6.1 Typical distances

[Proof of Theorem 6.5]

Theorem 6.5 is a consequence of Theorem 6.2. Indeed, almost surely, one of vertices 1 and 2 is a hub, so a positive proportion of vertices is a neighbour of vertex 1 or of vertex 2. This implies that with probability bounded away from zero, both U_n and V_n are a neighbour of either vertex 1 or 2, so that the distance between them is two and the distance from U_n to 1 is at most two. More formally, by Theorem 6.2 there is an $i \in \{1,2\}$ and an $\epsilon > 0$ so that $\mathbf{P}\{Z_i > \epsilon\} > \epsilon$. Then, for n large enough, $\mathbf{P}\{D_n(i) > \epsilon n/2\} > \epsilon/2$. For such n,

$$\mathbf{P}\{d_n(U_n,1) \le 2\} \ge \mathbf{P}\{D_n(i) > \epsilon n/2, d_n(i,U_n) = 1\} \ge \epsilon^2/4$$

and

$$\mathbf{P}\{d_n(U_n,V_n) = 2\} \ge \mathbf{P}\{D_n(i) > \epsilon n/2, d_n(i,U_n) = 1, d_n(i,V_n) = 1\} \ge \epsilon^3/8,$$

which proves the statement.

6.6.2 Diameter

[Proof of Theorem 6.6] We show that Diam_n grows logarithmically almost surely, with explicit asymptotic lower and upper bounds. We start by proving a lower bound.

Lemma 6.15. *Almost surely*

$$\liminf_n \frac{\mathrm{Diam}_n}{\log(n)} \ge 1 \,.$$

Proof. Among all the paths of length Diam_n present at time n, let us choose one. Denote it by $(i_0 \to i_1 \to \cdots \to i_{\mathrm{Diam}_n})$. Let us remark that for $n \ge 3$, Diam_n is always at least 2. Vertices i_1 and $i_{\mathrm{Diam}_n - 1}$ are such that at most one of their neighbours is not a leaf (otherwise there would be a path of length $\mathrm{Diam}_n + 1$). This implies that, at time n, conditioned on $V_n = i_1$, with probability at least $1/2$ we have that $W_n \in \mathcal{L}(i_1; T_n)$ (the same holds for $i_{\mathrm{Diam}_n - 1}$). But, if $W_n \in \mathcal{L}(i_1; T_n) \cup \mathcal{L}(i_{\mathrm{Diam}_n - 1}; T_n)$ then the diameter increases by 1. Because $\mathbf{P}\left\{ V_n \in \{i_1, i_{\mathrm{Diam}_n - 1}\} \,\middle|\, \mathrm{Diam}_n \ge 3 \right\} = 2/n$ and $\mathbf{P}\left\{ V_n \in \{i_1, i_{\mathrm{Diam}_n - 1}\} \,\middle|\, \mathrm{Diam}_n = 2 \right\} = 1/n$ (note that if $\mathrm{Diam}_n = 2$, then $i_1 = i_{\mathrm{Diam}_n - 1}$), then

$$\mathbb{E}\{\Delta \mathrm{Diam}_n \mid T_n\} \ge \frac{1}{n} \mathbb{I}\mathrm{Diam}_n \ge 3 + \frac{1}{2n} \mathbb{I}\mathrm{Diam}_n = 2 \,. \tag{6.6.1}$$

To prove the lemma, we first need to show that, almost surely, Diam_n reaches 3 in finite time. Using (6.6.1), there exists a coupling between Diam_n and

$$S_n := 2 + \sum_{i=4}^{n} Z_i \,,$$

where Z_i are independent Bernoulli random variables with parameter $1/(2i)$, such that $\mathrm{Diam}_n \geq S_n$ for $n \geq 3$. Let M be the first time when $\mathrm{Diam}_n = 3$ and M' the first time when $S_n = 3$. By our coupling of Diam_n and S_n, $M \leq M'$. A direct application of [**?** , Exercise 2.9] shows that $\liminf S_n = \infty$ almost surely and therefore M' (and in turn M) are finite almost surely. We can now introduce a coupling of Diam_n from time M onward. Conditionally on $M = m$, (6.6.1) implies that Diam_n can be coupled from m onward to

$$H_n^{(m)} := 3 + \sum_{i=m}^{n} X_i \,,$$

where X_i are independent Bernoulli random variables with parameters $1/i$, such that $\mathrm{Diam}_n \geq H_n^{(m)}$. Another direct application of [**?** , Exercise 2.9] implies that, for fixed m,

$$\liminf_n \frac{H_n^{(m)}}{\log(n)} \geq 1 \text{ a.s.}$$

By our coupling, if $M = m$, $\liminf \frac{\mathrm{Diam}_n}{\log(n)} \geq 1$ almost surely. Using that M is finite almost surely concludes the proof.

■

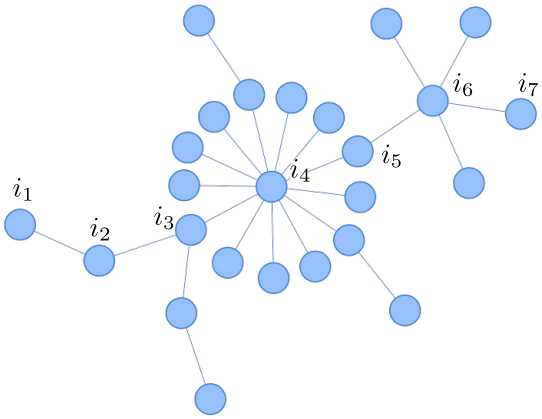

Figure 6.3: Illustration of a RFT of size 25 and diameter 6, with vertices of one of the paths of length 6 highlighted.

Next, we prove an upper bound for the diameter.

Lemma 6.16. *Almost surely*

$$\limsup_n \frac{\mathrm{Diam}_n}{\log(n)} \leq 4e\ .$$

To prove this lemma we couple the random friend tree with the URRT process. This coupling is the subject of the following lemma.

Lemma 6.17. *The random friend tree $(T_n, n \geq 2)$ and the uniform random recursive tree $(T'_n, n \geq 2)$ can be coupled in such a way that for any $i, j \in [n]$,*

$$d_n(i,j) \leq 2d'_n(i,j)\ ,$$

where $d_n(i,j)$ is the graph distance between vertex i and j in T_n, and $d'_n(i,j)$ is the graph distance between vertex i and j in T'_n.

Proof. Note that $T_2 = T'_2$ so the statement holds for $n = 2$. Fix $m \geq 2$ and suppose that we coupled $(T_2, \ldots, T_m)$ and $(T'_2, \ldots, T'_m)$ such that for all $i, j \in [m]$, $d_m(i,j) \leq 2d'_m(i,j)$. Now, sample uniformly at random $V_m \in [m]$ and let W_m be a uniform neighbour of V_m in T_m. Let T_{m+1} be the tree obtained by including vertex $m+1$ and edge $\{W_m, m+1\}$ in T_m and let T'_{m+1} be the tree obtained by including vertex $m+1$ and edge $\{V_m, m+1\}$ in T'_m. Observe that, for $i, j \in [m]$, $d_{m+1}(i,j) = d_m(i,j)$ and $d'_{m+1}(i,j) = d'_m(i,j)$. Now, let $i \in [m]$ and compute

$$\begin{aligned} d_{m+1}(i, m+1) &\leq d_{m+1}(i, V_m) + d_{m+1}(V_m, m+1) = d_m(i, V_m) + 2 \\ &\leq 2d'_m(i, V_m) + 2 \leq 2d'_m(i, m), \end{aligned}$$

where we use the triangle inequality, the induction hypothesis and the fact that for all $i \in [m]$, $d'_m(i,m) \leq 1 + d'_m(i, V_m)$. ∎

Proof. [Proof of Lemma 6.16] Couple $(T_n, n \geq 1)$ to the uniform random recursive tree $(T'_n, n \geq 1)$ as in Lemma 6.17 and observe that

$$\mathrm{Diam}_n = \max_{i,j \leq n} d_n(i,j) \leq 2 \max_{i,j \leq n} d'_n(i,j) \leq 4 \max_{i \leq n} d'_n(1,i).$$

Moreover, by Corollary 1.3 of Addario-Berry and Ford [1],

$$\frac{\max_{i \leq n} d'_n(1,i)}{\log n} \to e \text{ almost surely,}$$

which concludes the proof of the lemma. ∎

6.6.3 Leaf-depth

[Proof of Theorems 6.7 and 6.8] Theorem 6.2 implies that, asymptotically, almost surely each vertex is at distance at most 1 from a hub. By Theorem 6.1, each hub has mostly leaf neighbours. This suggests that at large times, most vertices that are not leaves are close to a leaf (distance 1 or 2). In this section, we show that at large times, there are exceptional vertices that are much further away from the nearest leaf, namely at distance $\Theta(\log n/\log\log n)$. We recall that

$$M_n = \max_{i\le n} \min_{\ell: D_n(\ell)=1} d_n(i,\ell)$$

is the maximal distance of any vertex to the closest leaf at time n, which we refer to as the *leaf-depth* at time n. To prove Theorem 6.7, we first need the following lemma to transfer upper bounds on the leaf-depth in the uniform random recursive trees to upper bounds on the leaf-depth in random friend trees.

Lemma 6.18. *The coupling defined in Lemma 6.17 between the random friend tree $(T_n, n\ge 1)$ and the uniform random recursive tree $(T'_n, n\ge 1)$ also satisfies that for any leaf ℓ' in T'_n, in T_n, the vertex ℓ' is at distance at most 1 from a leaf.*

Proof. The statement clearly holds for $n\le 2$. Next, suppose that for some m the statement is satisfied in T_{m-1} and T'_{m-1}. Fix an $\ell'\le m$ so that ℓ' is a leaf in T'_m. We claim that ℓ' is at distance at most 1 from a leaf in T_m. First, if $\ell'=m$, then ℓ' is also a leaf in T_m and the claim follows. If $\ell'\le m-1$, then ℓ' is also a leaf in T'_{m-1}, so by the induction hypothesis, ℓ' is at distance at most 1 from a leaf ℓ in T_{m-1}. If ℓ is also a leaf in T_m, the claim follows. Otherwise, for ℓ to be a leaf in T_{m-1}, but not T_m it is necessary that $W_m=\ell$. If $W_m=\ell$ and $d_{m-1}(\ell,\ell')=1$, then ℓ' is the unique neighbour of ℓ in T_{m-1} (because ℓ is a leaf in T_{m-1}), so $V_m=\ell'$, contradicting that ℓ' is a leaf in T'_m. Thus, if $W_m=\ell$, then $d_{m-1}(\ell,\ell')=0$, meaning that $\ell=\ell'$. Therefore, in T_m, vertex m is a leaf that is at distance 1 from ℓ'.

■

The next lemma gives an upper bound on the leaf-depth in the uniform random recursive tree. Together with Lemma 6.18 we deduce an upper bound for the leaf-depth in random friend trees, given in Proposition 6.20.

Lemma 6.19. *Let M'_n be the leaf-depth in the uniform random recursive tree at time n. Then, for any $\varepsilon>0$,*

$$\mathbf{P}\left\{M'_n \ge (1+\varepsilon)\frac{\log n}{\log\log n}\right\} = o(1).$$

Proof. Let T'_n be the uniform random recursive tree at time n. For any vertex $v \in T'_n$, we define a canonical path $\mathcal{P}_n(v)$ in T'_n that ends in a leaf. If v is a leaf in T'_n, set $\mathcal{P}_n(v) = v$. Otherwise, let $w = w_n(v)$ be the neighbour of v in T'_n with the largest label and let $\mathcal{P}_n(v)$ be v concatenated with $\mathcal{P}_n(w)$. So, to obtain the path $\mathcal{P}_n(v)$, start from v and sequentially move to the largest labelled neighbour until reaching a leaf. Define $T'_n(v)$, as the connected component of v in T'_n if the edge between v and its parent was removed (or, equivalently, $T'_n(v)$ is the subtree of T'_n that consists of all vertices that are connected to v by a path on which v is the lowest labelled vertex). Let $|\mathcal{P}_n(v)|$ be the number of edges on the path. To prove the lemma, it is sufficient to show that $\max_{v\in[n]} |\mathcal{P}_n(v)| \le (1+\epsilon)\log n/\log\log n$.
First, we check that, for any $\ell \ge 1$,

$$\mathbf{P}\left\{|T'_n(w_n(v))| \ge \ell \;\middle|\; |T'_n(v)| = m\right\} = \begin{cases} \frac{1}{\ell} \text{ if } \ell < m \\ 0 \text{ otherwise,} \end{cases} \tag{6.6.2}$$

where $m \le n$. Observe that, conditionally on $|T'_n(v)| = m$, if the vertices in $T'_n(v)$ are assigned labels in $[m]$ that respect the order of the original labels, the resulting tree has the same law as T'_m. Therefore, (6.6.2) follows if, for any m, we have that for all $\ell \ge 1$,

$$\mathbf{P}\left\{|T'_m(w_m(1))| \ge \ell\right\} = \begin{cases} \frac{1}{\ell} \text{ if } \ell \le m \\ 0 \text{ otherwise,} \end{cases}$$

where we recall that $T'_m(w_m(1))$ is the subtree rooted at the youngest child of 1.

The proof is by induction on m. The statement clearly holds for $m = 1$. Now, suppose the statement holds for $m = k-1$. For $m = k$, the statement is obvious for $\ell = 1$ and $\ell > k$ since $1 \le |T'_k(w_k(1))| \le k$. Observe that, if $V_k = 1$ (i.e, if vertex k connects to vertex 1), then $T'_k(w_k(1))$ consists only of the vertex k in which case $|T'_k(w_k(1))| = 1$. If $V_k \in T'_{k-1}(w_{k-1}(1))$, then $T'_k(w_k(1))$ is composed of the vertices of $T'_{k-1}(w_{k-1}(1))$ and vertex k, so $|T'_k(w_k(1))| = |T'_{k-1}(w_{k-1}(1))|+1$. If $V_k \notin \{1\} \cup T'_{k-1}(w_{k-1}(1))$ then $T'_k(w_k(1)) = T'_{k-1}(w_{k-1}(1))$, giving $|T'_k(w_k(1))| = |T'_{k-1}(w_{k-1}(1))|$. Therefore, for $1 < \ell \le k$,

$$\begin{aligned}\left\{|T'_k(w_k(1))| \ge \ell\right\} = &\left\{\left\{|T'_{k-1}(w_{k-1}(1))| \ge \ell\right\} \cap \{V_k \ne 1\}\right\} \\ &\cup \left\{\left\{|T'_{k-1}(w_{k-1}(1))| = \ell - 1\right\} \cap \left\{V_k \in T'_{k-1}(w_{k-1}(1))\right\}\right\}.\end{aligned}$$

By the induction hypothesis, for $\ell < k$,

$$\mathbf{P}\left\{|T'_k(w_k(1))| \ge \ell\right\} = \frac{1}{\ell}\frac{k-1}{k} + \frac{1}{\ell(\ell-1)}\frac{\ell-1}{k} = \frac{1}{\ell},$$

and for $\ell = k$,

$$\mathbf{P}\left\{|T'_k(w_k(1))| \ge k\right\} = 0 + \frac{1}{k-1}\frac{k-1}{k} = \frac{1}{k}.$$

The equality (6.6.2) follows for all m. Now, define $w^{(1)} = w_n(v)$ the vertex at distance 1 from v on $\mathcal{P}_n(v)$ and $w^{(\ell)} = w_n(w^{(\ell-1)})$ the vertex at distance ℓ from v on $\mathcal{P}_n(v)$. Then,

$$\{|\mathcal{P}_n(v)| \geq k\} = \{|T'_n(w^{(1)})| \geq k\} \cap \{|T'_n(w^{(2)})| \geq k-1\} \cap \cdots \cap \{|T'_n(w^{(k)})| \geq 1\}.$$

Together with (6.6.2), this implies that

$$\mathbf{P}\{|\mathcal{P}_n(v)| \geq k\} \leq \frac{1}{k!} \leq \frac{e^k}{k^k}.$$

Then, fix $\varepsilon > 0$ and subsitute $(1+\varepsilon)\frac{\log n}{\log\log n}$ to k. The above equation directly implies that

$$\mathbf{P}\left\{|\mathcal{P}_n(v)| \geq (1+\varepsilon)\frac{\log n}{\log\log n}\right\} = o(n^{-1}).$$

Finally, a union bound implies Lemma 6.19. ∎

From Lemmas 6.18and 6.19 we obtain an upper bound on the leaf depth in random friend trees.

Proposition 6.20. *For any* $\varepsilon > 0$

$$\mathbf{P}\left\{M_n \geq (2+\varepsilon)\frac{\log n}{\log\log n}\right\} = o(1).$$

Proof. Couple the random friend tree $(T_n, n \geq 1)$ and the uniform random recursive tree $(T'_n, n \geq 1)$ as in Lemma 6.18. Then, fix $i \leq n$ such that for $N'_n(i)$, the degree of vertex i in T'_n, we see that

$$\begin{aligned}\min_{\ell:D_n(\ell)=1} d_n(i,\ell) &\leq \min_{\ell:D_n(\ell)=1} \min_{\ell':N'_n(\ell')=1} (d_n(i,\ell') + d_n(\ell',\ell)) \\ &\leq \min_{\ell':N'_n(\ell')=1} d_n(i,\ell') + 1 \\ &\leq \min_{\ell':N'_n(\ell')=1} 2d'_n(i,\ell') + 1,\end{aligned}$$

where the last two inequalities follow from the properties of the coupling. By taking the maximum over $i \in [n]$, we have $M_n \leq 2M'_n + 1$ and so Lemma 6.19 implies the proposition. ∎

To conclude the proof of Theorem 6.7 it remains to prove an asymptotic lower bound for M_n. In order to show that the leaf depth is at least of order $\log n/\log\log n$ we first present a proof of the corresponding result for the URRT. To the best of our knowledge,

this result does not appear elsewhere. Moreover, the proof is less technical but has the same structure as its counterpart for random friend trees, so it is a good way to introduce the ideas needed for our main proof. In doing so, we hope that the technicalities in the proof of Lemma 6.22 are easier to understand.

Proposition 6.21. *Let M'_n be the leaf-depth in the uniform random recursive tree at time n. Then, $M'_n = \Omega_p(\log n/\log\log n)$.*

Proof. Let T'_n be the uniform random recursive tree at time n. In a uniform random recursive tree, the number of vertices with degree at least 3 goes to infinity almost surely (see Janson [78]). We call such a vertex a branch point. Therefore, we can choose n sufficiently large such that there is at least one branch point. Let P'_n be the maximal distance of any leaf in T'_n to the nearest branch point, that is,

$$P'_n = \max_{\ell:N'_n(\ell)=1} \min_{j:N'_n(j)\geq 3} d'_n(\ell, j).$$

We remark that $M'_n \geq P'_n/2$ because a midpoint of a longest leaf-to-branchpoint path is at distance at least $P'_n/2$ from the nearest leaf. It is well known (see Janson [78]) that the proportion of leaves in an URRT tends to $1/2$, consequently $X'^{\geq 2}_m/m \to 1/2$ almost surely. Thus, for $\varepsilon > 0$ and n sufficiently large,

$$\mathbf{P}\left\{1 - \frac{X'^{\geq 2}_k}{k} \leq \frac{1}{3};\ \forall k \geq n\right\} > 1 - \varepsilon.$$

Now, condition on the number of leaves at time $n/2$ being at least $n/6$, that is $X'^{1}_{n/2} \geq n/6$, and let $\{v_1, \cdots, v_P\}$ be an arbitrary set of $P = n/6$ leaves of $T'_{n/2}$. We study the subtrees rooted at v_i and will show that at time n, with high probability, for at least one $i \in [P]$ the subtree of v_i contains a path from v_i ending in a leaf and solely consisting of $\Omega(\log n/\log\log n)$ vertices of degree two. To this end, we say that a path consisting of degree two vertices that ends in a leaf *grows* at time m if vertex $m+1$ attaches to the leaf at the end of the path. This increases the length of the path by 1. We say that a path consisting of degree two vertices that ends in a leaf *dies* at time m if vertex $m+1$ connects to a degree-two vertex on the path that is at distance at most K from the leaf. That is, we only keep track of paths up to distance K from the leaf. Note that at each time step only one path can grow or die. We will only track paths until their first death. Note that, at time m, for $i \in [P]$, conditionally on the path rooted at v_i has not died yet, the probability of the path growing is $1/m$ and the probability of the path dying is at most K/m. Observe that P'_n stochastically dominates the minimum of K and the length of the longest path at time n rooted at some v_i for $i \in [P]$ that has not died. We can then couple these path growth

processes to a balls-in-bins model as follows. Let P be the number of bins. The process is started at time $n/2$ with P empty bins. For each time $m \in [n/2, n]$, with probability $P \cdot K/m$ add a black ball to a uniform random bin, or with probability P/m add a white ball to a uniform random bin. By doing so, a black ball is added to a given bin with probability K/m and a white ball is added to a given bin with probability $1/m$. We further see that P'_n stochastically dominates the smallest number between K and the maximum number of white balls at time n in a bin with zero black balls. It therefore suffices to show that for some $c > 0$ and $K := c \log n / \log \log n$, for n large enough, with high probability, one of the P bins contains at least K white balls and no black balls. Observe that, for n sufficiently large, between times $n/2$ and n, with probability at least $1 - \epsilon$, at most $B = 2PK$ black balls and at least $W = P/8$ white balls are added. Conditioned on this event, the probability that a specific bin contains at least K white balls is at least

$$\binom{W}{K} P^{-K} (1 - P^{-1})^{W-K} > e^{-1} \left(\frac{W}{PK}\right)^K = e^{-1} \left(\frac{1}{8K}\right)^K,$$

for large n, where we bound $(1 - P^{-1})^{W-K} \geq (1 - P^{-1})^W = ((1 - 6/n)^{n/6})^{1/8} > e^{-1}$. Therefore the expected number of bins containing at least K white balls can be bounded from below by

$$P \cdot e^{-1} (8K)^{-K} = e^{-1} \frac{n}{6} \left(\frac{\log \log n}{8c \log n}\right)^{c \log n / \log \log n} > n^{1-c}$$

for large n.

