UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Grands graphes et grands arbres aléatoires : Analyse du comportement asymptotique

THÈSE

présentée et soutenue publiquement le 11 Mai 2016

pour l'obtention du

Doctorat de l'Université de Lorraine

(Spécialité : Mathématiques Appliquées)

par

Lucas Mercier

Composition du jury

Présidente :	Frédérique Bassino, Professeure, Laboratoire d'Informatique de Paris-Nord
Rapporteurs :	Itai Benjamini, Professeur, Weizmann Institute of Science Jean Bertoin, Professeur, Institut für Mathematik, Universität Zürich
Examinateurs :	Brigitte Chauvin, Professeure, Université de Versailles Saint Quentin en Yvelines Nicolas Champagnat, Chargé de Recherche, Institut Élie Cartan de Lorraine Marc Lelarge, Directeur de Recherche, INRIA
Directeur :	Philippe Chassaing, Professeur, Institut Élie Cartan de Lorraine

Institut Élie Cartan, Laboratoire de Mathématiques,
B.P. 239, 54506 Vandoeuvre-lès-Nancy Cedex

Table des matières

Chapitre 1

Introduction

Contents

1.1 Graphes aléatoires et limites locales

Cette partie est consacrée à l'étude de graphes aléatoires, et des limites locales.

1.1.1 Graphes : notation et vocabulaire

Définition 1.1. Un *graphe non-orienté* est la donnée :

- d'un ensemble V, fini ou dénombrable, appellé *ensemble des sommets*,
- d'un ensemble E, fini ou dénombrable, appellé *ensemble des arêtes*,
- d'une application φ de E dans l'ensemble des parties de V à un ou deux élements de V. Si e est une arête, les élements de l'ensemble $\varphi(e)$ sont appelés les *extrémités* de e, appellée *fonction d'incidence*.

L'ensemble des parties de V à deux éléments sera noté $\mathscr{P}_2(V)$.

Définition 1.2. Deux graphes $G_1 = (V_1, E_1, \varphi_1)$ et $G_2 = (V_2, E_2, \varphi_2)$ sont *isomorphes* s'il existe deux bijections $\Psi_V : V_1 \mapsto V_2$ et $\Psi_E : E_1 \mapsto E_2$ satisfaisant : $\forall e \in E_1$, $\Psi_V(\varphi_1(e)) = \varphi_2(\Psi_E(e))$.

Dans la suite, les graphes seront considérés à isomorphisme près. Dans le cadre de ce manuscrit, tous les graphes seront non-orientés, et l'adjectif *non-orienté* sera donc omis.

Définition 1.3. Si $\varphi(e)$ est de cardinal 1, alors e est appelée une *boucle*. Si m arêtes ont les mêmes extrémités, on parle d'*arête multiple* de *multiplicité* m. Un graphe *simple* est un graphe sans boucle ni arête multiple c'est-à-dire, avec les notations de la définition 1.1, un graphe tel que :

- φ est injective.
- $\forall e \in E, \varphi(e) \in \mathscr{P}_2(V)$.

Dans le cas d'un graphe simple $G = (V, E, \varphi)$, le graphe $\tilde{G} = (V, \mathrm{Im}\varphi, Id)$ est isomorphe à G. Une définition alternative d'un graphe simple est donc la donnée :

- d'un ensemble V, fini ou dénombrable, appellé ensemble des sommets,
- d'un ensemble $E \subset \mathscr{P}_2(V)$ appellé ensemble des arêtes.

L'application Id étant implicite. Le terme *multigraphe* est utilisé pour désigner un graphe lorsque l'on souhaite insister sur le fait qu'il n'est pas forcément simple.

Définition 1.4. Pour tout ensemble de sommets V, le *graphe complet* sur V est le graphe simple ayant V comme ensemble de sommet et $\mathscr{P}_2(V)$ comme ensemble d'arêtes.

1.1.2 L'exemple du modèle d'Erdős et Rényi

Cette partie, consacrée aux modèles d'Erdős et Rényi, permettra d'introduire les outils utilisés dans la suite du manuscrit pour l'étude d'autres modèles de graphes aléatoires.

1.1.2.a Les différents modèles d'Erdős et Rényi

Le premier modèle de graphe aléatoire a été introduit par Erdős et Rényi en 1959 [ER59]. Le terme de modèle d'Erdős et Rényi recouvre en réalité plusieurs modèles proches, sur lesquels beaucoup de résultats sont connus. Tous ces modèles, bien que non formellement équivalents, satisfont en général les mêmes propriétés. Pour cette raison, lorsqu'une propriété est vérifiée par tous les modèles d'Erdős et Rényi présentés ici, nous dirons qu'elle est vérifiée par le graphe d'Erdős et Rényi.

Le premier modèle, tel qu'introduit dans l'article fondateur d'Erdős et Rényi [ER59], est défini de la manière suivante :

Définition 1.5. Soient n et m deux entiers naturels non nuls, tels que $m \leq \binom{n}{2}$. Soit E un sous-ensemble de $\mathscr{P}_2(\{1,\dots,n\})$ choisi uniformément parmi les sous-ensembles de $\mathscr{P}_2(\{1,\dots,n\})$ à m éléments. Le *premier modèle d'Erdős et Rényi* $\bar{G}(n,m)$ est le graphe simple dont l'ensemble des sommets est $\{1,\dots,n\}$ et l'ensemble des arêtes est E.

$\bar{G}(n,m)$ est donc un graphe à n sommets et m arêtes.

Le second modèle d'Erdős et Rényi, introduit par Gilbert in 1959 [Gil59], et également appelé modèle de Gilbert, est défini de la manière suivante :

Définition 1.6. Soit p un réel entre 0 et 1. Le *second modèle d'Erdős et Rényi* $G(n,p)$ est le graphe à n sommets où chacune des $\binom{n}{2}$ arêtes du graphe complet est présente indépendamment avec probabilité p.

Le nombre d'arêtes de $G(n,p)$ est une loi binomiale de paramètres $\binom{n}{2}$ et p. Conditionnellement à son nombre d'arêtes m, le second graphe d'Erdős et Rényi $G(n,p)$ a même loi que le premier graphe d'Erdős et Rényi $\bar{G}(n,m)$. Par la loi des grands nombres, le nombre d'arêtes de $G(n,p)$ est proche de son espérance $p\binom{n}{2}$ et donc $\bar{G}(n,m)$ et $G(n,p)$ sont proches lorsque $m = p\binom{n}{2}$. Le second modèle possède des propriétés d'indépendance intéressantes facilitant son étude, et donc pour cette raison est plus souvent utilisé .

