Instituto Nacional de Matemática Pura e Aplicada

Topics in discrete probability: Analysis of the past and the future

Alan Pereira

supervised by
Prof. Roberto Oliveira and Prof. Gábor Lugosi

Rio de Janeiro, 2018

Abstract

This thesis is concerned with problems about discrete structures that change with time. We study two problems in discrete models: random trees with uniform attachment and concentration in random partitions.

The first problem is about *Archaeology of Random Tree with Uniform Attachment*. Given a initial tree (seed tree) let us attach at each time a new vertex to a randomly chosen vertex of the current tree. If we let time grow until a big value n, and look the tree after this process, can we decide which vertices were in the seed tree? This is analysed for three possible seed trees: a path, a star and a random tree. Techniques of Polya Urns and concentration inequalities were the main ingredients of the solution of these problems.

The second problem is about *Generalized Chinese Restaurant Process*: suppose we have a restaurant with infinitely many tables. At each time a new costumer enters in the restaurant and sits a table. The probability of choosing some previously occupied table depends on the number of costumers at the table and of some parameters α and θ, and the probability of choose a new table is the complementary probability. We study (in terms of the parameters α and θ), the growth of the number of occupied tables and the number of tables with k costumers.We showed that the the normalized number of occupied tables and the number of table with a fixed number of costumers concentrates near a convenient random variable.

Acknowledgements

A thesis is the expression of a academic history of an individual. Many bricks and many shoulders. There is no self-made man, and I, particularly, am so far from being one.

First of all, thank for my mother Socorro, my father Francisco, my grandmother Lourdes and my grandfather Manoel, who always helped me and invested in my education and for my girlfriend Ayane and my friend Artur for their support. Mãe, pai, vô, vó, Ayá, Tutu muito obrigado por tudo, se cheguei aqui é porque vocês estiveram ao meu lado quando mais precisei.

It is obvious I would thank my advisor, Roberto Imbuzeiro (Oliveira) for the Mathematics, but once everybody can deduce this theorem, I will state another. Thank you Roberto for all the conversations about patience and about hope. Half of a thesis is mathematics, the other half is a psychological challenge, thank you for the advice and support you gave me in both.

I would like to thank to my co-advisor, Gábor Lugosi, for his receptiveness, enthusiasm and inspiration. Gábor is the personification of the sentence "focus on solution, not in the problem". Thank you Gábor, for the patience and the support you gave me during the time in Barcelona.

I would like to thank Rodrigo Ribeiro, with whom I worked in the second problem in this thesis, for the mathematical support, goodwill and readiness to talk about the life as a young mathematician.

I thank Luc Devroye, Miklós Rácz, and Tommy Reddad for their suggestion on the topic of chapter 2. I thank Anna Ben-Hamou for important reading suggestions on the topic of the chapter 3.

To André Contiero, Carlos Argolo, Davi Lima, Diogo Santos, Elivaldo Bezerra and Krerley Oliveira for the very significant help building the path from primary school to the Masters Program. Thanks to Adriana Sanchez, Cayo Dória, David Evangelista, Daniel Marroquin, Luiz Paulo Moreira, Mateus Souza, Maurício Collares and professor Augusto Teixeira for the very significant help building the path from the Masters until the end of PhD. Thank to professors Milton Jara and Robert Morris for the inspiration.

I would like to thank to CNPq and the Brazilian people for funding this research work.

Thank to all the people who directly or indirectly helped the construction of this journey, since my childhood until now.

Contents

Chapter 1

Introduction

Dynamically growing discrete processes represent complex relationships in numerous areas of science. In many interesting applications, one does not observe the entire dynamical growth procedure but merely a present-day snapshot of the discrete process is available for observation. Based on this snapshot, one wishes to infer various properties of the *past* and *future* of the process. We will consider two problems here, the first about the past of a process and the second about the future of another process.

Chapter 2:

In this chapter we investigate a problem in network archaeology. The simplest dynamically grown networks are trees that are grown by attaching vertices sequentially to the existing tree at random, according to a certain rule. In the *uniform attachment* model, at each step, an existing vertex is selected uniformly ar random, and a new vertex is attached to it by an edge. When the process is initialized from a single vertex, this procedure gives rise to the well-studied *uniform random recursive tree*, see Drmota [13]. In *preferential attachment* models (such as plane-oriented recursive trees) existing vertices with higher degrees are more likely to be chosen to be attached to. In this text we consider randomly growing uniform attachment trees that are grown from a fixed *seed*. Thus, initially, the tree is a given fixed (small) tree and further vertices are attached according to the uniform attachment process.

"Archeology" of randomly growing trees has received increasing attention recently, see Brautbar and Kearns [3], Borgs, Brautbar, Chayes, Khanna, and Lucier [2], Bubeck, Devroye, and Lugosi [5], Bubeck, Mossel, and Rácz [7], Bubeck, Eldan Mossel, and Rácz [6], Curien, Duquesne, Kortchemski, and Manolescu [11], Frieze and Pegden [18], Jog and Loh [22,23], Shah and Zaman [28,29] for a sample of the growing literature.

Several papers consider the problem of finding the initial vertex (or root) in a randomly growing tree started from a single vertex, see Brautbar and Kearns [3], Borgs, Brautbar, Chayes, Khanna, and Lucier [2], Frieze and Pegden [18], Shah and Zaman [28,29], Bubeck, Devroye, and Lugosi [5], Jog and Loh [22,23] for various models. Randomly growing trees started from an initial seed tree were considered by Bubeck, Mossel, and Rácz [7], Bubeck, Eldan Mossel, and Rácz [6], and Curien, Duquesne, Kortchemski, and Manolescu [11]. These papers prove that in uniform and preferential attachment models, for any pair of possible seed trees, one may construct a hypothesis test that decides which of the two seeds generated the observed tree, with a probability of error strictly smaller than 1/2, regardless of the size of the observed tree.

In this text we consider the problem of *finding* the seed tree (of known structure) in a large observed tree. This work was made joint with Gábor Lugosi, and the paper can be found in `https://arxiv.org/pdf/1801.01816.pdf`.

The questions we seek to answer are: (1) to what extent is it possible to identify the seed tree? (2) what is the role of the structure of the seed in the difficulty of the reconstruction problem? While we are far from completely answering these questions, this text contributes to the understanding of these problems. In particular, we consider three types of possible seed trees,

namely paths, stars, and random uniform recursive trees. For each of these examples, we present algorithms to recover, at least partially, the seed tree. In all cases, partial recovery is possible, with any prescribed probability of error, regardless of the size of the observed tree. However, the difficulty of the recovery depends heavily on the structure of the tree. Paths and stars are considerably easier to find than uniform random recursive trees.

Chapter 3:

In this chapter we talk about a model of random partitions. This type of model have attracted much attention in Probability and Statistics. Here we study the specific family of models of random partitions called *generalized Chinese Restaurant processes (GCRP)*. These models were introduced by Pitman [25], [26] as two-parameter generaliation of Ewens' sampling formula [14]. They are also important building blocks in topic models [19] and other Bayesian nonparametric methods [9].

The GRCP generates a sequence of random partitions $\mathcal{P}_n$ of $[n] := \{1, \ldots, n\}$ for $n = 1, 2, 3, \ldots$. We focus on a specific setting for the model where the number of parts in $\mathcal{P}_n$ grows like n^α for a parameter $\alpha \in (0, 1)$. Our main goal is to prove concentration for the total number of the number of parts with size k in each $\mathcal{P}_n$, that is:

$$N_n(k) := |\{A \in \mathcal{P}_n \,:\, |A| = k\}|.$$

As we explain below, the $\mathcal{P}_n$ are *mixtures of i.i.d. models*, and the above random variables do not concentrate around any fixed value. Nevertheless, we show that they do concentrate around random values. Our main result – Theorem 3.2.2 below – shows that, for large n, with high probability,

$$N_n(k) = c\, V_* \, \frac{\Gamma(k-\alpha)}{\Gamma(k+1)}\, n^\alpha + o\left(\frac{\Gamma(k-\alpha)}{\Gamma(k+1)}\, n^\alpha\right)$$

where V_* is a random variable with $V_* > 0$ a.s. and and $c > 0$ is a constant depending on model parameters. This result holds simultaneously for all k in a range that grows polynomially in k. Since

$$\Gamma(k-\alpha)/\Gamma(k+1) = \Theta(k^{-(1+\alpha)}) \text{ for large } k,$$

we verify that the power-law-type behavior in k that is known to hold asymptotically for the $N_n(k)$ is already visible for finite n. Moreover, in our proof we also obtain finite-n bounds on the number of parts in $\mathcal{P}_n$ (cf. Theorem 3.2.1 below).

Our proof method is based on martingale inequalities and is inspired by the analysis of preferential-attachment-type models [8]. However, there are some important technical differences, which we discuss in subsection 3.2.2. A salient feature of our approach is that the concentration-of-measure arguments we employ are fairly delicate, and rely on Freedman's concentration inequality [17].

Chapter 4

The main technical tools in this text are concentration inequalities for martingales and inequalities on gamma functios. We recall the inequalities by Mc Diarmid and by Freedman in this chapter. We also provide several technical estimates on Γ functions we use in chapter 3.

Chapter 2

Archaeology in Random Growing Trees

The results in this chapter are from a paper made jointly with Gábor Lugosi, which can be found in `https://arxiv.org/pdf/1801.01816.pdf`.

In Section 2.1 we introduce the mathematical model and state the main results. The proofs of all results are presented in Section 2.2.

2.1 Setup and results

Let $\ell \geq 1$ be a positive integer and let S_ℓ be a tree (i.e., a connected acyclic graph) on the vertex set $\{1, \ldots, \ell\}$. Let $n > \ell$ be another positive integer. We say that a random tree T_n on the vertex set $\{1, \ldots, n\}$ is a *uniform attachment tree with seed* S_ℓ if it is generated as follows:

1. $T_\ell = S_\ell$;
2. For $\ell < i \leq n$, T_i is obtained from T_{i-1} by joining vertex i to a vertex of T_{i-1} chosen uniformly at random, independently of all previous choices.

The problem we study in this chapter is the following. Suppose one observes a tree T_n generated by the uniform attachment process with seed S_ℓ but with the vertex labels hidden. The goal is to find the seed tree S_ℓ in the observed unlabeled tree. More precisely, given a target accuracy $\epsilon \in (0, 1)$ a seed-finding algorithm of *first kind* outputs a set $H_1(T_n, \epsilon)$ of vertices of size $k_\ell \leq \ell$, such that, with probability at least $1 - \epsilon$, $H_1(T_n, \epsilon) \subset S_\ell$, that is, all elements of $H_1(T_n, \epsilon)$ are vertices of the seed tree S_ℓ. (Here, with a slight abuse of notation, we identify the seed S_ℓ with its vertex set $\{1, \ldots, \ell\}$.)

Similarly, a seed-finding algorithm of *second kind* outputs a set $H_2(T_n, \epsilon)$ of vertices of size $k_\ell \geq \ell$, such that, with probability at least $1 - \epsilon$, $S_\ell \subset H_2(T_n, \epsilon)$, that is, $H_2(T_n, \epsilon)$ contains all vertices of the seed tree S_ℓ.

In both cases, one would like to have k_ℓ as close to ℓ as possible, even for small values of ϵ.

Bubeck, Devroye, and Lugosi [5] considered the case $\ell = 1$, that is, when the seed tree is a single vertex and seed-finding algorithms of the second kind. Thus, the aim of the seed-finding algorithm is to find the root of the observed tree. Their main finding is that, for all ϵ, the optimal value of k_1 stays bounded as the size n of the observed tree goes to infinity. They also show that there exist seed-finding algorithms of the second kind such that $k_1 = o(\epsilon^{-a})$ for all $a > 0$.

In this text we show that, if ℓ is sufficiently large (depending on ϵ), then k_ℓ may be made *proportional* to ℓ for seed-finding algorithms of second kind, and we make similar statements for k_ℓ for certain seed-finding algorithms of first kind. How the required value of ℓ depends on ϵ and what the achievable proportions are depend heavily on the structure of the seed. We consider three prototypical examples of seeds:

• A *path* P_ℓ on ℓ vertices is a tree that has exactly two vertices of degree one and $\ell - 2$ vertices of degree two.

• A *star* E_ℓ on ℓ vertices is a tree that has $\ell-1$ vertices of degree one and one vertex of degree $\ell-1$.

• The third example we consider is when the seed S_ℓ is a uniform random recursive tree on ℓ vertices. In this case the proposed seed finding algorithm does not need to know the structure of the tree. Thus, this example may be considered as a generalization of the root-finding problem studied in [5]. Here, instead of trying to locate the root of the tree, the goal is to find the first ℓ generations of the observed uniform random recursive tree T_n.

In what follows we present the main findings of the thesis that establish the existence of seed-finding algorithms that are able to recover a constant fraction of the seed if it is a uniform random recursive tree. If the seed is either a path or a star, then the situation is even better as one can recover almost the entire seed.

Importantly, all bounds established below are independent of the size n of the observed tree, meaning that (partial) reconstruction of the seed is possible regardless of how large the observed tree T_n is.

2.1.1 Finding the seed when it is a path

We begin with the case when the seed is a path:

Theorem 2.1.1. *Let $\epsilon \in (0,1)$ and $\gamma \in (0,1)$ and let $\ell \geq \max\left\{\frac{2e^2}{\gamma}\log\frac{1}{\epsilon}, \frac{2e^2}{\gamma}\log(4e^2)\right\}$ be a positive integer. Then for all $n \geq \ell$ sufficiently large, if T_n is a uniform attachment tree with seed $S_\ell = P_\ell$ (a path of ℓ vertices), then there exists a seed-finding algorithm that outputs a vertex set $H_n \subset \{1,\dots,n\}$ with $|H_n| \geq (1-\gamma)\ell$ such that*

$$\mathbb{P}\{H_n \subset P_\ell\} \geq 1-\epsilon\ .$$

The theorem states that, for any fixed $\gamma > 0$, if the size of the seed path ℓ is at least of the order of $\log(1/\epsilon)$, then there exists an algorithm that finds all but a γ-fraction of the seed path, regardless of how large the observed tree T_n is. Note that the required length of the path is merely logarithmic in $1/\epsilon$. In fact, this dependence is essentially best possible. The following result shows that if the seed path has less than $\frac{\log(1/\epsilon)}{\log\log(1/\epsilon)}$ vertices, then *any* seed finding algorithm must miss at least half of the seed, with probability greater than ϵ.

Theorem 2.1.2. *Let $\epsilon \in (0, e^{-e^2})$. Suppose that T_n is a uniform attachment tree with seed $S_\ell = P_\ell$ for $\ell \leq \frac{\log(1/\epsilon)}{\log\log(1/\epsilon)}$. Then, for all $n \geq 2\ell$, any seed-finding algorithm that outputs a vertex set H_n of size ℓ has*

$$\mathbb{P}\left\{|H_n \cap P_\ell| \leq \frac{\ell}{2}\right\} \geq \epsilon\ .$$

2.1.2 Finding the seed when it is a star

Next we state our results for the case when the seed tree is a star E_ℓ on ℓ vertices.

Theorem 2.1.3. *There exists a numerical positive constant C such that the following holds. Let $\epsilon \in (0,1)$ and $\gamma \in (0,1)$ and let $\ell \geq \max(C, 8/\gamma)\log(1/\epsilon)$ be a positive integer. Then for all $n \geq \ell$ sufficiently large, if T_n is a uniform attachment tree with seed $S_\ell = E_\ell$ (a star of ℓ vertices), then there exists a seed-finding algorithm that outputs a vertex set $H_n \subset \{1,\dots,n\}$ with $|H_n| \leq (1+\gamma)\ell$ such that*

$$\mathbb{P}\{E_\ell \subset H_n\} \geq 1-\epsilon\ .$$

Once again, the order of magnitude for the required size of the seed star is essentially optimal as a function of ϵ. The proof of the next theorem is similar to that of Theorem 2.1.2 and thus it is omitted.

Theorem 2.1.4. *Let $\epsilon \in (0, e^{-e^2})$. Suppose that T_n is a uniform attachment tree with seed $S_\ell = E_\ell$ for $\ell \leq \frac{\log(1/\epsilon)}{\log\log(1/\epsilon)}$. Then, for all $n \geq 2\ell$, any seed-finding algorithm that outputs a vertex set H_n of size ℓ has*

$$\mathbb{P}\left\{|H_n \cap E_\ell| \leq \frac{\ell}{2}\right\} \geq \epsilon \ .$$

2.1.3 Finding the first generations

Finally, we consider the case when the seed tree is a uniform random recursive tree in ℓ vertices. Unlike in the previous two examples, here the seed finding algorithm does now "know" the exact structure of the seed. This model may be equivalently formulated as follows: starting from a single vertex, one grows a uniform random recursive tree T_n of n vertices. Upon observing T_n (without vertex labels), one's aim is to recover as much of the tree T_ℓ (containing vertices attached in the first ℓ generations) as possible. The next theorem establishes the existence of a seed-finding algorithm of the first kind that identifies an $\Omega(1/\log(1/\epsilon))$ fraction of the vertices of the seed T_ℓ with probability at least $1-\epsilon$, whenever ℓ is at least proportional to $\log^3(1/\epsilon)$. One should note that this result is weaker than the one obtained for seed paths and seed stars above in various ways. First, unlike in the cases of Theorems 2.1.1 and 2.1.3, here we cannot guarantee that almost all of the seed tree is identified, but only a fraction of it whose size depends on ϵ–although in a mild manner. Second, the size of the seed tree needs to be somewhat larger as a function of ϵ as before. While in the previous cases ℓ needed to be logarithmic in $1/\epsilon$, now it needs to scale as $\log^3(1/\epsilon)$. Below we show that to some extent these weaker results are inevitable and that finding the seed tree T_ℓ is inherently harder than finding more structured seed trees such as stars and paths.

Our main positive result is as follows.

Theorem 2.1.5. *Let T_n be a uniform random recursive tree on n vertices and let $\epsilon > 0$ and $\ell \geq 1$. Let $a = 2\log(4\ell^2/\epsilon) + 1$. If ℓ is so large that*

$$\ell \geq 64a^2 \log(22a\ell^2/\epsilon) \ ,$$

then there exists a seed-finding algorithm that outputs a vertex set $H_n \subset \{1, \ldots, n\}$ with $|H_n| \geq \ell/(3a)$ such that

$$\liminf_{n\to\infty} \mathbb{P}\{H_n \subset T_\ell\} \geq 1 - \epsilon \ .$$

Note that the condition for ℓ is satisfied for $\ell \geq C\log^2(1/\epsilon)$ for a constant C.

Next we show that, regardless how large ℓ is, for n sufficiently large any seed-finding algorithm of first kind needs to output a set of vertices whose size is at most $c\ell$ where c is strictly smaller than 1. Similarly, any seed-finding algorithm of second kind needs to output a set of vertices whose size is at least $C\ell$ where $C > 1$.

In other words, when the seed tree is a uniform random recursive tree, the problem of finding it is strictly harder than finding a seed path or a seed star in the sense that no algorithm can have a performance as the one established in Theorem 2.1.1 or Theorem 2.1.3. Note however, that there remains a gap between the performance bound of Theorem 2.1.5 and the impossibility bound of Theorem 2.1.6 below, as the size of the vertex set in the seed found by the algorithm of Theorem 2.1.5 is only guaranteed to be of the order of $\ell/\log(1/\epsilon)$, a linear fraction but depending on ϵ.

The impossibility results mentioned above follow from the fact that, at time 2ℓ, a linear fraction of the vertices of the seed T_ℓ become indistinguishable from vertices that arrive between time $\ell+1$ and 2ℓ. To make the statement precise, we need a few definitions.

In a uniform random recursive tree T_ℓ, we call a vertex a *singleton* if it is a leaf and it is the only descendant of its parent vertex.