Then, if $c < 1/2$, since the numbers of white balls in two distinct bins have negative covariance, Chebyshev's inequality bounds the probability that the number of paths with at least K growth events is less than n^{1-2c} tends to 0, so in particular is smaller than ϵ for n sufficiently large. Any of these has 0 black balls with probability at least $(1 - P^{-1})^B \geq e^{-3K} = \omega(n^{1-2c})$, so another straightforward application of the second moment method implies that for n large enough, for any $c < 1/2$, with probability at least $1 - 3\epsilon$ at time n, there is a bin containing $c \log(n)/\log\log(n)$ white balls and no black balls. Recalling the stochastic dominance, that is valid on an event of probability at least $1 - \epsilon$, with probability at least $1 - 4\epsilon$ there is a leaf at distance at least $c \log n / \log \log n$ from the nearest branch point. ■

Using a similar proof, we can prove an asymptotic lower bound for the leaf-depth in random friend trees.

Proposition 6.22. *Let M_n be the leaf-depth in the random friend tree at time n. Then, $M_n = \Omega(\log n / \log \log n)$ in probability.*

Proof.

By Theorem 6.11, $X_n^{\geq 3} \to \infty$ almost surely, so we may pick n large enough such that $X_n^{\geq 3} \geq 1$. Then, define

$$P_n = \max_{\ell : D_n(\ell)=1} \min_{j : D_n(j) \geq 3} d_n(\ell, j)$$

to be the largest distance from a leaf to the nearest branch point in T_n. We see that $M_n \geq P_n/2$ because the midpoint of the longest leaf-to-branchpoint path is at distance at least $P_n/2$ from the nearest leaf. Therefore, it suffices to show that $P_n = \Omega(\log n / \log\log n)$ in probability.

As in Proposition 6.21 the proof is divided into two steps. We first show that, for n large enough, with high probability at time $n/2$ there are at least $n^{\delta/2}$ leaves attached to a degree two vertex. Remark that this step requires slightly more work than the URRT case. For the URRT, one only needs to ensure the presence of many leaves at time $n/2$ to guarantee that that a long path of subsequent degree-two vertex will emerge from one of these leaves. For the random friend tree, one needs to ensure that, at time $n/2$, there exists many leaves attached to a degree two vertex, to guarantee that new vertices attach to these leaves with some sufficient probability. This difference is due to the attachment rule in random friend trees. Denote by $\mathcal{L}_n$ the set of leaves attached to a vertex of degree two at time n and let $H_n = |\mathcal{L}_n|$ be the number of leaves attached to a degree two vertex in T_n. In the second part of the proof, we show that it is likely that at time n, a path of subsequent degree two vertices longer than $c \log n / \log\log n$ grows from at least one of the leaves of $\mathcal{L}_{n/2}$, that is,

$$\max_{\ell \in \mathcal{L}_n} \min_{j : D_n(j) \geq 3} d_n(\ell, j) > c \frac{\log n}{\log\log n} ,$$

with high probability. Lemma 6.29 states that

$$\mathbb{E}\left\{ \Delta X_n^{\geq 2} \;\middle|\; T_n \right\} \geq \frac{1}{3n} X_n^{\geq 3} ,$$

and by Theorem 6.11, $n^{-\delta} X_n^{\geq 3} \to \infty$ almost surely, for $0.1 < \delta < 0.9$. Therefore, for fixed $\varepsilon > 0$ and sufficiently large n,

$$\mathbf{P}\left\{ \frac{1}{3m} X_m^{\geq 3} > m^{-(1-\delta)} \ \forall m \geq n/4 \right\} \geq 1 - \epsilon .$$

It follows that, for all $n/4 \leq m \leq n/2$,

$$\mathbf{P}\left\{ \Delta X_m^{\geq 2} = 1 \;\middle|\; \frac{1}{3m} X_m^{\geq 3} > m^{-(1-\delta)} \right\} \geq n^{-(1-\delta)}$$

because $m^{-(1-\delta)} \geq n^{-(1-\delta)}$ if $m \leq n/2$.

We will show that H_n grows polynomially in probability. Note that for every vertex $v \in \mathcal{L}_m$, $\mathbf{P}\{W_m = v\} = 1/(2m)$ and $\Delta H_n = 1$ if and only if W_m is a leaf in T_m that is not in $\mathcal{L}_m$. Further, $\Delta X_m^{\geq 2} = 1$ if and only if W_m is a leaf in T_m. By summing over all possible values of W_m we therefore get the lower bound

$$\begin{aligned}&\mathbf{P}\left\{\Delta H_m = 1 \mid H_m = k,\ \frac{1}{3m} X_m^{\geq 3} > m^{\delta-1}\right\} \\ &\geq \mathbf{P}\left\{\Delta X_m^{\geq 2} = 1 \;\middle|\; \frac{1}{3m} X_m^{\geq 3} > m^{-(1-\delta)}\right\} - \frac{k}{2m} \geq n^{\delta-1} - \frac{2k}{n}\,,\end{aligned}$$

for $n/4 \leq m \leq n/2$. Therefore, either $H_m \geq n^\delta/16$ or $\mathbf{P}\left\{\Delta H_m = 1 \mid \frac{1}{3m} X_m^{\geq 3} > m^{-(1-\delta)}\right\} \geq \frac{7}{8} n^{-(1-\delta)}$ for $n/4 \leq m \leq n/2$.

Finally, $\Delta H_m = -1$ if W_m is a degree two vertex attached to a leaf, and there are exactly H_m such vertices in T_m. For v a given vertex of degree two attached to a leaf, $\mathbf{P}\{W_m = v\} \leq 2/m \leq 8/n$ for $n/4 \leq m \leq n/2$. By a union bound, for $n/4 \leq m \leq n/2$,

$$\mathbf{P}\{\Delta H_m = -1 \mid H_m = k\} \leq 8k/n\,.$$

By the two arguments above, for $n/4 \leq m \leq n/2$,

$$\mathbf{P}\left\{\Delta H_m = -1 \mid H_m \leq \frac{1}{16} n^\delta\right\} \leq \frac{1}{2} n^{-(1-\delta)}\,,$$

and

$$\mathbf{P}\left\{\Delta H_m = 1 \mid H_m \leq \frac{1}{16} n^\delta,\ \frac{1}{3m} X_m^{\geq 3} > m^{\delta-1}\right\} \geq \frac{7}{8} n^{-(1-\delta)}\,,$$

Then, it follows that, on the event $\left\{\forall m \in [n/4, n/2],\ \frac{1}{3m} X_m^{\geq 3} > m^{\delta-1}\right\}$, with probability at least $1 - \varepsilon$, for n sufficiently large, H_m grows to at least $\frac{1}{17} n^\delta$ between times $n/4$ and $n/2$. Finally, for n large enough, $n^\delta/17 > n^{\delta/2}$, so

$$\mathbf{P}\left\{H_{n/2} \geq n^{\delta/2}\right\} \geq 1 - 2\epsilon\,,$$

which concludes the first part of the proof.

We now condition on the event $\{H_{n/2} \geq n^{\delta/2}\}$. We show that, on this event, it is likely that at time n at least one of the leaves of $\mathcal{L}_{n/2}$ is at distance at least $\log n/\log\log n$ from the nearest branch point. The second part of the proof is identical from the URRT proof, but we present it again for clarity.

Let $\{v_1, \ldots, v_P\}$ be an arbitrary set of $P = n^{\delta/2}$ leaves in $\mathcal{L}_{n/2}$. We study the subtrees rooted at v_i and will show that at time n, with high probability, for at least one $i \in [P]$ the subtree of v_i contains a path ending in a leaf and solely consisting of $\Omega(\log n/\log\log n)$

vertices of degree two. To this end, we say that a path consisting of degree two vertices that ends in a leaf *grows* at time m if vertex $m+1$ attaches to the leaf at the end of the path. This increases the length of the path by 1. We say that a path consisting of degree two vertices that ends in a leaf *dies* at time m if vertex $m+1$ connects to a degree two vertex on the path that is at distance at most K from the leaf. That is, we only keep track of paths of length at most K. Note that, at each time step, only one path can grow or die. We will only track paths until their first death. Also note that, at time $m \in [n/2, n]$, for $i \in [P]$, conditionally on the path rooted at v_i has not died yet, the probability of the path growing is $1/(2m)$ and the probability of the path dying is at most K/m. Observe that P_n stochastically dominates the minimum of K and the longest path at time n rooted at some v_i for $i \in [P]$ that has not died. We can then couple these path growth processes to a balls-in-bins model. Let P be the number of bins and start the process with P empty bins. For each time $m \in [n/2, n]$, with probability $P \cdot K/m$ add a black ball to a uniform random bin, or with probability $P/(2m)$ add a white ball to a uniform random bin. By doing so, a black ball is added to a given bin with probability K/m and a white ball is added to a given bin with probability $1/(2m)$. We further see that P_n stochastically dominates the smallest number between K and the maximum number of white balls at time n in a bin with zero black balls. It therefore suffices to show that, for some $c > 0$ and $K := c \log n / \log\log n$, with high probability at least one of the P bins contains at least K white balls and no black balls. Observe that, for n sufficiently large, between time $n/2$ and n, with probability at least $1-\epsilon$, at most $B = 2PK$ black balls and at least $W = P/8$ white balls are added. Conditioned on this event, the probability that a specific bin has at least K white balls is at least

$$\binom{W}{K} P^{-K}(1-P^{-1})^{W-K} > e^{-1}\left(\frac{W}{PK}\right)^K = e^{-1}\left(\frac{1}{8K}\right)^K,$$

for large n, where we bound $(1-P^{-1})^{W-K} \geq (1-P^{-1})^W = ((1-n^{-\delta/2})^{n^{-\delta/2}})^{1/8} > e^{-1}$. The expected number of bins containing at least K white balls is bounded from below by

$$P \cdot e^{-1}(8K)^{-K} = e^{-1} n^{\delta/2} \left(\frac{\log\log n}{c \log n}\right)^{c \log n / \log\log n} > n^{\delta/2 - c},$$

for large n.

Then, if $c < \delta/8$, since the numbers of white balls in two distinct bins have negative covariance, Chebyshev's associaton inequality gives that the probability that the number of paths with at least K growth events is less than $n^{\delta/2-2c}$ tends to 0, so in particular, is smaller than ϵ for n sufficiently large. Any of these contains no black balls with probability

at least $(1-P^{-1})^B \geq e^{-3K} = \omega(n^{\delta/2-2c})$, so another straightforward application of the second moment method implies that for n large enough, for any $c < \delta/8$, with probability at least $1-3\epsilon$ there is a bin with $c\log(n)/\log\log(n)$ white balls and no black balls at time n. Recalling the statistical dominance, that is valid on an event of probability at least $1-2\epsilon$, with probability at least $1-5\epsilon$, there is a leaf at distance at least $c\log n/\log\log n$ from the nearest branch point. ■

6.6.4 High-degree vertices

[Proof of Theorems 6.9 and 6.10]

Recall that $Z_v = \liminf_{n\to\infty} \frac{D_n(v)}{n}$, and that by Theorem 6.1, in fact

$$Z_v = \lim_{n\to\infty} \frac{D_n(v)}{n} \text{ almost surely.}$$

The following lemma, combined with Theorem 6.11, gives a lower bound on the number of hubs and proves Theorem 6.9.

Lemma 6.23. *Almost surely,*

$$\#\{v \in [n] : Z_v > 0\} > \tfrac{1}{2} X_n^{\geq 2}.$$

We prove this lemma using Lemma 6.24, below, but we need some additional definitions for its statement. For a graph $G = (V, E)$, we say $V' \subset V$ is an *edge cover* of G if for each $e \in E$, there is a $v \in V'$ such that $v \in e$. Define the *minimal edge cover number* of a graph $G = (V, E)$, denoted by $EC(G)$, as follows

$$EC(G) := \min\{|V'| : V' \text{ is an edge cover of } G\}. \tag{6.6.3}$$

Lemma 6.23 is a direct consequence of Theorem 6.2, which states that each edge contains a hub, and the following lemma.

Lemma 6.24. *For any tree t, we have that*

$$|EC(t)| \geq \tfrac{1}{2} X^{\geq 2}(t),$$

where $X^{\geq 2}(t)$ denotes the number of non-leaves in the tree t.

Proof. We can assume that t is a rooted tree by declaring an arbitrary vertex in t the root. Decompose the tree t into the following vertex-disjoint paths. Let $\ell_1, \ldots, \ell_{X^1(t)}$ be

the leaves of t. For a leaf ℓ of t, let $P(\ell)$ be the path from ℓ to the root of t. For each $i \in [X^1(t)]$, let $P'(\ell_i) = P(\ell_i) \setminus \bigcup_{j=0}^{i-1} P'(\ell_j)$, that is, $P'(\ell_i)$ is the path $P(\ell_i)$ stripped of the vertices in $\bigcup_{j=0}^{i-1} P'(\ell_j)$. This decomposition gives us $X^1(t)$ disjoint paths $P'_{\ell_1}, \ldots, P'_{\ell_{X^1(t)}}$. Note that if $V' \subset V$ is an edge cover of t it must also be an edge cover of $P'(\ell_1) \cup \ldots \cup P'(\ell_{X^1(t)})$ (indeed, while removing edges, the requirement for a collection of edges to be an edge cover is weakened). An edge cover of a disconnected graph is a disjoint union of edge covers of the components, and an edge cover of a path of m vertices contains at least $\lfloor m/2 \rfloor$ vertices, so

$$|EC(t)| \geq \sum_{i=1}^{X^1(t)} \left\lfloor \frac{|P'_i|}{2} \right\rfloor \geq \sum_{i=1}^{X^1(t)} \left(\frac{|P'_i| - 1}{2} \right) = \frac{|V(t)| - X^1(t)}{2} = \frac{X^{\geq 2}(t)}{2}.$$

■

We now prove Theorem 6.10, which in particular implies that for any k, the number of vertices with degree at least k goes to infinity almost surely.

Proof. [Proof of Theorem 6.10] Fix a constant $M \in \mathbb{N}$ and let $(m_n, n \geq 1)$ be a sequence satisfying $m_n = o(n)$. We will prove that $\liminf_{n\to\infty} X_n^{\geq m_n} \geq M$ almost surely. By Lemma 6.15, there exists an almost surely finite time τ such that the diameter of the tree at time τ exceeds $2M$. Fix an arbitrary path of $2M+1$ vertices in T_τ and, for $n \geq \tau$, let $N_n^{(1)} \geq \cdots \geq N_n^{(2M+1)}$ be the degrees of the vertices on this path in decreasing order. Then, by Theorem 6.2, almost surely, at least M vertices on this path are hubs, so

$$\liminf_{n\to\infty} \frac{N_n^{(M)}}{n} > 0$$

and in particular, there is a finite time τ' such that $N_n^{(M)} > m_n$ for all $n \geq \tau'$. This implies that from time τ' onwards, there are at least M vertices with degree at least m_n, so

$$\mathbf{P}\left\{ \liminf_{n\to\infty} X_n^{\geq m_n} \geq M \right\} = 1.$$

■

6.6.5 Low-degree vertices

[Proof of Theorems 6.11 and 6.12] In this subsection we prove polynomial upper and lower bounds of $X_n^{\geq k}$, for k bounded and show that $X_n^{\geq k} = \Theta(X_n^{\geq 2})$ almost surely. We begin by

stating a general result on adapted processes, which will be of use in the proof of Theorem 6.11. Its proof can be found in the appendix. For each $k \in \mathbb{N}$, recall that X_n^k is the number of vertices of degree k in T_n and $X_n^{\geq k}$ is the number of vertices of degree at least k in T_n.

Proposition 6.25. *Let $(X_k)_{k\geq 0}$ and $(Y_k)_{k\geq 0}$ be integer-valued non-decreasing processes adapted to some filtration $(\mathcal{F}_k)_{k\geq 0}$ such that $0 \leq \Delta X_k + \Delta Y_k \leq 1$ for all k. Suppose there exists $\alpha > 0$ such that $\mathbb{E}\{\Delta X_k \mid \mathcal{F}_k\} \geq \alpha \mathbb{E}\{\Delta Y_k \mid \mathcal{F}_k\}$ on the event $\{X_k < \alpha Y_k\}$, except at finitely many times almost surely. Then for any $\beta \in (0, \alpha)$ there exists $C > 0$ such that*

$$\mathbf{P}\{X_n < \beta Y_n - C \log n \text{ infinitely often}\} = 0 ,$$

and for all n, $\mathbf{E}[X_n] \geq \beta \mathbf{E}[Y_n] - C \log n$.

Lemma 6.26. *For $n \geq 4$,*

$$\mathbb{E}\left\{\Delta X_n^{\geq 2} \;\middle|\; T_n\right\} \leq \frac{1}{2n} X_n^{\geq 2} + \frac{1}{2n} X_n^{\geq 3} , \tag{6.6.4}$$

and for $n \geq 5$,

$$\mathbb{E}\left\{\Delta X_n^{\geq 3} \;\middle|\; T_n\right\} \leq \frac{4}{3n} X_n^2 . \tag{6.6.5}$$

Proof. To prove (6.6.4), note that $\Delta X_n^{\geq 2} > 0$ precisely if W_n is a leaf of T_n, so

$$\mathbb{E}\left\{\Delta X_n^{\geq 2} \;\middle|\; T_n\right\} = \frac{1}{n} \sum_{v \in T_n} \frac{L_n(v)}{D_n(v)}.$$

When $n \geq 3$, leaves do not have leaves as neighbours, and when $n \geq 4$, any vertex v of degree two in T_n has at most one leaf neighbour, thus if v has degree two, $L_n(v)/D_n(v) = L_n(v)/2 \leq 1/2$. Together with the previous equality, this implies that

$$\mathbb{E}\left\{\Delta X_n^{\geq 2} \;\middle|\; T_n\right\} \leq \frac{1}{n}\left(\sum_{\{v : D_n(v) = 2\}} \frac{1}{2} + \sum_{\{v : D_n(v) \geq 3\}} \frac{L_n(v)}{D_n(v)} \right) \leq \frac{X_n^2}{2n} + \frac{X_n^{\geq 3}}{n} .$$

Since $X_n^2 = X_n^{\geq 2} - X_n^{\geq 3}$, this implies (6.6.4).

To prove (6.6.5), for $1 \leq i \leq j$ let

$$S_{ij} = \{v \in T_n : D_n(v) = 2, \text{ the neighbours of } v \text{ in } T_n \text{ have degrees } i \text{ and } j\}.$$

Then $X_n^2 = \sum_{1\le i\le j} |S_{i,j}|$ and

$$\mathbb{E}\Big\{\Delta X_n^{\ge 3} \;\Big|\; T_n\Big\} = \sum_{1\le i\le j} \mathbf{P}\Big\{W_n \in S_{ij} \;\Big|\; T_n\Big\} = \frac{1}{n}\sum_{1\le i\le j} |S_{ij}|\left(\frac{1}{i}+\frac{1}{j}\right).$$

We bound this sum by splitting it into three sums. First, for terms with $i = 1$ and $j \ge 3$ we have $1/i + 1/j \le 4/3$. For $v \in \bigcup_{2\le i\le j} S_{ij}$ we have $1/i + 1/j \le 1$, while for $v \in S_{12}$ we have $1/i + 1/j = 3/2$. We claim that at most half of the vertices in $S_{12} \cup \bigcup_{2\le i\le j} S_{ij}$ can be in S_{12}. Indeed, provided that $n \ge 5$, if $v \in S_{12}$ then for u its unique neighbour with degree 2 it holds that $u \in \bigcup_{2\le i\le j} S_{ij}$. Moreover, $n \ge 5$ implies that u has at most one neighbour in S_{12}. Therefore, $|S_{12}| \le |\bigcup_{2\le i\le j} S_{ij}|$. This implies that

$$\sum_{1\le i\le j} |S_{ij}|\left(\frac{1}{i}+\frac{1}{j}\right) \le \frac{4}{3}\sum_{j\ge 3}|S_{1j}| + \frac{1}{2}\left(1+\frac{3}{2}\right)\left(|S_{12}| + \sum_{2\le i\le j}|S_{ij}|\right) \le \frac{4}{3}X_n^2,$$

therefore $\mathbb{E}\Big\{\Delta X_n^{\ge 3} \;\Big|\; T_n\Big\} \le \frac{4}{3n}X_n^2$ as claimed. ∎

Before stating the next lemma we introduce the notation $X_n^{k,\le k}$, the number of vertices of degree k having at most one neighbour of degree at least $k+1$, and $X_n^{k,>k}$, the number of vertices of degree k with at least two neighbours of degree at least $k+1$.

Lemma 6.27. *For any positive integer k, $n \ge 3$,*

$$X_n^k = X_n^{k,>k} + X_n^{k,\le k}, \tag{6.6.6}$$

$$X_n^{\ge k+1} \ge X_n^{k,>k}, \tag{6.6.7}$$

$$\mathbb{E}\Big\{\Delta X_n^{\ge k+1} \;\Big|\; T_n\Big\} \ge \frac{k-1}{kn} X_n^{k,\le k}. \tag{6.6.8}$$

Proof. The equality (6.6.6) follows directly from the definition of $X_n^{k,>k}$ and $X_n^{k,\le k}$. To prove the second statement, remark that any vertex v contributing to $X_n^{k,>k}$ has at least two neighbours of degree at least $k+1$, and at least one is a child of v. Since every vertex is the child of at most 1 vertex, (6.6.7) follows.