Il est possible de coupler chacun de ces modèles lorsque m (resp. p) varie, à nombre de sommets fixés, au moyen d'un graphe aléatoire dynamique d'Erdős et Rényi. La version dynamique du premier modèle est définie de la manière suivante :

- Initialement, le graphe possède n sommets et aucune arête.
- À chaque étape, une arête est choisie uniformément parmi les arêtes du graphe complet qui ne sont pas encore présentes et est ajoutée au graphe.

La version dynamique associée au second modèle est le suivant :

- Chaque arête e parmi les $\binom{n}{2}$ arêtes possibles est étiquetée par une variable aléatoire Y_e de loi uniforme sur $[0, 1]$. La collection d'étiquettes est i.i.d.
- Le modèle $G(n, p)$ est alors le graphe à n sommets dont l'ensemble des arêtes est $\{e : Y_e \leq p\}$.

Ces deux couplages donnent un processus croissant (au sens de l'inclusion pour l'ensemble des arêtes) dont les marginales sont les deux modèles d'Erdős et Rényi présentés au début de la section.

Ces modèles sont parfois plus simples à étudier en autorisant les multi-arêtes, et éventuellement les boucles. Deux variantes possibles du premier modèle sont alors :

- Initialement, le graphe a n sommets et aucune arête.
- À chaque étape, une paire de sommets est choisie uniformément parmi les $\binom{n}{2}$ paires de sommets, et une arête les reliant est ajoutée au graphe.

ou bien

- Initialement, le graphe a n sommets et aucune arête.
- À chaque étape, deux sommets sont choisis indépendamment et une arête les reliant est ajoutée (une boucle si le même sommet est choisi deux fois).

Une variante multigraphe du second modèle est alors :

- Chaque arête e parmi les $\binom{n}{2}$ arêtes possibles est munie d'un processus ponctuel de Poisson Π_e d'intensité 1 sur $\mathbb{R}_+$. La collection de processus de Poisson est i.i.d ;
- Au temps t, chaque arête e est présente avec la multiplicité $|\Pi_e \cap [0, t]|$.

Cette variante sera utilisée dans la partie 3.

1.1.2.b Composantes géantes

La taille des composantes connexes (en nombre de sommets), en particulier la taille des plus grandes composantes connexes, apporte des informations utiles sur la structure d'un graphe. Pour un graphe G, et un entier i, on note par $C_i(G)$ la taille de la i-ème plus grande composante de G (avec $C_i(G) = 0$ si G a moins de i composantes). Pour alléger les notations, le G est omis de la notation C_i lorsqu'aucune confusion n'est possible.

Dans le cadre du premier modèle d'Erdős et Rényi, une transition de phase remarquable se produit lorsque le nombre d'arêtes est proche de $m = \frac{n}{2}$. Si $m = cn$ avec c une constante réelle positive, Erdős et Rényi [ER60] ont prouvé le résultat suivant :

Théorème 1.1.

- *Si* $0 < c < \frac{1}{2}$, $C_1(\bar{G}(n, cn)) = \Theta_p(\ln n)$.
- *Si* $c = \frac{1}{2}$, *pour tout entier* i, $C_i(\bar{G}(n, \frac{1}{2}n)) = \Theta_p(n^{\frac{2}{3}})$.
- *Si* $c > \frac{1}{2}$, $\frac{C_1(\bar{G}(n,cn))}{n} \xrightarrow[n\to\infty]{p} \alpha_c > 0$ *et* $C_2(\bar{G}(n, cn)) = \Theta_p(\ln(n))$.

Où α_c est une constante explicite dépendant de c.

Notation. Soient X_n est une suite de variables aléatoires positives, et $f(n)$ une suite réelle positive. La notation $X_n = \Theta_p(f(n))$ signifie que l'on a simultanément :

$$\lim_{a\to\infty} \limsup_n \mathbb{P}(X_n \geq af(n)) = 0 \quad (1.1)$$

$$\lim_{a\to\infty} \limsup_n \mathbb{P}(f(n) \geq aX_n) = 0 \quad (1.2)$$

En conditionnant le théorème 1.1 par le nombre d'arêtes multiples, on obtient le même théorème pour les variantes multigraphes du premier modèle d'Erdős et Rényi. En conditionnant par le nombre d'arêtes (resp. le nombre d'arêtes et le nombre d'arêtes multiples), on obtient des propriétés similaires pour le second modèle d'Erdős et Rényi (resp. sa variante multigraphe) lorsque la probabilité de chaque arête (resp. l'espérance de la multiplicité de chaque arête) est de la forme $\frac{c}{n}$, avec une transition de phase lorsque c dépasse 1.

Une composante de taille $\Theta(n)$ avec n le nombre de sommets est appelée *composante géante*. Dans les modèles d'Erdős et Rényi a donc lieu une transition de phase, où une faible modification de la probabilité d'existence d'une arête (ou du nombres d'arêtes) change significativement l'aspect du graphe. Ce phénomène est appelé « apparition de la composante géante ». De plus au seuil critique ($p = \frac{1}{n}$ ou $m = \frac{n}{2}$), le graphe contient un grand nombre de composantes de taille $\Theta(n^\alpha)$ pour un certain $\alpha > 0$, alors qu'une seule telle composante est présente lorsque que $c > 1$ et aucune quand $c < 1$. Des résultats plus précis obtenus par Aldous [Ald97] dans le cas critique seront présentés à la partie 1.1.2.g.

1.1.2.c Limite locale

Une technique utile pour l'étude de graphes aléatoires est de considérer le graphe *vu depuis un de ses sommets*. Pour cela, il est nécessaire de choisir un sommet, appelé racine :

Définition 1.7. Un *graphe enraciné* est la donnée de $(V, E, \varphi, \emptyset)$ où (V, E, φ) est un graphe et $\emptyset \in V$. Le sommet $\emptyset$ est appelé *racine* de $(V, E, \varphi, \emptyset)$.

Remarque. Lorsque le graphe est simple, φ peut-être omis s'il s'agit de l'identité.

On dit que deux graphes enracinés sont *isomorphes* s'il existe un isomorphisme de graphes de G vers $\tilde{G}$ qui envoie v sur $\tilde{v}$. On ne s'intéressera dans la suite qu'aux classes d'équivalence de cette relation, c'est-à-dire qu'un graphe enraciné sera défini à isomorphisme de graphes enracinés près.