Now consider a vertex v in T_ℓ and its position in the tree $T_{2\ell}$. We say that v is a *camouflaging* vertex if

1. In T_ℓ, v is a parent of a singleton d;

2. Between time $\ell+1$ and 2ℓ a vertex w is attached to v such that w is a leaf of $T_{2\ell}$

3. d is a leaf of $T_{2\ell}$.

Clearly, at time 2ℓ, and therefore at any time $n \geq 2\ell$, the two descendants d and w of any camouflaging vertex v are indistinguishable. Let G_ℓ denote the number of camouflaging vertices. Then if a seed-finding algorithm outputs a vertex set that contains an $(1-\gamma)\ell$ vertices of the seed, then one must have $G_\ell < \gamma\ell$. The next proposition shows that $\gamma \geq 1/384$ with high probability.

Theorem 2.1.6. *For any $\ell \geq 1$,*

$$\mathbb{E}G_\ell \geq \frac{\ell}{384}$$

and for any $t \geq 0$,

$$\mathbb{P}\left\{G_\ell \leq \frac{\ell}{384} - t\right\} \leq e^{\frac{-t^2}{2\ell}} \ .$$

2.2 Proofs

In this section we present the proofs of all theorems. The construction of all seed-finding algorithms uses a simple notion of centrality that we recall first.

2.2.1 Centrality

Let T be a tree with vertex set $V(T)$. A *rooted tree* (T, v) is the tree T with a distinguished vertex $v \in V(T)$. For a vertex $u \in V(T)$, denote by $(T, v)_{u\downarrow}$ the rooted subtree of T whose root is u and whose vertex set contains all vertices w of $V(T)$ such that the (unique) path connecting w and v in T contains u.

Given tree T, the *anti-centrality* of a vertex $v \in V(T)$ is defined by

$$\psi(v) = \max_{u \in V(T)\setminus\{v\}} |(T, v)_{u\downarrow}| \ .$$

Thus, $\psi(v)$ is the size of the largest subtree of the tree T rooted at v. Note that leaves of a tree T have the largest anti-centrality with $\psi(v) = |V(T)| - 1$. We say that v is *at least as central as* w if $\psi(v) \leq \psi(w)$.

For a positive integer k, we denote by $H_\psi(k)$ the set of k vertices of with smallest anti-centrality, where ties may be broken arbitrarily.

This notion of centrality played a crucial role in some of the root-finding algorithms of [5]. We refer to Jog and Loh [22, 23] for a study of this notion in various random tree models, including uniform random recursive trees.

2.2.2 Proof of Theorem 2.1.1

Let ϵ, γ, and ℓ be as in the assumptions of the theorem. We may assume, without loss of generality, that $\gamma\ell/2$ is an integer. We analyze a simple seed-finding algorithm that achieves the performance stated in the theorem. The proposed algorithm simply takes the $(1-\gamma)\ell$ most central vertices, as measured by the function ψ defined in Section 2.2.1.

Formally, let $k_\ell = (1-\gamma)\ell$ and define $H_n = H_\psi(k_\ell)$ be the set of k_ℓ most central vertices of the observed tree T_n.

It suffices to prove that, for all sufficiently large n, with probability at least $1-\epsilon$, all vertices of T_n not in the seed P_ℓ are less central than any vertex in P_ℓ whose distance to the leaves of P_ℓ is at least $\gamma\ell/2$, that is,

$$\mathbb{P}\left\{\min_{\ell < i \leq n} \psi(i) > \max_{\ell\gamma/2 \leq j \leq \ell(1-\gamma/2)} \psi(j)\right\} \geq 1-\epsilon \ . \tag{2.2.1}$$

(Recall that the vertex set of the seed P_ℓ is $\{1,\ldots,\ell\}$.)

Let $C_1,\ldots,C_\ell$ denote the components of the forest obtained by removing the edges of P_ℓ from T_n such that $k \in C_k$ for $k = 1,\ldots,\ell$. Then

$$\begin{aligned}\mathbb{P}\left\{\min_{\ell<i\le n}\psi(i) \le \max_{\ell\gamma/2\le j\le \ell(1-\gamma/2)}\psi(j)\right\} &\le \sum_{j=\gamma\ell/2}^{(1-\gamma/2)\ell}\mathbb{P}\left\{\min_{\ell<i\le n}\psi(i)\le\psi(j)\right\}\\ &\le \sum_{j=\gamma\ell/2}^{(1-\gamma/2)\ell}\sum_{k=1}^{\ell}\mathbb{P}\left\{\exists v\in C_k\setminus\{k\}:\psi(v)\le\psi(j)\right\}\ .\end{aligned}$$

To bound the probabilities on the right-hand side, suppose, without loss of generality, that $k \le j$. (The case $k > j$ is analogous.) If $v \in C_k\setminus\{k\}$ is such that $\psi(v) \le \psi(j)$. Let u be a vertex connected to v such that $|(T,v)_{u\downarrow}|$ is maximal (i.e., $\psi(v) = |(T,v)_{u\downarrow}|$). Then there are two possibilities:

(a) $(T,v)_{u\downarrow}$ is contained in C_k. In this case $|C_k| \ge \sum_{i\ne k}|C_i|$;

(b) $(T,v)_{u\downarrow} = \left(\bigcup_{i=1,i\ne k}^{\ell} C_i\right)\cup C_k'$ for some $C_k'\subset C_k$. In this case

$$\left|\bigcup_{i\ne k}C_i\right| \le \psi(v)\le\psi(j)\le\left|\bigcup_{i=1}^{j}C_i\right|$$

which implies $\sum_{i=j+1}^{\ell}|C_i|\le|C_k|$.

By this observation, we have

$$\begin{aligned}\mathbb{P}\left\{\exists v\in C_k\setminus\{k\}:\psi(v)\le\psi(j)\right\} &\le \mathbb{P}\left\{|C_k|\ge\sum_{i\ne k}|C_i|\right\}+\mathbb{P}\left\{\sum_{i=j+1}^{\ell}|C_i|\le|C_k|\right\}\\ &\le \mathbb{P}\left\{|C_k|\ge\sum_{i\ne k}|C_i|\right\}+\mathbb{P}\left\{\sum_{i=(1-\gamma/2)\ell}^{\ell}|C_i|\le|C_k|\right\}\end{aligned}$$

Now let $t = \gamma/e^2$. Then the right-hand side of the inequality above may be bounded further by

$$\mathbb{P}\left\{\sum_{i=1,i\ne k}^{\ell}|C_i|\le nt\right\}+\mathbb{P}\left\{\sum_{i=1}^{\gamma\ell}|C_i|\le nt\right\}+2\mathbb{P}\left\{|C_k|\ge nt\right\}$$

Thus, we have

$$\begin{aligned}&\mathbb{P}\left\{\min_{\ell<i\le n}\psi(i)\le\max_{\ell\gamma/2\le j\le\ell(1-\gamma/2)}\psi(j)\right\}\\ &\le (1-\gamma)\ell^2\left(\mathbb{P}\left\{\sum_{i=1,i\ne k}^{\ell}|C_i|\le nt\right\}+\mathbb{P}\left\{\sum_{i=1}^{\gamma\ell}|C_i|\le nt\right\}+2\mathbb{P}\left\{|C_k|\ge nt\right\}\right)\end{aligned}$$

To understand the behavior of the probabilities on the right-hand side, note that, for any $k = 1,\ldots,\ell-1$, $\sum_{i=1}^{k}|C_i|$ is just the number of red balls after taking n samples in a standard Pólya urn initialized with k red and $\ell-k$ blue balls. This implies that $\sum_{i=1}^{k}|C_i|/n$ converges, in distribution, to a Beta$(k,\ell-k)$ random variable. Hence,

$$\lim_{n\to\infty}\mathbb{P}\left\{|C_k|/n\ge t\right\} = (1-t)^{\ell-1}\le e^{-t(\ell-1)}$$

and

$$\lim_{n\to\infty} \mathbb{P}\left\{ \sum_{i=1, i\neq k}^{\ell} |C_i|/n \leq t \right\} \leq \lim_{n\to\infty} \mathbb{P}\left\{ \sum_{i=1}^{\gamma\ell} |C_i|/n \leq t \right\}$$
$$= (\ell-1)\binom{\ell-1}{\gamma\ell-1} \int_0^t x^{\gamma\ell-1}(1-x)^{\ell-\gamma\ell-1} dx \ .$$

We may bound the expression on the right-hand side by

$$\frac{\ell^{\gamma\ell}}{(\gamma\ell-1)!} \int_0^t x^{\gamma\ell-1} dx = \frac{(t\ell)^{\gamma\ell}}{(\gamma\ell)!} \leq \left(\frac{e\ell t}{\gamma\ell}\right)^{\gamma\ell} \leq e^{-\gamma\ell} \ ,$$

where we used Stirling's formula and the choice $t = \gamma/e^2$. Putting everything together, we have that

$$\limsup_{n\to\infty} \mathbb{P}\left\{ \min_{\ell<i\leq n} \psi(i) \leq \max_{\ell\gamma/2\leq j\leq \ell(1-\gamma/2)} \psi(j) \right\} \leq 2\ell^2 \left(e^{-\gamma\ell} + e^{-\gamma(\ell-1)/e^2} \right) \leq \epsilon$$

under our conditions for ℓ, as desired. □

2.2.3 Proof of Theorem 2.1.2

Let E be the event that either (1) vertex i attaches to vertex $i-1$ for all $i = \ell+1, \ldots, 2\ell$ or (2) vertex $\ell+1$ attaches to vertex 1 and for all $i = \ell+2, \ldots, 2\ell$, vertex i attaches to vertex $i-1$. On this event, $T_{2\ell}$ is a path of 2ℓ vertices such that the seed P_ℓ is on one of the two extremes of $T_{2\ell}$. The probability of this event is

$$\frac{2}{\ell} \cdot \frac{1}{\ell+1} \cdot \cdots \cdot \frac{1}{2\ell-1} \geq 2\frac{\ell!}{(2\ell)!} \geq 2(2\ell)^{-\ell} \ .$$

On this event, for $n \geq 2\ell$, for any seed-finding algorithm, the first and second halves of the path $T_{2\ell}$ are indistinguishable. At least one of the two halves of $T_{2\ell}$ is such that H_n intersects that half in at most $\ell/2$ vertices. Thus, (conditionally on E), the algorithm misses at least half of the seed path, with probability 1/2. Hence

$$\mathbb{P}\left\{ |H_n \cap P_\ell| \leq \frac{\ell}{2} \right\} \geq \frac{\mathbb{P}\{E\}}{2} \geq (2\ell)^{-\ell} \geq \epsilon$$

whenever $\ell \leq \frac{\log(1/\epsilon)}{\log\log(1/\epsilon)}$ and $\epsilon \leq e^{-e^2}$.

2.2.4 Proof of Theorem 2.1.3

Let $k_\ell = (1+\gamma)\ell$. Again, we may assume that k_ℓ is an integer. The seed finding algorithm we propose is slightly different. It is specifically tailored to the case when the seed tree to be found is a star. Let $v_n^* = \text{argmin}_{i=1,\ldots,n} \psi(i)$ be the most central vertex of T_n. We define H_n as the set of vertices that includes v_n^* and $k_\ell - 1$ other vertices j with largest value of $|(T_n, v_n^*)_{j\downarrow}|$ among the neighbors of v_n^* in T_n. In other words, the algorithm outputs the most central vertex v_n^* and those neighbors whose subtree away from v_n^* is largest.

First we recall that by Jog and Loh [23, Theorem 4], there exists a numerical constant C such that, if $\ell \geq C\log(1/\epsilon)$ and the uniform attachment tree is initialized with a star E_ℓ as seed of ℓ vertices and central vertex 1, then

$$\mathbb{P}\{v_n^* = 1 \text{ for all } n = \ell+1, \ell+2, \ldots\} \geq 1 - \frac{\epsilon}{2} \ ,$$

that is, with probability at least $1-\epsilon/2$, the center of the seed star remains the most central vertex of T_n for all n.

Let $v_1 \le v_2 \le \cdots$ be the vertices that are attached to vertex 1 (i.e., to the center of the seed star E_ℓ) in the uniform attachment process. (Thus, $v_1 > \ell$.) In view of the above-mentioned result of Jog and Loh, it suffices to show that for all n sufficiently large, all vertices v_j with $j > \gamma\ell$ have $\left|(T_n, 1)_{v_j\downarrow}\right|$ smaller than $|(T_n, 1)_{i\downarrow}|$ for all vertices i in the seed star E_ℓ, with probability at least $1 - \epsilon/2$. Thus, writing $g(i) = |(T_n, 1)_{i\downarrow}|$, we need to prove that

$$\limsup_{n\to\infty} \mathbb{P}\left\{\max_{j>\gamma\ell} g(v_j) < \min_{i=2,\ldots,\ell} g(i)\right\} > 1 - \frac{\epsilon}{2} \ . \tag{2.2.2}$$

To prove (2.2.2), first we write

$$\mathbb{P}\left\{\max_{j>\gamma\ell} g(v_j) \ge \min_{i=2,\ldots,\ell} g(i)\right\} \le \mathbb{P}\left\{v_{\gamma\ell+1} \le m\right\} + \mathbb{P}\left\{\max_{v_j>m} g(v_j) \ge \min_{i=2,\ldots,\ell} g(i)\right\} \ , \tag{2.2.3}$$

where we take $m = \lfloor e^{\gamma\ell/4} \rfloor$. The first term on the right-hand side is the probability that more than $\gamma\ell$ vertices are attached to vertex 1 up to time m. In order to bound this probability, denote by X_t, for $t \ge \ell$, the number of vertices attached to vertex 1 between time $\ell + 1$ and t. Thus, $X_\ell = 0$ and

$$\mathbb{P}\left\{v_{\gamma\ell+1} \le m\right\} = \mathbb{P}\left\{X_m > \gamma\ell\right\} \ .$$

Since

$$\mathbb{E}[X_t | X_{t-1}] = X_{t-1} + \frac{1}{t} \ ,$$

$$Y_t = X_t - \sum_{k=\ell+1}^{t} \frac{1}{k}, \quad t \ge \ell + 1$$

is a martingale with respect to the filtration generated by $X_\ell, X_{\ell+1}, \ldots$. Denote the corresponding martingale difference sequence by $Z_t = Y_t - Y_{t-1} = X_t - X_{t-1} - 1/t$. By Markov's inequality,

$$\mathbb{P}\left\{X_m > \gamma\ell\right\} = \mathbb{P}\left\{\sum_{j=\ell+1}^{m} Z_j + \sum_{j=\ell+1}^{m} \frac{1}{j} > \gamma\ell\right\} \le \frac{e^{\sum_{j=\ell+1}^{m} \frac{1}{j}} \cdot \mathbb{E}\left[e^{\sum_{j=\ell+1}^{m} Z_j}\right]}{e^{\gamma\ell}} \ . \tag{2.2.4}$$

In order to bound the right-hand side, observe that

$$\begin{aligned}
\mathbb{E}\left[e^{Z_m} | X_\ell, \ldots, X_{m-1}\right] &= \mathbb{E}\left[e^{X_m - X_{m-1} - \frac{1}{m}} |\ X_\ell, \ldots, X_{m-1}\right] \\
&= e^{-X_{m-1} - \frac{1}{m}} \mathbb{E}\left[e^{X_m} |\ X_\ell, \ldots, X_{m-1}\right] \\
&= e^{-X_{m-1} - \frac{1}{m}} \left(\frac{1}{m} e^{X_{m-1}+1} + \frac{(m-1)}{m} e^{X_{m-1}}\right) \\
&= \frac{e^{-\frac{1}{m}}}{m}(e + m - 1) \\
&\le \frac{(m+2)e^{-\frac{1}{m}}}{m} \ ,
\end{aligned}$$

and therefore

$$\begin{aligned}
\mathbb{E}\left[e^{\sum_{j=\ell+1}^{m} Z_j}\right] &= \mathbb{E}\left[\mathbb{E}\left[e^{\sum_{j=\ell+1}^{m} Z_j} | X_\ell, \ldots, X_{m-1}\right]\right] \\
&= \mathbb{E}\left[e^{\sum_{j=\ell+1}^{m-1} Z_j} \mathbb{E}\left[e^{Z_m} | X_\ell, \ldots, X_{m-1}\right]\right] \\
&\le \frac{(m+2)e^{-\frac{1}{m}}}{m} \mathbb{E}\left[e^{\sum_{j=\ell+1}^{m-1} Z_j}\right] \ .
\end{aligned}$$

Thus, by induction we obtain

$$\mathbb{E}\left[e^{\sum_{j=\ell+1}^{m} Z_j}\right] \le \frac{(m+2)^2}{\ell^2} e^{-\sum_{j=\ell+1}^{m} \frac{1}{j}}.$$

Substituting into (2.2.4), we get

$$\mathbb{P}\{v_{\gamma\ell+1} \le m\} = \mathbb{P}\{X_m > \gamma\ell\} \le \frac{(m+2)^2}{\ell^2 e^{\gamma\ell}} \le \frac{\epsilon}{4}$$

by our choice of m and by the condition on the value of ℓ. Hence, by (2.2.3), it suffices to show that

$$\mathbb{P}\left\{\max_{v_j>m} g(v_j) \ge \min_{i=2,\dots,\ell} g(i)\right\} \le \frac{\epsilon}{4}\ .$$

We proceed by writing

$$\mathbb{P}\left\{\max_{v_j>m} g(v_j) \ge \min_{i=2,\dots,\ell} g(i)\right\} \le \sum_{i=2}^{\ell} \mathbb{P}\left\{\max_{v_j>m} g(v_j) \ge g(i)\right\}\ .$$

Now fix $i \in \{2,\dots,\ell\}$ and notice that $\max_{v_j>m} g(j)$ is bounded by the number of vertices A attached to the tree formed by vertex 1 and all vertices in the subtrees $(T_n,1)_{j\downarrow}$ for $j > m$ such that vertex j is attached to vertex 1.

Denoting $B = g(i)$ and $C = n - A - B$, note that, conditioned on the tree T_m, the triple (A,B,C) behaves as the number of red, blue, and white balls in a Pólya urn in which initially (i.e., at time m) there is one red ball, $B_m = |(T_m,1)_{i\downarrow}|$ blue balls, and $m - 1 - |(T_m,1)_{i\downarrow}|$ white balls. Hence, for each $i = 2,\dots,\ell$, we have

$$\begin{aligned}\mathbb{P}\left\{\max_{v_j>m} g(v_j) \ge g(i)\right\} &\le \mathbb{P}\{A > B\} \\ &\le \mathbb{P}\left\{A > B | B_m \ge \frac{m\epsilon}{32\ell^2}\right\} + \mathbb{P}\left\{B_m < \frac{m\epsilon}{32\ell^2}\right\}\ .\end{aligned}$$

In order to bound the second term on the right-hand side, note that by the standard theory of Pólya urns, B_m has a beta-binomial distribution with parameters $(m,1,\ell-1)$. Thus, B_m is distributed as a binomial random variable $\mathrm{Bin}(m,\pi)$ where the parameter π is an independent $\mathrm{Beta}(1,\ell-1)$ random variable. Thus,

$$\begin{aligned}&\mathbb{P}\left\{B_m < \frac{m\epsilon}{32\ell^2}\right\} \\ &\le \mathbb{P}\left\{\mathrm{Bin}(m,\epsilon/16\ell^2) < \frac{m\epsilon}{32\ell^2}\right\} + \mathbb{P}\left\{\pi < \frac{\epsilon}{16\ell^2}\right\} \\ &\le e^{-m\epsilon/(128\ell^2)} + 1 - \left(1 - \frac{\epsilon}{16\ell^2}\right)^{\ell-1} \\ &\quad \text{(by a standard binomial estimate and expressing the beta distribution)} \\ &\le e^{-m\epsilon/(128\ell^2)} + \frac{\epsilon}{16\ell} \\ &\quad \text{(by the Bernoulli inequality)} \\ &\le \frac{\epsilon}{8\ell}\end{aligned}$$

whenever $\ell > (4\gamma)\left(\log(1/\epsilon) + \log\log(8\ell/\epsilon) + \log(128\ell^2)\right)$. To finish the proof it remains to show that

$$\limsup_{n\to\infty} \mathbb{P}\left\{A > B | B_m \ge \frac{m\epsilon}{32\ell^2}\right\} \le \frac{\epsilon}{8\ell}\ .$$

But this follows from the fact that this limiting probability is bounded by the the probability that a $\mathrm{Beta}(1, m\epsilon/32\ell^2)$ random variable is greater than $1/2$ which is at most $2^{-m\epsilon/32\ell^2}$. Since $m = \lfloor e^{\gamma\ell/4} \rfloor$, this is bounded by $\epsilon/(8\ell)$ for $\ell > (8/\gamma \vee C)\log(1/\epsilon)$, as desired. □

2.2.5 Proof of Theorem 2.1.5

Fix $\epsilon \in (0,1)$ and define $a = 2\log(\ell^2/\epsilon)+1$ and $k_\ell = \frac{\ell}{3a}$. A seed-finding algorithm with the desired property simply selects the k_ℓ most central vertices. (Again, for simplicity of the presentation,

we assume that k_ℓ is an integer.) With the notation introduced at the beginning of this section, we define $H_n = H_\psi(k_\ell)$. We need to show that the k_ℓ most central vertices of T_n are in T_ℓ with probability at least $1-\epsilon$ for all sufficiently large n.