Finally, to prove (6.6.8), one must understand how vertices of degree $k+1$ are created. In order to have $\Delta X_n^{\ge k+1} = 1$, it is sufficient (but not necessary) that vertex $n+1$ attaches to a vertex counted by $X_n^{k,\le k}$, or in other words, that $W_n = w$ for some vertex w with degree k which has at least $k-1$ neighbours of degree at most k. For each such vertex w, this happens with probability at least $\frac{k-1}{kn}$. This proves (6.6.7). ∎

Lemma 6.28. *For any integer k, whenever $n \geq k+2$.*

$$\mathbb{E}\left\{\Delta X_n^{\geq k+1} \;\middle|\; T_n\right\} \leq \frac{k-1/2}{n} X_n^k \,. \tag{6.6.9}$$

Proof. A vertex of degree $k+1$ is created at time n if W_n has degree k in T_n. For a vertex $w \in T_n$ of degree k, the probability of $W_n = w$ is maximized if the neighbours of w have lowest possible degree; that is, if w has $k-1$ leaf neighbours and one neighbour of degree two. In this case, the probability that $W_n = w$ equals $\frac{k-1/2}{n}$; the lemma follows. ∎

We use the following two lemmas to show that the number of non-leaves grows at least polynomially.

Lemma 6.29. *For $n \geq 3$,*

$$\mathbb{E}\left\{\Delta X_n^{\geq 2} \;\middle|\; T_n\right\} \geq \frac{1}{3n} X_n^{\geq 3}.$$

Proof. Note that $\Delta X_n^{\geq 2} = 1$ if and only if W_n is a leaf. Therefore, as observed before, if $n \geq 3$ then

$$\mathbb{E}\left\{\Delta X_n^{\geq 2} \;\middle|\; T_n\right\} = \frac{1}{n} \sum_{\{v : D_n(v) \geq 2\}} \frac{L_n(v)}{D_n(v)}. \tag{6.6.10}$$

Let T_n' be equal to T_n with all of its leaves removed. We consider different sets of vertices in T_n' and we study their contribution to the sum above.

Let V_1 be the set of vertices of T_n that are leaves in T_n' and that were vertices of degree 2 in T_n, so that for each $v \in V_1$, $\frac{L_n(v)}{D_n(v)} = \frac{1}{2}$. Let V_2 be the vertices of T_n that are leaves in T_n' that were vertices of degree at least 3 in T_n so that for each $v \in V_2$, $\frac{L_n(v)}{D_n(v)} = \frac{D_n(v)-1}{D_n(v)} \geq \frac{2}{3}$. Let V_3 be the vertices that have degree 2 in T_n' and had degree at least 3 in T_n, so that for each $v \in V_3$, $\frac{L_n(v)}{D_n(v)} = \frac{D_n(v)-2}{D_n(v)} \geq \frac{1}{3}$. Finally, let V_4 be the vertices that have degree at least 3 in T_n'. Therefore,

$$\sum_{\{v : D_n(v) \geq 2\}} \frac{L_n(v)}{D_n(v)} \geq \frac{1}{2}|V_1| + \frac{2}{3}|V_2| + \frac{1}{3}|V_3| \,. \tag{6.6.11}$$

To lower bound this sum note that $|V_1| + |V_2|$ is the number of leaves in T_n'. Since T_n' is a tree, the number of leaves in T_n' is given by

$$\sum_{\{v : |\mathcal{N}(v, T_n')| \geq 3\}} (|\mathcal{N}(v, T_n')| - 2) + 2$$

and so

$$|V_1| + |V_2| = \sum_{v \in V_4} (|\mathcal{N}(v, T_n')| - 2) + 2 \,.$$

Finally, since $\sum_{v \in V_4} (|\mathcal{N}(v, T_n')| - 2) + 2 \geq |V_4|$ we obtain that

$$|V_1| + |V_2| \geq |V_4| \,,$$

hence

$$\frac{1}{2}|V_1| + \frac{2}{3}|V_2| \geq \frac{1}{3}|V_2| + \frac{1}{3}|V_4| \,.$$

It also holds that $X_n^{\geq 3} = |V_2| + |V_3| + |V_4|$, so we conclude that

$$\frac{1}{2}|V_1| + \frac{2}{3}|V_2| + \frac{1}{3}|V_3| \geq \frac{1}{3} X_n^{\geq 3} \,.$$

Combined with (6.6.10) and (6.6.11), this completes the proof. ∎

Lemma 6.30. *Let $\alpha = (\sqrt{13} - 3)/2 \approx 0.303$ be the unique positive solution of $x = \frac{1-2x}{1+x}$. Then, for any $\beta \in (0, \alpha)$ there exists $c > 0$ such that*

$$\mathbf{P}\left\{X_n^{\geq 3} < \beta X_n^{\geq 2} - c \log n \text{ infinitely often}\right\} = 0,$$

and for all n, $\mathbf{E}\left[X_n^{\geq 3}\right] \geq \beta \mathbf{E}\left[X_n^{\geq 2}\right] - c \log n$.

Proof. The statements follow directly by applying Proposition 6.25, once we show that for α as in the Lemma statement, for $n \geq 4$, either $X_n^{\geq 3} \geq \alpha X_n^{\geq 2}$ or $\mathbb{E}\left\{\Delta X_n^{\geq 3} \;\middle|\; T_n\right\} \geq \alpha \mathbb{E}\left\{\Delta X_n^{\geq 2} \;\middle|\; T_n\right\}$. Suppose that $X_n^{\geq 3} < \alpha X_n^{\geq 2}$. Then, by (6.6.4),

$$\mathbb{E}\left\{\Delta X_n^{\geq 2} \;\middle|\; T_n\right\} \leq \frac{1}{2n} X_n^{\geq 2} + \frac{1}{2n} X_n^{\geq 3} \leq \frac{1+\alpha}{2n} X_n^{\geq 2}. \tag{6.6.12}$$

Moreover, observe that the case $k = 2$ of Lemma 6.27 gives that

$$X_n^2 = X_n^{2,>2} + X_n^{2,\leq 2}, \tag{6.6.13}$$

$$X_n^{\geq 3} \geq X_n^{2,>2}, \text{ and} \tag{6.6.14}$$

$$\mathbb{E}\left\{\Delta X_n^{\geq 3} \;\middle|\; T_n\right\} \geq \frac{1}{2n} X_n^{2,\leq 2}. \tag{6.6.15}$$

Since $X_n^{\geq 3} < \alpha X_n^{\geq 2}$, (6.6.14) implies that $X_n^{2,>2} < \alpha X_n^{\geq 2}$. Note that $X_n^{\geq 2} = X_n^{2,>2} + X_n^{2,\leq 2} + X_n^{\geq 3}$, so the bounds $X_n^{2,>2} < \alpha X_n^{\geq 2}$ and $X_n^{\geq 3} < \alpha X_n^{\geq 2}$ together imply that $X_n^{2,\leq 2} > (1-2\alpha)X_n^{\geq 2}$. Combining this bound with (6.6.15) and (6.6.12), we conclude that

$$\mathbb{E}\left\{\Delta X_n^{\geq 3} \;\middle|\; T_n\right\} \geq \frac{1-2\alpha}{2n} X_n^{\geq 2} \geq \frac{1-2\alpha}{1+\alpha}\mathbb{E}\left\{\Delta X_n^{\geq 2} \;\middle|\; T_n\right\} = \alpha \mathbb{E}\left\{\Delta X_n^{\geq 2} \;\middle|\; T_n\right\},$$

as required.

■

Lemma 6.31. *As $n \to \infty$, $X_n^{\geq 3} \to \infty$ almost surely.*

Proof. By Theorem 6.6, the diameter of T_n is $\Theta(\log n)$ almost surely. Thus,

$$X_n^{\geq 2} \to \infty \text{ almost surely.} \tag{6.6.16}$$

Suppose for a contradiction that $X_n^{\geq 3}$ does not go to infinity almost surely. That is, there exists a positive constant c such that

$$\mathbf{P}\left\{\forall n \in \mathbb{N},\ X_n^{\geq 3} < \infty\right\} = c > 0.$$

By continuity of probability, this implies that there exists some constant K such that

$$\mathbf{P}\left\{\forall n \in \mathbb{N},\ X_n^{\geq 3} \leq K\right\} \geq \frac{c}{2} > 0. \tag{6.6.17}$$

Using the fact that $X_n^{\geq 2}$ goes to infinity almost surely and $X_n^2 = X_n^{\geq 2} - X_n^{\geq 3}$ for all n, we have

$$\mathbf{P}\left\{X_n^2 \to \infty \;\middle|\; \forall n \in \mathbb{N},\ X_n^{\geq 3} \leq K\right\} = 1. \tag{6.6.18}$$

Define τ to be the smallest time after which, for all $n \geq \tau$, T_n always contains at least two neighbouring vertices each of degree two,

$$\tau := \inf\{m\colon \forall n \geq m, \exists\, u, v \in T_n,\ u \sim v,\ d_n(u) = d_n(v) = 2\}.$$

It follows from (6.6.18) that $\mathbf{P}\left\{\tau < \infty \mid \forall n \in \mathbb{N},\ X_n^{\geq 3} \leq K\right\} = 1$. Note that $\Delta X_n^{\geq 3} = 1$ if and only if vertex $n+1$ attaches to a vertex of degree two. At time $n \geq \tau$ this occurs with probability at least $1/n$. Thus, except for finitely many n,

$$\mathbb{E}\left\{\Delta X_n^{\geq 3} \;\middle|\; T_n, X_n^{\geq 3} \leq K\right\} \geq \frac{1}{n}\,;$$

this implies that $\mathbf{P}\left\{\lim_n X_n^{\geq 3} > K\right\} = 1$, which contradicts our hypothesis. ■

Lemma 6.32. *As $n \to \infty$, $X_n^{\geq 2}/\log n \to \infty$ almost surely.*

Proof. Fix $C > 0$. We will show that

$$\liminf_{n\to\infty} \frac{X_n^{\geq 2}}{\log n} \geq C \text{ almost surely,}$$

which implies the statement. Conditionally on T_n, $\Delta X_n^{\geq 2}$ is a Bernoulli random variable with parameter $\mathbb{E}\left\{\Delta X_n^{\geq 2} \;\middle|\; T_n\right\}$. We prove that $\mathbb{E}\left\{\Delta X_n^{\geq 2} \;\middle|\; T_n\right\} < C/n$ only finitely many times, in order to couple $(\Delta X_n^{\geq 2})_{n\geq 1}$ to a sequence of independent Bernoulli random variables with parameter C/n.

Let $(U_i, i \geq 1)$ be independent uniform random variables on $[0,1]$. Conditionally on T_n, construct T_{n+1} from T_n by setting $\Delta X_n^{\geq 2} = 1$ if and only if $U_n < \mathbb{E}\left\{\Delta X_n^{\geq 2} \;\middle|\; T_n\right\}$, and then sampling the additional randomness required to construct T_{n+1} conditionally on T_n and on the value of $\Delta X_n^{\geq 2}$. Define a coupling between $(\Delta X_n^{\geq 2})_{n\geq 1}$ and $(B_n)_{n\geq 1}$, a sequence of independent Bernoulli random variables with parameter C/n, by setting $B_n = 1$ if $U_n < C/n$ and $B_n = 0$ otherwise. It is immediate that $\Delta X_n^{\geq 2} \geq B_n$ whenever $\mathbb{E}\left\{\Delta X_n^{\geq 2} \;\middle|\; T_n\right\} \geq C/n$.

Lemma 6.29 states that $\mathbb{E}\left\{\Delta X_n^{\geq 2} \;\middle|\; T_n\right\} \geq \frac{1}{3n} X_n^{\geq 3}$, and by Lemma 6.31, $X_n^{\geq 3} \to \infty$ almost surely, thus

$$\mathbb{E}\left\{\Delta X_n^{\geq 2} \;\middle|\; T_n\right\} = \omega(1/n) \tag{6.6.19}$$

almost surely. Therefore, almost surely, $\mathbb{E}\left\{\Delta X_n^{\geq 2} \;\middle|\; T_n\right\} < C/n$ only finitely many times, and also $\Delta X_n^{\geq 2} < B_n$ only finitely many times. In particular

$$\liminf_{n\to\infty} \frac{X_n^{\geq 2}}{\log n} \geq \liminf_{n\to\infty} \frac{\sum_{i=1}^n B_i}{\log n}$$

almost surely. We claim that

$$\frac{\sum_{i=1}^n B_i}{\log n} \to C \text{ almost surely.} \tag{6.6.20}$$

Indeed, let $(Y_j)_{j\geq 1}$ be a sequence of independent random variables satisfying $Y_j = \sum_{i=\lfloor e^{j-1}\rfloor+1}^{\lfloor e^j \rfloor} B_i$. Then $\lim_{j\to\infty} \mathbf{E}[Y_j] = C$ and $\mathbf{E}\left[Y_j^2\right] \leq 10C^2$. By Kolmogorov's strong law of large numbers [59, Theorem 3.2.],

$$\frac{\sum_{j=1}^n Y_j}{n} \to C \text{ almost surely.}$$

This implies the convergence of (6.6.20) along the subsequence $(\lfloor e^j \rfloor)_{j\geq 1}$, and, by monotonicity, (6.6.20) follows.

■

Corollary 6.33. *For $\alpha = (\sqrt{13}-3)/2$ the unique positive solution of $x = \frac{1-2x}{1+x}$, for any $\beta \in (0,\alpha)$ and for all n sufficiently large, $\mathbf{E}\left[X_n^{\geq 3}\right] \geq \beta \mathbf{E}\left[X_n^{\geq 2}\right]$.*

By Lemma 6.30 we have $\mathbf{E}X_n^{\geq 3} \geq \beta \mathbf{E}X_n^{\geq 2} - c\log(n)$, and by Lemma 6.32 $X_n^{\geq 2} = \omega(\log n)$ almost surely; Corollary 6.33 follows.

Proposition 6.34. *For $\alpha = (\sqrt{13}-3)/2$ the unique positive solution of $x = \frac{1-2x}{1+x}$, for any $0 < \delta < \alpha/3 \approx 0.101$ we have $n^{-\delta}X_n^{\geq 2} \to \infty$ almost surely.*

Proof. By Lemmas 6.29 and 6.30, for all $\beta \in (0,\alpha)$ there exists a $c > 0$ such that

$$\mathbf{P}\left\{\mathbb{E}\left\{\Delta X_n^{\geq 2} \;\middle|\; T_n\right\} < \frac{\beta}{3n}(X_n^{\geq 2} - c\log n) \text{ infinitely often}\right\} = 0\,. \tag{6.6.21}$$

Fix δ such that $0 < \delta < \beta/3$ and fix γ rational such that $\delta < \gamma < \beta/3$. Lemma 6.32 states that $X_n^{\geq 2} = \omega(\log n)$ almost surely, therefore, almost surely there are only finitely many n such that

$$\frac{\beta}{3}(X_n^{\geq 2} - c\log n) < \gamma X_n^{\geq 2},$$

Define the time τ as

$$\tau = 3 \vee \sup\left\{n \geq 1 : \mathbb{E}\left\{\Delta X_n^{\geq 2} \;\middle|\; T_n\right\} < \frac{\gamma}{n} X_n^{\geq 2}\right\}.$$

As a consequence of (6.6.21), $\tau < \infty$ almost surely. For $n > \tau$, by the definition of τ, the probability that $X_n^{\geq 2}$ increases at time n is bounded from below by $\gamma X_n^{\geq 2}/n$. It is therefore natural to compare $X_n^{\geq 2}$ to a generalised Pólya urn. We first introduce an urn process, containing B_n black balls, and show that $n^{-\delta}B_n \to \infty$ almost surely. Conditionally on $\tau = t$, we then couple the sequences $(X_n^{\geq 2})_{n\geq t}$ and $(B_n)_{n\geq t}$ and conclude the proof of Proposition 6.34.

To introduce the urn process, let $M > 0$ be an integer such that $\gamma^{-1}M$ is a positive integer, and let $t > 0$ be another integer. Consider the urn process started at time t with M black balls and $t\gamma^{-1}M - M$ white balls. At every time step, draw a ball from the urn uniformly at random and return it to the urn together with $\gamma^{-1}M$ additional balls. If the drawn ball is white, all of the additional balls are white. If the drawn ball is black, M of the additional balls are black and the other $(\gamma^{-1}-1)M$ balls are white. Denote by B_n the number of black balls in the urn at time n. Then at time n, the number of black balls B_n increases by M with probability $B_n/(n\gamma^{-1}M)$. The described urn is *triangular* since if a white ball is drawn only white balls are added to the urn. The asymptotic behaviour of triangular

urn processes has been studied by Janson [79]. Theorem 1.3.(v) in [79] implies that $n^{-\gamma}B_n$ converges almost surely to some random variable Z, and Theorem 8.7. in [79] shows that Z puts no mass on 0, so $n^{-\delta}B_n \to \infty$ almost surely as $\delta < \gamma$.

We now introduce a coupling satisfying that $(X_n^{\geq 2})_{n\geq t}$ grows at least as fast as $(B_n)_{n\geq t}$, on the event $\tau < t$. To formalise this coupling, note that conditionally on $\tau = t$, if $n \geq t$, then $\mathbb{E}\left\{\Delta X_n^{\geq 2} \;\middle|\; T_n\right\} \geq \frac{\gamma}{n} X_n^{\geq 2}$. Since $\Delta X_n^{\geq 2} \in \{0,1\}$,

$$\mathbf{P}\left\{\Delta(MX_n^{\geq 2}) = M \;\middle|\; T_n\right\} \geq \frac{MX_n^{\geq 2}}{n\gamma^{-1}M}\,.$$

Remark that $MX_t^{\geq 2}$ is at least M. Let $(U_n)_{n\geq 1}$ be a sequence of independent uniform random variables on $[0,1]$. Conditionally on T_n, construct T_{n+1} from T_n by setting $\Delta X_n^{\geq 2} = 1$ if and only if $U_n \leq \mathbb{E}\left\{\Delta X_n^{\geq 2} \;\middle|\; T_n\right\}$ and sampling the remaining randomness in T_{n+1} conditional on T_n and the value of $\Delta X_n^{\geq 2}$. Let $B_t = 0$ and for $n \geq t$ let $\Delta B_n = M$ if $B_n/(n\gamma^{-1}M) \leq U_n$ and $\Delta B_n = 0$ otherwise. Then $(B_n)_{n\geq t}$ is distributed as the number of black balls in the Pólya urn described above. We already noted that $n^{-\delta}B_n \to \infty$ almost surely, and by our coupling we have that on the event $\tau < t$, $MX_n^{\geq 2} \geq B_n$ for all $n \geq t$. Thus, for $\varepsilon > 0$

$$\mathbf{P}\left\{n^{-\delta}X_n^{\geq 2} \to \infty\right\} > \mathbf{P}\left\{\tau < t\right\} \geq 1 - \varepsilon\,.$$

Since τ is almost surely finite, ε can be chosen arbitrarily small, by taking t large, which concludes the proof.

■

Lemma 6.35. *Let $\gamma = 3 - 2\sqrt{2} \approx 0.172$ be the unique positive solution to $x = \frac{1-5x}{1-x}$. Then for any $\beta \in (1-\gamma, 1)$ there exists $C > 0$ such that*

$$\mathbf{P}\left\{X_n^{\geq 3} > \beta X_n^{\geq 2} + C\log n \text{ infinitely often}\right\} = 0$$

and for all n, $\mathbf{E}\left[X_n^{\geq 3}\right] \leq \beta\mathbf{E}\left[X_n^{\geq 2}\right] + C\log n$.

Proof. We apply Proposition 6.25 to the sequences $X_k = X_n^{\geq 2}$ and $Y_k = X_n^{\geq 3}$. Note that $\Delta X_n^{\geq 2} + \Delta X_n^{\geq 3} \in \{0,1\}$. It remains to show that there exists $\gamma > 0$ such that for all n, either

$$X_n^{\geq 2} \geq \frac{1}{1-\gamma} X_n^{\geq 3} \text{ or}$$
$$\mathbb{E}\left\{\Delta X_n^{\geq 2} \;\middle|\; T_n\right\} \geq \frac{1}{1-\gamma}\mathbb{E}\left\{\Delta X_n^{\geq 3} \;\middle|\; T_n\right\}.$$

Lemma 6.35 follows then directly by applying Proposition 6.25. Suppose that $X_n^{\geq 2} < \frac{1}{1-\gamma} X_n^{\geq 3}$. Since $X_n^2 = X_n^{\geq 2} - X_n^{\geq 3}$, (6.6.5) gives that

$$\mathbb{E}\left\{\Delta X_n^{\geq 3} \;\middle|\; T_n\right\} \leq \frac{4}{3n}(X_n^{\geq 2} - X_n^{\geq 3}) < \frac{4\gamma}{3n} X_n^{\geq 2}.$$

From Lemma 6.29 we see that

$$\mathbb{E}\left\{\Delta X_n^{\geq 2} \;\middle|\; T_n\right\} \geq \tfrac{1}{3n} X_n^{\geq 3} > \tfrac{1-\gamma}{3n} X_n^{\geq 2}.$$

Therefore,

$$\mathbb{E}\left\{\Delta X_n^{\geq 2} \;\middle|\; T_n\right\} > \frac{1-\gamma}{4\gamma} \mathbb{E}\left\{\Delta X_n^{\geq 3} \;\middle|\; T_n\right\}.$$

The statement follows from the choice of γ. ∎

Proposition 6.36. *Let $\delta = 1 - \gamma/2$, where $\gamma = 3 - 2\sqrt{2}$. Then, for any $\delta < \lambda < 1$,*

$$n^{-\lambda} X_n^{\geq 2} \to 0,$$

almost surely.