Lorsque $(G^n)_{n\geq 1}$ est une suite de graphes aléatoires à n sommets, il est souvent utile d'étudier la loi de G^n enraciné en l'un de ses sommets. Une notion naturellement associée au concept de graphe « *vu depuis un sommet* » est la notion de limite locale, introduite par ni et Schramm [BS01].

Définition 1.8. Soient (G, v) et $(\tilde{G}, \tilde{v})$ deux graphes G et $\tilde{G}$ enracinés respectivement en v et $\tilde{v}$. On définit la distance entre ces deux graphes enracinés par :

$$d_{\mathrm{loc}}\big((G, v), (\tilde{G}, \tilde{v})\big) = \frac{1}{1 + \sup\big\{R \geq 0 : (B_G(v, R), v) \simeq (B_{\tilde{G}}(\tilde{v}, R), \tilde{v})\big\}}.$$

où $B_G(v, R)$ est la boule de rayon R centrée sur v et $\simeq$ est la relation d'isomorphisme des graphes enracinés. Cette distance est nulle si et seulement si la composante de v dans G, enracinée en v est isomorphe à la composante de $\tilde{v}$ dans $\tilde{G}$, enracinée en $\tilde{v}$. d_{loc} définit donc une pseudo-métrique sur l'ensemble des classes d'équivalence des graphes enracinés et une métrique si l'on se limite aux classes d'équivalence de graphes connexes.

Cette métrique est un cas particulier de la limite locale définie par Aldous et Steele [AS04], qui autorise des arêtes de longueur non unitaire.

Graphe multi-enraciné Il est possible de généraliser cette notion en autorisant k racines, pour n'importe quel entier k fixé. Dans ce cas, (G, v) est un graphe k-enraciné si v est un k-uplet de sommets de G. (G, v) est isomorphe à $(\tilde{G}, \tilde{v})$ s'il existe un isomorphisme de graphes de G vers $\tilde{G}$ qui envoie chaque v_i sur $\tilde{v}_i$.

La distance entre graphes k-enracinés devient alors :

$$d^k_{\mathrm{loc}}\Big((G, v), (\tilde{G}, \tilde{v})\Big) = \frac{1}{1 + \sup\big\{R \geq 0 : (\cup_{i=1}^k B_G(v_i, R), v) \simeq (\cup_{i=1}^k B_{\tilde{G}}(\tilde{v}_i, R), \tilde{v})\big\}}.$$

Comme précédemment, cette distance est compatible avec l'isomorphisme de graphes k enracinés, et est nulle si et seulement si les graphes k-enracinés restreints aux composantes de leurs racines sont isomorphes en tant que graphes k-enracinés.

1.1.2.d Processus d'exploration

Pour nombre de modèles de graphe aléatoire, la loi du graphe enraciné en un sommet converge localement vers un processus plus simple à étudier, comme un processus de branchement, ce qui peut souvent se voir en explorant la composante des racines de proche en proche. Soit v un sommet du graphe, on note par $C(v, G)$ la composante connexe de v dans le graphe G. L'algorithme, similaire à celui utilisé par Aldous dans [Ald97], utilise trois ensembles :

- Un ensemble $\mathcal{S}$, muni d'une règle FIFO (*First In, First Out*, parfois nommé PEPS en français, pour Premier Entré, Premier Sorti). Initialement, $\mathcal{S}$ ne contient que le sommet v. Les éléments de $\mathcal{S}$ seront appelés sommets *en réserve*.

- Un ensemble $\mathcal{U}$, contenant les sommets qui ont été dans $\mathcal{S}$ à un moment, mais n'y sont plus. Initialement, $\mathcal{U}$ est vide.

- Un ensemble $\mathcal{T}$, contenant tous les autres sommets. Initialement, tous les sommets sauf v sont donc dans l'ensemble $\mathcal{T}$.

Remarque. La règle FIFO sur $\mathcal{S}$ définit un ordre sur les éléments de $\mathcal{S}$, qui sont ordonnés par leur temps d'ajout à $\mathcal{S}$. Lorsque l'algorithme aura besoin de sélectionner un élément de $\mathcal{S}$, il choisira celui qui est présent depuis le plus longtemps (en cas d'égalité, l'ordre est tiré aléatoirement uniformément indépendamment du graphe).

Une itération de l'algorithme agit de la manière suivante :

-Instruction 1 : Soit w le premier élément de $\mathcal{S}$. Enlever w de $\mathcal{S}$ et l'ajouter à $\mathcal{U}$.

-Instruction 2 : Dénotons par $V_w^{\mathcal{T}}$ l'ensemble des éléments de $\mathcal{T}$ qui sont voisins de w. Pour tout élément v de $V_w^{\mathcal{T}}$, ôter v de $\mathcal{T}$ et l'ajouter à $\mathcal{S}$. Tracer l'arête reliant v à w (éventuellement multiple dans le cas d'un multigraphe).

-Instruction 2' : Dénotons par $V_w^{\mathcal{S}}$ l'ensemble des éléments de $\mathcal{S}$ qui sont voisins de w. Tracer les arêtes reliant w aux éléments de $V_w^{\mathcal{S}}$.

-Instruction 3 : Si $\mathcal{S}$ est vide, enlever un sommet choisi uniformément au hasard à $\mathcal{T}$ et l'ajouter à $\mathcal{S}$. Si $\mathcal{T}$ est vide, arrêter l'algorithme.

On notera par w_l le sommet w lors de la l-ème itération de l'algorithme, et par $\mathcal{S}_l$ (resp. $\mathcal{T}_l$, $\mathcal{U}_l$) l'ensemble $\mathcal{S}$ (resp. $\mathcal{T}$, $\mathcal{U}$) après la l-ème itération de l'algorithme.

Remarque.

- L'algorithme trace les différentes composantes de G, une par une, en commençant par $C(v, G)$ et s'arrête après avoir exploré toutes les composantes.
- Il s'agit d'une exploration en largeur, c'est-à-dire que l'on regarde tous les sommets à distance k de l'origine v avant de regarder les sommets à distance $k+1$. Il serait bien sûr possible possible d'explorer $C(v, G)$ dans un autre ordre, et l'on obtiendrait le même résultat final, mais cette méthode est généralement plus efficace pour démontrer des résultats de limites locales.
- Si l'on est uniquement intéressé par l'ensemble des sommets de chaque composante (et pas par l'ensemble des arêtes), l'instruction 2' peut-être supprimée. Cette instruction ne découvrant des arêtes qu'entre deux sommets de $\mathcal{S}$, la supprimer ne fait que supprimer les arêtes créant des cycles. L'algorithme trace alors un arbre couvrant pour chaque composante, donc une forêt couvrante de G.
- L'instruction 3 sert à continuer l'exploration lorsqu'une composante a été entièrement explorée. En effet, $\mathcal{S}$ ne devient vide que lorsque l'algorithme termine d'explorer une composante. Il choisit alors un autre sommet non encore découvert, puis explore sa composante.
- Si l'on ne s'intéresse qu'à $C(v, G)$, il est possible de remplacer l'instruction 3 par l'instruction suivante :

  -Instruction 3bis : Si $\mathcal{S}$ est vide, arrêter l'algorithme.