The strategy of our proof is as follows. First we show that, with probability at least $1-\epsilon/2$, the seed T_ℓ contains at least k_ℓ "deep" vertices. Then we prove that for all n sufficiently large, all deep vertices of T_ℓ are more central in T_n than any vertex outside of the seed T_ℓ.

We call a vertex $v \in T_\ell$ *deep* if it has at least a descendants, that is, if

$$|(T_\ell, 1)_{v\downarrow}| \geq a+1 \ .$$

Denote by $\mathcal{A}_\ell$ the set of all deep vertices of T_ℓ. Noticing that

$$\mathbb{P}\left\{H_n \not\subset T_\ell\right\} \leq \mathbb{P}\left\{|\mathcal{A}_\ell| \leq k_\ell\right\} + \mathbb{P}\left\{\exists v \in V(T_n)\backslash V(T_\ell), \exists u \in \mathcal{A}_\ell : \psi_n(v) \leq \psi_n(u)\right\} \ ,$$

it suffices to show that

$$\mathbb{P}\left\{|\mathcal{A}_\ell| \leq k_\ell\right\} \leq \frac{\epsilon}{2} \ . \tag{2.2.5}$$

and

$$\limsup_{n\to\infty} \mathbb{P}\left\{\exists v \in V(T_n)\backslash V(T_\ell), \exists u \in \mathcal{A}_\ell : \psi_n(v) \leq \psi_n(u)\right\} \leq \frac{\epsilon}{2} \ . \tag{2.2.6}$$

(2.2.5) follows from inequality (4.2.1) in the Appendix under the condition $\ell \geq 64a^2\log(22a/\epsilon)$.

It remains to prove (2.2.6). To this end, for $i \in \{1,\ldots,\ell\}$, denote by C_i the component of vertex i in the forest obtained by removing the edges of T_ℓ from T_n. Then

$$\begin{aligned}&\mathbb{P}\left\{\exists v \in V(T_n)\backslash V(T_\ell), \exists u \in \mathcal{A}_\ell : \psi(v) \leq \psi(u) | T_\ell\right\} \\ &\quad\leq \sum_{u\in\mathcal{A}_\ell}\sum_{k=1}^{\ell} \mathbb{P}\left\{\exists v \in C_k\backslash\{k\} : \psi(v) \leq \psi(u) | T_\ell\right\} \ .\end{aligned}$$

Now fix T_ℓ and vertices $k \in \{1,\ldots,\ell\}$ and $u \in \mathcal{A}_\ell$. For any vertex $v \in C_k\backslash\{k\}$ such that $\psi(v) \leq \psi(u)$, there are two possibilities:
(1) either the largest subtree of T_n rooted at v is inside C_k, in which case $|C_k| \geq \sum_{i\neq k}|C_i|$;
(2) or the largest subtree of T_n rooted at v is $\left(\bigcup_{i=1,i\neq k}^{\ell} C_i\right) \cup C'_k$ for some $C'_k \subset C_k$. In this case, $\psi(v) \leq \psi(u)$ implies that

$$\sum_{i\in T_n\backslash(T_\ell,v)_{u\downarrow}} |C_i| \leq |C_k| \ .$$

Since $u \in A_\ell$, this means that the left-hand side is dominated by the number of red balls in a standard Pólya urn with after $n-\ell$ draws initialized with at least a red, one blue, and $n-a-\ell-1$ white balls; while $|C_k|$ behaves like the number of blue balls in the same urn.

By the same calculations as in the proof of Theorem 2.1.1, the probability of case (1) may be bounded by

$$\limsup_{n\to\infty}\mathbb{P}\left\{\left.|C_k| \geq \sum_{i\neq k}|C_i|\right| T_\ell\right\} = \limsup_{n\to\infty}\mathbb{P}\left\{|C_k| \geq (n-\ell)/2 | T_\ell\right\} \leq e^{-(\ell-1)/2} \leq \frac{\epsilon}{4\ell^2} \ .$$

Similarly, the probability of case (2) satisfies

$$\limsup_{n\to\infty}\mathbb{P}\left\{\left.\sum_{i\in T_n\backslash(T_\ell,v)_{u\downarrow}} |C_i| \leq |C_k|\right| T_\ell\right\} \leq e^{-(a-1)/2} \leq \frac{\epsilon}{4\ell^2}$$

by our choice $a = 2\log(\ell^2/\epsilon)+1$. This concludes the proof of (2.2.6) and hence that of Theorem 2.1.5.

2.2.6 Proof of Theorem 2.1.6

We prove the lower bound for the expected number of camouflaging vertices by induction. To this end, fix a singleton d and its parent v in T_ℓ. For $j \geq \ell$, let

$$E_j^{(v)} = \{\exists d' \in V(T_j)\backslash\{d\} : d' \sim v \text{ and } d', d \text{ are leaves in } T_j\} \ .$$

Observe that $E_{2\ell}^{(v)}$ is the event that v is a camouflaging vertex. Consider the sequences

$$\begin{aligned} a_j &= \mathbb{P}\left\{E_j^{(v)} | T_\ell\right\} \\ c_j &= \mathbb{P}\left\{d \text{ is a singleton in } T_j | T_\ell\right\} \ . \end{aligned}$$

Now, observe that the event $E_{j+1}^{(v)}$ occurs if $E_j^{(v)}$ occurs and the vertex $j+1$ is neither attached to d nor to d', or if d is a singleton of T_j and the $j+1$ is attached to v. Thus

$$a_{j+1} = a_j \cdot \left(1 - \frac{2}{j}\right) + c_j \cdot \frac{1}{j} \ .$$

Multiplying both sides by $j(j-1)$, we get

$$j(j-1)a_{j+1} = (j-1)(j-2)a_j + (j-1)c_j \ .$$

Summing over $j = \ell+1, \ldots, 2\ell-1$,

$$(2\ell-1)(2\ell-2)a_{2\ell} = \ell(\ell-1)a_{\ell+1} + \sum_{j=\ell+1}^{2\ell-1}(j-1)c_j \ ,$$

which implies that

$$a_{2\ell} \geq \frac{1}{(2\ell-1)(2\ell-2)} \sum_{j=\ell+1}^{2\ell-1}(j-1)c_j \geq \frac{1}{4(\ell-1)} \sum_{j=\ell+1}^{2\ell-1} c_j \ .$$

Note that, for $j \in \{\ell+1, \ldots, 2\ell-1\}$,

$$\begin{aligned} c_j &= \prod_{k=\ell}^{j-1}\left(1 - \frac{2}{k}\right) \\ &\geq \exp\left(-4\sum_{k=\ell}^{j-1}\frac{1}{k}\right) \quad (\text{since } 1 - x \geq e^{-2x} \text{ for } x < 3/4) \\ &\geq \exp\left(4\log\ell - 4\log j\right) \\ &> \frac{\ell^4}{(2\ell)^4} = \frac{1}{16} \ , \end{aligned}$$

and therefore

$$a_{2\ell} \geq \frac{1}{4(\ell-1)} \sum_{j=\ell+1}^{2\ell-1} c_j \geq \frac{1}{64} \ .$$

Let P_ℓ be the set of vertices in T_ℓ that are parents of a singleton. Then

$$\begin{aligned} \mathbb{E}[G_\ell | T_\ell] &= \mathbb{E}\left[\sum_{v \in P_\ell} 1_{E_{2\ell}^{(v)}} | T_\ell\right] \\ &= \sum_{v \in P_\ell} \mathbb{P}\left\{E_{2\ell}^{(v)} | T_\ell\right\} \\ &\geq \frac{1}{64}|P_\ell| \ , \end{aligned}$$

which implies that $\mathbb{E}G_\ell \geq \frac{1}{64}\mathbb{E}|P_\ell|$.

It remains to bound the expected number of singletons $\mathbb{E}|P_\ell|$ in the uniform random recursive tree T_ℓ. Write $S_k = |P_k|$ and note that S_k equals the number of parents of singletons in T_k.

When a new vertex is attached to the tree T_k, we lose one singleton if the new vertex is attached to the parent of a singleton. This happens with probability S_k/k. If a the new vertex is attached to a singleton, then the number remains the same. If the new vertex is attached to some vertex that is not a leaf nor a parent of a singleton, then, the number of singletons also remains unchanged. Finally, if the new vertex is attached to a leaf that is not a singleton, the number of singletons increases by 1. Thus, denoting the number of leaves of T_k by L_k,

$$\begin{aligned}\mathbb{E}[S_{k+1}|T_k] &= (S_k - 1)\frac{S_k}{k} + S_k\left(\frac{S_k}{k} + 1 - \frac{S_k}{k} - \frac{L_k}{k}\right) + (S_k + 1)\left(\frac{L_k}{k} - \frac{S_k}{k}\right) \\ &= \left(1 - \frac{2}{k}\right)S_k + \frac{L_k}{k}\ .\end{aligned}$$

Taking expectations and using the fact that $\mathbb{E}L_k = k/2$, we have that $\mathbb{E}S_\ell = \ell/6$. Summarizing, the expected number of camouflaging vertices satisfies

$$\mathbb{E}G_\ell \geq \frac{1}{64}\cdot\frac{\ell}{6} = \frac{\ell}{384}\ .$$

We prove the second inequality of Theorem 2.1.6 using the *bounded differences inequality* of McDiarmid, Theorem 4.1.1.

Observe that given T_ℓ, there is a bijection between the set of recursive trees of size 2ℓ containing T_ℓ as subgraph and the set $\mathcal{S} = [\ell]\times\cdots\times[2\ell-1]$. The bijection is simply given by associating the vector $\kappa = (a_{\ell+1},\cdots,a_{2\ell})$ to the recursive tree $T(\kappa)$ where the vertex $k \in [\ell+1, 2\ell]$ is attached to the vertex a_k, starting by T_ℓ until obtaining $T_{2\ell}$. Then we may consider the set $\mathcal{S}$ as the set of recursive trees with 2ℓ vertices that contain T_ℓ as subtree.

Importantly, the components of κ that represent the uniform random recursive tree $T_{2\ell}$ are independent random variables.

Given T_ℓ, consider the function $g : \mathcal{S} \to \mathbb{R}$ such that $g(T_{2\ell})$ is the number of camouflaging vertices.

By the bounded differences inequality, it suffices to show that, given $T, T' \in \mathcal{S}$, if T and T' differ by exactly one coordinate, then $|g(T) - g(T')| \leq 2$.

To this end, let $v \in V(T_n)$ be a parent of a singleton d. v is a camouflaging vertex of a tree $T = (a_{\ell+1},\cdots,a_{2\ell})$ if and only if

1. $d \notin \{a_{\ell+1},\cdots,a_{2\ell}\}$;
2. $\exists k \in \{\ell+1,\cdots,2\ell\}\backslash\{a_{k+1},\cdots,a_{2\ell}\}$ such that $a_k = v$.

Now, consider $T = (a_{\ell+1},\cdots,a_{2\ell})$, $T' = (b_{\ell+1},\cdots,b_{2\ell})$ two trees with $a_r \neq b_r$ for some r and $a_j = b_j$ for $j \neq r$. For a camouflaging vertex v in T (with corresponding singleton d in T_ℓ) not to be a camouflaging vertex in T', it is necessary (but not sufficient) that either

1. b_r is a child of v,
2. or $a_r = v$.

Similarly, for a not camouflaging vertex v in T (with corresponding singleton d in T_ℓ), to be a camouflaging vertex in T' it is necessary that either

1. a_r is a descendant of v,
2. or $b_r = v$.

Thus, $|g(T) - g(T')| \leq 2$, and the bounded differences condition is satisfied, proving the second inequality of Theorem 2.1.6.

Chapter 3

Generalized Chinese Restaurant Process

In this chapter we decribe joint work with Roberto Oliveira and Rodrigo Ribeiro.

The chapter is organized as follows. We fix some notation in the next paragraph. In section 3.1, we introduce the model, discuss its regimes, and give some background on its theory and applications. Section 3.2 states our main theorems. We will also outline their proofs and compare them with previous results. Section 4.1 contains the main concentration-of-meausre results we will need, including Freedman's inequality. Actual proofs start in Section 3.3 with the analysis of the number of parts in $\mathcal{P}_n$. The arguments for $N_n(k)$ is slightly more convoluted and takes three sections. Section 3.4 gives some preliminary results, including a recursive formula. Section 3.5 obtains high-probability upper and lower bounds for $N_n(k)$. The proof of our main Theorem is wrapped up in Section 3.6. The final section contains some concluding remarks. The appendix collects several technical estimates

Notation: In this chapter $\mathbb{N} = \{1, 2, 3 \dots\}$ is the set of positive integers. Given $n \in \mathbb{N}$, we let $[n] := \{1, \dots, n\}$ denote the set of all numbers from 1 to n. Given a nonempty set S, a *partition* $\mathcal{P}$ of S is a collection of pairwise disjoint and nonempty subsets of S whose union is all of S. The elements of $\mathcal{P}$ are called the *parts*. We denote the cardinality of a finite set S by $|S|$. In particular, for a finite partition $\mathcal{P}$, $|\mathcal{P}|$ denotes the number of parts in S. Finally, when we talk about sequences $\{x_n\}_{n=0}^{+\infty}$ of random or deterministic values, we will write $\Delta x_n := x_n - x_{n-1}$.

3.1 The model

3.1.1 Definitions

Fix two parameters $\theta, \alpha \in \mathbb{R}$; extra conditions will be imposed later. GCRP(α, θ) – shorthand for the Generalized Chinese Restaurant Process with parameters (α, θ) – is a Markov chain

$$\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3, \mathcal{P}_4, \dots$$

where, for each $n \in \mathbb{N}$, $\mathcal{P}_n$ is a partition of $[n] := \{1, \dots, n\}$. We let

$$V_n := |\mathcal{P}_n| \tag{3.1.1}$$

denote the number of parts in $\mathcal{P}_n$ and write

$$\mathcal{P}_n = \{A_{i,n} \,:\, i = 1, \dots, V_n\}, \tag{3.1.2}$$

where the $A_{i,n}$ are the parts of $\mathcal{P}_n$. In the colorful metaphor of the "Chinese restaurant", the $A_{i,n}$ are the tables occupied by customers $1, \dots, n$, who arrive sequentially, with V_n being the number of occupied tables. So $\mathcal{P}_n$ describes the table arrangements of the first n customers.

The evolution of the process is as follows.

- **Initial state:** customer 1 sits by herself i.e. $\mathcal{P}_1 = \{\{1\}\}$.

- **Evolution:** Given $\mathcal{P}_1, \ldots, \mathcal{P}_n$, with $\mathcal{P}_n$ as in (3.1.2), we define $\mathcal{P}_{n+1}$ via a random choice:
  - For each $i = 1, \ldots, V_{n-1}$, with probability
$$\frac{|A_{i,n}| - \alpha}{n + \theta},$$
customer $n+1$ sits at the ith table. That is,
$$\mathcal{P}_{n+1} = \{A_{j,n} : j \in [V_n] \backslash \{i\}\} \cup \{A_{i,n} \cup \{n+1\}\}.$$
Notice that $V_{n+1} = V_n$ in this case.
  - With probability
$$\frac{\alpha\, V_n + \theta}{n + \theta},$$
customer $n+1$ sits by herself at a new table. That is, we set
$$\mathcal{P}_n = \{A_{i,n} : i = 1, \ldots, V_n\} \cup \{\{n+1\}\}.$$
In this case $V_{n+1} = V_n + 1$.

Our focus in this chapter is on V_n and the random variables

$$N_n(k) := |\{A \in \mathcal{P}_n : |A| = k\}| = |\{i \in [V_n] : |A_{i,n}| = k\}| \ \ (k \in [n]) \tag{3.1.3}$$

that count how many of the parts in $\mathcal{P}_n$ have size k.

3.1.2 Choices of parameters and different regimes

The attentive reader will have noticed that the above process only makes sense for certain values of θ and α. Specifically, there are different assumptions one can make, which lead to different behavior [25, 26].

- *Bounded number of parts*: if $\alpha < 0$ and $\theta = -m\alpha$ for some $m \in \mathbb{N}$, then $V_n \to m$ almost surely. After V_n reaches value m, the process behaves like an urn model with m urns.

- *Logarithmically growing number of parts*: if $\theta > 0$, $\alpha = 0$, then
$$\frac{V_n}{\log n} \to \theta \text{ almost surely}$$
and V_n has Gaussian fluctuations at the scale of $\sqrt{\log n}$.

- *Polynomially growing number of parts*: if $\alpha > 0$ and $\theta > -\alpha$,
$$\frac{V_n}{n^\alpha} \to V^o \text{ almost surely} \tag{3.1.4}$$
where V^o is a nondegenerate random variable with a density over $(0, +\infty)$. In particular, $0 < V^o < +\infty$ almost surely.

This last regime is the focus of the present chapter.

3.1.3 Some background

We discuss here a bit of the history and applications of the GCRP. Those interested only in results may skip to the next section.

The GCRP is an exchangeable model in the sense that the law of $\mathcal{P}_n$ is invariant under permutations of $[n]$. One consequence of this is that the natural infinite limit $\mathcal{P}_\infty$ of $\mathcal{P}_n$ is an exchangeable random partition of the natural numbers $\mathbb{N}$. That is, the law of $\mathcal{P}_\infty$ is invariant under any finite permutation of $\mathbb{N}$.

A well-known result of Kingman [24] says that exchangeable random partitions of $\mathbb{N}$ can always be built from mixtures of paintbox partitions. Suppose P is a random probability distribution over $\mathbb{N} \cup \{\star\}$ where $\star \notin \mathbb{N}$. Conditionally on P, let $\{X_i\}_{i\in\mathbb{N}}$ be an i.i.d.-P sequence. Form a partition of $\mathbb{N}$ by placing each $i \in \mathbb{N}$ with $X_i = \star$ in a singleton, and (for each $k \in \mathbb{N}$) putting all j with $X_j = k$ in the same part. Clearly, such a construction always leads to an exchangeable random partition, and Kingman's theorem says that this is the *only way* to build such partitions. In the specific case of the infinite GCRP(α, θ), the law of P is the two-parameter Poisson-Dirichlet distribution PD(α, θ). This can be used to derive explicit formulae for the distribution of $\mathcal{P}_n$ for each n.

The GCRP was first mentioned in print by Aldous [1]. It was studied by Pitman [25], [26] as an example of a partially exchangeable model where many explicit calculations are possible. In particular, the exact distribution of the random variables $N_n(k)$ we consider can be computed explicitly. Based on these formulae, [15], [16] obtained large and moderate deviation results for these variables. These results are briefly described in subsection 3.2.1 below.