Proof. By (6.6.4), $\mathbb{E}\left\{\Delta X_n^{\geq 2} \;\middle|\; T_n\right\} \leq \frac{1}{2n} X_n^{\geq 2} + \frac{1}{2n} X_n^{\geq 3}$. Combining this with Lemma 6.35 gives that for any $\beta \in (1-\gamma, 1)$, there exists a $C > 0$ such that

$$\mathbf{P}\left\{\mathbb{E}\left\{\Delta X_n^{\geq 2} \;\middle|\; T_n\right\} > \frac{1+\beta}{2n} X_n^{\geq 2} + C \log n \text{ infinitely often}\right\} = 0.$$

By mimicking the proof of Proposition 6.34, we can compare $X_n^{\geq 2}$ to a generalised Pólya urn and obtain an upper bound on $\Delta X_n^{\geq 2}$. Omitting the details of the coupling, we conclude that for $\lambda \in (\frac{1+\beta}{2}, 1)$ it holds that $\mathbf{P}\left\{n^{-\lambda} X_n^{\geq 2} \to 0\right\} = 1$. ∎

Proof. [Proof of Theorem 6.11 and 6.12] The second part of Theorem 6.11 follows from Proposition 6.36, by noting that $X_n^{\geq k} \leq X_n^{\geq 2}$ and therefore, for any $k \geq 2$,

$$n^{-\lambda} X_n^{\geq k} \to 0 \text{ a.s.}$$

The upper bound in Theorem 6.12 follows directly since $X_n^{\geq k+1} \leq X_n^{\geq k}$ for all k.

By Proposition 6.34, $\lim_n n^{-\delta} X_n^{\geq 2} = \infty$ almost surely. We prove the remaining cases in the first part of Theorem 6.11 and the lower bound of Theorem 6.12, by using induction to prove that, almost surely, for all $k \geq 2$,

(i) there exists a positive constant c_k such that $\liminf_n X_n^{\geq k+1} / X_n^{\geq k} > c_k$, and

(ii) $\lim_n n^{-\delta} X_n^{\geq k+1} = \infty$.

The fact that $\lim_n n^{-\delta} X_n^{\geq 2} = \infty$ almost surely and Lemma 6.30 imply that (i) holds for $k = 2$ for any $0 < c_2 < \alpha$. This then also implies (ii) for $k = 2$.

Now, fix $k \geq 3$ and suppose that the induction hypothesis holds for all $2 \leq \ell \leq k-1$. Let $b_k \in (0,1)$ be the solution to

$$\frac{b_k}{1-2b_k} = \frac{c_{k-1}(k-1)}{k(k-3/2)},$$

and fix $0 < c_k < b_k$. We claim that for all n, either

$$X_n^{\geq k+1} \geq b_k X_n^{\geq k} \text{ or}$$
$$\mathbb{E}\left\{\Delta X_n^{\geq k+1} \;\middle|\; T_n\right\} \geq b_k \mathbb{E}\left\{\Delta X_n^{\geq k} \;\middle|\; T_n\right\}$$

almost surely, except at finitely many times. If this holds, then applying Proposition 6.25 gives us that there exists $C > 0$ such that

$$\mathbf{P}\left\{X_n^{\geq k+1} < c_k X_n^{\geq k} - C \log n \text{ infinitely often}\right\} = 0.$$

By the induction hypothesis, $n^{-\delta} X_n^{\geq k} \to \infty$ and so

$$\mathbf{P}\left\{X_n^{\geq k+1} / X_n^{\geq k} < c_k \text{ infinitely often}\right\} = 0\,.$$

This implies that (i) holds at step k, which in turn implies part (ii).

It remains to prove the claim. Suppose that

$$X_n^{\geq k+1} < b_k X_n^{\geq k}\,.$$

From (6.6.7), $b_k X_n^{\geq k} > X_n^{k,>k}$, which combined with (6.6.6) implies that

$$X_n^k < b_k X_n^{\geq k} + X_n^{k,\leq k}\,.$$

Now, using (6.6.8) gives

$$\mathbb{E}\left\{\Delta X_n^{\geq k+1} \;\middle|\; T_n\right\} > \frac{k-1}{kn}\left(X_n^k - b_k X_n^{\geq k}\right) = \frac{k-1}{kn}\left((1-b_k) X_n^{\geq k} - X^{\geq k+1}\right)\,,$$

where the equality holds since $X_n^k = X_n^{\geq k} - X_n^{\geq k+1}$. Our assumption $X_n^{\geq k+1} < b_k X_n^{\geq k}$ then gives

$$\mathbb{E}\left\{\Delta X_n^{\geq k+1} \;\middle|\; T_n\right\} \geq \frac{k-1}{kn}(1-2b_k)X_n^{\geq k}\,.$$

The induction hypothesis for $k-1$ implies that almost surely $X_n^{\geq k} > c_{k-1}X_n^{\geq k-1} \geq c_{k-1}X_n^{k-1}$, except at finitely many times. Therefore,

$$\mathbb{E}\left\{\Delta X_n^{\geq k+1} \;\middle|\; T_n\right\} \geq \frac{c_{k-1}(k-1)}{kn}(1-2b_k)X_n^{k-1}$$

except at finitely many times. But (6.6.9) implies that for all n sufficiently large we have that $X_n^{k-1} \geq \frac{k-3/2}{n}\mathbb{E}\left\{\Delta X_n^{\geq k} \;\middle|\; T_n\right\}$, so we conclude that, except at finitely many times

$$\mathbb{E}\left\{\Delta X_n^{\geq k+1} \;\middle|\; T_n\right\} \geq \frac{c_{k-1}(k-1)}{k(k-3/2)}(1-2b_k)\mathbb{E}\left\{\Delta X_n^{\geq k} \;\middle|\; T_n\right\}\,.$$

This proves the claim by our choice of b_k, which concludes the proof. ∎

6.7 Open questions and future directions

We conclude with some open questions about the random friend tree.

1. In Theorem 6.11 we prove that, for some $0.1 < \delta < \lambda < 0.9$, almost surely $n^\delta \ll X_n^{\geq 2} \ll n^\lambda$. A question of interest would be whether the gap between the upper and lower bound can be closed and whether, for some $\mu > 0$ and some random variable X with non-trivial support, $n^{-\mu}X_n^{\geq 2} \to X$ almost surely. Simulations by Krapivsky and Redner [93] suggest that $X_n^{\geq 2}$ grows as n^μ, with $\mu \approx 0.566$.

2. We prove that, for fixed k, the number of vertices with degree at least $k+1$ is of the same order as the number of vertices with degree at least k, see Theorem 6.12. Can we prove, for fixed k, that the number of degree-k nodes is of the same order as the the number of degree-$(k+1)$ nodes? Does it hold that $\limsup_{n\to\infty} \frac{X_n^{\geq k}}{X_n^{\geq 2}}$ goes to 0 as k goes to infinity? Or, informally, are most of the non-leaves vertices of bounded degree? Krapivsky and Redner [93] conjecture that for each k, $\frac{X_n^k}{X_n^{\geq 2}}$ has an almost sure limit that is $\Theta(k^{-(1+\mu)})$.

3. Is $\mathbf{P}\{Z_u > 0\}$ decreasing in u? More generally, does Z_u stochastically dominate Z_v for $u < v$?

4. Does it hold that $\sum_{i\geq 1} Z_i = 1$ almost surely?

5. We know that every edge contains a vertex of linear degree, but the diameter of the tree grows logarithmically, so there must be connected subtrees consisting of low-degree vertices whose linear growth has not kicked in yet. This is illustrated by the proof of Proposition 6.22, which shows that there are paths of length $\Theta(\log(n)/\log\log(n))$ that consist of just degree 2 vertices. It would be interesting to get a better understanding of the law of these exceptional substructures that contain most of the low-degree vertices. What does the forest induced by the vertices of degree at most N, for large N, look like? Do these subtrees look like 'young' friend trees?

6. A natural extension of the model is to attach the new vertex to multiple, say m, vertices. There are two variants: either V_n is a uniformly random vertex and the new vertex $n+1$ attaches to m independently sampled random neighbours of V_n, or we let $V_n^{(1)}, \dots, V_n^{(m)}$ to be independent random vertices and we let $n+1$ connect to a uniform neighbour of each of the $V_n^{(i)}$.

7. A second variation is to choose $0 < p < 1$ and connect to V_n with probability p and to W_n with probability $1-p$. This modification makes it much easier for neighbours of high-degree vertices to grow their degree, and in particular, the degree of every vertex goes to infinity almost surely as the tree grows. It would be interesting to see how much of the structure of the random friend tree remains after this modification.

8. Another final modification of the model, as described in the introduction, is to let W_n be the endpoint of a random walk with k steps rather than 1 step from V_n. In the case of $k = 0$, we obtain an URRT and if k is sufficiently large such that the random walk is perfectly mixed, we get a PA tree. One could study how properties such as the size of the largest degree depend on k.

6.8 Appendix

We prove Proposition 6.25. We start by stating and proving a technical lemma that is needed for its proof.

We make use of the following straightforward fact. Fix $a, b > 0$ and let $(Y_k)_{k\geq 0}$ be a random walk with steps in $\{-a, b\}$ such that $\mathbf{E}[\Delta Y_k] = c > 0$. Then, $\mathbf{P}\{\Delta Y_k = b\} = (a+c)/(a+b)$ and by writing $\tau = \inf\{k \geq 0 : Y_k < 0\}$, we have $\mathbf{P}\{\tau = \infty\} > 0$.

Lemma 6.37. *Let $B = (B_k)_{k\geq 0}$ be a random process adapted to a filtration $(\mathcal{F}_k)_{k\geq 0}$. Suppose that there exist $a, b > 0$ such that, almost surely for each k, $\Delta B_k \in \{-a, 0, b\}$. Suppose further*

that there exists a constant $c > 0$ such that

$$\mathbf{E}\left[\Delta B_k \mid \mathcal{F}_k, B_k < 0, \Delta B_k \neq 0\right] \geq c\,.$$

Then, there exists a constant $C = C(a, b, c)$ such that

$$\mathbf{P}\{B_n < -C \log n \text{ infinitely often}\} = 0,$$

and $\mathbf{E}[B_n] \geq -C \log n$.

Proof. We bound B from below by another, simpler process $S = (S_n)_{n\geq 0}$. The conditions of the lemma imply that we may couple B with a sequence $(Y_k)_{k\geq 0}$ of independent, random variables taking values in $\{-a, b\}$ with

$$\mathbf{P}\{Y_k = b\} = \frac{a+c}{a+b} = 1 - \mathbf{P}\{Y_k = -a\},$$

such that for all $k \geq 0$, on the event that $B_k < 0$ and $\Delta B_k \neq 0$ we have $\Delta B_k \geq Y_k$. We define S via the following coupling with B:

1. if $B_n \geq 0$, then $S_{n+1} = -a$,
2. if $B_n < 0$ and $\Delta B_n = 0$ then $\Delta S_n = 0$,
3. if $B_n < 0$ and $\Delta B_n \neq 0$ then $\Delta S_n = Y_n$.

An illustration of the coupling can be found in Figure 6.4. Since $B_n \geq 0$ implies $B_{n+1} \geq -a$, it is immediate that $S_n \leq B_n$ for all $n \geq 0$. The process S is a sequence of independent negative (incomplete) excursions of a random process that, restricted to the non-constant steps, has independent increments with positive mean except for finitely many times almost surely. There are at most $n/2$ such excursions by time n, and if we collapse the constant steps, they are independent realisations of a random walk with step size in $\{-a, b\}$ and drift c, started at $-a$ and ended before (or when) reaching 0. To understand their minimum, let us denote by $R = (R_n)_{n\geq 0}$ a random walk starting at 0, with steps in $\{-a, b\}$ and strictly positive drift c. We define

$$\tau_1 = \inf\{k\colon\ R_k < 0\},$$

and

$$\tau_\ell = \inf\left\{k > \tau_{\ell-1} \colon R_k < R_{\tau_{\ell-1}}\right\}.$$

With positive probability, R stays positive forever, and in particular, using the fact stated just before Lemma 6.37, there exists $0 < p < 1$ such that

$$0 < 1 - p = \mathbf{P}\{\tau_\ell = \infty | \tau_{\ell-1} < \infty\}.$$

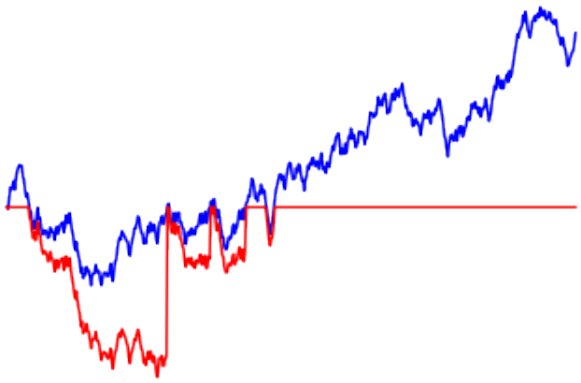

Figure 6.4: Illustration of a coupling between B (in blue) and S (in red).

Since the increments are bounded from below by $-a$, we know that $R_{\tau_\ell} - R_{\tau_{(\ell-1)}} \geq -a$, which together with the previous identity implies that

$$\min_{k\geq 1}\{R_k\} \geq_{st} -a\cdot \mathrm{Geom}(1-p)\,.$$

From the definition of the coupling, we know that $B_n \geq S_n$ and that $S_n + a$ stochastically dominates the minimum of n realisations of $(R_k)_{k\geq 0}$. Let $(A_n)_{n\geq 1}$ be independent $\mathrm{Geom}(1-p)$ random variables. Then

$$\mathbf{P}\{S_n \leq -a(k+1)\} \leq \mathbf{P}\left\{\max_{i\in[n]} A_i \geq k\right\} \leq np^k.$$

The upper bound is at most n^{-2} if $k \geq -3\log n/\log p$ and the first assertion follows from the Borel–Cantelli lemma. Taking $k = -3\log n/\log p + \ell$, the above bound likewise implies that

$$\mathbf{P}\left\{-\frac{S_n}{a} + 3\log n/\log p \geq \ell + 1\right\} \leq np^{-3\log n/\log p}p^\ell,$$

and summing over $\ell \geq 0$ gives the bound

$$\mathbf{E}\left[-\frac{S_n}{a} + 3\log n/\log p\right] \leq np^{-3\log n/\log p}\frac{1}{1-p},$$

which establishes the second assertion of the lemma. ∎

We are now ready to prove the proposition. **Proof.** [Proof of Proposition 6.25] We start by proving Proposition 6.25 in the case where the hypothesis holds for all n (and not all but finitely many n). Let $\beta \in (0, \alpha)$. We prove the proposition by applying Lemma 6.37 to the process $B = (B_k)_{k\geq 0}$ with $B_k := X_k - \beta Y_k$. Note that B has increments in $\{-\beta, 0, 1\}$. We need to show that there exists a constant c such that for all k,

$$\mathbb{E}\{\Delta B_k \mid \mathcal{F}_k, B_k < 0, \Delta B_k \neq 0\} \geq c\,. \tag{6.8.1}$$

Define

$$p_k^+ = \mathbf{P}\{\Delta X_k = 1 \mid \mathcal{F}_k, (\Delta X_k, \Delta Y_k) \neq (0,0)\} = \mathbf{P}\{\Delta B_k = 1 \mid \mathcal{F}_k, \Delta B_k \neq 0\}$$

and so

$$1 - p_k^+ = \mathbf{P}\{\Delta Y_k = 1 \mid \mathcal{F}_k, (\Delta X_k, \Delta Y_k) \neq (0,0)\} = \mathbf{P}\Big\{\Delta B_k = -\beta \;\Big|\; \mathcal{F}_k, \Delta B_k \neq 0\Big\}.$$

Note that, if $\{B_k < 0\}$, then $\{X_k < \beta Y_k\}$, and so $\mathbb{E}\{\Delta X_k \mid \mathcal{F}_k\} \geq \alpha \mathbb{E}\{\Delta Y_k \mid \mathcal{F}_k\}$, which implies that $p_k^+ \geq \alpha(1-p_k^+)$. We split the event $\{B_k < 0\}$ into two cases and show that $\mathbb{E}\{\Delta B_k \mid \mathcal{F}_k, \Delta B_k \neq 0\}$ is bounded below by some positive constant in both cases. If $B_k < 0$ and $p_k^+ \leq \frac{1/2+\beta}{1+\beta}$, then

$$\mathbb{E}\{\Delta B_k \mid \mathcal{F}_k, \Delta B_k \neq 0\} = p_k^+ - \beta(1-p_k^+) \geq (\alpha-\beta)(1-p_k^+) \geq \frac{(\alpha-\beta)}{2+2\beta}.$$

On the other hand, if $B_k < 0$ and $p_k^+ > \frac{1/2+\beta}{1+\beta}$, then

$$\mathbb{E}\{\Delta B_k \mid \mathcal{F}_k, \Delta B_k \neq 0\} = p_k^+(1+\beta) - \beta \geq 1/2.$$

Therefore, (6.8.1) holds, with a lower bound of $\min\left\{\frac{(\alpha-\beta)}{2+2\beta}, 1/2\right\}$ for c, and the claim then directly follows from Lemma 6.37.

Finally, we prove that the first statement in the proposition still holds when the assumptions fail at a finite number of times.

We call a time k *bad* when $\mathbf{E}[\Delta X_k \mid \mathcal{F}_k] < \alpha \mathbf{E}[\Delta Y_k \mid \mathcal{F}_k]$ and $\{X_k < \alpha Y_k\}$; otherwise we call it *good*. We couple (X, Y) to a slightly modified process (X', Y') that has the same increments as (X, Y) except at bad times, and that satisfies the assumptions at all times. Observe that for each k, given $\mathcal{F}_k$, we know whether k is bad or not. If k is bad, set $(\Delta X_k', \Delta Y_k') = (1, 0)$. If k is good, set $(\Delta X_k', \Delta Y_k') = (\Delta X_k, \Delta Y_k)$. We claim that (X', Y') satisfies the assumptions at all times. The requirement $0 \leq \Delta X_k' + \Delta Y_k' \leq 1$ for all k is obviously satisfied. Moreover, for bad k, $\mathbf{E}\left[\Delta X_k' \mid \mathcal{F}_k\right] = 1$ and $\mathbf{E}\left[\Delta Y_k' \mid \mathcal{F}_k\right] = 0$, so at bad times the second requirement is also satisfied. Finally, by construction, $X_k' \geq X_k$ and $Y_k' \leq Y_k$ for all k, so if k is good and $\{X_k' < \alpha Y_k'\}$, then also $\{X_k < \alpha Y_k\}$, and therefore

$$\mathbf{E}\left[\Delta X_k' \mid \mathcal{F}_k\right] = \mathbf{E}[\Delta X_k \mid \mathcal{F}_k] \geq \alpha \mathbf{E}[\Delta Y_k \mid \mathcal{F}_k] = \alpha \mathbf{E}\left[\Delta Y_k' \mid \mathcal{F}_k\right].$$

Then, the first part of the proof implies that for any $\beta \in (0, \alpha)$ there exists $C > 0$ such that $\mathbf{P}\{X'_n < \beta Y'_n - C \log n \text{ i.o.}\} = 0$. Finally, for B the total number of bad times, for each k it holds that $X'_k \leq X_k + B$ and $Y'_k \geq Y_k - B$. This implies that if $X_k < \beta Y_k - 2C \log k$ then either $X'_k < \beta Y'_k - C \log n$ or $2B > C \log k$. Therefore,

$$\mathbf{P}\{X_n < \beta Y_n - 2C \log n \text{ i.o.}\} \leq \mathbf{P}\{X'_n < \beta Y'_n - C \log n \text{ i.o.}\} + \mathbf{P}\{B = \infty\} = 0.$$

■

Chapter 7

On the quality of randomized approximations of Tukey's depth

Contents

Abstract

Tukey's depth (or halfspace depth) is a widely used measure of centrality for multivariate data. However, exact computation of Tukey's depth is known to be a hard problem in high dimensions. As a remedy, randomized approximations of Tukey's depth have been proposed. In this chapter we explore when such randomized algorithms return a good approximation of Tukey's depth. We study the case when the data are sampled from a log-concave isotropic distribution. We prove that, if one requires that the algorithm runs in polynomial time in the dimension, the randomized algorithm correctly approximates the maximal depth $1/2$ and depths close to zero. On the other hand, for any point of intermediate depth, any good approximation requires exponential complexity.

This Chapter is based on a joint work with Gábor Lugosi and Roberto Imbuzeiro Oliveira (Briend, Lugosi, and Oliveira [28]).

7.1 Introduction

Ever since Tukey introduced a notion of data depth [137], it has been an important tool of data analysts to measure centrality of data points in multivariate data. Apart from Tukey's depth (also called halfspace depth), many other depth measures have been developed, such as simplical depth (Liu [99, 100]), projection depth (Liu [101], Zuo and Serfling [143]), a notion of "outlyingness" (Stahel [134], Donoho [56]), and the zonoid depth (Dyckerhoff, Mosler, and Koshevoy [61], Koshevoy and Mosler [92]). Each of these notions offer distinct stability and computability properties that make them suitable for different applications (Mosler and Mozharovskyi [111]). For surveys of depth measures and their applications we refer the reader to Mosler [110], Aloupis [5], Dyckerhoff and Mozharovskyi [60], and Nagy et al. [114].