  Dans ce cas, l'algorithme s'arrête après avoir exploré $C(v, G)$.

Utilisation de l'algorithme d'exploration pour la limite locale.

Cette partie a pour but de montrer un exemple d'utilisation de l'algorithme d'exploration pour montrer un limite locale. On se place dans le second modèle d'Erdős et Rényi, où la probabilité d'existence d'une arête est de la forme $\frac{c}{n}$.

Théorème 1.2. *Soit T_c^{GW} l'arbre de Galton-Watson de loi de reproduction* Poi(c) *où* Poi(c) *est une variable aléatoire de Poisson de paramètre c.*

Soit k un entier. Soient $v_1, v_2, \ldots, v_k$ k sommets de $G(n, \frac{c}{n})$ choisis indépendamment uniformément au hasard et indépendamment de $G(n, \frac{c}{n})$. Alors le graphe $G(n, \frac{c}{n})$ k-enraciné en $(v_1, \ldots v_k)$ converge localement en distribution vers k copies indépendantes de T_c^{GW}.

Une preuve de ce résultat peut être trouvée dans [DM10, proposition 2.6]. Démontrons d'abord le résultat pour $k = 1$. Pour tout entier l, on note par $\mathcal{U}_l$ (resp. $\mathcal{S}_l$, $\mathcal{T}_l$) l'ensemble $\mathcal{U}$ (resp. $\mathcal{S}$, $\mathcal{T}$) après l itérations de l'algorithme d'exploration, et par w_l le sommet w de l'instruction 1 de la l-ème itération.

Par construction, après l itérations de l'algorithme d'exploration, la liste des arêtes adjacentes aux sommets de U_l est connue. On note par E_l^{inconnu} l'ensemble des arêtes du graphe complet à n sommets n'ayant pas d'extrémités dans U_l.

Lemme 1.3. *Pour tout entier l fixé, conditionnellement aux l premières itérations de l'algorithme appliqué à $G(n,p)$, chaque arête de $E_l^{inconnu}$ est présente indépendamment avec probabilité p.*

Ce lemme se montre par récurrence sur l. Pour $l = 0$, l'algorithme d'exploration n'a pas commencé, et donc le lemme découle de la définition de $G(n,p)$.

On suppose que le lemme est vérifié pour un certain entier l. L'itération $l+1$ ajoute le sommet w à l'ensemble $\mathcal{U}$, et regarde la présence de toutes les arêtes entre w et des sommets de $\mathcal{S}_l \cup T_l$. En utilisant l'hypothèse de récurrence, les autres arêtes, c'est-à-dire les arêtes appartenant à E_{l+1}^{inconnu}, sont présentes indépendamment avec probabilité p conditionnellement aux $l+1$ premières itérations de l'algorithme.

Lemme 1.4. *Pour tout entier l fixé, conditionnellement aux $l-1$ premières itérations de l'algorithme, la loi du nombre d'arêtes entre w_l et des éléments de T_l (resp. $\mathcal{S}_l$) converge en distribution vers une variable aléatoire de Poisson de paramètre c (resp. 0).*

Ce résultat se montre par récurrence forte. On suppose que le lemme est vérifié pour tout $l' < l$. Par construction, après l'instruction 1 de la l-ème itération, $|U| = l$. D'après le lemme 1.3, chacune des arêtes possibles entre w et un élément de $\mathcal{T}$ (resp. $\mathcal{S}$) est présente dans le graphe indépendamment avec probabilité $\frac{c}{n}$, donc, conditionnellement aux $l-1$ premières itération de l'algorithme, la loi du nombre d'arêtes est une binomiale de paramètres $|\mathcal{T}|$ (resp. $|\mathcal{S}|$) et $\frac{c}{n}$. À l fixé, en utilisant le fait que $\mathcal{S}_l \cup \mathcal{U}_l \setminus \{v\}$ est l'ensemble des sommets découverts aux instructions 2 et 3 des $l-1$ premières itérations et l'hypothèse de récurrence, on obtient que $|\mathcal{S}_l| = O_p(1)$ et $|\mathcal{T}_l| = n - |\mathcal{S}_l| - |\mathcal{U}_l| \sim n$. La loi du nombre d'arêtes entre w et des sommets de $\mathcal{T}_l$ (resp. $\mathcal{S}_l$) converge donc en distribution vers $\text{Poi}(c)$ (resp. 0).

Par conséquent, la loi du graphe tracé par l'algorithme (c'est-à-dire $C(G(n, \frac{c}{n}, v))$ converge localement vers un arbre de Galton-Watson de loi de reproduction $\text{Poi}(c)$.

Pour montrer le résultat pour $k \geq 2$, nous arrêtons l'exploration de la composante de chaque racine dès que la boule de rayon R est connue,

et nous commençons l'exploration de la composante de la racine suivante. Par le même raisonnement que précédemment, le nombre d'enfants (resp. de cycles) découverts à chaque étape converge en distribution vers Poi(c) (resp. 0) pour tous k et R fixés. Par conséquent les boules $(B_G(v_i, R))_{1\leq i\leq k}$ sont asymptotiquement presque sûrement disjointes, et la loi jointe converge vers la loi de k copies indépendantes de la boule de rayon R de l'arbre de Galton-Watson de loi de reproduction Poi(c).