The class of models we consider is also important in many applications. On the one hand, it is a generalization of Ewens' neutral allele sampling model in population Genetics [14]. On the other hand, the GCRP and its variants are important building blocks for topic models [19] and many other Bayesian nonparametric methods. We refer to Crane's recent survey [9] for much more information on our model, its extensions and the many contexts where it has appeared.

3.2 Results

Let $n \in \mathbb{N}$ and recall the definitions of V_n and $N_n(k)$ in (3.1.1) and (3.1.3), respectively. Our theorem describes these random variables in the setting where $\alpha \in (0, 1)$ and $\theta + \alpha > 0$. Recall from Section 3.1.2 that in this setting the random variables $n^{-\alpha}V_n$ have a nontrivial limit $V^o > 0$ (cf. (3.1.4)) . For our purposes, it is more convenient to work with the random variables V_n/ϕ_n, where

$$\phi_n := \frac{\Gamma(1+\theta)}{\Gamma(1+\theta+\alpha)} \frac{\Gamma(n+\alpha+\theta)}{\Gamma(n+\theta)}.$$

Note that ϕ_n/n^α converges to a constant $c > 0$ when $n \to +\infty$. In particular, the limit

$$V^* := \lim_{n\to+\infty} \frac{V_n}{\phi_n} \text{ almost surely} \tag{3.2.1}$$

exists and is a.s. positive (it is a rescaling of V^o). Our first result quantifies the convergence in this statement.

Theorem 3.2.1 (Proven in subsection 3.3.3)**.** *Consider a realization $\{\mathcal{P}_n\}_{n\in\mathbb{N}}$ of the Generalized Chinese Restaurant Process GCRP(α, θ) with parameters $\alpha \in (0, 1)$ and $\theta > -\alpha$. Then there exist constants $K = K(\alpha, \theta) > 0$ and $c_* = c_*(\alpha, \theta) > 0$ such that for $\delta < e^{-K}$ the following holds with probability $\geq 1 - \delta$:*

$$\forall m \in \mathbb{N} : \left| \frac{V_m}{\phi_m} - V_* \right| \leq \frac{c_* \left[\log\log(m+2) + \log\left(\frac{1}{\delta}\right)\right]}{(m+\theta)^{\alpha/2}}.$$

Our second and main result gives concentration of the random variables $N_n(k)$ simultaneously for all $k = o(n^{\alpha/(2\alpha+4)}/(\log n)^{1/(\alpha+2)})$.

Theorem 3.2.2 (Main; proven in section 3.6)**.** *Consider a realization $\{\mathcal{P}_n\}_{n\in\mathbb{N}}$ of the Generalized Chinese Restaurant Process GCRP(α,θ) with parameters $\alpha\in(0,1)$ and $\theta>-\alpha$. Then there exist constants $n_0=n_0(\alpha,\theta)$, $C=C(\alpha,\theta)$ such that the following holds. Assume $n\in\mathbb{N}$ with $n\geq n_0$. Take $A\geq 0$, $\varepsilon\in(0,1/2)$ and define $k_{\varepsilon,n}:=\lceil \varepsilon\, n^{\alpha/(2\alpha+4)}/(\log n)^{1/(\alpha+2)}\rceil$. Then the following holds with probability $1-e^{-A}$:*

$$\forall k\in[k_{\epsilon,n}]\,:\,\left|N_n(k)-c(\alpha,\theta)\,\frac{\Gamma(k-\alpha)}{\Gamma(k+1)}\,V_*\,n^\alpha\right|\leq C\frac{\Gamma(k-\alpha)}{\Gamma(k+1)}n^\alpha\varepsilon^{\alpha+2}\left(1+\frac{A}{\log n}\right)$$

where

$$c(\alpha,\theta):=\frac{\alpha\,\Gamma(1+\theta)}{\Gamma(1-\alpha)\,\Gamma(1+\alpha+\theta)}>0.$$

The following immediate corollary is perhaps somewhat easier to parse.

Corollary 3.2.1. *In the setting of Theorem 3.2.2, let $\varepsilon=\varepsilon_n\to 0$ with n. Then there exist sequences $C_n\to+\infty$, $\xi_n\to 0$ such that the the probability that for some we have*

$$\mathbb{P}\left(\forall k\in[k_{\varepsilon,n}]\,:\,V_*-\xi_n<\frac{N_n(k)}{c(\alpha,\theta)\,\frac{\Gamma(k-\alpha)}{\Gamma(k+1)}\,n^\alpha}<V_*+\xi_n\right)\geq 1-n^{-C_n},$$

for large enough $n\in\mathbb{N}$.

Proof. [Proof sketch] Apply Theorem 3.2.2 with $A=C_n\log n$, where $C_n\to+\infty$ but $\varepsilon_n^{\alpha+2}\,C_n\to 0$. Then take:

$$\xi_n=\frac{C}{c(\alpha,\theta)}\,\varepsilon_n^{\alpha+2}\,(1+C_n).$$

■

3.2.1 Related work

One consequence of our results is the a.s. asymptotics for $N_n(k)/V_n$:

$$\frac{N_n(k)}{V_n}\to c(\alpha,\theta)\,\frac{\Gamma(k-\alpha)}{\Gamma(k+1)}.$$

This kind of Law of Large Numbers was first obtained by Pitman [26, Chapter 3] with no explicit convergence rates.

Much more recently, Favaro, Feng and Gao [15, 16] have used Pitman's explicit formulae to obtain large and moderate deviation results for the $N_n(k)$. Reference [16], which is the closest to our work, focuses on precise estimates for probabilities like

$$\mathbb{P}\left(\frac{N_n(k)}{n^\alpha\,\beta_n}>c\right)\text{ when }\beta_n\gg(\log n)^{1-\alpha}.\tag{3.2.2}$$

The paper [15] considers even larger sequences β_n. By contrast, we obtain finite-n estimates for deviations at smaller scales, which (as expected) are not as precise. There is also a difference in proof methods: whereas they rely on explicit formulae, our argument is based on recursions and martingales.

Another important conceptual difference between our work and that of Favaro et al. is that, for their purposes, the lack of concentration in V_n/ϕ_n is not an issue. Indeed, if one goes "deep enough" into the tail of the $N_n(k)$, as in (3.2.2), the nontrivial distribution of $V_*=\lim V_n/\phi_n$ becomes irrelevant. Our theorems operate at a finer scale and complement these previous papers by giving tail bounds for V_* and $\sup_n V_n/\phi_n$ matter (cf. Theorem 3.3.1). As a result, we find in Theorem 3.2.2 that the sequence $\{N_n(k)\}_k$ is essentially a deterministic function of V_*.

3.2.2 Proof outline

The general methodology in our proof is based in the study of degree distributions in preferential attachment random graphs, as in the book by Chung and Lu [8, Chapter 3]. However, a new phenomenon arises. In the graph setting, the total number of vertices at time n is usually linear n (at least with high probability). By contrast, the analogue of the total number of vertices is V_n – the number of parts –, which is sublinear and not concentrated.

One consequence of this point in our analysis is that the martingale arguments are much more delicate, and rely on Friedman's martingale inequality (cf. section 4.1), instead of the more usual (and less precise) Azuma-Höffding bound. Another point is that we must first obtain results on the number of parts V_n, which we do in section 3.3.

We then consider the random variables $N_n(k)$. The general strategy is to write these variables in terms of "recursions + martingales" depending on $N_{n-1}(i)$ for $i = k-1, k$, and then observe how the "martingale" part concentrates. These first steps, which are taken in section 3.4, are similar to the analysis in [8, Chapter 3]. However, the results obtained are not directly employable to prove the main theorem. Section 3.5 then turns these arguments into actionable bounds. This leads to the proof of the main result in section 3.6.

3.3 Estimates on the number of parts

In this section we obtain results on the number of parts V_n of $\mathcal{P}_n$. In particular, we prove Theorem 3.2.1 above.

In subsection 3.3.1 we prove a recurrence relation for V_n. We use this in subsection 3.3.2 to derive concentration for the whole sequence. Finally subsection 3.3.3 proves Theorem 3.2.2.

The following normalizing factor will appear in our proofs:

$$\phi_n := \prod_{j=1}^{n-1} \left(1 + \frac{\alpha}{j+\theta}\right) = \frac{\Gamma(1+\theta)}{\Gamma(1+\theta+\alpha)} \frac{\Gamma(n+\alpha+\theta)}{\Gamma(n+\theta)}. \tag{3.3.1}$$

Note that by Lemma 4.3.6 we have $\phi_n = \Theta(n^\alpha)$.

3.3.1 A recurrence relation

The first result in this section is the following Lemma.

Lemma 3.3.1 (Recurrence relation for V_n)**.** *For all $n, m \in \mathbb{N}$ the recurrence relation holds*

$$\frac{V_n}{\phi_n} = \frac{V_m}{\phi_m} + (M_n - M_m) + \frac{O(1)}{(m+\theta)^\alpha}, \tag{3.3.2}$$

where $(M_n, \mathcal{F}_n)$ is a martingale satisfying $M_0 = 0$,

1. $|\Delta M_j| \leq \frac{2\Gamma(1+\theta+\alpha)}{\Gamma(1+\theta)\cdot(1+\theta)^\alpha}$;
2. $\mathbb{E}[(\Delta M_j)^2 | \mathcal{F}_{j-1}] \leq \dfrac{2\Gamma(1+\theta+\alpha)\alpha}{\Gamma(1+\theta)} \cdot (j+\theta)^{-\alpha-1} \left(\dfrac{V_{j-1} + \frac{\theta}{\alpha}}{\phi_{j-1}}\right)$,

for all $j \in \mathbb{N}$.

Proof. Recall $\Delta V_n = V_n - V_{n-1}$. On the other hand, we also know that

$$\mathbb{P}\left(\Delta V_n = 1 | \mathcal{F}_{n-1}\right) = \mathbb{E}\left[\Delta V_n | \mathcal{F}_{n-1}\right] = \frac{\alpha V_{n-1} + \theta}{n-1+\theta}. \tag{3.3.3}$$

In other words, conditioned on $\mathcal{F}_{n-1}$, the random variable ΔV_n is distributed as $\text{Be}\left(\frac{\alpha V_{n-1}+\theta}{n-1+\theta}\right)$. In order to obtain mean zero martingale, it will be useful to centralize the random variable ΔV_n. Thus we may write V_n as

$$\begin{aligned} V_n &= V_{n-1} + \Delta V_n \\ V_n &= \left(1 + \frac{\alpha}{n-1+\theta}\right) V_{n-1} + \left(\Delta V_n - \frac{\alpha V_{n-1}+\theta}{n-1+\theta}\right) + \frac{\theta}{n-1+\theta}. \end{aligned} \tag{3.3.4}$$

Thus, dividing the above identity by ϕ_n, we obtain

$$\frac{V_n}{\phi_n} = \frac{V_{n-1}}{\phi_{n-1}} + \zeta_n + \frac{\theta}{(n-1+\theta)\phi_n}, \tag{3.3.5}$$

where

$$\zeta_n := \frac{\Delta V_n - \frac{\alpha V_{n-1}+\theta}{n-1+\theta}}{\phi_n}. \tag{3.3.6}$$

Observe that

$$\mathbb{E}[\zeta_n | \mathcal{F}_n] = 0. \tag{3.3.7}$$

Iterating this argument $n-m$ steps leads to

$$\frac{V_n}{\phi_n} = \frac{V_m}{\phi_m} + (M_n - M_m) + (\theta_n - \theta_m), \tag{3.3.8}$$

where

$$M_n := \sum_{j=2}^{n} \zeta_j \text{ and } \theta_n = 1 + \sum_{j=1}^{n-1} \frac{\theta}{(j+\theta)\phi_{j+1}}. \tag{3.3.9}$$

Notice that identity (3.3.7) implies that M_n is a zero mean martingale.

Now we estimate the order of the deterministic contribution of $\theta_n - \theta_m$ on identity (3.3.8). By Lemma 4.3.6, the following upper bound holds

$$\frac{1}{(j+\theta)\phi_{j+1}} < \frac{2\Gamma(1+\theta+\alpha)}{\Gamma(1+\theta)\cdot(j+\theta)^{1+\alpha}}. \tag{3.3.10}$$

Thus, bounding the sum by the integral, we obtain

$$\theta_n - \theta_m = \sum_{j=m}^{n-1} \frac{\theta}{\phi_{j+1}(j+\theta)} \leq \frac{4\Gamma(1+\theta+\alpha)\theta}{\alpha\Gamma(1+\theta)} \frac{1}{(m+\theta)^{\alpha}}. \tag{3.3.11}$$

which proves the first statement of the lemma.

In the remainder of the proof we estimate the increments of the martingale M_n as well as its conditioned quadratic variation. By the definition of M_j and recalling that ΔV_j is at most one and the bound on (3.3.11) we obtain that

$$|\Delta M_j| \leq \frac{1}{\phi_j} \leq \frac{2\Gamma(1+\theta+\alpha)}{\Gamma(1+\theta)\cdot(j+\theta)^{\alpha}} \tag{3.3.12}$$

and also

$$\begin{aligned} \mathbb{E}[(\Delta M_j)^2 | \mathcal{F}_{j-1}] &\leq \frac{\alpha}{(j-1+\theta)\phi_j} \frac{\phi_{j-1}}{\phi_j} \frac{V_{j-1}+\frac{\theta}{\alpha}}{\phi_{j-1}} \\ &\leq \frac{2\Gamma(1+\theta+\alpha)\alpha}{\Gamma(1+\theta)} \cdot (j+\theta)^{-\alpha-1} \frac{V_{j-1}+\frac{\theta}{\alpha}}{\phi_{j-1}}, \end{aligned} \tag{3.3.13}$$

which proves the lemma. ■

3.3.2 Concentration and tail bounds

We combine the recurrence relation we have proven with Freedman's inequality to obtain the following theorem

Theorem 3.3.1. *In the (α,θ)-GCRP there are constants $K = K(\alpha,\theta) > 0$ and $c_V = c_V(\alpha,\theta) > 0$ such that for all $m \geq 0$ integer and $A \geq K$ we have*

$$\mathbb{P}\left(\sup_{j\geq m}\left(\frac{V_j}{\phi_j} - \frac{V_m}{\phi_m}\right) \geq \frac{A}{(m+\theta)^{\alpha/2}}\right) \leq \exp(-c_V A).$$

In particular, for $m = 0$, considering $V_0 = 0$ and $\phi_0 = 1$ we have

$$\mathbb{P}\left[\sup_{j\in\mathbb{N}}\left(\frac{V_j}{\phi_j}\right) \geq A\right] \leq \exp(-c_V A).$$

Proof. We start with the particular case $m = 0$ and then use it to prove the general result. *Case $m = 0$.* From Lemma 3.3.1 we know that the V_n may be written as a mean zero martingale M_n plus a deterministic factor θ_n, where $\{\theta_n\}_{n\in\mathbb{N}}$ is a increasing positive and bounded sequence of real numbers. Thus, $\{\theta_n\}_{n\in\mathbb{N}}$ converges to some positive number θ_∞. For a positive real number A, consider the following stopping time

$$T_A := \inf\left\{i \in \mathbb{N} : \frac{V_i}{\phi_i} \geq A + \theta_i\right\}. \tag{3.3.14}$$

Observe that

$$\begin{aligned}\mathbb{P}\left(\sup_{j\in\mathbb{N}}\left(\frac{V_j}{\phi_j}\right) \geq A + \theta_\infty\right) &\leq \mathbb{P}\left(\exists j \in \mathbb{N} : \frac{V_j}{\phi_j} \geq A + \theta_j\right) \\ &= \lim_n \mathbb{P}\left(\frac{V_{T_A\wedge n}}{\phi_{T_A\wedge n}} \geq A + \theta_{T_A\wedge n}\right) \\ &= \lim_n \mathbb{P}\left(M_{T_A\wedge n} \geq A\right).\end{aligned} \tag{3.3.15}$$

By the above inequality, the first case is proven if we obtain a proper upper bound for the tail of the stopped martingale $\{M_{T_A\wedge n}\}_{n\in\mathbb{N}}$. We will do this via Lemma 4.1.1, which requires bounds on the increment and quadratic variation of $\{M_{T_A\wedge n}\}_{n\in\mathbb{N}}$. We obtain these bounds on the next lines. For the increment a direct application of Lemma 3.3.1 gives us

$$|M_{T_A\wedge(j+1)} - M_{T_A\wedge j}| \leq R,$$

where $R = \dfrac{2\Gamma(1+\theta+\alpha)}{\Gamma(1+\theta)}(1+\theta)^{-\alpha}$. For the quadratic variation $W_{n\wedge T_A}$ we have that, also by Lemma 3.3.1,

$$\begin{aligned}W_{n\wedge T_A} &= \sum_{j=2}^{n\wedge T_A} \mathbb{E}[(\Delta M_j)^2|\mathcal{F}_{j-1}] \\ &\leq R^2 + \sum_{j=2}^{n\wedge T_A - 1} \frac{2\Gamma(1+\theta+\alpha)\alpha}{\Gamma(1+\theta)} \cdot (j+\theta)^{-\alpha-1}\frac{V_{j-1} + \frac{\theta}{\alpha}}{\phi_{j-1}} \\ &\leq R^2 + \sum_{j=2}^{n\wedge T_A} \frac{2\Gamma(1+\theta+\alpha)\alpha}{\Gamma(1+\theta)} \cdot (j+\theta)^{-\alpha-1}\left(A + \theta_j + \frac{\theta}{\alpha\phi_{j-1}}\right).\end{aligned} \tag{3.3.16}$$

Choosing $A \geq K(\alpha,\theta)$, which is defined below:

$$K(\alpha,\theta) := \max\left\{\frac{\theta}{\alpha} + \sup_{j\in\mathbb{N}}\{\theta_j\}, R\right\}; \tag{3.3.17}$$

on (3.3.16), we obtain

$$W_{n\wedge T_A} \leq R^2 + \frac{4\Gamma(1+\theta+\alpha)}{\Gamma(1+\theta)}(1+\theta)^{-\alpha}\cdot A = \frac{3}{R}\cdot(R^2A).$$

Finally, applying Lemma 4.1.1, with

$$c_1 := \frac{3}{R} = \frac{3\Gamma(1+\theta)}{2\Gamma(1+\theta+\alpha)}(1+\theta)^{\alpha}$$

we obtain

$$\mathbb{P}\left(M_{T_A\wedge n} \geq A\right) \leq \exp\left(\frac{-A}{\frac{3\Gamma(1+\theta)}{2\Gamma(1+\theta+\alpha)}(1+\theta)^{\alpha}+\frac{2}{3}}\right), \tag{3.3.18}$$

and

$$\mathbb{P}\left(M_{T_A\wedge n} \geq A\right) \leq \exp\left(-c_2A\right), \tag{3.3.19}$$

for

$$c_2 = \left(\frac{3\Gamma(1+\theta)}{2\Gamma(1+\theta+\alpha)}(1+\theta)^{\alpha}+\frac{2}{3}\right)^{-1}. \tag{3.3.20}$$

The above inequality combined with (3.3.15) gives us

$$\mathbb{P}\left(\sup_{m\in\mathbb{N}}\left(\frac{V_m}{\phi_m}\right) \geq A\right) \leq \exp(-c_2A),$$

proving the result for $m = 0$.
Case $m > 0$. The proof of the case $m > 0$ is similar to the first case, but it requires another stopping time and the case $m = 0$ itself. So, consider the following stopping time:

$$\begin{aligned}\hat{T}_B :&= \inf\left\{j \geq m : \frac{V_j}{\phi_j} - \frac{V_m}{\phi_m} \geq B\right\}\\ &= \inf\left\{j\in\mathbb{N} : (M_j - M_m) + (\theta_j - \theta_m) \geq B\right\}.\end{aligned}$$