Tukey's depth is defined as follows: for $x \in \mathbb{R}^d$ and unit vector $u \in S^{d-1}$ (where S^{d-1} is the unit sphere of $\mathbb{R}^d$ under the euclidean norm), introduce the closed half space

$$H(x,u) = \left\{ y \in \mathbb{R}^d \ : \ \langle y, u \rangle \leq \langle x, u \rangle \right\} ,$$

where $\langle \cdot, \cdot \rangle$ is the usual scalar product on $\mathbb{R}^d$. Given a set of n data points $\{x_1, \ldots, x_n\}$ in $\mathbb{R}^d$, for each $x \in \mathbb{R}^d$, define the directional depth

$$r_n(x,u) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{x_i \in H(x,u)} .$$

The depth of x in the point set $\{x_1, \ldots, x_n\}$ is defined as

$$d_n(x) = \inf_{u \in S^{d-1}} r_n(x,u) .$$

Note that, due to the normalization in our definition, $d_n(x) \in [0, 1/2]$ for all $x \in \mathbb{R}^d$. Tukey's depth possesses properties expected of a depth measure. It is affine invariant, it vanishes at infinity, and it is monotone decreasing on rays emanating from the deepest point. It is also robust under a symmetry assumption (Donoho and Gasko [57]).

A well-known disadvantage of Tukey's depth is that even its approximate computation is known to be a np-hard problem (Amaldi and Kann [6], Bremner et al. [24], Johnson

and Preparata [85]), presenting challenges for applications (as it is conjectured no polynomial time algorithms exist to solve np-hard problems). While fast algorithms exist for computing the depth of the deepest point in two dimensions (Chan [37]), the computational complexity grows exponentially with the dimension. Chan [37] gives a maximum-depth computation algorithm of complexity $\mathcal{O}(n^{d-1})$.

The curse of dimensionality affects several other depth measures, posing significant challenges in multivariate analysis. To address these challenges, focus has been put on developing approximation algorithms. Dyckerhoff, Mozharovskyi, and Nagy [62] emphasize the importance of finding such algorithms and Shao et al. [131] propose mcmc methods for approximating the projection depth. Zuo [142] suggests an approximate version of Tukey's depth and provides an algorithm with linear time complexity in the dimension, though the proposed version may be a poor approximation of Tukey's depth.

A natural way of approximating Tukey's depth, proposed by Cuesta-Albertos and Nieto-Reyes [45], is a randomized version in which the infimum over all possible directions $u \in S^{d-1}$ in the definition of $d_n(x)$ is replaced by the minimum over a number of randomly chosen directions. More precisely, let $U_1, \ldots, U_k$ be independent identically distributed vectors sampled uniformly on the unit sphere S^{d-1}, and define the *random Tukey depth* (with respect to the point set $\{x_1, \ldots, x_n\}$) as

$$D_{n,k}(x) = \min_{i=1,\ldots,k} r_n(x, U_i) \, .$$

It is easy to see that for every $x \in \mathbb{R}^d$, $\lim_{k\to\infty} D_{n,k}(x) = d_n(x)$ with probability 1. However, this randomized approach is only useful if the number of random directions k is reasonably small so that computation is feasible. The purpose of this chapter is to explore the tradeoff between computational complexity and accuracy. In particular, we may ask how large k has to be in order to guarantee that, for given accuracy and confidence parameters $\epsilon \in (0, 1/2)$ and $\delta \in (0, 1)$, $|D_{n,k}(x) - d_n(x)| \leq \epsilon$ with probability at least $1 - \delta$.

It is easy to see that the value of k required to satisfy the property above may be arbitrarily large. To see this, consider the two-dimensional example in which, for $i = 1, \ldots, n$, the points $x_i = (x_{i,1}, x_{i,2})$ are defined by

$$x_{i,1} = \frac{i}{n} \, , \qquad x_{i,2} = a\left(\frac{i}{n}\right)^2$$

where $a > 0$ is a parameter. For any k, as $a \to 0$, the random depth fails to approximate $d_n(x_{n/2}) = 1/n$ (see the Introduction for more details about this example).

In order to exclude the anomalous behaviour of the example above, we assume that the points x_i are drawn randomly from an isotropic log-concave distribution μ. Recall that a distribution μ is log-concave if it is absolutely continuous with respect to the

Lebesgue measure, with density f of the form $f(x) = e^{-g(x)}$ where $g : \mathbb{R}^d \to \mathbb{R}$ is a convex function. μ is isotropic if for a random vector X distributed by μ, the covariance matrix $\mathbb{E}(X - \mathbb{E}X)(X - \mathbb{E}X)^T$ is the identity matrix. Examples of log-concave distributions include Gaussian distributions and the uniform distribution on a convex body in $\mathbb{R}^d$.

For random data, one may introduce the "population" counterpart of r_n defined by

$$\overline{r}(x,u) = \mu(H(x,u)) \,.$$

Similarly, the population versions of the Tukey depth and randomized Tukey depth are defined by

$$\overline{d}(x) = \inf_{u \in S^{d-1}} \overline{r}(x,u) \quad \text{and} \quad \overline{D}_k(x) = \min_{i=1,\dots,k} \overline{r}(x,U_i) \,.$$

As it was observed by Cuesta-Albertos and Nieto-Reyes [45] and Chen, Gao, and Ren [39], as long as $n \gg d$, the population versions of the Tukey depth $\overline{d}(x)$ and randomized Tukey depth $\overline{D}_k(x)$ are good approximations of $d_n(x)$ and $D_{n,k}(x)$, respectively. This follows from standard uniform convergence results of empirical process theory based on the vc dimension. The next lemma quantifies this closeness. For completeness we include its proof in the Appendix.

Lemma 7.1. *Let $\delta > 0$. If $X_1, \dots, X_n$ are independent, identically distributed random vectors in $\mathbb{R}^d$, then*

$$\mathbb{P}\left\{ \sup_{x \in \mathbb{R}^d} |\overline{d}(x) - d_n(x)| \geq c\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}} \right\} \leq \delta$$

where c is a universal constant. Also, given any fixed values of $U_1, \dots, U_k$,

$$\mathbb{P}\left\{ \sup_{x \in \mathbb{R}^d} |\overline{D}_k(x) - D_{n,k}(x)| \geq c\sqrt{\frac{\min(d, \log(k))}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}} \,\middle|\, U_1, \dots, U_k \right\} \leq \delta$$

Thanks to Lemma 7.1, in the rest of the chapter we restrict our attention to the population quantities $\overline{d}(x)$ and $\overline{D}_k(x)$ and we may forget the data points $X_1, \dots, X_n$. In particular, we are interested in finding out for what points $x \in \mathbb{R}^d$ and $k \geq 0$ the random Tukey depth $\overline{D}_k(x)$ is a good approximation of $\overline{d}(x)$. To this end, we fix an accuracy $\epsilon > 0$ and a confidence level $\delta > 0$ and ask that

$$\overline{D}_k(x) - \overline{d}(x) \leq \epsilon \quad \text{holds with probability at least } 1 - \delta. \tag{7.1.1}$$

(Note that, by definition, $\overline{D}_k(x) \geq \overline{d}(x)$ for all x and k.) The main results of the chapter show an interesting trichotomy: for most "shallow" points (i.e., those with $\overline{d}(x) \leq \epsilon$), we have $\overline{D}_k(x) \leq \epsilon$ with probability at least $1 - \delta$ even for k of *constant* order, depending only

on ϵ and δ. When x has near maximal depth in the sense that $\overline{d}(x) \approx 1/2$ (note that such points may not exist unless the density of μ is symmetric), then for values of k that are slightly larger than a linear function of d, (7.1.1) holds. However, in sharp contrast with this, for points x of intermediate depth, k needs to be exponentially large in d in order to guarantee (7.1.1). Hence, roughly speaking, the depth of very shallow and very deep points can be efficiently approximated by the random Tukey depth but for all other points, any reasonable approximation by the random Tukey depth requires exponential complexity.

7.1.1 Related literature

Cuesta-Albertos and Nieto-Reyes [45] explore various properties of the random Tukey depth and report good experimental behavior. The maximum discrepancy between d_n and its randomized approximation has also been studied by Nagy et al. [115]. They establish conditions under which $\sup_{x\in\mathbb{R}^d}\left(\overline{D}_k(x) - \overline{d}(x)\right) \to 0$ as $k \to \infty$ and provide bounds for the rate of convergence. As opposed to the global view of [115], our aim is to identify the points x for which the random Tukey depth approximates well $\overline{d}(x)$ for values of k that are polynomial in the dimension.

Brazitikos, Giannopoulos, and Pafis [23] show that the average depth $\int \overline{d}(x)d\mu(x)$ is exponentially small in the dimension when μ is log-concave.

Brunel [31] studies convergence of the empirical level sets when the data points are drawn independently from the same distribution.

Chen et al. [38] study the quality of other randomized approximations of the Tukey depth for point sets in general position.

7.1.2 Contributions and outline

As mentioned above, the main results of this chapter show that, for isotropic log-concave distributions, the quality of approximation of the random Tukey depth varies dramatically, depending on the depth of the point x.

Most points have a small random Tukey depth

In Section 7.2 we establish results related to shallow points. It follows from results of Brazitikos et al. [23] and Markov's inequality that all but an exponentially small (in the

dimension) fraction of points are shallow in the sense that, for all $\epsilon > 0$,

$$\mu\left(\left\{x \in \mathbb{R}^d : \overline{d}(x) > \epsilon\right\}\right) \leq \frac{e^{-cd}}{\epsilon},$$

where $c > 0$ is a universal constant. The main result of Section 7.2 is that, in high dimensions, not only most points are shallow but most points even have a small random Tukey depth for k of *constant* order, only depending on the desired accuracy. In particular, Theorem 7.5 implies the following.

Corollary 7.2. *Assume that μ is an isotropic log-concave measure on $\mathbb{R}^d$. There exist universal constants $c, \kappa, C > 0$ such that for any $\epsilon, \delta, \gamma > 0$, if*

$$k = \left\lceil \max\left(C, \frac{4}{\epsilon}\log\frac{3}{\gamma}, \frac{2}{c}\log\frac{4}{\delta}\right)\right\rceil,$$

and the dimension d is so large that

$$d \geq \max\left(\left(\frac{3(k+1)}{\gamma}\right)^{1/\kappa}, \frac{64\log(1/\epsilon)k}{\pi}\log\frac{3k}{\gamma}, \left(\frac{1}{c}\log\frac{6k}{\delta}\right)^2, \left(\frac{2}{\epsilon}\right)^{\kappa}\right),$$

then, with probability at least $1-\delta$,

$$\mu\left(\left\{x \in \mathbb{R}^d : \overline{D}_k(x) > \epsilon\right\}\right) < \gamma.$$

Of course, $\overline{D}_k(x) \leq \epsilon$ implies that $\overline{d}(x) \leq \epsilon$ and, in particular, that $\overline{D}_k(x) - \overline{d}(x) \leq \epsilon$. Thus, Corollary 7.2 implies that the random Tukey depth of *most* points (in terms of the measure μ) is a good approximation of the Tukey depth after taking just a constant number of random directions. All of these points are shallow in the sense that $\overline{d}(x) \leq \epsilon$.

It is natural to ask whether the Tukey depth of every shallow point is well approximated by its random version. However, this is false as the following example shows.

Example. Let μ be the uniform distribution on $[-(3/2)^{1/3}, (3/2)^{1/3}]^d$ so that μ is isotropic and log-concave on $\mathbb{R}^d$. If $x = ((3/2)^{1/3}, 0, \ldots, 0)$, then $\overline{d}(x) = 0$, but it is a simple exercise to show that $\overline{D}_k \geq 1/4$ with high probability, unless k is exponentially large in d.

Intermediate depth is hard to approximate

Arguably the most interesting points are those whose depth is in the intermediate range, bounded away from 0 and $1/2$. Unfortunately, for all such points, the random Tukey depth

is an inefficient approximation of the Tukey depth. In Section 7.3 we show that for all points in this range, the random Tukey depth $\overline{D}_k(x)$ is close to $1/2$, with high probability, unless k is exponentially large in the dimension. Hence, in high dimensions, $\overline{D}_k(x)$ fails to efficiently approximate the true depth $\overline{d}(x)$. In particular, Theorem 7.8 implies the following.

Corollary 7.3. *Assume that μ is an isotropic log-concave measure on $\mathbb{R}^d$ and let $\delta \in (0,1)$. For any $\gamma \in (0,1/2)$, there exists a positive constant $c = c(\gamma)$ such that if $x \in \mathbb{R}^d$ is such that $\overline{d}(x) = \gamma$, then for every $\epsilon < c$, if $k \leq \delta e^{d\epsilon^2 \log^2(1/\epsilon)/c}$, then, with probability at least $1-\delta$,*

$$\overline{D}_k(x) - \overline{d}(x) \geq \epsilon \, .$$

Points of maximum depth are easy to localize

As mentioned above, the Tukey depth $\overline{d}(x)$ of any $x \in \mathbb{R}^d$ is at most $1/2$. If $\overline{d}(x) = 1/2$, then for every $u \in S^{d-1}$, the median of the projection $\langle X, u\rangle$ equals $\langle x, u\rangle$ (where the random vector X is distributed as μ). Such points are quite special and may not exist at all. If there exists an $x \in \mathbb{R}^d$ with $\overline{d}(x) = 1/2$, then the measure μ is called *halfspace symmetric* (see Nagy et al. [114], Zuo and Serfling [144]). It is easy to see that if μ is halfspace symmetric, there is a unique $m \in \mathbb{R}^d$ with $\overline{d}(m) = 1/2$. We call m the *Tukey median* of μ. Centrally symmetric measures are halfspace symmetric though the converse does not hold in general. Remarkably, if μ is the uniform distribution over a convex body and it is halfspace symmetric, then it is also centrally symmetric, see Funk [70], Schneider [128].

We note that for any log-concave measure, $1/e \leq \sup_{x \in \mathbb{R}^d} \overline{d}(x) \leq 1/2$, see Nagy et al. [114, Theorem 3].

If $m \in \mathbb{R}^d$ is such that $\overline{d}(m) = 1/2$, then clearly $\overline{D}_k(m) = 1/2$ for all $k \geq 1$. In Section 7.4 we show that, for values of k that are only polynomial in d, points with $\overline{D}_k(x) \approx 1/2$ must be close to x. Hence, the random Tukey depth efficiently estimates the Tukey median for halfspace symmetric isotropic log-concave distributions. More precisely, Theorem 7.9, combined with Lemma 7.1 implies the following.

Corollary 7.4. *Assume that μ is an isotropic log-concave, halfspace symmetric measure on $\mathbb{R}^d$. Let $X_1, \ldots, X_n$ be independent random vectors distributed as μ. Let $m_{n,k} \in \mathbb{R}^d$ be an empirical random Tukey median, that is, $m_{n,k}$ is such that $D_{n,k}(m_{n,k}) = \max_{x \in R^d} D_{n,k}(x)$. There exist universal constants $c, C > 0$ such that for any $\delta \in (0,1)$ and $\gamma \in (0, c)$, if $n \geq Cd/\gamma^2$ and*

$$k \geq c\left(d \log d + \log(1/\delta)\right) ,$$

then $\|m_{n,k} - m\| \leq C\gamma\sqrt{d}$ with probability at least $1 - \delta$.

By taking γ of the order of $1/\sqrt{d}$, the corollary above shows that, as long as $n \gg d^2$, it suffices to take $O(d \log d)$ random directions so that the empirical random Tukey median is within distance of constant order of the Tukey median. Note that, due to the "thin-shell" property of log-concave measures (see, e.g., [63]), the measure μ is concentrated around a sphere of radius $\sqrt{d}$ centered at the Tukey median m and hence localizing m to within a constant distance is a nontrivial estimate.

One may even take γ to be smaller order than $1/\sqrt{d}$ and get a better precision with the same value of k. However, for better precision, one requires the sample size n to be larger.

7.2 Random Tukey depth of typical points

In this section we show that for isotropic log-concave distributions, in high dimensions, a constant number k of random directions suffice to make the random Tukey depth $\overline{D}_k$ small for most points. In other words, the curse of dimensionality is avoided in a strong sense. In particular, we prove the following theorem that implies Corollary 7.2 in a straightforward manner.

Theorem 7.5. *Assume that μ is an isotropic log-concave measure on $\mathbb{R}^d$. There exist universal constants $c, \kappa > 0$ such that the following holds. Let $\epsilon > 0$ and suppose that d is so large that $d^{-\kappa} \leq \epsilon/2$. Then for every $k \leq cd^{\kappa}$,*

$$\mu\Big(\{x \in \mathbb{R}^d : \overline{D}_k(x) > \epsilon\}\Big) \leq (1 - \epsilon/4)^k + (k+1)d^{-\kappa} + ke^{\frac{-d\pi}{64\log(1/\epsilon)k}}$$

with probability at least $1 - ke^{-ck} - 3ke^{-c\sqrt{d}}$ over the choice of directions $U_1, \ldots, U_k$.

Our main tool is the following extension of Klartag's celebrated central limit theorem for convex bodies (Klartag [87]). Let $G_{d,k}$ denote the grassmannian of all k-dimensional

subspaces of $\mathbb{R}^d$ and let $\sigma_{d,k}$ be the unique rotationally invariant probability measure on $G_{d,k}$.

Proposition 7.6. *(Klartag [88].) Let the random vector X take values in $\mathbb{R}^d$ and assume that X has an isotropic log-concave distribution. Let S_k be a random k-dimensional subspace of $\mathbb{R}^d$ drawn from the distribution $\sigma_{d,k}$. There exist universal constants $c, \kappa > 0$ such that the following holds: if $k \le cd^\kappa$, then with probability at least $1 - e^{-c\sqrt{d}}$, for every measurable set $A \subset S_k$,*

$$|\mathbb{P}\{\pi_k(X) \in A\} - \mathbb{P}\{N \in A\}| \le d^{-\kappa}$$

where N is a k-dimensional normal vector in S_k with zero mean and identity covariance matrix, and π_k is the orthogonal projection on S_k.

Proof of Theorem 7.5: First note that the random subspace of $\mathbb{R}^d$ spanned by the independent uniform vectors $U_1, \ldots, U_k$ has a rotation-invariant distribution and therefore it is distributed by $\sigma_{d,k}$ over the grassmannian $G_{d,k}$.

For any $u \in S^{d-1}$, define $q(\epsilon, u)$ as the ϵ-quantile of the distribution of $\langle X, u\rangle$, that is,

$$\mu(\{x : \langle x, u\rangle \le q(\epsilon, u)\}) = \epsilon \, .$$

Observe that, by Proposition 7.6 (applied with $k = 1$) and the union bound, with probability at least $1 - ke^{-c\sqrt{d}}$,

$$\text{for all } i = 1, \ldots, k, \quad q(\epsilon, U_i) \ge \Phi^{-1}(\epsilon/2)$$

whenever d is so large that $d^{-\kappa} \ge \epsilon/2$ where $\Phi(z) = \int_{-\infty}^{z} (2\pi)^{-1/2} e^{-x^2/2} dx$ denotes the standard Gaussian cumulative distribution function.

Then, with probability at least $1 - ke^{-c\sqrt{d}}$,

$$\begin{aligned} \mu\Big(\Big\{x : \overline{D}_k(x) > \epsilon\Big\}\Big) &= \mu\left(\left\{x : \min_{i=1,\ldots,k} \mu(H(x, U_i)) > \epsilon\right\}\right) \\ &= \mu(\{x : \langle x, U_i\rangle > q(\epsilon, U_i) \text{ for all } i = 1, \ldots, k\}) \\ &\le \mu\Big(\Big\{x : \langle x, U_i\rangle > \Phi^{-1}(\epsilon/2) \text{ for all } i = 1, \ldots, k\Big\}\Big) \end{aligned}$$

If the U_i were orthogonal, we could now use Proposition 7.6. This is not the case but almost. In order to handle this issue, we perform Gram-Schmidt orthogonalization defined, recursively, by $V_1 = U_1$ and, for $i = 2, \ldots, k$,

$$R_i = \sum_{j=1}^{i-1} \langle U_i, V_j\rangle V_j \quad \text{and} \quad V_i = \frac{U_i - R_i}{\|U_i - R_i\|} \, .$$

Then $V_1, \ldots, V_k$ are orthonormal vectors, spanning the same subspace as $U_1, \ldots, U_k$.