1.1.2.e Utilisation de la limite locale pour étudier la composante géante

La limite locale permet d'obtenir des informations sur les propriétés macroscopiques du graphe, comme l'existence et la taille d'une composante géante. Plus précisément on a la majoration suivante :

Théorème 1.5. *Soit G^n une suite de graphes aléatoires à n sommets.*

- *Si G^n enraciné en un sommet uniforme v converge localement en distribution vers le graphe enraciné aléatoire $\mathscr{T}$, alors*

$$\limsup_{n\to\infty} \frac{\mathbb{E}(C_1(G^n))}{n} \leq \mathbb{P}(|\mathscr{T}| = \infty).$$

- *Si G^n enraciné en deux sommets uniformes indépendants converge localement en distribution vers deux copies indépendantes de $\mathscr{T}$, alors*

$$\forall \epsilon > 0, \mathbb{P}\left(\frac{C_1(G^n)}{n} \leq \mathbb{P}(|\mathscr{T}| = \infty) + \epsilon\right) \xrightarrow[n\to\infty]{} 1.$$

Pour un grand nombre de modèles de graphes aléatoires, on a en réalité la limite suivante :

$$\frac{C_1(G^n)}{n} \xrightarrow[n\to\infty]{p} \mathbb{P}(|\mathscr{T}| = \infty). \tag{1.3}$$

Par exemple, ce phénomène apparaît dans le modèle d'Erdős et Rényi, où l'on obtient :

$$\frac{C_1(G(n, \frac{c}{n}))}{n} \xrightarrow[n\to\infty]{p} \mathbb{P}(|T_c^{GW}| = \infty) \tag{1.4}$$

Par exemple, si $c \leq 1$, l'arbre de Galton-Watson T_c^{GW} est critique ou sous-critique, donc est fini presque sûrement. La limite (1.4) implique que $C_1(G(n, \frac{c}{n})) = o_p(n)$ a.p.s, alors que si $c > 1$, T_c^{GW} a une probabilité positive de survivre indéfiniment, et donc $C_1(G(n, \frac{c}{n})) = \Theta_p(n)$. Ce résultat permet de plus de décrire la constante α_c comme la probabilité de survie d'un arbre de Galton-Watson.

Heuristiquement, le lien entre composante géante et limite locale peut se voir de la manière suivante : $\mathbb{P}(|T| = \infty)$ est la probabilité que la racine de T soit dans une composante infinie, ce qui correspond à la probabilité asymptotique que la racine v de G^n soit dans une *grande* composante de G^n. S'il existe une unique grande composante, par exemple dans le cadre du graphe d'Erdős-Rényi surcritique, alors la probabilité que v soit dans une grande composante est proche de la probabilité que v soit dans la plus grande composante de G^n, c'est-à-dire $\mathbb{E}(\frac{C_1(G^n)}{n})$.

Pour tout entier i, dénotons par N_i^n le nombre de sommets appartenant à des composantes de taille au moins i dans le graphe à G^n.

Lemme 1.6. *Soit G^n une suite de graphes aléatoires à n sommets et $(b_n)_{n\in\mathbb{N}}$ une suite d'entiers tendant vers $+\infty$.*

- *Si G^n enraciné en un sommet uniforme converge localement vers le graphe enraciné aléatoire $\mathscr{T}$, alors*

$$\limsup_n \frac{\mathbb{E}(N_{b_n}^n)}{n} \leq \mathbb{P}(|\mathscr{T}| = \infty).$$

- *Si G^n enraciné en deux sommets uniformes indépendants converge localement vers deux copies indépendantes de $\mathscr{T}$, alors*

$$\forall \epsilon > 0, \mathbb{P}\left(\frac{N_{b_n}^n}{n} \leq \mathbb{P}(|\mathscr{T}| = \infty) + \epsilon\right) \xrightarrow[n\to\infty]{} 1.$$

Preuve du lemme. On suppose pour l'instant la convergence locale en distribution de G enraciné en un sommet uniforme. Pour tout entier i, l'événement « il y a au moins i sommets dans la composante de la racine » est continu pour la limite locale, donc $\mathbb{P}(|C(v, G^n)| \geq i) \xrightarrow[n\to+\infty]{} \mathbb{P}(|T| \geq i)$ avec v un sommet choisi uniformément parmi les sommets de G^n.

Le sommet v est choisi uniformément parmi les n sommets de G^n, donc $\mathbb{P}(|C(v, G^n)| \geq i) = \frac{\mathbb{E}(N_i^n)}{n}$.

Comme b_n tend vers $+\infty$ et que N_i^n est décroissant en i, cela implique que pour tout i, $\limsup_n \frac{\mathbb{E}(N_{b_n}^n)}{n} \leq \mathbb{P}(|\mathscr{T}| \geq i)$. Il suffit ensuite de passer à la limite en i pour obtenir le résultat.

Dans le cas de la limite enracinée en deux sommets, en notant v_1 et v_2 deux sommets uniformes, pour tout $i \in \mathbb{N}$, par la limite locale :

$$\begin{aligned}
\mathbb{E}\left(\frac{(N_i^n)^2}{n^2}\right) = \mathbb{P}\Big(|C(v_1, G^n)| \geq i \text{ et } |C(v_2, G^n)| \geq i\Big) &\to \mathbb{P}(|\mathscr{T}| \geq i)^2 \\
\mathbb{E}\left(\left(\frac{N_i^n}{n}\right)^2\right) - \left(\mathbb{E}\left(\frac{N_i^n}{n}\right)\right)^2 &\to \mathbb{P}(|\mathscr{T}| \geq i)^2 - P(|\mathscr{T}| \geq i)^2 = 0
\end{aligned}$$

Donc

$$\frac{N_i^n}{n} \xrightarrow[n\to\infty]{p} \mathbb{P}(|\mathscr{T}| \geq i)$$

Passer à la limite en i permet d'obtenir le résultat du lemme. □

De plus, pour tout entier l, $|C_1(G_n)| \leq \max(l, N_l^n) \leq l + N_l^n$. En effet, si $C_1(G_n) > l$, alors par définition de N_l^n, $|C_1(G_n)| \leq N_l^n$. En prenant $l_n = \sqrt{n}$, on obtient l'inégalité :

$$\left|\frac{C_1(G^n)}{n}\right| \leq \frac{N_{\sqrt{n}}^n}{n} + \frac{\sqrt{n}}{n}$$

ce qui permet d'obtenir le théorème 1.5 à partir du lemme 1.6.

Dans le cas où $|\mathscr{T}| < \infty$ p.s., le théorème 1.5 implique la limite (1.3). Dans le cas où $\mathbb{P}(|\mathscr{T}| = \infty) > 0$, le théorème 1.5 est insuffisant pour montrer une limite de type (1.3). Il faut également montrer que presque tous les sommets appartenant à une *grande* composante dans G^n sont en réalité dans la même composante. Dans le cas des modèles d'Erdős et Rényi, une telle démonstration peut être trouvée dans [JŁR00], p108 sqq. Il est parfois possible d'utiliser le lemme suivant pour prouver le résultat :

Lemme 1.7. *Soit G^n une suite de graphes aléatoires, et (v_1, v_2) deux sommets choisis indépendamment uniformément parmi les sommets de G^n. On considère E_{v_1,v_2} l'événement que v_1 et v_2 soient dans la même composante de G^n. Si*

1. *(G^n, v_1, v_2) converge en loi vers deux copies indépendantes de $\mathscr{T}$ pour la limite locale et*

2. *$\mathbb{P}(E_{v_1,v_2}) \xrightarrow[n\to\infty]{} \mathbb{P}(|\mathscr{T}| = \infty)^2$*

alors $\frac{C_1(G^n)}{n} \xrightarrow[n\to\infty]{p} \mathbb{P}(|\mathscr{T}| = \infty)$.