Observe that, as showed in the proof of Lemma 3.3.1,

$$\theta_n - \theta_m \leq \frac{4\Gamma(1+\theta+\alpha)\theta}{\alpha\Gamma(1+\theta)}\frac{1}{(m+\theta)^{\alpha}}. \tag{3.3.21}$$

Now, let $B = \dfrac{A}{(m+\theta)^{\alpha/2}}$ and suppose $A \geq 2\theta_\infty$. Thus,

$$\begin{aligned}\mathbb{P}\left(\sup_{j\geq m}\left(\frac{V_j}{\phi_j} - \frac{V_m}{\phi_m}\right) \geq \frac{A}{(m+\theta)^{\alpha/2}}\right) &\leq \mathbb{P}\left(\exists j \leq n : (M_j - M_m) + (\theta_j - \theta_m) \geq \frac{A}{(m+\theta)^{\alpha/2}}\right)\\ &= \mathbb{P}\left((M_{\hat{T}_B\wedge n} - M_m) + (\theta_{\hat{T}_B\wedge n} - \theta_m) \geq B\right)\\ \text{(use that } \theta_{\hat{T}_B\wedge n} \geq \theta_m) &\leq \lim_n \mathbb{P}\left(M_{\hat{T}_B\wedge n} - M_m \geq \frac{A}{2(m+\theta)^{\alpha/2}}\right).\end{aligned}$$

Let T_A be the same as defined in (3.3.14). Then:

$$\begin{aligned}
\mathbb{P}\left(M_{\hat{T}_B\wedge n} - M_m \geq \frac{A}{2(m+\theta)^{\alpha/2}}\right) &\leq \mathbb{P}\left(M_{\hat{T}_B\wedge n} - M_m \geq \frac{A}{2(m+\theta)^{\alpha/2}}, T_A \geq n\right) \\
&\quad + \mathbb{P}(T_A < n) \\
&\leq \mathbb{P}\left(M_{\hat{T}_B\wedge T_A\wedge n} - M_m \geq \frac{A}{2(m+\theta)^{\alpha/2}}, T_A \geq n\right) \\
&\quad + \mathbb{P}\left(\sup_{j\in\mathbb{N}} \frac{V_j}{\phi_j} \geq A\right) \\
&\leq \mathbb{P}\left(M_{\hat{T}_B\wedge T_A\wedge n} - M_m \geq \frac{A}{2(m+\theta)^{\alpha/2}}\right) \\
&\quad + \mathbb{P}\left(\sup_{j\in\mathbb{N}} \frac{V_j}{\phi_j} \geq A\right).
\end{aligned}$$

As in the case $m = 0$, by Lemma 3.3.1, the increment of $\{M_{j\wedge\hat{T}_b\wedge T_a}\}$ satisfies the following upper bound

$$|(M_{(j+1)\wedge\hat{T}_B\wedge T_A} - M_m) - (M_{j\wedge\hat{T}_B\wedge T_A} - M_m)| \leq \frac{2\Gamma(1+\theta+\alpha)}{\Gamma(1+\theta)}(m+\theta)^{-\frac{\alpha}{2}},$$

whereas its quadratic variation satisfies

$$W_{n\wedge T_A} \leq \frac{4\Gamma(1+\theta+\alpha)}{\Gamma(1+\theta)}(m+\theta)^{-\alpha}\cdot A.$$

Thus, again by Lemma 4.1.1 it follows that

$$\mathbb{P}\left(M_{\hat{T}_B\wedge T_A\wedge n} - M_m \geq \frac{A}{2(m+\theta)^{\alpha/2}}\right) \leq \exp(-c_3 A),$$

for some constant c_3, which implies

$$\mathbb{P}\left(\sup_{j\geq m}\left(\frac{V_j}{\phi_j} - \frac{V_m}{\phi_m}\right) \geq \frac{A}{(m+\theta)^{\alpha/2}}\right) \leq \exp(-c_2 A) + \exp(-c_3 A) \leq \exp(-c_V A),$$

for $c_V = \log 2 \cdot \min\{c_2, c_3\}$. ■

3.3.3 Proof of Theorem 3.2.1

A consequence of Theorem 3.3.1 is to give estimates of how large the deviation of V_j/ϕ_j from its limit V_* can be uniformly in time.

Proof. [Proof of theorem 3.2.1]

Given δ define

$$\delta_j = \frac{\delta}{(j+1)(j+2)}. \tag{3.3.22}$$

Let E_j denote the following event

$$E_j := \left\{\forall m \geq 2^j : \left|\frac{V_m}{\phi_m} - \frac{V_{2^j}}{\phi_{2^j}}\right| \leq \frac{\log\frac{1}{\delta_j}}{c_V(2^j+\theta)^{\frac{\alpha}{2}}}\right\}. \tag{3.3.23}$$

Assuming $\log\frac{2}{\delta} \geq K_1$ we have by Theorem 3.3.1

$$\mathbb{P}(E_j^c) \leq \exp\left(-\log\frac{1}{\delta_j}\right) \leq \frac{\delta}{(j+1)(j+2)},$$

which implies, by union bound,

$$\begin{aligned}\mathbb{P}\left(\bigcap_{j\geq 0} E_j\right) &\geq 1-\sum_{j\geq 0}\mathbb{P}(E_j^c)\\ &\geq 1-\sum_{j\geq 0}\delta_j\\ &\geq 1-\delta.\end{aligned}$$

Now, observe that, when E_j occurs, we have for all $m\in[2^j,2^{j+1}]$

$$\begin{aligned}\left|\frac{V_m}{\phi_m}-V_*\right| &\leq \left|\frac{V_m}{\phi_m}-\frac{V_{2^j}}{\phi_{2^j}}\right|+\left|\frac{V_{2^j}}{\phi_{2^j}}-V_*\right|\\ &\leq 2\sup_{m\geq 2^j}\left|\frac{V_m}{\phi_m}-\frac{V_{2^j}}{\phi_{2^j}}\right|\\ &\leq \frac{2\log\frac{1}{\delta_j}}{(2^j+\theta)^{\frac{\alpha}{2}}}\\ &\leq \frac{1}{c_V(2^j+\theta)^{\frac{\alpha}{2}}}\left[4\log(j+2)+2\log\left(\frac{1}{\delta}\right)\right],\end{aligned}$$

and once $m\in[2^j,2^{j+1}]$ it follows that

$$\left|\frac{V_m}{\phi_m}-V_*\right| \leq \frac{32}{c_V(m+\theta)^{\frac{\alpha}{2}}}\left[\log\log(m+2)+\log\left(\frac{1}{\delta}\right)\right],$$

for any $j\in\{0,1,2,\cdots\}$. To finish take $c_*=\frac{32}{c_V}$. ■

3.4 Preliminary estimates for the number of parts of size k

This section is devoted to give estimates for the number of classes with fixed number of elements at time n, $N_n(k)$. As in the case for V_n, we investigate the behaviour of $N_n(k)$ properly normalized. In this sense, we let $\psi_n(k)$ be the normalization factor for $N_n(k)$ given by the expression below

$$\psi_n(k) := \prod_{j=1}^{n-1}\left(1-\frac{k-\alpha}{j+\theta}\right)=\frac{\Gamma(k+\theta)\Gamma(n-k+\alpha+\theta)}{\Gamma(\alpha+\theta)\Gamma(n+\theta)}. \tag{3.4.1}$$

We note that, for each k fixed, $\psi_n(k)=\Theta(n^{\alpha-k})$. The proof of this result may be done similarly to that one given to ϕ_n. We also let $X_n(k)$ be

$$X_n(k) := \frac{N_n(k)}{\psi_n(k)}. \tag{3.4.2}$$

The first step in the analysis of the non-asymptotic behavior of $N_n(k)$ is to prove that $X_n(k)$ also satisfies a recurrence relation (Subsection 3.4.1). We then present a martingale concentration argument that will be useful in analyzing the recurrence (Subsection 3.4.2). Subsequent sections will use these results to give upper and lower bounds on $N_n(k)$.

3.4.1 Recurrence relation for $X_n(k)$

The goal of this part is to derive a recurrence relation for $X_n(k)$. The proof is essentially the same we have given for V_n.

Lemma 3.4.1. *For all $n, k \in \mathbb{N}$ the sequence $\{X_n(k)\}_{n\in\mathbb{N}}$ satisfies*

$$X_n(1) = M_n(1) + \sum_{j=1}^{n-1} \frac{\alpha V_j}{(j+\theta)\psi_{j+1}(1)} + \theta_n; \tag{3.4.3}$$

$$X_n(k) = M_n(k) + X_k(k) + \frac{k-1-\alpha}{k-1+\theta} \sum_{j=k}^{n-1} X_j(k-1), \forall k > 1, \tag{3.4.4}$$

where $\{M_n(k)\}_{n\in\mathbb{N}}$ are zero mean martingales defined in (3.4.8) and (3.4.11) for all $k \in \mathbb{N}$ and

$$\theta_n(1) := N_1(1) + \sum_{j=1}^{n-1} \frac{\theta}{(j+\theta)\psi_{j+1}(1)}. \tag{3.4.5}$$

Proof. We treat the case $k = 1$ separately since $X_n(1)$ satisfies a recurrence relation slight different from the other cases. However, the proof for both cases follow the recipe given by the proof of Lemma 3.3.1, so we do not fill all the details here.
Case $k = 1$. Note that $\Delta N_n(1) \in \{-1, 0, 1\}$. Thus, conditioned to $\mathcal{F}_{n-1}$ we know its distribution, which is given by

$$\begin{aligned}
\mathbb{P}\left(\Delta N_n(1) = -1|\mathcal{F}_{n-1}\right) &= \frac{(1-\alpha)N_{n-1}(1)}{n-1+\theta}; \\
\mathbb{P}\left(\Delta N_n(1) = 1|\mathcal{F}_{n-1}\right) &= \frac{\alpha V_{n-1}+\theta}{n-1+\theta}; \\
\mathbb{P}\left(\Delta N_n(1) = 0|\mathcal{F}_{n-1})\right) &= 1 - \mathbb{P}\left(\Delta N_n(1) = -1|\mathcal{F}_{n-1}\right) - \mathbb{P}\left(\Delta N_n(1) = 1|\mathcal{F}_{n-1}\right)
\end{aligned} \tag{3.4.6}$$

Again, as in Lemma 3.3.1 but normalizing properly, define

$$\zeta_n(1) := \frac{1}{\psi_n(1)} \left(\Delta N_n(1) - \frac{\alpha V_{n-1} + \theta - (1-\alpha)N_{n-1}(1)}{n-1+\theta}\right), \tag{3.4.7}$$

$$M_n(1) := \sum_{j=2}^{n} \zeta_j(1), \tag{3.4.8}$$

and observe that the identities (3.4.6) imply that the sequence $\{M_n(1)\}_n \in \mathbb{N}$ is a zero mean martingale. Thus

$$\begin{aligned}
&N_n(1) = N_{n-1}(1) + \Delta N_n(1) \\
\Rightarrow\ &N_n(1) = \left(1 - \frac{1-\alpha}{n-1+\theta}\right) N_{n-1}(1) + \left(\Delta N_n(1) - \frac{\alpha V_{n-1}+\theta-(1-\alpha)N_{n-1}}{n-1+\theta}\right) + \frac{\alpha V_{n-1}+\theta}{n-1+\theta} \\
\Rightarrow\ &\frac{N_n(1)}{\psi_n(1)} = \frac{N_{n-1}(1)}{\psi_{n-1}(1)} + \zeta_n(k) + \frac{\alpha V_{n-1}+\theta}{(n-1+\theta)\psi_n(1)}.
\end{aligned}$$

We recognize above the terms $X_m(1) = N_m(1)/\psi_m(1)$ for $m = n-1, n$. We conclude

$$X_n(1) = M_n(1) + \sum_{j=1}^{n-1} \frac{\alpha V_j}{(j+\theta)\psi_{j+1}(1)} + \theta_n,$$

where

$$\theta_n(1) := N_1(1) + \sum_{j=1}^{n-1} \frac{\theta}{(j+\theta)\psi_{j+1}(1)}. \tag{3.4.9}$$

Case $k > 1$. As before we calculate the conditional distribution of $\Delta N_n(k)$, which is given below.

$$\begin{aligned}
\mathbb{P}\left(\Delta N_n(k) = -1|\mathcal{F}_{n-1}\right) &= \frac{(k-\alpha)N_{n-1}(k)}{n-1+\theta}; \\
\mathbb{P}\left(\Delta N_n(k) = 1|\mathcal{F}_{n-1}\right) &= \frac{(k-1-\alpha)N_{n-1}(k-1)}{n-1+\theta}; \\
\mathbb{P}\left(\Delta N_n(k) = 0|\mathcal{F}_{n-1}\right) &= 1 - \mathbb{P}\left(\Delta N_n(k) = -1|\mathcal{F}_{n-1}\right) - \mathbb{P}\left(\Delta N_n(k) = 1|\mathcal{F}_{n-1}\right);
\end{aligned}$$

Again we centralize and normalize it and define our martingale from its sum:

$$\zeta_n(k) := \frac{\Delta N_n(k) - \frac{(k-1-\alpha)N_{n-1}(k-1)-(k-\alpha)N_{n-1}(k)}{n-1+\theta}}{\psi_n(k)}; \tag{3.4.10}$$

$$M_n(k) := \sum_{j=k+1}^{n} \zeta_j(k). \tag{3.4.11}$$

The relation below between $\psi_n(k-1)$ and $\psi_{n+1}(k)$ will be useful to our purposes:

$$\frac{\psi_n(k-1)}{\psi_{n+1}(k)} = \frac{n+\theta}{k-1+\theta}. \tag{3.4.12}$$

This follows from the definition of $\psi_n(k)$ given at (3.4.1). This relation allows us to derive the desired recurrence relation as follows

$$\begin{aligned} N_n(k) &= \left(1 - \frac{k-\alpha}{n+\theta}\right) N_{n-1}(k) \\ &\quad + \left(\Delta N_n(k) - \frac{(k-1-\alpha)N_{n-1}(k-1) - (k-\alpha)N_{n-1}(k)}{n-1+\theta}\right) \\ &\quad + \frac{(k-1-\alpha)N_{n-1}(k-1)}{n-1+\theta} \\ \Rightarrow \frac{N_n(k)}{\psi_n(k)} &= \frac{N_{n-1}(k)}{\psi_{n-1}(k)} + \zeta_n(k) + \frac{(k-1-\alpha)N_{n-1}(k-1)}{(n-1+\theta)\psi_n(k)}. \end{aligned}$$

We have above the terms $X_m(k) = N_m(k)/\psi_m(k)$ for $m = n, n+1$. The last term in the right-hand side is:

$$\frac{N_{n-1}(k-1)}{\psi_n(k)} = \frac{\psi_{n-1}(k-1)}{\psi_n(k)} X_{n-1}(k-1) = \frac{n-1+\theta}{k-1+\theta} X_{n-1}(k-1) \text{ by (3.4.12).}$$

We deduce:

$$X_n(k) = X_{n-1}(k) + \zeta_n(k) + \frac{k-1-\alpha}{k-1+\theta} X_{n-1}(k-1),$$

from which the recursion follows. ■

We may obtain an upper bound for $\theta_n(1)$ using the bounds for ratios of gamma functions in the Appendix:

$$\begin{aligned} \theta_n(1) &= 1 + \sum_{j=1}^{n-1} \frac{\theta}{(j+\theta)\psi_{j+1}(1)} \\ &= 1 + \frac{\theta\Gamma(\alpha+\theta)}{\Gamma(1+\theta)} \sum_{j=1}^{n-1} \frac{\Gamma(j+\theta)}{\Gamma(j+\theta+\alpha)} \\ \theta_n(1) &\leq 1 + \frac{2\theta\Gamma(\alpha+\theta)}{(1-\alpha)\Gamma(1+\theta)} (n+\theta)^{1-\alpha}. \end{aligned} \tag{3.4.13}$$

This upper bound will be useful latter.

3.4.2 The martingale component of $X_n(k)$

In this subsection we prove a concentration inequality result for a martingale sequence whose increment and quadratic variation satisfy certain hypothesis. Then we prove that the martingale component of $\{X_n(k)\}_n$ satisfies these conditions, for all k, proving then that the martingale component of the $\{X_n(k)\}_n$ is well behaved.

Lemma 3.4.2. *Let $d > 0$ and $k \in \mathbb{N}$ be constants and $\{M_n\}_{n\in\mathbb{N}}$ be a martingale sequence satisfying*

1. $|\Delta M_j| \leq \dfrac{d}{\Gamma(k+\theta)} \cdot (j-1+\theta)^{k-\alpha}$,

2. $\mathbb{E}[(\Delta M_j)^2 | F_{j-1}] \leq \dfrac{d^2 \cdot (2k-\alpha)}{\Gamma(k+\theta)^2} \cdot (j-1+\theta)^{2k-\alpha-1} \cdot \left(\dfrac{V_{j-1}}{\phi_{j-1}} + b_{j-1}\right)$,

then there exists a constant c_M such that

$$\mathbb{P}\left(|M_n - M_m| \geq \frac{\sqrt{2}d}{\Gamma(k+\theta)}(n+\theta)^{k-\frac{\alpha}{2}} A\right) \leq e^{-c_M A}.$$

for all $A \geq \max_j\{b_j\}$.

Proof. Let W_n the quadratic variation of the martingale $\{M_n - M_m\}_n$. By our assumptions:

$$W_n := \sum_{j=m+1}^{n} \mathbb{E}[(\Delta M_j)^2 | \mathcal{F}_j] \leq \frac{d^2 \cdot |2k-\alpha|}{\Gamma(k+\theta)^2} \sum_{j=m+1}^{n} (j-1+\theta)^{2k-\alpha-1} \left(\frac{V_{j-1}}{\phi_{j-1}} + b_{j-1}\right).$$

Moreover, in the occurrence of the event $\left\{\sup_{j\in\mathbb{N}} \left(\frac{V_j}{\phi_j}\right) \leq A\right\}$, and using that $b_j \leq A$ we have

$$W_n \leq \frac{2d^2 \cdot A}{\Gamma(k+\theta)^2}(n+\theta)^{2k-\alpha},$$

in symbols, the following inclusion of events holds

$$\left\{W_n \geq \frac{2d^2 \cdot A}{\Gamma(k+\theta)^2}(n+\theta)^{2k-\alpha}\right\} \subset \left\{\sup_{j\in\mathbb{N}} \left(\frac{V_j}{\phi_j}\right) \geq A\right\},$$

which combined with Theorem 3.3.1 yields

$$\mathbb{P}\left(W_n \geq \frac{2d^2 \cdot A}{\Gamma(k+\theta)^2}(n+\theta)^{2k-\alpha}\right) \leq \exp(-c_V A).$$

Finally, applying Lemma 4.1.1 with $R = \frac{d^2 \cdot A}{\Gamma(k+\theta)^2}(n+\theta)^{2k-\alpha}$ and $c_1 = 1$ we obtain

$$\mathbb{P}\left(|M_n - M_m| \geq \frac{\sqrt{2}d}{\Gamma(k+\theta)}(n+\theta)^{k-\frac{\alpha}{2}} A\right) \leq \exp\left(\frac{-A}{2+\frac{2}{3}}\right) + \exp(-c_V A) \leq \exp(-c_M A)$$

for some constant c_M. ∎

Lemma 3.4.3. *Let $\{M_n(k)\}_n$, $k \geq 1$, be the martingale defined in (3.4.8) and (3.4.11) and $A \geq 0$ a constant. Then there is a constant $h_{\alpha,\theta}$ such that*

$$\mathbb{P}\left(|M_n(k)| \geq \frac{h_{\alpha,\theta}}{\Gamma(k+\theta)}(n+\theta)^{k-\frac{\alpha}{2}} (A + 2\log n)\right) \leq \frac{e^{-A}}{n^2}.$$

Proof. We will prove that the martingales in (3.4.8) and (3.4.11) satisfy the hypotheses of Lemma 3.4.2, and the result will follow from that lemma.