Now, we may write

$$\begin{aligned}
&\mu\left(\left\{x : \overline{D}_k(x) > \epsilon\right\}\right) \\
&\quad \leq \quad \mu\left(\left\{x : \langle x, U_i\rangle > \Phi^{-1}(\epsilon/2) \text{ for all } i = 1, \ldots, k\right\}\right) \\
&\quad \leq \quad \mu\left(\left\{x : \langle x, V_i\rangle > \Phi^{-1}(\epsilon/4) \text{ for all } i = 1, \ldots, k\right\}\right) \\
&\qquad + \mu\left(\left\{x : \langle x, U_i - V_i\rangle > \Phi^{-1}(\epsilon/2) - \Phi^{-1}(\epsilon/4) \text{ for some } i = 1, \ldots, k\right\}\right) \\
&\quad \leq \quad \mu\left(\left\{x : \langle x, V_i\rangle > \Phi^{-1}(\epsilon/4) \text{ for all } i = 1, \ldots, k\right\}\right) \\
&\qquad + \sum_{i=1}^{k} \mu\left(\left\{x : \langle x, U_i - V_i\rangle > \frac{\sqrt{2\pi}}{4\log(1/\epsilon)}\right\}\right), \qquad (7.2.1)
\end{aligned}$$

where the last inequality follows from the union bound and the inequality

$$\Phi^{-1}(\epsilon/2) - \Phi^{-1}(\epsilon/4) \geq \frac{\sqrt{2\pi}}{4\log(1/\epsilon)}\,. \qquad (7.2.2)$$

Indeed, since Φ^{-1} is concave on $[0, 1/2]$, we have

$$\frac{\Phi^{-1}(\epsilon/2) - \Phi^{-1}(\epsilon/4)}{\epsilon/4} \geq (\Phi^{-1})'(\epsilon/2)\,.$$

Using the fact that $(\Phi^{-1})' = 1/(\Phi' \Phi^{-1})$ and $\Phi'(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$,

$$\Phi^{-1}(\epsilon/2) - \Phi^{-1}(\epsilon/4) \geq \frac{\epsilon}{4}\sqrt{2\pi} e^{\Phi^{-1}(\epsilon/2)^2/2}\,. \qquad (7.2.3)$$

By Gordon's inequality for the Mills' ratio (Gordon [73]), for $t \leq 0$,

$$\Phi(t) \geq -\frac{1}{\sqrt{2\pi}} \frac{t}{t^2+1} e^{-t^2/2}\,,$$

and therefore

$$t \geq \Phi^{-1}\left(-\frac{1}{\sqrt{2\pi}} \frac{t}{t^2+1} e^{-t^2/2}\right),$$

leading, for $t < -1$, to

$$t \geq \Phi^{-1}\left(-\frac{e^{-t^2/2}}{10t}\right). \qquad (7.2.4)$$

Choosing $t_\epsilon = -\sqrt{2\log(1/\epsilon)}\sqrt{1 - \frac{\log\log(1/\epsilon)}{\log(1/\epsilon)}}$ for $\epsilon < e^{-2}$ and noting that

$$-\frac{e^{-t_\epsilon^2/2}}{10 t_\epsilon} \geq \epsilon/2,$$

(7.2.4) implies that

$$-\sqrt{2\log(1/\epsilon)}\sqrt{1-\frac{\log\log(1/\epsilon)}{\log(1/\epsilon)}} \geq \Phi^{-1}(\epsilon/2)\,.$$

Plugging this inequality into (7.2.3)

$$\Phi^{-1}(\epsilon/2) - \Phi^{-1}(\epsilon/4) \geq \frac{\sqrt{2\pi}}{4\log(1/\epsilon)}\,,$$

proving (7.2.2).

As $\langle x, V_1\rangle, \ldots, \langle x, V_k\rangle$ are coordinates of the orthogonal projection of x on the random subspace spanned by $U_1, \ldots, U_k$, we may use Proposition 7.6 to bound the first term on the right-hand side of (7.2.1). Let $N_1, \ldots, N_k$ be independent standard normal random variables. Then by Proposition 7.6, with probability at least $1 - e^{-c\sqrt{d}}$,

$$\begin{aligned}
&\mu\Big(\Big\{x : \langle x, V_i\rangle > \Phi^{-1}(\epsilon/4) \text{ for all } i = 1, \ldots, k\Big\}\Big) \\
&\quad\leq \mathbb{P}\{N_i > \Phi^{-1}(\epsilon/4) \text{ for all } i = 1, \ldots, k\} + d^{-\kappa} \\
&\quad= \mathbb{P}\{N_1 > \Phi^{-1}(\epsilon/4)\}^k + d^{-\kappa} \\
&\quad= (1 - \epsilon/4)^k + d^{-\kappa}
\end{aligned}$$

It remains to bound the second term on the right-hand side of (7.2.1). Once again, we use Proposition 7.6. By rotational invariance, the distribution of $U_i - V_i/\|U_i - V_i\|$ is uniform on S^{d-1} and therefore the distribution of $\mu\left(\left\{x : \langle x, U_i - V_i\rangle > \frac{\sqrt{2\pi}}{4\log(1/\epsilon)}\right\}\right)$ is the same as that of

$$\mu\left(\left\{x : \langle x, W\rangle > \frac{\sqrt{2\pi}}{4\log(1/\epsilon)\|U_i - V_i\|}\right\}\right)$$

(if $\epsilon \leq 1/2$) where W is uniformly distributed on S^{d-1}, independent of $U_1, \ldots, U_n$. .By Lemma 7.7 below, with probability at least $1 - ke^{-ck}$,

$$\max_{i=1,\ldots,k} \|U_i - V_i\| \leq \sqrt{4k/d}.$$

Combining this with Proposition 7.6, we have that, with probability at least $1 - ke^{-ck} - ke^{-c\sqrt{d}}$,

$$\begin{aligned}
\sum_{i=1}^{k} \mu\left(\left\{x : \langle x, U_i - V_i\rangle > \frac{\sqrt{2\pi}}{4\log(1/\epsilon)}\right\}\right) &\leq kd^{-\kappa} + k\mathbb{P}\left\{N > \frac{\sqrt{2\pi}}{4\log(1/\epsilon)}\sqrt{\frac{d}{4k}}\right\} \\
&\leq kd^{-\kappa} + ke^{\frac{-d\pi}{64\log(1/\epsilon)k}}\,.
\end{aligned}$$

In order to complete the proof of Theorem 7.5, it remains to prove the following simple inequality.

Lemma 7.7. *For every $i = 1, \ldots, k$, with probability at least $1 - e^{-ck}$,*

$$\|U_i - V_i\| \leq \sqrt{\frac{4k}{d}}$$

where c is a universal constant.

Proof. Note that, since $\|R_i\|^2 = \langle U_i, R_i \rangle$,

$$\langle U_i, V_i \rangle = \frac{1 - \langle U_i, R_i \rangle}{\|U_i - R_i\|} = \sqrt{1 - \|R_i\|^2} \leq 1 - \|R_i\|^2$$

and therefore

$$\|U_i - V_i\|^2 = 2(1 - \langle U_i, V_i \rangle) \leq 2\|R_i\|^2 = 2 \sum_{j=1}^{i-1} \langle U_i, V_j \rangle^2 .$$

We may write $U_i = Z_i / \|Z_i\|$ where Z_i is a Gaussian vector in $\mathbb{R}^d$ with zero mean and identity covariance matrix. Since Z_i is independent of $V_1, \ldots, V_{i-1}$ and the V_j are orthonormal, $\sum_{j=1}^{i-1} \langle Z_i, V_j \rangle^2$ is a χ^2 random variable with $i-1$ degrees of freedom. Thus, $\|U_i - V_i\|^2$ is the ratio of a $\chi^2(i-1)$ and a $\chi^2(d)$ random variable (which are not independent). By standard tail bounds of the χ^2 distribution (see, e.g., [19]), with probability at least $1 - e^{-ck}$,

$$\|U_i - V_i\|^2 \leq \frac{4k}{d} .$$

■

7.3 Estimating intermediate depth is costly

In this section we prove that, even though the random Tukey depth is small for most points $x \in \mathbb{R}^d$ (according to the measure μ), whenever the depth $\overline{d}(x)$ of a point is not small, its random Tukey depth $\overline{D}_k(x)$ is close to $1/2$, unless k is exponentially large in d. This implies that for points whose depth is bounded away from $1/2$, the random Tukey depth is a poor approximation of $\overline{d}(x)$.

The main result of the section is the following theorem that immediately implies Corollary 7.3 stated in Section 6.1.

Theorem 7.8. *Assume that μ is an isotropic log-concave measure on $\mathbb{R}^d$ and let $0 < \gamma < 1/2$. Let $x \in \mathbb{R}^d$ be such that $\overline{d}(x) = \gamma$ and let $\epsilon > 0$. Then*

$$\mathbb{P}\left\{ \overline{D}_k(x) \leq \frac{1}{2} - C_\gamma \frac{\epsilon}{\log\left(\frac{1}{\epsilon}\right)} \right\} \leq 2ke^{-(d-1)\epsilon^2/2} ,$$

where $C_\gamma > 0$ is a constant depending only on γ.

Proof. Without loss of generality, we may assume that the origin has maximal depth, that is, $\overline{d}(0) = \sup_{x \in \mathbb{R}^d} \overline{d}(x)$. Fix $x \in \mathbb{R}^d$ with $\overline{d}(x) = \gamma$, and note that $\overline{d}(0) \geq \gamma$.

The main tool of this proof is Lévy's isoperimetric inequality (Schmidt [127], Lévy [98], see also Ledoux [96]). It states that if the random vector U is uniformly distributed on the sphere S^{d-1} and A a is Borel-measurable set such that $\mathbb{P}\{U \in A\} \geq 1/2$, then for any $\epsilon > 0$,

$$\mathbb{P}\left\{ \inf_{v \in A} \|U - v\| \geq \epsilon \right\} \leq 2e^{-(d-1)\epsilon^2/2} . \tag{7.3.1}$$

Lévy's inequality may be used to prove concentration inequalities for smooth functions of the random vector U. Our goal is to prove that the measure $\mu(H(x,U))$ of the random halfspace $H(x,U)$ is concentrated around its median $1/2$.

In order to prove smoothness of the function $\mu(H(x,u))$ (as a function of $u \in S^{d-1}$), fix $u, v \in S^{d-1}$, $u \neq v$. Consider the 2-dimensional cone spanned by the segments (x,u) and (x,v) defined by

$$C(x,u,v) = \{x + au + bv : \ a,b \in \mathbb{R}^+\} .$$

Denote by $\mathcal{H}$ the only two-dimensional affine space containing $x,\ x+u,\ x+v$.

We also define $P_{\mathcal{H}}$ as the orthogonal projection onto $\mathcal{H}$. Denoting by $\widetilde{\mu} = P_{\mathcal{H}}\#\mu$ and $\widetilde{H}(x,u) = P_{\mathcal{H}}(H(x,u))$, we have

$$\mu(H(x,u)) = \widetilde{\mu}\left(\widetilde{H}(x,u)\right) .$$

Thus, after projecting on the plane $\mathcal{H}$, it suffices to control

$$\begin{aligned} |\mu(H(x,u)) - \mu(H(x,v))| &= \left|\widetilde{\mu}\left(\widetilde{H}(x,u)\right) - \widetilde{\mu}\left(\widetilde{H}(x,v)\right)\right| \\ &= \left|\widetilde{\mu}\left(C(x,u^\perp,v^\perp)\right) - \widetilde{\mu}\left(C(x,-u^\perp,-v^\perp)\right)\right| \\ &\leq \widetilde{\mu}\left(C(x,u^\perp,v^\perp)\right) + \widetilde{\mu}\left(C(x,-u^\perp,-v^\perp)\right) , \end{aligned} \tag{7.3.2}$$

that is, the measure of two cones in a 2 dimensional affine space. Here, given an arbitrary orientation to the plane $\mathcal{H}$, $u^\perp$ and $v^\perp$ are the only unit vectors orthogonal to u and v,

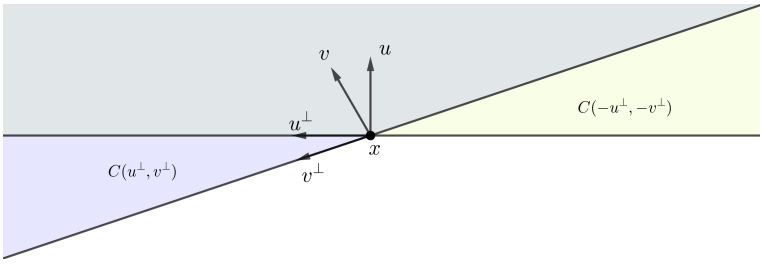

Figure 7.1: Illustration of the cones $C(x, u^\perp, v^\perp)$ and $C(x, -u^\perp, -v^\perp)$.

respectively, in $\mathcal{H}$ such that $u^\perp$ and $v^\perp$ are rotated 90 degrees counter-clockwise from u and v, see Figure 7.1.

Since the measure $\widetilde{\mu}$ is itself an isotropic log-concave measure (see Saumard and Wellner [126, Section 3], Prékopa [120]), the problem becomes two dimensional. Next, we show that neither $\|x\|$ nor $|m_v|$ are too large, where m_v denotes the median of the random variable $\langle X, v\rangle$. (Note that m_v is uniquely defined since $\langle X, v\rangle$ is log-concave and therefore has a unimodal density.)

In the Appendix we gather some useful facts on log-concave densities. In particular, Lemma 7.13 shows that any one dimensional log-concave density with unit variance is upper bounded by an exponential function centered at the median of the log-concave density. Since $\overline{d}(0) \le \overline{r}(0, v)$ for all $v \in S^{d-1}$, Lemma 7.13 implies that there exist universal constants $c_1, c_2 > 0$ such that

$$\overline{d}(0) \le c_1 e^{-c_2|m_v|} .$$

Since $\overline{d}(0) \ge \gamma$, we have

$$c_2|m_v| \le \log(c_1/\gamma) . \tag{7.3.3}$$

Moreover, since $\overline{d}(x) \ge \gamma$, the same argument leads to

$$\gamma \le c_1 e^{-c_2|\langle x, v\rangle - m_v|} .$$

Using the above with $v = x/\|x\|$ and the inequality $|a - b| \ge |a| - |b|$ yields

$$c_2\|x\| \le \log(c_1/\gamma) + c_2|m_{x/\|x\|}| ,$$

which, put together with (7.3.3), implies

$$\|x\| \le c \log(c_1/\gamma) , \tag{7.3.4}$$

for a positive constant c. In particular, $\|P_{\mathcal{H}}(x)\| \le c \log(c_1/\gamma)$. We use this inequality to control the measure of half spaces around x. Using Lemma 7.13 we can uniformly upper

bound the measure of every half space around the median by

$$\widetilde{\mu}\Big(\widetilde{H}(m_v v + tv, v)\Big) \le c_1 e^{-c_2|t|} ,$$

where c_1 and c_2 are as in Lemma 7.13. Now using (7.3.3) and (7.3.4), we may uniformly bound the measure of half spaces around x. In particular, there exist constants $c_\gamma, c'_\gamma > 0$ such that for all $t \in \mathbb{R}$ and $u \in S^1$,

$$\widetilde{\mu}\Big(\widetilde{H}(x + tu, u)\Big) \le c_\gamma e^{-c'_\gamma|t|} . \tag{7.3.5}$$

Next we use the fact that the density of an isotropic log-concave density in $\mathbb{R}^2$ is upper bounded by a universal constant. Obtaining upper bounds for log-concave densities is an important problem in high-dimensional geometry. In particular, the so-called *isotropic constant* of a log-concave density f defined by

$$L_f := \sqrt{\frac{\sup f}{\int f}} \sqrt[4]{\det(\mathrm{Cov}(f))} .$$

has a deep connection to Bourgain's "slicing problem" and the Kannan-Lovász-Simonovits conjecture, see, e.g., Klartag and Lehec [89], Lutwak [104]. Here we only need the simple fact that in a fixed dimension ($d = 2$ in our case) one has $\sup_f L_f \le K$ for a constant K. For an isotropic log-concave density, $L_f = \sqrt{\sup f}$, so indeed there exists an universal constant K which upper bounds any log-concave isotropic density in dimension 2.

Now we are ready to derive upper bounds for the right-hand side of (7.3.2). To this end, we decompose the cone $C(x, u^\perp, v^\perp)$ into two parts. For any $t > 0$ we may write

$$\widetilde{\mu}\Big(C(x, u^\perp, v^\perp)\Big) \le \widetilde{\mu}\Big(C(x, u^\perp, v^\perp) \cap B(x, t)\Big) + \widetilde{\mu}\Big(C(x, u^\perp, v^\perp) \cap \widetilde{H}(x + tu, u)\Big) ,$$

where $B(x, t)$ denotes the closed ball of radius t centered at x. Thus, from (7.3.5) and the upper bound on the density, we obtain

$$\widetilde{\mu}\Big(C(x, u^\perp, v^\perp)\Big)\pi \le Kt^2\theta + c_\gamma e^{-c'_\gamma t} ,$$

where $\theta \in [0, \pi]$ denotes the angle formed by vectors u and v. Choosing $t = \log(1/\theta)/c'_\gamma$, (7.3.2) implies

$$|\mu(H(x, u)) - \mu(H(x, v))| \le C'_\gamma \frac{\theta}{\log^2\left(\frac{1}{\theta}\right)}$$

for a constant C'_γ depending only on γ. Since $\theta \le \frac{\pi}{2}\|u - v\|$, we conclude that there exists a positive constant C_γ such that

$$|\mu(H(x, u)) - \mu(H(x, v))| \le C_\gamma \frac{\|u - v\|}{\log^2\left(\frac{1}{\|u-v\|}\right)} . \tag{7.3.6}$$

Now we are prepared to use Lévy's isoperimetric inequality. Choosing $A = \left\{v \in S^{d-1} : \ \mu(H(x,v)) \geq 1/2\right\}$, we clearly have $\mathbb{P}\{A\} = 1/2$ and therefore by (7.3.1)

$$\mathbb{P}\left\{\inf_{v\in A} \|U - v\| \geq \epsilon\right\} \leq 2e^{-(d-1)\epsilon^2/2} .$$

But for any $u \in S^{d-1}$ such that $\inf_{v\in A} \|u - v\| \geq \epsilon$, (7.3.6) implies that

$$\mu(H(x,u)) \geq \frac{1}{2} - C_\gamma \frac{\epsilon}{\log^2\left(\frac{1}{\epsilon}\right)} ,$$

so

$$\mathbb{P}\left\{\mu(H(x,U)) \leq \frac{1}{2} - C_\gamma \frac{\epsilon}{\log^2\left(\frac{1}{\epsilon}\right)}\right\} \leq 2e^{-(d-1)\epsilon^2/2} .$$

Since $\overline{D}_k(x) = \min_{i=1\ldots k} \mu(H(x,U_i))$ for $U_1, \ldots, U_k$ independently sampled uniformly on S^{d-1}, the union bound yields

$$\mathbb{P}\left\{\overline{D}_k(x) \leq \frac{1}{2} - C_\gamma \frac{\epsilon}{\log^2\left(\frac{1}{\epsilon}\right)}\right\} \leq 2ke^{-(d-1)\epsilon^2/2} ,$$

concluding the proof. ∎

7.4 Detection and localization of Tukey's median

As explained in the introduction, a measure μ is called halfspace symmetric if there exists a point $m \in \mathbb{R}^d$ with $\overline{d}(m) = 1/2$. Such a point is necessarily unique and we call it the Tukey median. Clearly, for all $k \geq 1$, the random Tukey depth of the Tukey median equals $\overline{D}_k(m) = 1/2$ and therefore, it is trivially an exact estimate of the Tukey depth of m. Here we show that, for any positive γ bounded by some constant, already for values of k that are of the order of $d \log d$, all points that are at least a distance of order $\gamma\sqrt{d}$ away from m have a random Tukey depth less than $1/2 - \gamma$, with high probability. This result implies that the Tukey median of isotropic log-concave, halfspace symmetric distributions are efficiently estimated by the random Tukey median, as stated in Corollary 7.4.

Theorem 7.9. *Assume that μ is an isotropic log-concave, halfspace symmetric measure on $\mathbb{R}^d$. Let $\delta > 0$ and let $\gamma, r > 0$ be such that $r \geq 32e^4\gamma$ and $r \leq \min\left(e^{-4}/6, 8e^4\gamma\sqrt{d}/2\right)$. There exists a universal constant $C > 0$ such that, if*

$$k \geq C\left(d \log \frac{r}{\gamma} + \log(1/\delta)\right) \frac{\gamma\sqrt{d}}{r} e^{C\gamma^2 d/r^2} ,$$

then

$$\mathbb{P}\left\{\sup_{x\in\mathbb{R}^d:\|x\|\geq r} \overline{D}_k(x) \geq \frac{1}{2} - \gamma\right\} \leq \delta\,.$$

In particular, by taking $r = 8e^4\gamma\sqrt{d/2}$, there exist universal constants $c, C > 0$ such that for all $\gamma \leq c$, if

$$k \geq C\left(d\log d + \log(1/\delta)\right)\,,$$

then

$$\mathbb{P}\left\{\sup_{x\in\mathbb{R}^d:\|x\|\geq C\gamma\sqrt{d}} \overline{D}_k(x) \geq \frac{1}{2} - \gamma\right\} \leq \delta\,.$$

Proof. Without loss of generality, we may assume that $m = 0$, that is, $\overline{d}(0) = 1/2$.

The outline of the proof is as follows. First, we show that for a fixed $x \in \mathbb{R}^d$ of norm r, we have $\overline{D}_k(x) \leq \frac{1}{2} - 2\gamma$ with high probability.

Then we use an ϵ-net argument to extend the control to the sphere $r \cdot S^{d-1}$. To this end, we need to establish certain regularity of the function $x \mapsto \overline{D}_k(x)$. We then use a monotonicity argument to extend the control to all points outside of the ball of radius r.

Recall that f denotes the density of the measure μ and the random vector X has distribution μ. For any direction $u \in S^{d-1}$, we denote by $\Phi_u(t) = \mathbb{P}\{\langle X, u\rangle \leq t\}$ the cumulative distribution function of the projection of X in direction u.