Remarque. La première condition implique $\limsup_n \mathbb{P}(E_{v_1,v_2}) \leq \mathbb{P}(|\mathscr{T}| = \infty)^2$, donc la deuxième condition peut être remplacée par $\liminf_n \mathbb{P}(E_{v_1,v_2}) \geq \mathbb{P}(|\mathscr{T}| = \infty)^2$.

Preuve de la remarque. Conditionnellement à G^n, la probabilité que deux sommets choisis au hasard soient dans la même composante vaut $\frac{1}{n^2}\sum_i (C_i(G^n))^2$.

En prenant l'espérance :

$$\begin{aligned}
\mathbb{P}(E_{v_1,v_2}) &= \frac{1}{n^2}\mathbb{E}\left(\sum_i C_i^2\right) \\
&= \frac{1}{n^2}\mathbb{E}\left(\sum_{i:C_i<\sqrt{n}} C_i^2\right) + \frac{1}{n^2}\mathbb{E}\left(\sum_{i:C_i\geq\sqrt{n}} C_i^2\right) \\
&\leq \frac{1}{n^2}\sqrt{n}\mathbb{E}\left(\sum_{i:C_i<\sqrt{n}} C_i\right) + \frac{1}{n^2}\mathbb{E}\left((N^n_{\sqrt{n}})^2\right) \\
&\leq n^{-\frac{1}{2}} + \mathbb{E}\left(\left(\frac{N^n_{\sqrt{n}}}{n}\right)^2\right)
\end{aligned}$$

La démonstration de la remarque se termine en passant à la limite en n et en utilisant le lemme 1.6. □

Preuve du lemme 1.7. On suppose $\mathbb{P}(|\mathscr{T}| = \infty) > 0$ (sinon le lemme découle du théorème 1.5). Pour tout $\epsilon > 0$, si n est assez grand on a :

$$\begin{aligned}
\mathbb{P}(|\mathscr{T}| = \infty)^2 - \epsilon &\leq \mathbb{P}(E_{v_1,v_2}) \\
&= \frac{1}{n^2}\mathbb{E}(\sum_i C_i^2) \\
&= \frac{1}{n^2}\mathbb{E}\left(\sum_{i:C_i<\sqrt{n}} C_i^2 + \sum_{i:C_i\geq\sqrt{n}} C_i^2\right) \\
&\leq \frac{1}{n^2}\mathbb{E}\left(\sqrt{n}\sum_{i:C_i<\sqrt{n}} C_i + C_1\sum_{i:C_i\geq\sqrt{n}} C_i\right) \\
&\leq \frac{1}{n^2}\mathbb{E}\left(n\sqrt{n} + C_1 N^n_{\sqrt{n}}\right) \\
&= \mathbb{E}\left(\frac{C_1}{n}\frac{N^n_{\sqrt{n}}}{n}\right) + \frac{1}{\sqrt{n}}
\end{aligned}$$

Les variables aléatoires $\frac{C_1}{n}$ et $\frac{N^n_{\sqrt{n}}}{n}$ sont majorées par 1. De plus, avec probabilité tendant vers 1, $\frac{N^n_{\sqrt{n}}}{n} \leq \mathbb{P}(|\mathscr{T}| = \infty) + \epsilon$. Par conséquent, pour tout $\epsilon > 0$, pour n assez grand :

$$\mathbb{P}(|T| = \infty)^2 - \epsilon \leq \mathbb{E}\left(\frac{C_1}{n}\right)(\mathbb{P}(|\mathscr{T}| = \infty) + \epsilon) + \epsilon.$$

En conséquence, nous avons simultanément :

- $\liminf_n \mathbb{E}(\frac{C_1}{n}) \geq \mathbb{P}(|\mathscr{T}| = \infty)$ en faisant tendre ϵ vers 0 ;
- $\frac{C_1}{n} \leq 1$ p.s ;
- Pour tout $\epsilon > 0$, $\mathbb{P}(\frac{C_1}{n} \geq \mathbb{P}(|\mathscr{T}| = \infty) + \epsilon) \to 0$, d'après le théorème 1.5.

Par conséquent $\frac{C_1}{n}$ converge vers $\mathbb{P}(|\mathscr{T}| = \infty)$ en probabilité.

□

1.1.2.f Chemin de Lukasiewicz

L'algorithme d'exploration d'un graphe aléatoire permet également d'étudier certains graphes au seuil critique de la transition de phase. Dans le second modèle d'Erdős et Rényi, le seuil critique correspond à une probabilité d'existence de chaque arête égale à $\frac{1}{n}$ (ou parfois $\frac{1}{n}(1+o(1))$). Dans ce cas, la limite locale est un arbre de Galton-Watson critique, donc presque sûrement finie, et la taille de la plus grande composante est donc $o_p(n)$ par le théorème 1.5. Un résultat plus précis peut être obtenu en utilisant le chemin de Lukasiewicz associé à l'arbre créé par l'algorithme d'exploration.

La suite de cette partie présente l'utilisation de cette technique par Aldous [Ald97] pour étudier les tailles des composantes connexes du graphe d'Erdős et Rényi au voisinage du seuil critique.

Définition 1.9. Le *chemin de Lukasiewicz* $(L_k)_{k\in\mathbb{N}}$ associé au processus d'exploration est défini par :

- $L_0 = 0$.
- Pour tout entier k strictement positif , $L_k = L_{k-1} + \Delta_k - 1$ avec Δ_k le cardinal de $V_w^{\mathcal{T}}$ à la k-ème itération de l'algorithme, c'est-à-dire le nombre de voisins de w_k appartenant à T_k .

Si l'instruction 2' n'est pas utilisée, et que l'algorithme trace donc une forêt couvrante, alors Δ_k est par construction le nombre d'enfants qu'a le sommet w_k dans cette forêt.