Since $1/\psi_n(k)$ is increasing, by Lemma 4.3.7 in Appendix, the following bound holds

$$|\Delta M_j(1)| \leq \frac{e^{\frac{1}{12}}\Gamma(\alpha+\theta)}{\Gamma(1+\theta)}(n+\theta)^{1-\alpha}.$$

And by definition of $\Delta M_j = \zeta_j$, we also have that

$$\begin{aligned}\mathbb{E}[(\Delta M_j(1))^2|\mathcal{F}_{j-1}] &= \frac{1}{\psi_j(1)^2}\cdot\left[\frac{\alpha V_{j-1}+\theta}{j-1+\theta}\cdot\left(1-\frac{\alpha V_{j-1}+\theta-(1-\alpha)N_{j-1}(1)}{j-1+\theta}\right)^2\right.\\ &\quad\left.+\frac{(1-\alpha)N_{j-1}(1)}{j-1+\theta}\cdot\left(-1-\frac{\alpha V_{j-1}+\theta-(1-\alpha)N_{j-1}(1)}{j-1+\theta}\right)^2\right]\\ &\leq \frac{4}{\psi_j(1)^2}\frac{V_{j-1}+\theta}{j-1+\theta}.\end{aligned}$$

Multiplying and deviding the above expression by ϕ_{j-1} and using the bound

$$\frac{\phi_{j-1}}{(\psi_j(k))^2\cdot(j-1+\theta)} \leq \frac{e^{1/6}\Gamma(1+\theta)\Gamma(\alpha+\theta)^2}{\Gamma(1+\theta+\alpha)\Gamma(k+\theta)^2},(j-1+\theta)^{1-\alpha},$$

which may be deduced from see Lemma 4.3.8 in appendix, it follows that

$$\begin{aligned}\mathbb{E}[(\Delta M_j(1))^2|\mathcal{F}_{j-1}] &\leq \frac{4\phi_{j-1}}{(j-1+\theta)\psi_j(1)^2}\frac{V_{j-1}+\theta}{\phi_{j-1}}\\ &\leq \frac{4e^{1/6}\Gamma(1+\theta)\Gamma(\alpha+\theta)^2}{\Gamma(1+\theta+\alpha)\Gamma(1+\theta)^2}(j-1+\theta)^{1-\alpha}\frac{V_{j-1}+\theta}{\phi_{j-1}},\end{aligned}$$

and since $2-\alpha>1$, it also follows that

$$\mathbb{E}[(\Delta M_j(1))^2|\mathcal{F}_{j-1}] \leq \frac{4(2-\alpha)e^{1/6}\Gamma(1+\theta)\Gamma(\alpha+\theta)^2}{\Gamma(1+\theta+\alpha)\Gamma(1+\theta)^2}(j-1+\theta)^{1-\alpha}\frac{V_{j-1}+\theta}{\phi_{j-1}}.$$

Analogously, for $k>1$, we have

$$\begin{aligned}&\mathbb{E}[(\Delta M_j(k))^2|\mathcal{F}_{j-1}]\\ &= \frac{1}{\psi_j^2(k)}\cdot\left[\frac{N_{j-1}(k-1)(k-1-\alpha)}{j-1+\theta}\cdot\left(1-\frac{(k-1-\alpha)N_{j-1}(k-1)-(k-\alpha)N_n(k)}{j-1+\theta}\right)^2\right.\\ &\quad\left.+\frac{N_{j-1}(k)(k-\alpha)}{j-1+\theta}\cdot\left(-1-\frac{(k-1-\alpha)N_{j-1}(k-1)-(k-\alpha)N_{j-1}(k)}{j-1+\theta}\right)^2\right]\\ &\leq \frac{4}{\psi_j^2(k)(j-1+\theta)}[N_{j-1}(k-1)(k-1-\alpha)+N_{j-1}(k)(k-\alpha)].\end{aligned}$$

Since $N_j(k)$ is bounded from above by V_j, for all k and j, we obtain

$$\begin{aligned}\mathbb{E}[(\Delta M_j(k))^2|\mathcal{F}_{j-1}] &\leq \frac{4\phi_{j-1}}{(\psi_j(k))^2\cdot(j-1+\theta)}\left[\frac{V_{j-1}}{\phi_{j-1}}\cdot(k-1-\alpha)+\frac{V_{j-1}}{\phi_{j-1}}\cdot(k-\alpha)\right]\\ &\leq \frac{4(2k-\alpha)\Gamma(1+\theta)\Gamma(\alpha+\theta)^2}{\Gamma(1+\theta+\alpha)\Gamma(k+\theta)^2}(j-1+\theta)^{2k-\alpha-1}\frac{V_{j-1}}{\phi_{j-1}}.\end{aligned}$$

Finally, by Lemma 3.4.2 we have, for

$$h_{\alpha,\theta} = \frac{2\sqrt{2}e^{\frac{1}{12}}}{c_M}\cdot\Gamma(\alpha+\theta)\cdot\max\left\{1,2\sqrt{\frac{\Gamma(1+\theta)}{\Gamma(1+\theta+\alpha)}}\right\},$$

that

$$\mathbb{P}\left(|M_n(k)|\geq\frac{h_{\alpha,\theta}}{\Gamma(k+\theta)}(n+\theta)^{k-\frac{\alpha}{2}}(A+\log^2 n)\right)\leq\frac{e^{-A}}{n^2},$$

as we desired. ∎

3.5 Bounds for the number of parts with size k

Let us go through what we did in Section 3.4. In Subsection 3.4.1 we found recurrence relations relating the values $X_n(k) = N_n(k)/\psi_n(k)$ for different k and n. This is the content of Lemma 3.4.1, where we obtained that:

$$\begin{aligned} X_n(1) &= M_n(1) + \sum_{j=1}^{n-1} \frac{\alpha V_j}{(j+\theta)\psi_{j+1}(1)} + \theta_n(1); \\ X_n(k) &= M_n(k) + X_k(k) + \frac{k-1-\alpha}{k-1+\theta} \sum_{j=k}^{n-1} X_j(k-1), \forall k > 1. \end{aligned}$$

The terms $M_n(k)$ above are martingales. Subsection 3.4.2 proves that the martingale terms are all small. Since we already know $V_j/\phi_j \approx V_*$ for j large, this will lead to bounds of the form:

$$\begin{aligned} X_n(1) &\approx a_0(1)\, V_*; \\ X_n(k) &\approx \frac{k-1-\alpha}{k-1+\theta} \sum_{j=k}^{n-1} X_j(k-1), \forall k > 1 \end{aligned}$$

where

$$a_0(1) := \frac{\alpha}{\alpha+\theta}.$$

If we treat the above recursions as equalties, we then obtain by induction in k that

$$X_n(k) \approx a_0(k)\, V_*\, n^k$$

where

$$a_0(k) = \frac{(k-1-\alpha)\cdot a_0(k-1)}{(k-1+\theta)k} = \frac{\Gamma(k-\alpha)\Gamma(1+\theta)}{k!\cdot\Gamma(1-\alpha)\Gamma(k+\theta)} a_0(1).$$

The purpose of this section is to make the above approximations precise and to show that $X_n(k)$ does behave as expected up to leading order, in high probability. In particular, we will prove the following Theorem (recall the definition of $X_n(k)$ in (3.4.2)).

Theorem 3.5.1. *Given $A > K(\alpha,\theta)$, $n \in \mathbb{N}$ and $k \le n$, there are coefficients $a_0(k)$ (defined above) and $a_1(k)$ with $a_1(k) = O\left(a_0(k)\cdot k^{\alpha+2}\right)$, such that the following holds. Define the event where $X_m(s)$ is "well-controlled from above".*

$$F^{(up)}_{m,s} := \Big\{ X_m(s) \le a_0(s) V_* (m-1)^s + a_1(s)(m+\theta)^{s-\alpha/2}(A+\log n) \Big\}.$$

Similarly, define the event that $X_m(s)$ is "well-controlled from below".

$$F^{(dn)}_{m,s} := \Big\{ X_m(s) \ge a_0(s) V_* (m-s)^s - a_1(s)(m+\theta)^{s-\alpha/2}(A+\log n) \Big\}.$$

Finally, define the event where the above inequalities hold for all times $m \le n$ and part sizes $s \le k$:

$$E_{n,k} := \bigcap_{m\le n} \bigcap_{s\le k} (F^{(up)}_{m,s} \cap F^{(dn)}_{m,s}).$$

Then:

$$\mathbb{P}(E_{n,k}) \ge 1 - \frac{k}{n} e^{-A}.$$

As we will see, this theorem follows directly from the results in the remainder of this section.

Proof. [Proof of Theorem 3.5.1] The bound $a_1(k) = O\left(a_0(k)\cdot k^{\alpha+2}\right)$ is contained in Lemma 3.5.1 in subsection 3.5.1. The probability of $E_{n,k}$ is bounded in Lemmas 3.5.2 and 3.5.3 in subsection 3.5.2. ∎

3.5.1 The choice of coefficients

The coefficients $a_0(1)$ and $a_1(1)$ will arise from the analysis of the recursion (3.4.3) in the Lemma 3.4.1. As we have seen, $a_0(k)$ appears naturally when we work out the leading order terms for $X_n(k)$. The extra coefficient $a_1(k)$ controls the error, and comes from combining errors in estinating $X_s(k-1)$ (induction step); the error in setting $M_n(k) \approx 0$; and various other estimates in the proof (see (3.4.4) and Lemma 3.5.1).

We define:

$$a_0(1) := \frac{\alpha}{\alpha+\theta}, \tag{3.5.1}$$

$$a_1(1) := \frac{h_{\alpha,\theta}}{\Gamma(1+\theta)} + \frac{\alpha}{c_V(\alpha+\theta)(1-\frac{\alpha}{2})} + 1 + \frac{2\theta\Gamma(\alpha+\theta)}{(1-\alpha)\Gamma(1+\theta)}, \tag{3.5.2}$$

$$a_0(k) := \frac{(k-1-\alpha)\cdot a_0(k-1)}{(k-1+\theta)k} = \frac{\Gamma(k-\alpha)\Gamma(1+\theta)}{k!\cdot\Gamma(1-\alpha)\Gamma(k+\theta)}a_0(1), \tag{3.5.3}$$

$$a_1(k) := \frac{h_{\alpha,\theta}}{\Gamma(k+\theta)} + \frac{(k-1-\alpha)\cdot a_1(k-1)}{(k-1+\theta)(k-\frac{\alpha}{2})}. \tag{3.5.4}$$

From the analysis of recursions involving $X_n(k)$, it will arise naturally terms which are polynomials whose coefficients are the above coefficients. Thus, it will be useful to have estimates for such polynomials as well. We do this in the next lemma.

Lemma 3.5.1. *The coefficients $a_0(k)$ and $a_1(k)$ defined as in (3.5.3) and (3.5.4) satisfy the following relations:*

1. $\dfrac{k-1-\alpha}{k-1+\theta}\displaystyle\sum_{j=k}^{m-1} a_0(k-1)j^{k-1} \quad \le a_0(k)m^k,$

2. $\dfrac{k-1-\alpha}{k-1+\theta}\displaystyle\sum_{j=k}^{m-1} a_0(k-1)(j-(k-1))^{k-1} \quad \ge a_0(k)(m-k)^k,$

3. $\dfrac{k-1-\alpha}{k-1+\theta}\displaystyle\sum_{j=k}^{m-1} a_1(k-1)(j+\theta)^{k-1-\alpha/2} \le \left(a_1(k) - \frac{h_{\alpha,\theta}}{\Gamma(k+\theta)}\right)(m+\theta)^{k-\alpha/2},$

4. $a_1(k) \le C_U a_0(k)\cdot k^{\alpha+2}$, *for some constant C_U.*

Proof. Throughout this proof we will make use of the integral bound below

$$\frac{(m-k)^k}{k} = \int_0^{m-k} x^{k-1} \le \sum_{j=1}^{m-1} j^{k-1} \le \int_0^m x^{k-1} = \frac{m^k}{k}. \tag{3.5.5}$$

(1) For the first bound, observe that

$$\frac{k-1-\alpha}{k-1+\theta}\sum_{j=k}^{m-1} a_0(k-1)j^{k-1} \le \frac{k-1-\alpha}{k-1+\theta}a_0(k-1)\sum_{j=1}^{m-1} j^{k-1}.$$

Using the upper bound given by (3.5.5), yields

$$\frac{k-1-\alpha}{k-1+\theta}\sum_{j=\ell+1}^{m-1} a_0(k-1)j^{k-1} \le \frac{(k-1-\alpha)\cdot a_0(k-1)}{(k-1+\theta)k}$$

which is exactly the definition of $a_0(k)$ in (3.5.3).

(2) For the second relation, we have

$$\frac{k-1-\alpha}{k-1+\theta}\sum_{j=k}^{m-1} a_0(k-1)(j-(k-1))^{k-1} = \frac{k-1-\alpha}{k-1+\theta}\sum_{j=1}^{m-k} a_0(k-1)j^{k-1}$$
$$\geq \frac{k-1-\alpha}{k-1+\theta}a_0(k-1)\sum_{j=1}^{m-k} j^{k-1}.$$

By the lower bound given by (3.5.5), we obtain

$$\frac{k-1-\alpha}{k-1+\theta}\sum_{j=1}^{m-k} a_0(k-1)j^{k-1} \geq \frac{(k-1-\alpha)\cdot a_0(k-1)}{(k-1+\theta)k}(m-k)^k = a_0(k)(m-k)^k.$$

(3) If we proceed exactly as in the item (1) we obtain

$$\frac{k-1-\alpha}{k-1+\theta}\sum_{j=k}^{m-1} a_1(k-1)(j+\theta)^{k-1-\alpha/2} \leq \frac{(k-1-\alpha)\cdot a_1(k-1)}{(k-1+\theta)(k-\frac{\alpha}{2})}(m+\theta)^{k-\alpha/2},$$

but by definition (3.5.4)

$$\frac{(k-1-\alpha)\cdot a_1(k-1)}{(k-1+\theta)(k-\frac{\alpha}{2})} = \left(a_1(k) - \frac{h_{\alpha,\theta}}{\Gamma(k+\theta)}\right).$$

(4) We begin substituting the formulas for $a_0(k)$ and $a_1(0)$ in an analogous way we did above, to obtain an affine recurrence

$$\frac{a_1(k)}{a_0(k)} = d\cdot\frac{k!}{\Gamma(k-\alpha)} + \frac{k\cdot a_1(k-1)}{(k-\frac{\alpha}{2})\cdot a_0(k-1)}, \tag{3.5.6}$$

where d is defined as

$$d := d_{\alpha,\theta}\cdot\frac{\Gamma(1-\alpha)}{a_0(1)\cdot\Gamma(1+\theta)}.$$

We rearrange (3.5.6) by letting $s(k)$ to be

$$s(k) := \frac{a_1(k)}{a_0(k)}\frac{\Gamma(k-\frac{\alpha}{2}+1)}{\Gamma(k+1)}$$

and multiplying both sides by $\frac{\Gamma(k-\frac{\alpha}{2}+1)}{\Gamma(k+1)}$ to obtain the identity below

$$s(k) = s(k-1) + d\cdot\frac{\Gamma(k-\frac{\alpha}{2}+1)}{\Gamma(k-\alpha)},$$

so we can find the general formula to the recurrence

$$s(k) = s(1) + d\cdot\sum_{j=1}^{k}\frac{\Gamma(j-\frac{\alpha}{2}+1)}{\Gamma(j-\alpha)}.$$

Using the bound $\frac{\Gamma(j-\frac{\alpha}{2}+1)}{\Gamma(j-\alpha)} \leq e^{1/12}j^{1+\frac{\alpha}{2}}$ we have

$$s(k) = s(1) + e^{1/12}d\cdot\sum_{j=1}^{k} j^{\frac{\alpha}{2}+1} \leq s(1) + e^{1/12}d\cdot k^{\frac{\alpha}{2}+2} \leq d_1 k^{\frac{\alpha}{2}+2},$$

where $d_1 = 2\max\{s(1), e^{1/12}d\}$. Finally, we obtain

$$\frac{a_1(k)}{a_0(k)} \leq d_1\frac{\Gamma(k+1)}{\Gamma(k-\frac{\alpha}{2}+1)}k^{\frac{\alpha}{2}+2} = C_U k^{\alpha+2},$$

for some C_U.

■

3.5.2 Bound on $X_n(k)$

We now bound the probability of the events $E_{n,k}$ defined in the statement of Theorem 3.5.1. Our approach is induction on k. But before we go to the proof, let us recall the definition of the sequence of events $E_{n,k}$. The event $F_{m,s}^{(up)}$ is defined as the event where $X_m(s)$ is "well-controlled from above"

$$F_{m,s}^{(up)} = \Big\{ X_m(s) \leq a_0(s) V_* (m-1)^s + a_1(s)(m+\theta)^{s-\alpha/2}(A + \log n) \Big\}.$$

Analogously, $F_{m,s}^{(dn)}$ is the event where $X_m(s)$ is "well-controlled from below"

$$F_{m,s}^{(dn)} := \Big\{ X_m(s) \geq a_0(s) V_* (m-s)^s - a_1(s)(m+\theta)^{s-\alpha/2}(A + \log n) \Big\}.$$

Finally, the event $E_{n,k}$ is the event where the above inequalities hold for all times $m \leq n$ and part sizes $s \leq k$:

$$E_{n,k} := \bigcap_{m \leq n} \bigcap_{s \leq k} (F_{m,s}^{(up)} \cap F_{m,s}^{(dn)}).$$

Now, we start by the case $k = 1$.