Fix $x \in r \cdot S^{d-1}$. Since $\overline{D}_k(x) = \min_{i=1\ldots k} \Phi_{U_i}(\langle x, U_i\rangle)$,

$$\mathbb{P}\left\{\overline{D}_k(x) \geq \frac{1}{2} - 2\gamma\right\} = \mathbb{P}\left\{\Phi_U(\langle x, U\rangle) \geq \frac{1}{2} - 2\gamma\right\}^k\,. \tag{7.4.1}$$

Next we bound the probability on the right-hand side. Since $\overline{d}(0) = 1/2$, for all $u \in S^{d-1}$, $\Phi_u(0) = 1/2$. Clearly, the function $t \mapsto \Phi_u(t)$ is non-decreasing, as it is a cumulative distribution function. Since projections of an isotropic log-concave measure are also log-concave and isotropic (see Saumard and Wellner [126, Section 3] and Prékopa [120]). Lemma 7.11 in the Appendix implies that for all $t \in [-e^{-4}/6, e^{-4}/6]$,

$$\Phi_u'(t) \geq e^{-4}/4\,,$$

and therefore, for all such t, we have

$$\left|\Phi_u(t) - \frac{1}{2}\right| \geq \frac{e^{-4}}{4}t\,.$$

Since $\|x\| = r \le e^{-4}/6$, we have $|\langle x, U_i\rangle| \le e^{-4}/6$ and hence

$$\begin{aligned}\mathbb{P}\left\{\Phi_U(\langle x, U\rangle \ge \frac{1}{2} - 2\gamma\right\} &\le \mathbb{P}\left\{\frac{1}{4e^4}\langle x, U\rangle \ge -2\gamma\right\} \\ &= 1 - \mathbb{P}\left\{\langle \frac{1}{r}x, U\rangle \ge \frac{8e^4\gamma}{r}\right\}.\end{aligned}$$

Since $\|\frac{1}{r}x\| = 1$, the probability on the right-hand side corresponds to the (normalized) measure of a spherical cap of height $h = 8e^4\gamma/r$. Thus, we may further bound the expression on the right-hand side by applying a lower bound for the measure of a spherical cap. Brieden, Gritzmann, Kannan, Klee, Lovász, and Simonovits [25] provide such a lower bound for $\sqrt{2/d} \le h \le 1$ which is guaranteed by our condition on r. We obtain

$$\mathbb{P}\left\{\Phi_U(\langle x, U\rangle \ge \frac{1}{2} - 2\gamma\right\} \le 1 - \frac{1}{6h\sqrt{d}}(1-h^2)^{\frac{d-1}{2}}.$$

Hence, by (7.4.1) we have that for any x with $\|x\| = r \in [32e^4\gamma, e^{-4}/6]$,

$$\begin{aligned}\mathbb{P}\left\{\overline{D}_k(x) \ge \frac{1}{2} - 2\gamma\right\} &\le \left(1 - \frac{1}{6h\sqrt{d}}(1-h^2)^{\frac{d-1}{2}}\right)^k \\ &\le \left(1 - \frac{1}{6h\sqrt{d}}e^{-h^2(d-1)/4}\right)^k \quad \text{(since } 1-x \ge e^{-x/2} \text{ for } x \in (0,1/2)\text{)} \\ &\le \exp\left(-\frac{k}{6h\sqrt{d}}e^{-h^2(d-1)/4}\right) \quad \text{(since } 1-x \le e^{-x} \text{ for } x \ge 0\text{).} \qquad (7.4.2)\end{aligned}$$

It remains to extend this inequality for a fixed x to a uniform control over all $\|x\| \ge r$. To this end, we need to establish regularity of the function $x \mapsto \overline{D}_k(x)$.

Since $\|u\| = 1$, the mapping $x \mapsto \langle x, u\rangle$ is 1-Lipschitz. Φ_u is the cumulative distribution function of an isotropic, one-dimensional, log-concave measure, and therefore its derivative is a log-concave density with variance 1. As stipulated in Lemma 7.12 in the Appendix, such a density is upper bounded by e^4. Hence, for any $u \in S^{d-1}$, $x \mapsto \Phi_u(\langle x, u\rangle)$ is e^4-Lipschitz. Furthermore, since the minimum of Lipschitz functions is Lipschitz, $x \mapsto \overline{D}_k(x)$ is also e^4-Lipschitz.

For $\epsilon > 0$, an ϵ-net of the sphere $r \cdot S^{d-1}$ is a subset N of $r \cdot S^{d-1}$ of minimal size such that for all $x \in r \cdot S^{d-1}$ there exists $y \in N$ with $\|x - y\| \le \epsilon$. It is well known (see, e.g., [108]) that for all $\epsilon > 0$, $r \cdot S^{d-1}$ has an ϵ-net N_ϵ of size at most $|N_\epsilon| \le \left(\frac{2r}{\epsilon} + 1\right)^d$. Using the fact that $\overline{D}_k(x)$ is e^4-Lipschitz, by taking $\epsilon = e^{-4}\gamma$, using (7.4.2) and the union bound, we have

$$\mathbb{P}\left\{\sup_{x \in \mathbb{R}^d : \|x\| = r} \overline{D}_k(x) \ge \frac{1}{2} - \gamma\right\} \le \left(2re^4\sqrt{d} + 1\right)^d \exp\left(-\frac{k}{6h\sqrt{d}}e^{-h^2(d-1)/4}\right). \qquad (7.4.3)$$

It remains to extend the inequality to include all points outside $r \cdot S^{d-1}$. To this end, it suffices to show that for any $a \geq 1$,

$$\overline{D}_k(ax) \leq \overline{D}_k(x) \,.$$

To see this, note that the deepest point 0 has depth $1/2$, so every closed half-space with 0 on its boundary has measure $1/2$. Hence, $\mu(H(x,u)) < 1/2$ if and only if $0 \notin H(x,u)$, which is equivalent to $\langle x, u\rangle < 0$. On the event $\left\{\sup_{x\in\mathbb{R}^d:\|x\|=r} \overline{D}_k(x) < \frac{1}{2} - \gamma\right\}$, for every $x \in r \cdot S^{d-1}$ there exists an $i \in [k]$ such that $\mu(H(x,U_i)) < 1/2$. This implies that for such an i, $\langle x, U_i\rangle < 0$, so for any $a \geq 1$ $\langle ax, U_i\rangle \leq \langle x, U_i\rangle$. Since $\mu(H(x,U_i)) = \Phi_{U_i}(\langle x, U_i\rangle)$ and that Φ_{U_i} is non decreasing, we have

$$\mu(H(ax,U_i)) \leq \mu(H(x,U_i)) \,,$$

leading to $\overline{D}_k(ax) \leq \overline{D}_k(x)$ as desired. This extends (7.4.3) to the inequality

$$\mathbb{P}\left\{\sup_{x\in\mathbb{R}^d:\|x\|\geq r} \overline{D}_k(x) \geq \frac{1}{2} - \gamma\right\} \leq \left(2re^4\sqrt{d}+1\right)^d \exp\left(-\frac{k}{6h\sqrt{d}}e^{-h^2(d-1)/4}\right) .$$

Recalling that $h = 8e^4\gamma/r$ and that r is bounded, this implies the announced statement. ■

Acknowledgement

We would like to thank Imre Bárány, Shahar Mendelson, Arshak Minasyan, Bill Steiger, and Nikita Zhivotovsky for helpful discussions. We also thank Reihaneh Malekian for her thorough reading of the original manuscript and for pointing out some inaccuracies.

7.5 Appendix

In this section, we compile several properties of one-dimensional, isotropic, log-concave densities. For a survey on log-concave densities, see Samworth [124].

7.5.1 Lower bounds for log-concave densities

Lemma 7.10. *Let $f(t) = e^{-g(t)}$ be a log-concave probability density on $\mathbb{R}$ having variance 1 and let m denote its (unique) median. Then*

$$e^{-g(m)} \geq \frac{e^{-4}}{2} .$$

Proof. Without loss of generality, we may assume that $m = 0$ and g takes its minimum on $\mathbb{R}^-$.

Since a convex function on an open interval is continuous, the only discontinuous log-concave density is the uniform density over an interval of length $2\sqrt{3}$ for which the statement holds, and therefore we may assume that f is continuous. If $g(0) \leq 0$ the result is obvious, so suppose $g(0) > 0$. Since g is convex, by taking its minimum on $\mathbb{R}^-$ it is non decreasing on $\mathbb{R}^+$. By continuity, there exists $L > 0$ such that $g(L) = 2g(0)$.

From the convexity of g we have that $g'(L) \geq \frac{g(0)}{L}$, and therefore for all $t \geq L$,

$$g(t) \geq g(L) + \frac{g(0)}{L}(t - L) \geq \frac{g(0)}{L} t . \tag{7.5.1}$$

Since $\int f(x)dx = 1$ and 0 is the median,

$$\frac{1}{2} = \int_0^L e^{-g(t)}dt + \int_L^\infty e^{-g(t)}dt .$$

Using (7.5.1),

$$\int_L^\infty e^{-g(t)}dt \leq \int_L^\infty e^{-g(0)t/L}dt = \frac{L}{g(0)} e^{-g(0)} .$$

Moreover, since g is convex and reaches its minimum on $\mathbb{R}^-$ it is non-decreasing on $\mathbb{R}^+$, so

$$\int_0^L e^{-g(t)}dt \leq e^{-g(0)} L ,$$

leading to

$$\frac{1}{2} \leq \frac{L}{g(0)} e^{-g(0)} + e^{-g(0)} L = e^{-g(0)} L \left(1 + \frac{1}{g(0)}\right) . \tag{7.5.2}$$

Now we use the fact that the variance equals 1, that is,

$$1 = \int_{-\infty}^{+\infty} t^2 e^{-g(t)}dt - \left(\int_{-\infty}^{\infty} t e^{-g(t)}dt\right)^2 .$$

Since the difference between the expectation and the median of any distribution is at most the standard deviation, we have $|\int_{-\infty}^{\infty} te^{-g(t)}dt| \leq 1$. Moreover, since g is increasing on $\mathbb{R}^+$, for all $t \in [0, L]$ we have $g(t) \leq 2g(0)$, and therefore $1 \geq \int_0^{\infty} t^2 e^{-g(t)}dt - 1$ implies

$$2 \geq \int_0^L t^2 e^{-2g(0)}dt = \frac{L^3}{3}e^{-2g(0)} . \tag{7.5.3}$$

From (7.5.2) we have

$$e^{-2g(0)}L^3 e^{-g(0)}\left(1 + \frac{1}{g(0)}\right)^3 \geq \frac{1}{8} .$$

Hence, by plugging the inequality into (7.5.3), we get

$$e^{-g(0)}\left(1 + \frac{1}{g(0)}\right)^3 \geq \frac{1}{48} . \tag{7.5.4}$$

Note that the function $h : t \mapsto e^{-t}\left(1 + \frac{1}{t}\right)^3$ is non increasing on $\mathbb{R}^+$. To conclude, observe that

- if $g(0) \leq 4.5$, then $e^{-g(0)} \geq \frac{e^{-4}}{2}$.
- if $g(0) > 4.5$, then
$$h(g(0)) < \frac{1}{48},$$
contradicting (7.5.4).

■

The next result shows that an isotropic log-concave density is in fact bounded from below by a universal constant on an interval around the median.

Lemma 7.11. *Let $f(t) = e^{-g(t)}$ be a log-concave probability density on $\mathbb{R}$ having variance 1 and median $m = 0$. Then for all $t \in \left[-\frac{1}{6e^4}, \frac{1}{6e^4}\right]$,*

$$f(t) \geq \frac{1}{4e^4} .$$

Proof. Denote $\alpha = 1/(6e^4)$ and suppose that there exists $t \in [-\alpha, \alpha]$ such that $f(t) < 1/(4e^4)$. Since log-concave densities are unimodal, on $[-\alpha, \alpha]$ the density f reaches its minimum on an endpoint of the interval. Without any loss of generality, assume that

$$e^{-g(\alpha)} < \frac{1}{4e^4} ,$$

that is,

$$g(\alpha) > 4 + \log(4) .$$

By the convexity of g, for all $t \geq \alpha$,

$$g(t) \geq \frac{g(\alpha) - g(0)}{\alpha}(t - \alpha) + g(\alpha) .$$

Since by Lemma 7.10, $g(0) \leq 4 + \log(2)$, we get that for all $t \geq \alpha$

$$g(t) \geq \frac{\log(2)}{\alpha}(t - \alpha) + \log(4e^4) .$$

It follows that

$$\int_{\alpha}^{\infty} e^{-g(t)} dt \leq \frac{1}{4e^4} \cdot \frac{\alpha}{\log(2)} .$$

We also prove in Lemma 7.12 below that $\sup_{t \in \mathbb{R}} e^{-g(t)} \leq e^4$, so

$$\int_{0}^{\alpha} e^{-g(t)} dt \leq \alpha e^4 .$$

Using the fact that 0 is the median, we get

$$1 = \frac{1}{2} + \int_{\mathbb{R}^+} e^{-g(t)} dt \leq \frac{1}{2} + \alpha \left(e^4 + \frac{1}{4e^4} \frac{1}{\log(2)} \right) .$$

But

$$\alpha \left(e^4 + \frac{1}{4e^4} \frac{1}{\log(2)} \right) < \frac{1}{2} ,$$

which is a contradiction. This concludes the proof. ∎

7.5.2 Upper bounds for log-concave densities

Lemma 7.12. *Let $f(t) = e^{-g(t)}$ be a log-concave probability density on $\mathbb{R}$ having variance 1. Then*

$$\sup_{t \in \mathbb{R}} e^{-g(t)} \leq e^4.$$

Proof. Without loss of generality, we may assume that $g(0) = \inf_{t \in \mathbb{R}} g(t)$ and $\int_0^{\infty} t^2 e^{-g(t)} dt \geq 1/2$. We may also assume that g is continuous. (Otherwise f is the uniform density over an interval of length $2\sqrt{3}$ for which the statement holds.)

First note that if $g(0) \geq 0$, then there's nothing to prove, so suppose that $g(0) < 0$. By the intermediate value theorem there exists $L > 0$ such that $g(L/2) = g(0)/2$. Since g is convex and $\int \exp(-g(t))dt = 1$, we have

$$Le^{-g(0)/2} \leq 1 \,. \tag{7.5.5}$$

Since g has a non-decreasing derivative, for all $t \geq L/2$,

$$g'(t) \geq -\frac{g(0)}{2} \cdot \frac{2}{L} = -\frac{g(0)}{L} \,.$$

Then for all $t \geq L/2$, $g(t) \geq g(0) - \frac{g(0)}{L}(t - L)$, which implies

$$\int_{L/2}^{\infty} t^2 e^{-g(t)} dt \leq e^{-2g(0)} \int_{L/2}^{\infty} t^2 e^{\frac{g(0)}{L} t} dt \,.$$

Since for $c > 0$

$$\int_{L/2}^{\infty} t^2 e^{-ct} dt = \left(\frac{L^2}{4c} + \frac{L}{c^2} + \frac{2}{c^3} \right) e^{-cL/2},$$

Taking $c = -g(0)/L$, which is positive,

$$\int_{L/2}^{\infty} t^2 e^{-g(t)} dt \ \leq \left(\frac{-L^3}{4g(0)} + \frac{L^3}{g(0)^2} - \frac{2L^3}{g(0)^3} \right) e^{-3g(0)/2} \,. \tag{7.5.6}$$

Next we establish a lower bound for $\int_{L/2}^{\infty} t^2 e^{-g(t)} dt$. The fact that the second moment on $\mathbb{R}^+$ is greater than $1/2$ implies

$$\int_{L/2}^{\infty} t^2 e^{-g(t)} dt \geq \frac{1}{2} - \int_0^{L/2} t^2 e^{-g(t)} dt \,.$$

It is immediate from the fact that g reaches its minimum in 0 that

$$\int_0^{L/2} t^2 e^{-g(t)} dt \leq \frac{L^3}{4} e^{-g(0)} \,,$$

leading to

$$\int_{L/2}^{\infty} t^2 e^{-g(t)} dt \geq \frac{1}{2} - \frac{L^3}{4} e^{-g(0)} \,. \tag{7.5.7}$$

Comparing (7.5.6) and (7.5.7), we obtain

$$\frac{1}{2} - \frac{L^3}{4} e^{-g(0)} \leq L^3 \left(\frac{-1}{4g(0)} + \frac{1}{g(0)^2} - \frac{2}{g(0)^3} \right) e^{-3g(0)/2} \,,$$

leading to

$$\frac{1}{2} \leq L^3 \left(\frac{e^{g(0)/2}}{4} - \frac{1}{4g(0)} + \frac{1}{g(0)^2} - \frac{2}{g(0)^3} \right) e^{-3g(0)/2} . \tag{7.5.8}$$

From (7.5.5) we have $L^3 e^{-3g(0)/2} \leq 1$, which, plugged into (7.5.8) yields

$$1 \leq 2 \left(\frac{e^{g(0)/2}}{4} - \frac{1}{4g(0)} + \frac{1}{g(0)^2} - \frac{2}{g(0)^3} \right) .$$

Since $g(0) \leq 0$,

$$1 \leq \frac{1}{2} - \frac{1}{2g(0)} + \frac{2}{g(0)^2} - \frac{4}{g(0)^3} . \tag{7.5.9}$$

The function $h : t \mapsto \frac{1}{2} - \frac{1}{2t} + \frac{2}{t^2} - \frac{4}{t^3}$ is non-decreasing on $\mathbb{R}^-$. To conclude the proof, note that if $g(0) \geq -4$, then $e^{-g(0)} \leq e^4$. Otherwise, if $g(0) < -4$, then, since h is non-decreasing,

$$h(g(0)) \leq h(-4) = \frac{13}{16} < 1 ,$$

which contradicts (7.5.9). ∎

It is known (see, e.g., Cule and Samworth [46]) that for any log-concave density f on $\mathbb{R}^d$, there exist positive constants α, β such that $f(x) \leq e^{-\alpha \|x\| + \beta}$ for all $x \in \mathbb{R}^d$. The next lemma shows that for isotropic log-concave densities on $\mathbb{R}$ with median at 0, one may choose α and β independently of f.

Lemma 7.13. *Let $f(x) = e^{-g(t)}$ be a log-concave probability density on $\mathbb{R}$ having variance 1 and median $m = 0$. Then there exist universal constants $\alpha, \beta > 0$ such that for all $t \in \mathbb{R}$,*

$$f(t) \leq \alpha e^{-\beta |t|} .$$

Proof. By Lemma 7.10 we have $e^{-g(0)} \geq e^{-4}/2$. The log-concavity of the density implies that on any given interval, the minimum is reached at one of the endpoints of the interval. Thus,

$$\int_0^{2e^4} e^{-g(t)} dt \geq 2e^4 \min \left(e^{-g(2e^4)}, e^{-4}/2 \right) .$$

Since 0 is the median of f, $2e^4 \min(e^{-g(2e^4)}, e^{-4}/2) \leq 1/2$. Thus,

$$e^{-g(2e^4)} \leq \frac{e^{-4}}{4} . \tag{7.5.10}$$

A mirror argument proves that $e^{-g(-2e^4)} \leq \frac{e^{-4}}{4}$. By Lemma 7.10, $g(0) \leq \log(2) + 4$ and (7.5.10) implies $g(2e^4) \geq \log(4) + 4$. Using the convexity of g yields that for all $t \geq 2e^4$,

$$g(t) \geq 4 + \log(4) + (t - 2e^4) \frac{\log(2)}{2e^4} ,$$

so, using Lemma 7.12 which states that $g(0) \geq 4$, for all $t \in \mathbb{R}^+$,

$$g(t) \geq \log(4) + (t - 2e^4)\frac{\log(2)}{2e^4} .$$

A identical argument on $\mathbb{R}^-$ concludes the proof of the Lemma. ■

7.5.3 Proof of Lemma 7.1

Proof. To prove the first inequality, observe that

$$\begin{aligned} \sup_{x\in\mathbb{R}^d} |\overline{d}(x) - d_n(x)| &= \sup_{x\in\mathbb{R}^d} \left| \inf_{u\in S^{d-1}} \mu(H(x,u)) - \inf_{u\in S^{d-1}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i\in H(x,u)} \right| \\ &\leq \sup_{x\in\mathbb{R}^d} \sup_{u\in S^{d-1}} \left| \mu(H(x,u)) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i\in H(x,u)} \right| . \end{aligned}$$

The first inequality of the Lemma follows from the Vapnik-Chervonenkis inequality and the fact that the vc dimension of the class of all half spaces $H(x,u)$ equals $d+1$.

The second inequality is proved similarly, combining it with a simple union bound that gives a better bound when $\log(k) \ll d$. ■

Bibliography

[1] Louigi Addario-Berry and Kevin Ford. Poisson–Dirichlet branching random walks. *The Annals of Applied Probability*, 23(1):283 – 307, 2013. doi: 10.1214/12-AAP840.

[2] Louigi Addario-Berry, Luc Devroye, Gábor Lugosi, and Vasiliki Velona. Broadcasting on random recursive trees. *The Annals of Applied Probability*, 32(1):497–528, 2022.

[3] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47, 2002.

[4] Noga Alon, Michael Krivelevich, and Benny Sudakov. Finding a large hidden clique in a random graph. *Random Structures & Algorithms*, 13(3-4):457–466, 1998.

[5] Greg Aloupis. Geometric measures of data depth. *DIMACS series in discrete mathematics and theoretical computer science*, 72:147, 2006.

[6] Edoardo Amaldi and Viggo Kann. The complexity and approximability of finding maximum feasible subsystems of linear relations. *Theoretical Computer Science*, 147(1-2): 181–210, 1995.

[7] Omer Angel and Yinon Spinka. Geometric random graphs on circles. In Daniel Hernández-Hernández, Florencia Leonardi, Ramsés H. Mena, and Juan Carlos Pardo Millán, editors, *Advances in Probability and Mathematical Statistics*, pages 23–41, Cham, 2021. Springer International Publishing. ISBN 978-3-030-85325-9.

[8] Tonći Antunović, Elchanan Mossel, and Miklós Z Rácz. Coexistence in preferential attachment networks. *Combinatorics, Probability and Computing*, 25(6):797–822, 2016.

[9] Caelan Atamanchuk, Luc Devroye, and Gabor Lugosi. A note on estimating the dimension from a random geometric graph, arxiv preprint arxiv:2311.13059, 2023.

[10] Sayan Banerjee and Shankar Bhamidi. Persistence of hubs in growing random networks. *Probability Theory and Related Fields*, 180(3-4):891–953, 2021.

[11] Sayan Banerjee and Shankar Bhamidi. Root finding algorithms and persistence of Jordan centrality in growing random trees. *The Annals of Applied Probability*, 32(3): 2180–2210, 2022.