Lemme 1.8. *Le processus L vérifie les deux propriétés suivantes :*

1. *le nombre de composantes connexes entièrement explorées par l'algorithme lors des k premières itérations de l'algorithme est égal à $-\min_{k'\leq k} L_{k'}$;*
2. *pour tout k, $|\mathcal{S}_k| = 1 + L_k - \min_{k'\leq k} L_{k'}$.*

Pour tout k, $|\mathcal{U}_k| = k$. Par conséquent le processus L permet de de connaître à tous moments la taille des trois ensembles $\mathcal{S}$, $\mathcal{T}$ et $\mathcal{U}$ utilisés par l'algorithme, car $|\mathcal{T}| = n - |\mathcal{S}| - |\mathcal{U}|$.

Preuve. Ce lemme se démontre par récurrence sur k. Pour $k = 0$, aucune composante connexe n'a été explorée, $\min_{k' \leq 0} L_{k'} = L_0 = 0$ et $|S_k| = 1 = 1 + L_0 - \min_{k' \leq 0} L_{k'}$.

On suppose que le lemme est vérifié pour un certain entier $k-1$. Soit $\mathcal{S}_{k-}$ l'ensemble $\mathcal{S}$ après les instructions 1, 2 et 2' de la k-ème iteration de l'algorithme, mais avant l'instruction 3. Lors des instructions 1, 2 et 2', un sommet (w_k) est déplacé de $\mathcal{S}$ à $\mathcal{U}$, et Δ_k sommets sont déplacés de $\mathcal{T}$ à $\mathcal{S}$, donc $|\mathcal{S}_{k-}| - |\mathcal{S}_{k-1}| = \Delta_k - 1$.

En utilisant l'hypothèse de récurrence pour $k-1$,

$$\begin{aligned} |S_{k-}| &= \Delta_k - 1 + 1 + L_{k-1} - \min_{k' \leq k-1} L_{k'} \\ &= 1 + L_k - \min_{k' \leq k-1} L_{k'} \end{aligned}$$

Deux cas sont possibles. Si $|S_{k-}| \geq 1$, alors $L_k \geq \min_{k' \leq k-1} L_{k'}$, donc le minimum de L est inchangé. Comme $\mathcal{S}_{k-}$ n'est pas vide, l'itération k ne termine pas l'exploration d'une composante, et donc les deux quantités de la première partie du lemme sont inchangées. Comme aucun sommet n'est ajouté à $\mathcal{S}$ à l'instruction 3,

$$\begin{aligned} |\mathcal{S}_k| &= |\mathcal{S}_{k-}| \\ &= 1 + L_k - \min_{k' \leq k-1} L_{k'} \\ &= 1 + L_k - \min_{k' \leq k} L_{k'} \end{aligned}$$

ce qui prouve la deuxième partie du lemme dans ce cas.

Si $\mathcal{S}_{k-}$ est vide, alors $L_k = \min_{k' \leq k-1} L_{k'} - 1$, et l'exploration d'une composante se termine à l'itération k. Par conséquent les quantités de la première partie du lemme sont toutes les deux augmentées de 1. De plus un sommet est ajouté à $\mathcal{S}$ à l'instruction 3, donc :

$$\begin{aligned} |S_k| &= |S_{k-}| + 1 \\ &= 1 + L_k - \min_{k' \leq k-1} L_{k'} + 1 \\ &= 1 + L_k - \min_{k' \leq k} L_{k'}. \end{aligned}$$

□

1.1.2.g Utilisation du chemin de Lukasiewicz pour l'étude du cas critique

Pour tout entier i fixé, au seuil critique du modèle d'Erdős et Rényi, les i plus grandes composantes sont de taille similaire d'ordre $n^{\frac{2}{3}}$. La suite de cette partie décrit les résultats obtenus par Aldous [Ald97] pour la convergence de la suite $(n^{-\frac{2}{3}} C_i, i \geq 1)$ vers la longeur d'excursions d'un certain mouvement brownien lorsque n tend vers l'infini.

On se place dans l'espace $l^2_\searrow$ des suites décroissantes positives $x = (x_1, x_2, \dots)$ telles que $\sum_i x_i^2 < \infty$, muni de la distance euclidienne. Soit t un paramètre réel. On considère le graphe $G(n, \frac{1}{n} + tn^{-\frac{4}{3}})$, que l'on dénotera dans ce paragraphe par G_n^t, de chemin de Lukasiewicz L. On considère le chemin de Lukasiewicz renormalisé $\mathscr{L}$ défini par

$$\mathscr{L}_s = (n^{-\frac{1}{3}} L^n_{\lfloor n^{\frac{2}{3}} s \rfloor})$$

pour tout $s \geq 0$.

Définissons les processus aléatoires suivants : $(W(s))_{s\in\mathbb{R}_+}$ est un mouvement brownien standard. $W^t(s) = W(s) + ts - \frac{1}{2}s^2$ est un mouvement brownien avec dérive parabolique, et $B^t(s) = W^t(s) - \min_{s'\leq s} W^t(s')$ est le processus réfléchi en 0 associé à $W^t(s)$. On appelle excursion de B^t tout intervalle $[g, d]$ tel que $B^t(g) = B^t(d) = 0$ et $B^t(s) > 0$ pour tout s appartenant à $]g, d[$.

En étudiant la loi du degré de w_k, Aldous démontre que $\mathscr{L}$ converge en loi vers W^t. D'après le lemme 1.8, l'écart entre les instants de deux records consécutifs vers le bas de $\mathscr{L}$ correspond à la taille renormalisée des composantes de G_n^t. L'écart entre les instants de deux records consécutifs vers le bas de W^t correspond à la longueur des excursions de B^t. Ce lien permet à Aldous [Ald97] de démontrer le théorème suivant :

Théorème 1.9.

$$(n^{-\frac{2}{3}} C_i(G_n^t), i \geq 1) \xrightarrow[n\to\infty]{d} (\gamma_i, i \geq 1),$$

la convergence ayant lieu dans $l^2_\searrow$.

1.1.3 Modèle de Söderberg

Dans [SÖ2], Söderberg introduit une généralisation inhomogène multitype du modèle d'Erdős et Rényi. Le caractère inhomogène consiste à considérer qu'il existe différents types de sommets, et que la probabilité que deux sommets soient reliés dépend de leur type.

1.1.3.a Définition des modèles

Deux modèles proches seront définis dans cette partie. Les paramètres sont :

- un entier ℓ désignant le nombre de types ;
- un élement r du $(\ell - 1)$ simplexe, qui désigne la probabilité de chaque type ;
- Une matrice P symétrique de taille $\ell \times \ell$ de réels compris entre 0 et 1. Le coefficient (i, j) de la matrice P représente la probabilité qu'une arête entre un sommet de type i et un sommet de type j existe.