Lemma 3.5.2 (Case $k = 1$)**.** *Given $A > 0$ and $n \in \mathbb{N}$, let $E_{n,1}$ be as in the statement of Theorem 3.5.1. Then:*

$$\mathbb{P}(E_{n,1}) \geq 1 - \frac{e^{-A}}{n}.$$

Proof. The equation (3.4.3) says us that

$$X_n(1) = M_n(1) + \sum_{j=1}^{n-1} \frac{\alpha V_j}{(j+\theta)\psi_{j+1}} + \theta_n(1).$$

We will bound each term in the right-hand side to obtain a bound on $X_n(1)$. Before, we manipulate algebraically the above expression for $X_n(1)$ in such way it can be expressed in terms of the observables we already know how to control. In this direction, we start summing and subtracting the sum below

$$\sum_{j=1}^{n-1} \frac{\alpha \phi_j}{(j+\theta)\psi_{j+1}} V_*$$

in the second member of (3.4.3) to use that the ratio V_j/ϕ_j is approximated by V_*. This yields

$$X_n(1) = M_n(1) + \sum_{j=1}^{n-1} \frac{\alpha \phi_j}{(j+\theta)\psi_{j+1}} \left(\frac{V_j}{\phi_j} - V_* \right) + \sum_{j=1}^{n-1} \frac{\alpha \phi_j}{(j+\theta)\psi_{j+1}} V_* + \theta_n(1). \tag{3.5.7}$$

Using the relation below

$$\frac{\phi_j}{(j+\theta)\psi_{j+1}(1)} = \frac{1}{(\theta+\alpha)}$$

on identity (3.5.7) allows us to obtain

$$X_n(1) = M_n(1) + \sum_{j=1}^{n-1} \frac{\alpha}{(\theta+\alpha)} \left(\frac{V_j}{\phi_j} - V_* \right) + \frac{\alpha V_*}{(\theta+\alpha)}(n-1) + \theta_n(1).$$

Taking the absolute value on both sides of the above identity and using the triangle inequality yields

$$|X_n(1)| \leq |M_n(1)| + \left| \sum_{j=1}^{n-1} \frac{\alpha}{(\theta+\alpha)} \left(\frac{V_j}{\phi_j} - V_* \right) \right| + \frac{\alpha V_*}{(\theta+\alpha)}(n-1) + \theta_n(1). \tag{3.5.8}$$

and

$$|X_n(1)| \geq \frac{\alpha V_*}{(\theta+\alpha)}(n-1) - |M_n(1)| - \left|\sum_{j=1}^{n-1} \frac{\alpha}{(\theta+\alpha)}\left(\frac{V_j}{\phi_j} - V_*\right)\right|. \tag{3.5.9}$$

By Lemma 3.4.3, the probability of the event below

$$\left\{|M_n(1)| \geq \frac{h_{\alpha,\theta}}{\Gamma(1+\theta)}(n+\theta)^{1-\frac{\alpha}{2}}(A+\log n)\right\} \tag{3.5.10}$$

is bounded from above by

$$\mathbb{P}\left(|M_n(1)| \geq \frac{h_{\alpha,\theta}}{\Gamma(1+\theta)}(n+\theta)^{1-\frac{\alpha}{2}}(A+\log n)\right) \leq \frac{e^{-A}}{n^2}, \tag{3.5.11}$$

and by Corollary 3.2.1, with $\delta = \frac{e^{-A}}{n^2}$ and observing that $\log m \leq \log n$, for $1 \leq m \leq n-1$ we have

$$\mathbb{P}\left(\left|\frac{V_j}{\phi_j} - V_*\right| \geq \frac{A+\log n}{c_V(j+\theta)^{\alpha/2}}, \text{ for some } 1 \leq j \leq n-1\right) \leq \frac{e^{-A}}{n^2}. \tag{3.5.12}$$

On the occurrence of the event

$$\left\{\left|\frac{V_j}{\phi_j} - V_*\right| \leq \frac{A+\log n}{c_V(m+\theta)^{\alpha/2}}, \text{ for some } 1 \leq j \leq n-1\right\} \tag{3.5.13}$$

we have

$$\left|\sum_{j=1}^{n-1} \frac{\alpha}{(\theta+\alpha)}\left(\frac{V_j}{\phi_j} - V_*\right)\right| \leq \sum_{j=1}^{n-1} \frac{\alpha}{(\theta+\alpha)} \frac{A+\log n}{c_V(j+\theta)^{\alpha/2}} \tag{3.5.14}$$

$$\leq \frac{(n-1+\theta)^{1-\alpha/2}}{1-\alpha/2} \frac{\alpha(A+\log n)}{c_V(\theta+\alpha)}. \tag{3.5.15}$$

By (3.4.13), the term $\theta_n(1)$ is bounded in the following way

$$\begin{aligned}\theta_n(1) &\leq 1 + \frac{2\theta\Gamma(\alpha+\theta)}{(1-\alpha)\Gamma(1+\theta)}(n+\theta)^{1-\alpha} \\ &\leq \left(1 + \frac{2\theta\Gamma(\alpha+\theta)}{(1-\alpha)\Gamma(1+\theta)}\right)(n+\theta)^{1-\frac{\alpha}{2}}(A+\log n).\end{aligned}$$

Thus, on the occurrence of both events (3.5.10) and (3.5.13), we have

$$|X_n(1)| \leq a_0(1)V_*(j-1) + a_1(1)(j+\theta)^{1-\alpha/2}(A+\log n)$$

for $a_0(1)$ and $a_1(1)$ whose definition we recall below

$$a_0(1) = \frac{\alpha}{\alpha+\theta}, \tag{3.5.16}$$

$$a_1(1) = \frac{h_{\alpha,\theta}}{\Gamma(1+\theta)} + \frac{\alpha}{c_V(\alpha+\theta)(1-\frac{\alpha}{2})} + 1 + \frac{2\theta\Gamma(\alpha+\theta)}{(1-\alpha)\Gamma(1+\theta)}. \tag{3.5.17}$$

Therefore

$$\mathbb{P}((E_{n,1})^c) \leq \sum_{j=1}^{n} \mathbb{P}\left(|M_n(j)| \geq \frac{h_{\alpha,\theta}}{\Gamma(j+\theta)}(n+\theta)^{j-\frac{\alpha}{2}}(A+\log n)\right) \leq \frac{e^{-A}}{n},$$

which proves the first step of the induction. ■

Now we prove on the next lemma the inductive step.

Lemma 3.5.3 (The inductive step)**.** *Given $A > 0$, $n \in \mathbb{N}$ and $k \leq n$, let $E_{n,k}$ be as in the statement of Theorem 3.5.1. Then:*

$$\mathbb{P}(E_{n,k}) \geq 1 - \frac{k}{n}e^{-A}.$$

Proof. The key step of the proof is the following inclusion of events

$$E_{n,k} \supset E_{n,k-1} \cap \left\{ |M_j(k)| \leq \frac{h_{\alpha,\theta}}{\Gamma(k+\theta)}(j+\theta)^{k-\frac{\alpha}{2}}(A+\log n), \text{ for all } 1 \leq j \leq n \right\}, \tag{3.5.18}$$

for all $k \geq 2$. The result then follows by induction from our previous results and the inequality below

$$\begin{aligned}\mathbb{P}(E_{n,k}^c) &\leq \mathbb{P}(E_{n,k-1}^c) + \sum_{j=1}^{n} \mathbb{P}\left(|M_j(k)| \geq \frac{h_{\alpha,\theta}}{\Gamma(k+\theta)}(j+\theta)^{k-\frac{\alpha}{2}}(A+\log n) \right) \\ &\leq (k-1)\frac{e^{-A}}{n} + n\frac{e^{-A}}{n^2}.\end{aligned}$$

Let us then explain why (3.5.18) holds. At a high level, when event $E_{n,k-1}$ occurs,we have that all $X_j(s)$ are "well-behaved" for all values of $s \leq k-1$ and $1 \leq j \leq n$. Now we will prove that combining this with bounds on the martingale component of $X_n(k)$, $X_n(k)$ itself will be "well-behaved". To do that, we will just bound the recursion for $X_n(k)$ using the bounds given by the above events and Lemma 3.5.1.

We start by restating the recursion for a fixed m:

$$X_m(k) = M_m(k) + X_k(k) + \frac{k-1-\alpha}{k-1+\theta}\sum_{j=k}^{m-1} X_j(k-1). \tag{3.5.19}$$

We let $Z_m(k-1)$ denote the sum in the RHS of the previous display.

$$Z_m(k-1) := \frac{k-1-\alpha}{k-1+\theta}\sum_{j=k}^{m-1} X_j(k-1). \tag{3.5.20}$$

In the event

$$E_{n,k-1} \cap \left\{ |M_j(k)| \leq \frac{h_{\alpha,\theta}}{\Gamma(k+\theta)}(j+\theta)^{k-\frac{\alpha}{2}}(A+\log n), \quad \text{for all } 1 \leq j \leq n \right\}$$

we have that each $X_j(k-1)$ is bounded by

$$X_j(k-1) \leq a_0(k-1)V_* j^{k-1} + a_1(k-1)(j+\theta)^{k-1-\alpha/2}(A+\log n),$$

which implies the following bound

$$Z_m(k-1) \leq \frac{k-1-\alpha}{k-1+\theta}\sum_{j=k}^{m-1}\left[a_0(k-1)V_* j^{k-1} + a_1(k-1)(j+\theta)^{k-1-\alpha/2}(A+\log n)\right]. \tag{3.5.21}$$

Now, recall that Lemma 3.5.1 gives us bounds on the polynomials on j whose coefficients are $a_i(k-1)$. Thus, combining this with the above bound we obtain

$$Z_m(k-1) \leq a_0(k)V_* m^k + \left(a_1(k) - \frac{h_{\alpha,\theta}}{\Gamma(k+\theta)}\right)(m+\theta)^{k-\alpha/2}(A+\log n). \tag{3.5.22}$$

Arguing the same way, but applying the lower bound to $X_j(k-1)$ given by $E_{n,k-1}$ instead we may obtain that $Z_m(k-1)$ is bounded from below by

$$\frac{k-1-\alpha}{k-1+\theta}\sum_{j=k}^{m-1}\left[a_0(k-1)V_*(j-(k-1))^{k-1}-a_1(k-1)(A+2\log n)(j+\theta)^{k-1-\alpha/2}\right].$$

And again, by Lemma 3.5.1 we have

$$Z_m(k-1)\geq a_0(k)V_*(j-k)^k-\left(a_1(k)-\frac{h_{\alpha,\theta}}{\Gamma(k+\theta)}\right)(m+\theta)^{k-\alpha/2}(A+2\log n). \tag{3.5.23}$$

On the other hand, since, for all k we also have

$$X_m(k)\geq M_m(k)+\frac{k-1-\alpha}{k-1+\theta}\sum_{j=k}^{m-1}X_j(k-1),$$

the result then follows by joining (3.5.22) and (3.5.23) with the martingale bound given by the other event in the intersection. ■

3.6 Proof of Theorem 3.2.2

This section is devote to the proof of Theorem 3.2.2 which ensures bounds to the number of parts of size k itself. **Proof.** [Proof of Theorem 3.2.2] First observe that in the event $E_{n,k}$ we have

$$X_n(k)-a_0(k)V_*n^k\leq a_1(k)(n+\theta)^{k-\alpha/2}(A+\log n).$$

Consequently, by lemma 3.5.1

$$X_n(k)-a_0(k)V_*n^k\leq C_Ua_0(k)\cdot k^{\alpha+2}(n+\theta)^{k-\alpha/2}(A+\log n).$$

By the same argument we also have the lower bound

$$X_n(k)\geq a_0(k)V_*(n-k)^k-C_Ua_0(k)(n+\theta)^{k-\alpha/2}(A+\log n).$$

Moreover, note that

$$(n-k)^k=n^k\left(1-\frac{k}{n}\right)^k.$$

By Bernoulli's inequality,

$$(1+x)^m\geq 1+mx$$

for all $x\geq -1$ and $m\in\mathbb{N}$. Then, for $x=-k/n\geq -1$ and $m=k$ we have

$$n^k\left(1-\frac{k}{n}\right)^k\geq n^k\left(1-\frac{k^2}{n}\right)=n^k-k^2n^{k-1}.$$

Thus

$$X_n(k)\geq a_0(k)V_*(n^k-k^2n^{k-1})-C_Ua_0(k)(n+\theta)^{k-\alpha/2}(A+\log n),$$

which implies

$$X_n(k)-a_0(k)V_*n^k\geq -a_0(k)V_*k^2n^{k-1}-C_Ua_0(k)\cdot k^{\alpha+2}(n+\theta)^{k-\alpha/2}(A+\log n).$$

Moreover, on the occurrence of the event

$$E_* := \left\{ V_* \leq \frac{2(A+\log n)}{c_V} \right\}$$

we also have

$$\begin{aligned} X_n(k) - a_0(k)V_* n^k &\geq -a_0(k)\left(\frac{2}{c_V}k^2 n^{k-1} + C_U \cdot k^{\alpha+2}(n+\theta)^{k-\alpha/2}\right)(A+\log n) \\ &\geq -a_0(k)D \cdot k^{\alpha+2}(n+\theta)^{k-\alpha/2}(A+\log n), \end{aligned}$$

where $D := 2c_V^{-1} + C_U$. Thus, on the intersection of $E_{n,k}$ and E_*, we have

$$|X_n(k) - a_0(k)V_* n^k| \leq D a_0(k) k^{\alpha+2}(n+\theta)^{k-\alpha/2}(A+\log n). \tag{3.6.1}$$

To simplify our writing, define

$$f_n(k) := a_0(k) \cdot \psi_n(k) \cdot n^k. \tag{3.6.2}$$

Multiplying both sides of (3.6.1) by $\psi_n(k)$ we have

$$|N_n(k) - f_n(k)V_*| \leq D f_n(k) k^{\alpha+2} \frac{(n+\theta)^{k-\alpha/2}}{n^k}(A+\log n).$$

Now, using that $1 + x \leq e^x$, we have

$$(n+\theta)^\gamma \leq e^{\frac{\theta\gamma}{n}} n^\gamma, \tag{3.6.3}$$

which implies, for $k < n/\theta$

$$|N_n(k) - f_n(k)V_*| \leq eD f_n(k) \frac{k^{\alpha+2}}{n^{\alpha/2}}(A+\log n).$$

Recalling the definition of $a_0(k)$

$$a_0(k) = \frac{\Gamma(k-\alpha)\Gamma(1+\theta)}{k! \cdot \Gamma(1-\alpha)\Gamma(k+\theta)} a_0(1) \tag{3.6.4}$$

and replacing it and $\psi_n(k)$ on $f_n(k)$ it may be written as

$$f_n(k) = \left[\frac{\alpha\Gamma(1+\theta)}{\Gamma(1-\alpha)\Gamma(\alpha+\theta+1)}\frac{\Gamma(k-\alpha)}{\Gamma(k+1)}\right] \cdot \left[\frac{\Gamma(n-k+\alpha+\theta)}{\Gamma(n+\theta)}\right] n^k.$$

By Lemma 4.3.5 in the Appendix, for k of order $n^{\alpha/(2\alpha+4)}$, we have

$$\left[\frac{\Gamma(n-k+\alpha+\theta)}{\Gamma(n+\theta)}\right] = \frac{1}{n^{k-\alpha}}\left(1 + O\left(\frac{k^2}{n-k}\right)\right), \tag{3.6.5}$$

which implies

$$f_n(k) = \left[\frac{\alpha\Gamma(1+\theta)}{\Gamma(1-\alpha)\Gamma(\alpha+\theta+1)}\frac{\Gamma(k-\alpha)}{\Gamma(k+1)}\right]\left(1 + O\left(\frac{k^2}{n}\right)\right) n^\alpha. \tag{3.6.6}$$

Now, by the above identity, we have that

$$N_n(k_n) - c_{\alpha,\theta}\frac{\Gamma(k-\alpha)}{\Gamma(k+1)} \cdot n^\alpha \cdot V_* = N_n(k_n) - f_n(k)V_* + \frac{\Gamma(k-\alpha)}{\Gamma(k+1)} O\left(\frac{k^2}{n}\right) \cdot n^\alpha \cdot V_* \tag{3.6.7}$$

Also, observe that

$$\frac{\Gamma(k-\alpha)}{\Gamma(k+1)} \leq e^{\frac{1}{12}}\left(1 + \frac{1+\alpha}{k-\alpha}\right)^{1/2}\left(\frac{1}{k-\alpha}\right)^{1+\alpha} \leq \frac{4}{k^{1+\alpha}}. \tag{3.6.8}$$

Applying the triangle inequality on (3.6.7), recalling we are inside E_* and using the above upper bound, we obtain that

$$\left| N_n(k) - c_{\alpha,\theta} \frac{\Gamma(k-\alpha)}{\Gamma(k+1)} \cdot n^\alpha \cdot V_* \right| \leq D_2 \left(\frac{k^{\alpha+2}}{n^{\alpha/2}} f_n(k) + \left(\frac{k}{n} \right)^{1-\alpha} \right) (A + \log n).$$

for some positive constant D_2. Finally, for every k satisfying

$$k \leq \frac{\varepsilon n^{\frac{\alpha}{2\alpha+4}}}{(\log n)^{\frac{1}{\alpha+2}}}$$

there is another absolute constant C such that

$$\begin{aligned} \left| N_n(k) - c_{\alpha,\theta} \frac{\Gamma(k-\alpha)}{\Gamma(k+1)} \cdot n^\alpha \cdot V_* \right| &\leq C \frac{\Gamma(k-\alpha)}{\Gamma(k+1)} \cdot n^\alpha \frac{\left(\frac{\varepsilon n^{\frac{\alpha}{2\alpha+4}}}{(\log n)^{\frac{1}{\alpha+2}}} \right)^{\alpha+2}}{n^{\alpha/2}} (A + \log n) \\ &\leq C \frac{\Gamma(k-\alpha)}{\Gamma(k+1)} \cdot n^\alpha \cdot \varepsilon^{\alpha+2} \cdot \left(\frac{A}{\log n} + 1 \right), \end{aligned}$$

proving our main theorem. ■

3.7 Final remarks

The main open problem that could be addressed by our methods is to push the analysis to larger values of k. We conjecture that a tighter analysis would work for all $k = o(n^{\alpha/(1+\alpha)})$ or some similar range. This is in the spirit of the recent paper by Brightwell and Luczak [4]. There the authors analyze the degree distribution of a preferential attachment tree nearly all the way to the maximum degree. Proving something similar in our setting would require modifications in Lemma 3.4.2, where the quadratic variation of the martingale for $N_n(k)$ is controlled in a wasteful manner via V_n.

Another kind of question is to study the distribution of the largest part sizes in $\mathcal{P}_n$. We would like to obtain such results and apply them to the "Hollywood model" of complex networks recently proposed by Crane and Dempsey [10].

Chapter 4

Technical tools

In this chapter we recall some basic concentration results used in chapter 2 and some gamma function properties used in chapter 3.

4.1 Concentration inequalities

Theorem 4.1.1 (Bounded differences inequality)**.** *Let* $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ *be a family of independent random variables with* X_k *taking values in a set* A_k *for each* k*. Suppose that the real-valued function* g *defined on* $\prod A_k$ *satisfies*

$$|g(\mathbf{x}) - g(\mathbf{x}')| \leq c_k$$

whenever the vectors $\mathbf{x}$ *and* $\mathbf{x}'$ *differ only in the* k*-th co-ordinate. Then for any* $t \geq 0$*,*

$$\mathbb{P}(|g(\mathbf{X}) - \mathbb{E}[g(\mathbf{X})]| \geq t) \leq e^{-2t^2/\sum c_k^2}$$

We recall here Freedman's inequality and a particular corollary that will be important to our proofs.

Theorem 4.1.2 (Freedman's Inequality [17])**.** *Let* $(M_n, \mathcal{F}_n)_{n\geq 1}$ *be a martingale with* $M_0 = 0$ *and* $R > 0$ *a constant. Write*

$$W_n := \sum_{k=2}^{n} \mathbb{E}[(\Delta M_j)^2 | \mathcal{F}_j].$$

Suppose

$$|\Delta M_j| \leq R, \quad \text{for all } j.$$

Then, for all $\lambda > 0$ *we have*

$$\mathbb{P}(M_n \geq \lambda, W_n \leq \sigma^2) \leq \exp\left(\frac{-\lambda^2}{2\sigma + 2R\lambda/3}\right).$$

The lemma below is a straightforward consequence of Freedman's inequality. Since we will deal with the problem of bounding martingales under some constraints frequently, it will be convenient to have this precise statement.