[12] Sayan Banerjee and Xiangying Huang. Degree centrality and root finding in growing random networks. *Electronic Journal of Probability*, 28:1 – 39, 2023.

[13] Sayan Banerjee, Shankar Bhamidi, and Xiangying Huang. Co-evolving dynamic networks. *arXiv preprint arXiv:2203.11877*, 2022.

[14] A. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286 (5439):509–512, 1999. doi: 10.1126/science.286.5439.509.

[15] Louigi Addario Berry, Simon Briend, Luc Devroye, Serte Donderwinkel, Céline Kerriou, and Gábor Lugosi. Random friend trees, 2024.

[16] Jean Bertoin. Limits of Pólya urns with innovations, 2022.

[17] Puspal Bhabak, Hovhannes A. Harutyunyan, and Shreelekha Tanna. Broadcasting in Harary-like graphs. In *2014 IEEE 17th International Conference on Computational Science and Engineering*, pages 1269–1276, 2014.

[18] Ramesh Bhandari. *Survivable networks: algorithms for diverse routing*. Springer Science & Business Media, 1999.

[19] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[20] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[21] Anna M Brandenberger, Luc Devroye, and Marcel K Goh. Root estimation in Galton–Watson trees. *Random Structures & Algorithms*, 61(3):520–542, 2022.

[22] Michael Brautbar and Michael J. Kearns. Local algorithms for finding interesting individuals in large networks. In *Innovations in Theoretical Computer Science (ITCS)*, 2010.

[23] Silouanos Brazitikos, Apostolos Giannopoulos, and Minas Pafis. Half-space depth of log-concave probability measures, 2022.

[24] David Bremner, Dan Chen, John Iacono, Stefan Langerman, and Pat Morin. Output-sensitive algorithms for Tukey depth and related problems. *Statistics and Computing*, 18:259–266, 2008.

[25] A. Brieden, P. Gritzmann, R. Kannan, V. Klee, L. Lovász, and M. Simonovits. Deterministic and randomized polynomial-time approximation of radii. *Mathematika. A Journal of Pure and Applied Mathematics*, 48(1-2):63–105, 2001.

[26] Simon Briend, Francisco Calvillo, and Gábor Lugosi. Archaeology of random recursive dags and cooper-frieze random networks. *Combinatorics, Probability and Computing*, pages 1–15, 2023.

[27] Simon Briend, Luc Devroye, and Gabor Lugosi. Broadcasting in random recursive dags. *arXiv preprint arXiv:2306.01727*, 2023.

[28] Simon Briend, Gábor Lugosi, and Roberto Imbuzeiro Oliveira. On the quality of randomized approximations of tukey's depth. *arXiv preprint arXiv:2309.05657*, 2023.

[29] Simon Briend, Christophe Giraud, Gábor Lugosi, and Déborah Sulem. Estimating the history of a random recursive tree, 2024.

[30] Nicolas Broutin and Omar Fawzi. Longest path distance in random circuits. *Combinatorics, Probability and Computing*, 21(6):856–881, 2012.

[31] Victor-Emmanuel Brunel. Concentration of the empirical level sets of Tukey's halfspace depth. *Probability Theory and Related Fields*, 173(3):1165–1196, 2019.

[32] Sébastien Bubeck, Elchanan Mossel, and Miklós Rácz. On the influence of the seed graph in the preferential attachment model. *IEEE Transactions on Network Science and Engineering*, 2(1):30–39, 2015.

[33] Sébastien Bubeck, Luc Devroye, and Gábor Lugosi. Finding Adam in random growing trees. *Random Structures & Algorithms*, 50(2):158–172, 2017.

[34] Sébastien Bubeck, Ronen Eldan, Elchanan Mossel, and Miklós Rácz. From trees to seeds: on the inference of the seed from large trees in the uniform attachment model. *Bernoulli*, 23(4A):2887–2916, 2017.

[35] Chris Cannings and Jonathan Jordan. Random walk attachment graphs. *Electronic Communications in Probability*, 18(none):1 – 5, 2013. doi: 10.1214/ECP.v18-2518.

[36] George T Cantwell, Guillaume St-Onge, and Jean-Gabriel Young. Recovering the past states of growing trees. *arXiv preprint arXiv:1910.04788*, 2019.

[37] Timothy M Chan. An optimal randomized algorithm for maximum Tukey depth. In *SODA*, volume 4, pages 430–436, 2004.

[38] Dan Chen, Pat Morin, and Uli Wagner. Absolute approximation of Tukey depth: Theory and experiments. *Computational Geometry*, 46(5):566–573, 2013.

[39] Mengjie Chen, Chao Gao, and Zhao Ren. Robust covariance and scatter matrix estimation under Huber's contamination model. *The Annals of Statistics*, 46(5):1932–1960, 2018.

[40] Alice Contat, Nicolas Curien, Perrine Lacroix, Etienne Lasalle, and Vincent Rivoirard. Eve, adam and the preferential attachment tree, 2023.

[41] Colin Cooper and Alan M. Frieze. On a general model of web graphs. *Random Structures & Algorithms*, 22:311–335, 2003.

[42] Harry Crane and Min Xu. Inference on the history of a randomly growing tree. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(4):639–668, 2021.

[43] Harry Crane and Min Xu. Root and community inference on the latent growth process of a network using noisy attachment models. *arXiv preprint arXiv:2107.00153*, 2021.

[44] Irene Crimaldi, Pierre-Yves Louis, and Ida G. Minelli. An urn model with random multiple drawing and random addition. *Stochastic Processes and their Applications*, 147:270–299, 2022.

[45] Juan Antonio Cuesta-Albertos and Alicia Nieto-Reyes. The random Tukey depth. *Computational Statistics & Data Analysis*, 52(11):4979–4988, 2008.

[46] Madeleine Cule and Richard Samworth. Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electronic Journal of Statistics*, 4:254 – 270, 2010.

[47] Nicolas Curien, Thomas Duquesne, Igor Kortchemski, and Ioan Manolescu. Scaling limits and influence of the seed graph in preferential attachment trees. *Journal de l'École Polytechnique–Mathématiques*, 2:1–34, 2015.

[48] Colin Desmarais, Cecilia Holmgren, and Stephan Wagner. Broadcasting induced colourings of random recursive trees and preferential attachment trees. *arXiv preprint arXiv:2110.15050*, 2021.

[49] Luc Devroye. Branching processes in the analysis of the heights of trees. *Acta Informatica*, 24(3):277–298, 1987.

[50] Luc Devroye. Applications of the theory of records in the study of random trees. *Acta Informatica*, 26(1):123–130, 1988. doi: 10.1007/BF02915448.

[51] Luc Devroye and Svante Janson. Long and short paths in uniform random recursive dags. *Arkiv för Matematik*, 49(1):61–77, 2011.

[52] Luc Devroye and Jiang Lu. The strong convergence of maximal degrees in uniform random recursive trees and dags. *Random Structures & Algorithms*, 7(1):1–14, 1995.

[53] Luc Devroye and Tommy Reddad. On the discovery of the seed in uniform attachment trees. *Internet Mathematics*, pages 75–93, 2019.

[54] Josep Díaz Cort, María José Serna Iglesias, Paul George Spirakis, Jacobo Torán Romero, and Tatsuie Tsukiji. On the expected depth of boolean circuits. Technical report, Technical Report LSI-94-7-R, Universitat Politecnica de Catalunya, Dep. LSI, 1994.

[55] Sander Dommers, Remco van der Hofstad, and Gerard Hooghiemstra. Diameters in preferential attachment models. *Journal of Statistical Physics*, 139:72–107, 2010.

[56] David Donoho. Breakdown properties of multivariate location estimators. Technical report, Technical report, Harvard University, 1982.

[57] David L. Donoho and Miriam Gasko. Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness. *The Annals of Statistics*, 20(4): 1803 – 1827, 1992.

[58] Michael Drmota. *Random trees: an interplay between combinatorics and probability*. Springer Science & Business Media, 2009.

[59] Richard. Durrett. *Essentials of Stochastic Processes by Richard Durrett.* Springer Texts in Statistics. Springer International Publishing, Cham, 3rd edition 2016. edition, 2016. ISBN 9783319456140.

[60] Rainer Dyckerhoff and Pavlo Mozharovskyi. Exact computation of the halfspace depth. *Computational Statistics & Data Analysis*, 98:19–30, 2016.

[61] Rainer Dyckerhoff, Karl Mosler, and Gleb Koshevoy. Zonoid data depth: Theory and computation. In *COMPSTAT: Proceedings in Computational Statistics*, pages 235–240. Springer, 1996.

[62] Rainer Dyckerhoff, Pavlo Mozharovskyi, and Stanislav Nagy. Approximate computation of projection depths, 2020.

[63] Ronen Eldan and Joseph Lehec. Bounding the norm of a log-concave vector via thin-shell estimates. In *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 2011-2013*, pages 107–122. Springer, 2014.

[64] Janos Engländer, Giulio Iacobelli, Gábor Pete, and Rodrigo Ribeiro. Structural results for the tree builder random walk. *arXiv preprint arXiv:2311.18734*, 2023.

[65] Laura Eslava. Depth of vertices with high degree in random recursive trees. *ALEA*, 2022.

[66] T. S. Evans and J. P. Saramäki. Scale-free networks from self-organization. *Phys. Rev. E*, 72:026138, Aug 2005. doi: 10.1103/PhysRevE.72.026138.

[67] William Evans, Claire Kenyon, Yuval Peres, and Leonard J. Schulman. Broadcasting on trees and the Ising model. *The Annals of Applied Probability*, 10(2):410 – 433, 2000.

[68] Bertrand Even, Christophe Giraud, and Nicolas Verzelen. Computation-information gap in high-dimensional clustering, 2024.

[69] Alan Frieze and Michał Karoński. *Introduction to random graphs*. Cambridge University Press, 2016.

[70] Paul Funk. Über eine geometrische Anwendung der Abelschen Integralgleichung. *Mathematische Annalen*, 77(1):129–135, 1915.

[71] Edward N Gilbert. Random plane networks. *Journal of the society for industrial and applied mathematics*, 9(4):533–543, 1961.

[72] Christophe Giraud, Yann Issartel, and Nicolas Verzelen. Localization in 1d non-parametric latent space models from pairwise affinities. *Electronic Journal of Statistics*, 17(1):1587–1662, 2023.

[73] Robert D Gordon. Values of Mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *The Annals of Mathematical Statistics*, 12(3):364–366, 1941.

[74] John Haigh. The recovery of the root of a tree. *Journal of Applied Probability*, 7(1): 79–88, 1970.

[75] Hovhannes A. Harutyunyan and Zhiyuan Li. A new construction of broadcast graphs. *Discrete Applied Mathematics*, 280:144–155, 2020.

[76] S. Janson, T. Łuczak, and A. Rucinski. *Random graphs*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience, New York, 2000.

[77] Svante Janson. Functional limit theorems for multitype branching processes and

generalized pólya urns. *Stochastic Processes and their Applications*, 110(2):177–245, 2004.

[78] Svante Janson. Asymptotic degree distribution in random recursive trees. *Random Structures & Algorithms*, 26(1-2):69–83, 2005.

[79] Svante Janson. Limit theorems for triangular urn schemes. *Probability Theory and Related Fields*, 134:417–452, 03 2006. doi: 10.1007/s00440-005-0442-7.

[80] Svante Janson. Random replacements in Pólya urns with infinitely many colours. *Electronic Communications in Probability*, 24:1 – 11, 2019.

[81] Svante Janson and Elchanan Mossel. Robust reconstruction on trees is determined by the second eigenvalue. *Annals of Probability*, 32(3B):2630–2649, 2004.

[82] Jeannette Janssen and Aaron Smith. Reconstruction of line-embeddings of graphons. *Electronic Journal of Statistics*, 16(1):331 – 407, 2022. doi: 10.1214/21-EJS1940.

[83] Varun Jog and Po-Ling Loh. Analysis of centrality in sublinear preferential attachment trees via the crump-mode-jagers branching process. *IEEE Transactions on Network Science and Engineering*, 4(1):1–12, 2016.

[84] Varun Jog and Po-Ling Loh. Persistence of centrality in random growing trees. *Random Structures and Algorithms*, 52(1):136–157, 2018.

[85] D.S. Johnson and F.P. Preparata. The densest hemisphere problem. *Theoretical Computer Science*, 6(1):93–107, 1978.

[86] Justin Khim and Po-Ling Loh. Confidence sets for the source of a diffusion in regular trees. *IEEE Transactions on Network Science and Engineering*, 4(1):27–40, 2016.

[87] B. Klartag. A central limit theorem for convex sets. *Inventiones Mathematicae*, 168(1): 91–131, 2007.

[88] B. Klartag. Power-law estimates for the central limit theorem for convex sets. *Journal of Functional Analysis*, 245(1):284–310, 2007.

[89] Bo'az Klartag and Joseph Lehec. Bourgain's slicing problem and KLS isoperimetry up to polylog. *Geometric and Functional Analysis*, 32(5):1134–1159, 2022.

[90] J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Lecture notes in computer science. In *Proceedings of the International Conference on Combinatorics and Computing*, volume 1627, pages 1–18. Springer-Verlag, 1999.

[91] Margarete Knape and Ralph Neininger. Pólya urns via the contraction method. *Combinatorics, Probability and Computing*, 23(6):1148–1186, 2014.

[92] Gleb Koshevoy and Karl Mosler. Zonoid trimming for multivariate distributions. *The Annals of Statistics*, 25(5):1998–2017, 1997.

[93] P. L. Krapivsky and S. Redner. Emergent network modularity. *J. Stat. Mech. Theory Exp.*, (7):073405, 22, 2017. doi: 10.1088/1742-5468/aa7a3f.

[94] Markus Kuba and Hosam M. Mahmoud. Two-color balanced affine urn models with multiple drawings. *Advances in Applied Mathematics*, 90:1–26, 2017.

[95] Sanjay Kumar, Neeraj Bhat, and BS Panda. Analysis of social network metrics based on the model of random recursive tree. *Journal of Interdisciplinary Mathematics*, 23 (1):237–246, 2020.

[96] M. Ledoux. *The Concentration of Measure Phenomenon*. American Mathematical Society, 2001.

[97] Clement Lee and Darren J Wilkinson. A review of stochastic block models and extensions for graph clustering. *Applied Network Science*, 4(1):1–50, 2019.

[98] P. Lévy. *Problèmes conrets d'analyse fonctionelle*. Gauthier-Villars, 1951.

[99] Regina Y Liu. On a notion of simplicial depth. *Proceedings of the National Academy of Sciences*, 85(6):1732–1734, 1988.

[100] Regina Y Liu. On a notion of data depth based on random simplices. *The Annals of Statistics*, pages 405–414, 1990.

[101] Regina Y Liu. Data depth and multivariate rank tests. *L1-statistical analysis and related methods*, pages 279–294, 1992.

[102] Gábor Lugosi. Lecture on combinatorial statistics, July 2017.

[103] Gábor Lugosi and Alan S. Pereira. Finding the seed of uniform attachment trees. *Electronic Journal of Probability*, 24:1–15, 2019.

[104] Erwin Lutwak. Chapter 1.5 - selected affine isoperimetric inequalities. In P.M. GRUBER and J.M. WILLS, editors, *Handbook of Convex Geometry*, pages 151–176. North-Holland, Amsterdam, 1993.

[105] Hosam Mahmoud. *Pólya urn models*. Chapman and Hall/CRC, 2008. doi: 10.1201/9781420059847.

[106] Hosam M Mahmoud. The degree profile in some classes of random graphs that generalize recursive trees. *Methodology and Computing in Applied Probability*, 16(3): 527–538, 2014.

[107] Anuran Makur, Elchanan Mossel, and Yury Polyanskiy. Broadcasting on random directed acyclic graphs. *IEEE Transactions on Information Theory*, 66(2):780–812, 2020.

[108] J. Matoušek. *Lectures on Discrete Geometry*. Springer, 2002.

[109] John W Moon. On the centroid of recursive trees. *Australasian Journal of Combinatorics*, 25:211–220, 2002.

[110] Karl Mosler. *Multivariate dispersion, central regions, and depth: the lift zonoid approach*, volume 165. Springer Science & Business Media, 2002.

[111] Karl Mosler and Pavlo Mozharovskyi. Choosing among notions of multivariate depth statistics, 2021.

[112] Elchanan Mossel. Reconstruction on trees: beating the second eigenvalue. *The Annals of Applied Probability*, 11(1):285–300, 2001.

[113] Elchanan Mossel. Survey: information flow on trees. In *Graphs, morphisms and statistical physics*, volume 63 of *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*, pages 155–170. Amer. Math. Soc., Providence, RI, 2004.

[114] S. Nagy, C. Schuett, and E.M. Werner. Data depth and floating body. *Statistics Surveys*, 13, 2019.

[115] Stanislav Nagy, Rainer Dyckerhoff, and Pavlo Mozharovskyi. Uniform convergence rates for the approximated halfspace and projection depth. *Electronic Journal of Statistics*, 14(2):3939–3975, 2020.

[116] Saket Navlakha and Carl Kingsford. Network archaeology: uncovering ancient networks from present-day interactions. *PLoS Computational Biology*, 7(4):e1001119, 2011.

[117] Robin Pemantle. Nonconvergence to unstable points in urn models and stochastic approximations. *The Annals of Probability*, 18(2):698–712, 1990.

[118] Robin Pemantle. A survey of random processes with reinforcement. *Probability Surveys*, 4:9–12, 2007.

[119] Boris Pittel. Note on the heights of random recursive trees and random m-ary search trees. *Random Structures & Algorithms*, 5(2):337–347, 1994.

[120] A. Prékopa. On logarithmic concave measures and functions. *Acta Sci. Math.(Szeged)*, 34:335–343, 1973.

[121] Antoine Recanati, Thomas Brüls, and Alexandre d'Aspremont. A spectral algorithm for fast de novo layout of uncorrected long nanopore reads. *Bioinformatics*, 33(20): 3188–3194, 2017.

[122] Antoine Recanati, Thomas Kerdreux, and Alexandre d'Aspremont. Reconstructing latent orderings by spectral clustering, 2018.

[123] William S Robinson. A method for chronologically ordering archaeological deposits. *American antiquity*, 16(4):293–301, 1951.

[124] Richard J. Samworth. Recent Progress in Log-Concave Density Estimation. *Statistical Science*, 33(4):493 – 509, 2018.

[125] Jari Saramäki and Kimmo Kaski. Scale-free networks generated by random walkers. *Physica A: Statistical Mechanics and its Applications*, 341:80–86, 2004. ISSN 0378-4371. doi: https://doi.org/10.1016/j.physa.2004.04.110.

[126] Adrien Saumard and Jon A Wellner. Log-concavity and strong log-concavity: a review. *Statistics Surveys*, 8:45, 2014.

[127] Erhard Schmidt. Die Brunn-Minkowskische Ungleichung und ihr Spiegelbild sowie die isoperimetrische Eigenschaft der Kugel in der euklidischen und nichteuklidischen Geometrie. I. *Mathematische Nachrichten*, 1(2-3):81–157, 1948.

[128] Rolf Schneider. Functional equations connected with rotations and their geometric applications. *Enseignenment Math.(2)*, 16:297–305, 1970.

[129] Devavrat Shah and Tauhid Zaman. Finding rumor sources on random trees. *Operations Research*, 64(3):736–755, 2016.

[130] Devavrat Shah and Tauhid R. Zaman. Rumors in a network: Who's the culprit? *IEEE Transactions on Information Theory*, 57(8):5163–5181, 2011.

[131] Wei Shao, Yijun Zuo, and June Luo. Employing the MCMC technique to compute the projection depth in high dimensions. *Journal of Computational and Applied Mathematics*, 411:114278, 2022.

[132] Allan Sly. Reconstruction for the Potts model. *The Annals of Probability*, 39(4):1365 – 1406, 2011.

[133] Jithin Kazuthuveettil Sreedharan, Abram Magner, Ananth Y. Grama, and Wojtek Szpankowski. Inferring temporal information from a snapshot of a dynamic network. *Scientific Reports*, 9, 2019.

[134] Werner A Stahel. *Robuste schätzungen: infinitesimale optimalität und schätzungen von kovarianzmatrizen*. PhD thesis, ETH Zürich, 1981.

[135] Tatsuie Tsukiji and H Mahmoud. A limit law for outputs in random recursive circuits. *Algorithmica*, 31(3):403–412, 2001.

[136] Tatsuie Tsukiji and Fatos Xhafa. On the depth of randomly generated circuits. In *European Symposium on Algorithms*, pages 208–220. Springer, 1996.

[137] J. W. Tukey. Mathematics and the picturing of data. *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, 2:523–531, 1975.

[138] Remco Van Der Hofstad. *Random Graphs and Complex Networks*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2016. doi: 10.1017/9781316779422.

[139] Stephan Wagner and Kevin Durant. On the centroid of increasing trees. *Discrete Mathematics & Theoretical Computer Science*, 21, 2019.

[140] LJ Wei. The generalized Pólya's urn design for sequential medical trials. *The Annals of Statistics*, 7(2):291–296, 1979.

[141] Li-Xin Zhang. Convergence of randomized urn models with irreducible and reducible replacement policy. *arXiv preprint arXiv:2204.04810*, 2022.

[142] Yijun Zuo. A new approach for the computation of halfspace depth in high dimensions. *Communications in Statistics - Simulation and Computation*, 48(3):900–921, 2019.

[143] Yijun Zuo and Robert Serfling. General notions of statistical depth function. *Annals of Statistics*, pages 461–482, 2000.

[144] Yijun Zuo and Robert Serfling. On the performance of some robust nonparametric location measures relative to a general notion of multivariate symmetry. *Journal of Statistical Planning and Inference*, 84(1-2):55–79, 2000.