Lemma 4.1.1. *Let* M_j *is a martingale and* $R > 0$ *a constant such that,* $M_0 = 0$*,* $|M_{j+1} - M_j| \leq R \ \ \forall j \leq n$ *and* W_n *is its quadratic variation, then for any constant* $c_1 > 0$ *we have*

$$\mathbb{P}\left(|M_n| \geq R\lambda\right) \leq 2\exp\left(\frac{-\lambda}{2c_1 + \frac{2}{3}}\right) + \mathbb{P}\left(W_n \geq c_1 R^2 \lambda\right).$$

Proof. It follows of the union bound and Freedman's inequality to the martingales M_j and $-M_j$:

$$\begin{aligned}\mathbb{P}\left(|M_n| \geq R\lambda\right) &\leq \mathbb{P}\left(|M_n| \geq R\lambda, W_n \leq c_1 R^2\lambda\right) + \mathbb{P}\left(W_n \geq R^2\lambda\right) \\ &\leq 2\exp\left(\frac{-(R\lambda)^2}{2c_1R^2\lambda + \frac{2R}{3}(R\lambda)}\right) + \mathbb{P}\left(W_n \geq c_1R^2\lambda\right) \\ &\leq 2\exp\left(\frac{-\lambda}{2c_1 + \frac{2}{3}}\right) + \mathbb{P}\left(W_n \geq c_1R^2\lambda\right).\end{aligned}$$

■

4.2 Number of vertices with k descendants in a URRT

Devroye [12] proved a central limit theorem for the number of vertices with k descendants in a uniform random recursive tree. In particular, if $L_{k,n}$ denotes the the number of vertices with k descendants in a uniform random recursive tree of $n > k+1$ vertices, then Devroye shows that

$$\mathbb{E}L_{k,n} = \frac{n-k-1}{(k+1)(k+2)} + \frac{1}{k+1} = \frac{n+1}{(k+1)(k+2)}$$

and, for any fixed k, as $n \to \infty$,

$$\frac{L_{k,n} - \frac{n}{(k+1)(k+2)}}{\sqrt{n\sigma_k^2}}$$

converges, in distribution, to a standard normal random variable, where

$$\sigma_k^2 = \frac{1}{(k+1)(k+2)}\left(1 - \frac{1}{(k+1)(k+2)}\right) - \frac{2}{(k+1)(k+2)^2} + \frac{1}{(k+1)^2(2k+3)}\ .$$

Devroye's proof is based on representing $L_{k,n}$ as a sum of $(k+1)$-dependent indicator random variables and on a central limit theorem of Hoeffding and Robbins [20] for such sums. In this text we need a non-asymptotic version of Devroye's theorem. Quantitative, Berry-Esseen-type versions of the Hoeffding-Robbins limit theorem are available via Stein's method, see, for example, Rinott [27, Theorem 2.2]. On the other hand, a simple bound may be proved by combining Devroye's representation with a concentration inequality of Janson [21, Corollary 2.4] for sums of dependent random variables, to obtain the following:

Lemma 4.2.1. *If $L_{k,n}$ denotes the the number of vertices with k descendants in a uniform random recursive tree of $n > k+1$, then for all $t > 0$,*

$$\mathbb{P}\left\{L_{k,n} \geq \mathbb{E}L_{k,n} + t\right\} \leq \exp\left(\frac{-8t^2(k+2)}{25(n + (k+1)(k+2)t/3}\right)$$

and

$$\mathbb{P}\left\{L_{k,n} \leq \mathbb{E}L_{k,n} - t\right\} \leq \exp\left(\frac{-8t^2(k+2)}{25n}\right)\ .$$

Note that the number of vertices with *at least* k descendants $M_{k,n} = \sum_{i=k}^{n-1} L_{i,n} = n - \sum_{i=0}^{k-1} L_{i,n}$ has expected value

$$\mathbb{E}M_{k,n} = \mathbb{E}\sum_{i=k}^{n-1} L_{i,n} = n - \sum_{i=0}^{k-1}\mathbb{E}L_{i,n} = \frac{n+1}{k+1} - 1\ ,$$

and therefore

$$\begin{aligned}\mathbb{P}\left\{M_{k,n} \leq \frac{n+1}{k+1} - 1 - t\right\} &= \mathbb{P}\left\{\sum_{i=0}^{k-1} L_{i,n} \geq \sum_{i=0}^{k-1} \mathbb{E}L_{i,n} + t\right\} \\ &\leq \sum_{i=0}^{k-1} \mathbb{P}\left\{L_{i,n} \geq \mathbb{E}L_{i,n} + \frac{t}{k}\right\} \\ &\leq k \exp\left(\frac{-8t^2}{25k(n+(k+1)t/3}\right) .\end{aligned}$$

In particular, by generously bounding constants, we get

$$\mathbb{P}\left\{M_{k,n} \leq \frac{n}{3k}\right\} \leq k \exp\left(-\frac{1}{32}\frac{n}{k^2}\right) . \tag{4.2.1}$$

4.3 Some estimates on $\Gamma(x)$

In this appendix we prove some useful bounds regarding gamma functions and other relations involving them.

4.3.1 Preliminaries estimates

Lemma 4.3.1 (Stirling formula for Gamma function - see formula 6.1.42 in [?]). *For all $x > 0$ we have*

$$\frac{(2\pi)^{1/2}}{e^x} x^{x-\frac{1}{2}} \leq \Gamma(x) \leq \frac{(2\pi)^{1/2} e^{1/12x}}{e^x} x^{x-1/2}.$$

Lemma 4.3.2. *For all positive x, it follows that*

$$\Gamma(x) = \frac{(2\pi)^{1/2}}{e^x} x^{x-\frac{1}{2}} \left(1 + O\left(\frac{1}{x}\right)\right).$$

Proof. Observe that by the Lemma 4.3.1

$$0 \leq \Gamma(x) - \frac{(2\pi)^{1/2}}{e^x} x^{x-\frac{1}{2}} \leq \frac{(2\pi)^{1/2}}{e^x} x^{x-\frac{1}{2}} \left(e^{1/12x} - 1\right),$$

and the result follows by Taylor approximation. ∎

Lemma 4.3.3. *Let β, λ be two positive real numbers with $\beta > \lambda$ then*

1. $\dfrac{\Gamma(\beta-\lambda)}{\Gamma(\beta)} \leq e^{\frac{1}{12(\beta-\lambda)}} \left(\dfrac{\beta}{\beta-\lambda}\right)^{1/2} \left(\dfrac{1}{\beta-\lambda}\right)^{\lambda}$;
2. $\dfrac{\Gamma(\beta)}{\Gamma(\beta-\lambda)} \leq e^{\frac{1}{12\beta}} \left(\dfrac{\beta-\lambda}{\beta}\right)^{1/2} \beta^{\lambda}$.

Proof. For the first item, by Lemma 4.3.1 and the bound $(1-\frac{x}{n})^n \leq e^{-x}$ it follows that

$$\begin{aligned}\frac{\Gamma(\beta-\lambda)}{\Gamma(\beta)} &\leq \frac{e^{\frac{1}{12(\beta-\lambda)}} (\beta-\lambda)^{\beta-\lambda-1/2}}{e^{\beta-\lambda}} \frac{e^{\beta}}{\beta^{\beta-1/2}} \\ &\leq \frac{e^{\frac{1}{12(\beta-\lambda)}}}{e^{-\lambda}} \left(1-\frac{\lambda}{\beta}\right)^{\beta} \left(1+\frac{\lambda}{\beta-\lambda}\right)^{1/2} (\beta-\lambda)^{-\lambda} \\ &\leq e^{\frac{1}{12(\beta-\lambda)}} \left(\frac{\beta}{\beta-\lambda}\right)^{1/2} \left(\frac{1}{\beta-\lambda}\right)^{\lambda}.\end{aligned}$$

The second item follows analogously. ∎

Lemma 4.3.4. *For $0 < x < 1$ and $y > 0$ we have*

$$(1-x)^y = e^{-xy}(1 + O(y^2x^3)).$$

Proof.
Observe that $(1-x)^y = \exp(y \cdot \log(1-x))$. Recalling the Taylor expansion of log

$$\log(1-x) = -x - O(x^2)$$

we have

$$(1-x)^y = \exp(-xy - O(yx^2)) = (1 - O(yx^2)) \exp(-xy).$$

■

Lemma 4.3.5. *For $k = O(n^{\frac{\alpha}{2\alpha+4}})$ we have*

$$\frac{\Gamma(n+\theta-k+\alpha)}{\Gamma(n+\theta)} = \frac{1}{n^{k-\alpha}}\left(1 + O\left(\frac{k^2}{n-k}\right)\right).$$

Proof. Using the expression given by Lemma 4.3.2, we obtain

$$\frac{\Gamma(n+\theta-k+\alpha)}{\Gamma(n+\theta)} = \frac{e^{k-\alpha}}{(n+\theta-k+\alpha)^{k-\alpha}}\left(1 - \frac{k-\alpha}{n+\theta}\right)^{n+\theta-\frac{1}{2}} \frac{1 + O(\frac{1}{n+\theta-k+\alpha})}{1 + O(\frac{1}{n+\theta})}.$$

Now, multiplying and dividing by $n^{k-\alpha}$ the right-hand side of the above identity becomes

$$\frac{e^{k-\alpha}}{n^{k-\alpha}}\left(1 + \frac{k-\theta-\alpha}{n+\theta-k+\alpha}\right)^{k-\alpha}\left(1 - \frac{k-\alpha}{n+\theta}\right)^{n+\theta-\frac{1}{2}} \frac{1 + O(\frac{1}{n+\theta-k+\alpha})}{1 + O(\frac{1}{n+\theta})}.$$

Moreover, by the Lemma 4.3.4 it follows

$$1 \le \left(1 + \frac{k-\theta-\alpha}{n+\theta-k+\alpha}\right)^{k-\alpha} \le \exp\left(\frac{(k-\alpha)(k-\theta-\alpha)}{n+\theta-k+\alpha}\right) = 1 + O\left(\frac{k^2}{n-k}\right).$$

Also, by Lemma 4.3.4, for $x = \frac{k-\alpha}{n+\theta}$ and $y = n+\theta$ we have

$$\begin{aligned}\left(1 - \frac{k-\alpha}{n+\theta}\right)^{n+\theta} &= \exp\left(-(n+\theta)\frac{k-\alpha}{n+\theta}\right)\exp\left(O\left(\frac{(k-\alpha)^2}{n+\theta}\right)\right) \\ &= e^{-k+\alpha}\left(1 + O\left(\frac{k^2}{n}\right)\right),\end{aligned}$$

and for $k = O(n^{\frac{\alpha}{2\alpha+4}})$

$$\begin{aligned}&e^{-k+\alpha}\left(1 + O\left(\frac{k^2}{n}\right)\right)\left(1 - \frac{k-\alpha}{n+\theta}\right)^{-\frac{1}{2}}\left(1 + O\left(\frac{k^2}{n-k}\right)\right)\frac{1 + O(\frac{1}{n+\theta-k+\alpha})}{1 + O(\frac{1}{n+\theta})} \\ &\quad = e^{-k+\alpha}\left(1 + O\left(\frac{k^2}{n}\right)\right).\end{aligned}$$

Thus

$$\frac{\Gamma(n+\theta-k+\alpha)}{\Gamma(n+\theta)} = \frac{e^{k-\alpha}}{n^{k-\alpha}}e^{-k+\alpha}\left(1 + O\left(\frac{k^2}{n}\right)\right) = \frac{1}{n^{k-\alpha}}\left(1 + O\left(\frac{k^2}{n}\right)\right).$$

■

4.3.2 Order of ϕ_n and $\psi_n(k)$

This part is devoted to prove bounds for the two normalizing factors ϕ_n and $\psi_n(k)$ whose definition we recall latter.

$$\phi_n = \frac{\Gamma(1+\theta)}{\Gamma(1+\theta+\alpha)} \frac{\Gamma(n+\alpha+\theta)}{\Gamma(n+\theta)},$$

Lemma 4.3.6. *Let ϕ_n be as above, then the following bounds hold*

1. $\dfrac{1}{\phi_j} < \dfrac{2\Gamma(1+\theta+\alpha)}{\Gamma(1+\theta)\cdot(j+\theta)^{\alpha}}$;
2. $\dfrac{1}{(j+\theta)\phi_{j+1}} < \dfrac{2\Gamma(1+\theta+\alpha)}{\Gamma(1+\theta)\cdot(j+\theta)^{1+\alpha}}$;
3. *There exists a constant C_ϕ such that*

$$\phi_j \leq C_\phi j^{\alpha};$$

In particular $\phi_n = \Theta(n^{\alpha})$.

Proof. Let us prove the first two items and the third will follow analogously.
(1). By Lemma 4.3.3

$$\frac{\Gamma(j+\theta)}{\Gamma(j+\theta+\alpha)} \leq e^{\frac{1}{12(j+\theta)}} \left(1+\frac{\alpha}{j+\alpha+\theta}\right)^{1/2} (j+\theta)^{-\alpha} \leq 2(j+\theta)^{-\alpha}.$$

then

$$\frac{1}{\phi_j} < \frac{2\Gamma(1+\theta+\alpha)}{\Gamma(1+\theta)\cdot(j+\theta)^{1+\alpha}}.$$

(2). This part follows using the duplication property $\Gamma(x+1) = x\Gamma(x)$ and the previous item and the inequality

$$\frac{\Gamma(j+1+\theta)}{\Gamma(j+1+\theta+\alpha)} = \frac{(j+\theta)\Gamma(j+\theta)}{(j+\theta+\alpha)\Gamma(j+\theta+\alpha)} < \frac{\Gamma(j+\theta)}{\Gamma(j+\theta+\alpha)}.$$

■

The next lemma provides similar bounds for the normalization factor $\psi_n(k)$ whose definition is recalled bellow.

$$\psi_n(k) = \frac{\Gamma(k+\theta)\Gamma(n-k+\alpha+\theta)}{\Gamma(\alpha+\theta)\Gamma(n+\theta)}.$$

Lemma 4.3.7. *For $\psi_n(k)$ defined as above, the following bounds hold*

1. $\psi_n(k) \leq \dfrac{2\Gamma(k+\theta)}{\Gamma(\alpha+\theta)} \dfrac{1}{(n+\theta-k+\alpha)^{k-\alpha}}$, *for $n \geq 2k$;*
2. $\dfrac{1}{\psi_n(k)} \leq \dfrac{e^{\frac{1}{12}}\Gamma(\alpha+\theta)}{\Gamma(k+\theta)}(n+\theta)^{k-\alpha}$

Proof. *(1).* By Lemma 4.3.3 we have

$$\frac{\Gamma(n-k+\alpha+\theta)}{\Gamma(n+\theta)} \leq e^{\frac{1}{12(n+\theta-k+\alpha)}} \left(1+\frac{k-\alpha}{n+\theta-k+\alpha}\right)^{1/2} \frac{1}{(n+\theta-k+\alpha)^{k-\alpha}}.$$

And for $2k \leq n$ it follows that

$$\frac{\Gamma(n-k+\alpha+\theta)}{\Gamma(n+\theta)} \leq \frac{2}{(n+\theta-k+\alpha)^{k-\alpha}}.$$

Then

$$\psi_n(k) \leq \frac{2\Gamma(k+\theta)}{\Gamma(\alpha+\theta)} \frac{1}{(n+\theta-k+\alpha)^{k-\alpha}}.$$

2. Again, by Lemma 4.3.3 we have

$$\frac{\Gamma(n+\theta)}{\Gamma(n-k+\alpha+\theta)} \leq e^{\frac{1}{12(n+\theta)}} \left(1-\frac{k-\alpha}{n+\theta}\right)^{1/2} (n+\theta)^{k-\alpha} \leq e^{\frac{1}{12}} (n+\theta)^{k-\alpha}$$

and the result follows from the previous inequality. ■

Lemma 4.3.8. *For the ration of the factors ϕ_n and $\psi_n(k)$ the following upper bound holds*

$$\frac{\phi_j}{(\psi_{j+1}(k))^2 \cdot (j+\theta)} \leq \frac{\Gamma(1+\theta)\Gamma(\alpha+\theta)^2}{\Gamma(1+\theta+\alpha)\Gamma(k+\theta)^2} (j+\theta)^{2k-\alpha-1}.$$

Proof. Using the definition of both factors, we have

$$\frac{\phi_j}{(\psi_{j+1}(k))^2 (j+\theta)} = \frac{\Gamma(1+\theta)\Gamma(\alpha+\theta)^2}{\Gamma(1+\theta+\alpha)\Gamma(k+\theta)^2} \frac{\Gamma(j+\theta)\Gamma(j-1+\alpha+\theta)}{\Gamma(j+1-k+\alpha+\theta)^2} (j+\theta)^2 (j+\alpha+\theta)$$

and using the bounds on ratio of gamma functions in Lemma 4.3.3, we have

$$\frac{\phi_j}{(\psi_{j+1}(k))^2 \cdot (j+\theta)} \leq \frac{e^{\frac{1}{12}}\Gamma(1+\theta)\Gamma(\alpha+\theta)^2}{\Gamma(1+\theta+\alpha)\Gamma(k+\theta)^2} (j+\theta)^{2k-\alpha-1}.$$

■

Bibliography

[1] David Aldous, Ibragimov Ildar, and Jean Jacod. *Ecole d'Ete de Probabilites de Saint-Flour XIII, 1983*, volume 1117 of *Ecole d'Ete de Probabilites de Saint-Flour*. Springer-Verlag Berlin Heidelberg, 1985.

[2] Christian Borgs, Michael Brautbar, Jennifer Chayes, Sanjeev Khanna, and Brendan Lucier. The power of local information in social networks. In *Internet and Network Economics*, pages 406–419. Springer, 2012.

[3] Michael Brautbar and Michael J. Kearns. Local algorithms for finding interesting individuals in large networks. In *Innovations in Theoretical Computer Science (ITCS)*, 2010.

[4] Graham Brightwell and Malwina Luczak. Vertices of high degree in the preferential attachment tree. *Electron. J. Probab.*, 17:no. 14, 1–43, 2012.

[5] Sébastien Bubeck, Luc Devroye, and Gábor Lugosi. Finding Adam in random growing trees. *Random Structures & Algorithms*, 50(2):158–172, 2017.

[6] 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.

[7] Sébastien Bubeck, Elchanan Mossel, and Miklós Z 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.

[8] Fan Chung and Linyuan Lu. *Complex Graphs and Networks*, volume 107. AMS and CBMS, 2006.

[9] Harry Crane. The ubiquitous ewens sampling formula. *Statistical Science*, 31(1):1–19, 2016.

[10] Harry Crane and Walter Dempsey. Edge exchangeable models for interaction networks. *Journal of the American Statistical Association*, 2017.

[11] 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.

[12] Luc Devroye. Limit laws for local counters in random binary search trees. *Random Structures & Algorithms*, 2(3):303–315, 1991.

[13] Michael Drmota. *Random trees: an interplay between combinatorics and probability*. Springer Science & Business Media, 2009.

[14] Warren J. Ewens. The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, 3(1):87–132, 1972.

[15] Stefano Favaro and Shui Feng. Large deviation principles for the Ewens-Pitman sampling model. *Electron. J. Probab.*, 20(40):26 pp., 2015.

[16] Stefano Favaro, Shui Feng, and Fuqing Gao. Large deviation principles for the Ewens-Pitman sampling model. *Sankhya A - Springer India*, (13171):1–12, 2018.

[17] David A. Freedman. On tail probabilities for martingales. *Ann. Probab.*, 3(1):100–118, 02 1975.

[18] Alan Frieze and Wesley Pegden. Looking for vertex number one. *The Annals of Applied Probability*, 27(1):582–630, 2017.

[19] Thomas L. Griffiths, Michael I. Jordan, Joshua B. Tenenbaum, and David M. Blei. Hierarchical topic models and the nested chinese restaurant process. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 17–24. MIT Press, 2004.

[20] Wassily Hoeffding and Herbert Robbins. The central limit theorem for dependent random variables. *Duke Mathematical Journal*, 15(3):773–780, 1948.

[21] Svante Janson. Large deviations for sums of partly dependent random variables. *Random Structures & Algorithms*, 24(3):234–248, 2004.

[22] Varun Jog and Po-Ling Loh. Analysis of centrality in sublinear preferential attachment trees via the CMJ branching process. *IEEE Transactions on Network Science and Engineering*, 4(1):1–12, 2017.

[23] Varun Jog and Po-Ling Loh. Persistence of centrality in random growing trees. *Random Struct. Algorithms*, 52:136–157, 2018.

[24] John F.C. Kingman. Uses of exchangeability. *Ann. Probab.*, 6(2):183–197, 1978.

[25] James Pitman. Exchangeable and partially exchangeable random partitions. *Probab. Th. Rel. Fields*, 102(2):145–158, 1995.

[26] James Pitman. *Combinatorial Stochastic Processes*, volume 1875. Springer-Verlag Berlin Heidelberg, 2006.

[27] Yosef Rinott. On normal approximation rates for certain sums of dependent random variables. *Journal of Computational and Applied Mathematics*, 55(2):135–143, 1994.

[28] Devavrat Shah and Tauhid Zaman. Finding rumor sources on random trees. *Operations Research*, 64(3):736–755, 2016.

[29] Devavrat Shah and Tauhid R. Zaman. Rumors in a network: Who's the culprit? *IEEE Transactions on Information Theory*, 57(8):5163–5181, 2011.