Définition 1.10. *Le modèle de Söderberg* $\tilde{G}^S(n, r, P)$ est le graphe simple aléatoire à n sommets défini par :

- Un type est attribué à chaque sommet indépendamment et selon la loi r.
- Conditionnellement à l'ensemble des types, chacune des $\binom{n}{2}$ arêtes est présente indépendamment, avec probabilité $P_{i,j}$ si ses extrémités sont de type i et j.

Ce modèle est une généralisation du second modèle d'Erdős et Rényi, que l'on retrouve en prenant $l = 1$, $P = (p)$ et $r = (1)$.

Il est parfois intéressant de regarder le graphe $\tilde{G}^S(n, r, P)$ conditionnellement aux nombres de sommets de chaque type, pour étudier uniquement l'influence de l'aléa des arêtes. Soit N un ℓ-uplets d'entiers positifs et $n = \sum_i N_i$.

Définition 1.11. Le *modèle de Söderberg* $G^S(N, P)$ est le graphe simple défini par :

- L'ensemble des sommets est $\{(v, i) \in \mathbb{N} \times \{1, \ldots, \ell\}, 1 \leq v \leq N_i\}$. La seconde coordonnée d'un sommet désigne son type.
- Chacune des $\binom{n}{2}$ arêtes est présente indépendamment. L'arête entre (v, i) et (w, j) est présente avec probabilité $P_{i,j}$.

Si N suit une loi multinomiale de paramètres n et r, $G^S(N, P)$ suit la loi de $\tilde{G}^S(n, r, P)$.

1.1.3.b Apparition de la composante géante

Le modèle de Söderberg exhibe une transition de phase similaire à celle du modèle d'Erdős et Rényi (théorème 1.1) : la taille de la plus grande composante connexe dans le modèle de Söderberg change drastiquement lorsque la matrice des probabilités d'existence des arêtes dépasse un certain seuil.

On suppose que la matrice des probabilités d'existence des arêtes est de la forme

$$P_{i,j} = \min\left(\frac{\mathscr{P}_{i,j}}{n}, 1\right)$$

où $\mathscr{P}$ est une matrice symétrique positive irréductible (si $\mathscr{P}$ est réductible, le graphe peut être décomposé en deux ou plusieurs sous-graphes disjoints, étudiables séparément). Posons

$$H_{i,j} = r_j \mathscr{P}_{i,j}$$

pour tout i et j appartenant à $\{1, \ldots, \ell\}$. Notons qu'en vertu du théorème de Perron-Frobenius, la plus grande valeur propre λ de H est réelle positive.

Söderberg démontre [S̈02] le résultat suivant :

Théorème 1.10. *La taille des composantes connexes de $\tilde{G}^s(n,r,P)$ présente un transition de phase lorsque λ traverse 1 :*

- *Si $0<\lambda<1$, $C_1(\tilde{G}^s(n,r,P))=\Theta_p(\ln(n))$.*
- *Si $\lambda=1$, $\forall i\in\mathbb{N}^*$, $C_i(\tilde{G}^s(n,r,P))=\Theta_p(n^{\frac{2}{3}})$.*
- *Si $\lambda>1$, $C_1(\tilde{G}^s(n,r,P))=\Theta_p(n)$, et $C_2(\tilde{G}^s(n,r,P))=\Theta_p(\ln(n))$.*

1.1.3.c Interprétation de la transition de phase via la limite locale

Dans cette section nous allons interpréter H comme un paramètre naturel de la limite locale, via le processus d'exploration introduit à la partie 1.1.2.d. Nous allons ici considérer le graphe $G^S(N,P)$, qui possède des propriétés intéressantes d'indépendance entre les arêtes. Les résultats sur $\tilde{G}^S(n,r,P)$ peuvent ensuite être obtenus en conditionnant par N, le nombre de sommets de chaque type. Par la loi des grands nombres, pour tout i, N_i est proche de nr_i.

Soit k un entier fixé. On désigne par w le sommet considéré lors de la k-ème itération de l'algorithme, et l'on note par i son type. L'entier k étant fixé, le nombre de sommets déjà découverts par l'algorithme est négligeable devant n, donc le nombre de sommets de chaque type non encore découverts est proche de nr. Les arêtes entre w et les sommets non encore découverts sont présentes indépendamment, avec probabilité $P_{i,j}$ où j est le type de l'autre extrémité. En faisant tendre le nombre de sommets vers l'infini, la loi du nombre de voisins de w_k de chaque type découverts par l'algorithme d'exploration converge donc vers ℓ variables de Poisson indépendantes, de paramètres respectifs $\lim_n N_j P_{i,j} = r_j \mathscr{P}_{i,j} = H_{i,j}$.

Ces observations, permettent de montrer que la limite locale de $\tilde{G}^S(n,r,P)$ enraciné en un sommet uniforme est l'arbre de Galton-Watson multitype suivant :

- Le type de la racine est choisi aléatoirement selon la loi r.
- La loi de reproduction est définie par :
  - Le nombre d'enfants de chaque type est indépendant.
  - Le nombre d'enfants de type j d'un sommet de type i est une variable de Poisson de paramètre $H_{i,j}$.

Le caractère critique d'un tel arbre de Galton-Watson dépend de la plus grande valeur propre λ de H. Il est sous-critique si $\lambda<1$, critique si $\lambda=1$ et surcritique si $\lambda>1$, ce qui explique heuristiquement la transition de phase du théorème 1.10.

1.1.3.d Étude de la fenêtre critique : le modèle

Le reste de cette partie est consacrée aux résultats obtenus dans le chapitre 2 de ma thèse, à propos de l'étude de la fenêtre critique dans le modèle de Söderberg.

On s'intéresse au graphe $G^S(N^n, P^n)$ où N^n et P^n sont de la forme suivante :

$$\begin{aligned} \forall n, \sum_{i=1}^{\ell} N_i^n &= n. && (1.5)\\ \forall i \in \{1, \dots, \ell\},\ N_i^n &= r_i n + \tilde{r}_i n^{\frac{2}{3}} + o(n^{\frac{2}{3}}) && (1.6)\\ P_{i,j}^n &= \min\left(1, (\frac{1}{n} + t n^{-\frac{4}{3}}) \mathscr{P}_{i,j}\right) && (1.7) \end{aligned}$$

où r est un élément du $\ell - 1$ simplexe ouvert, $\tilde{r}$ est un vecteur de $\mathbb{R}^\ell$ de somme nulle, $\mathscr{P}$ est une matrice symétrique réelle positive irréductible de taille $\ell \times \ell$ et t un réel. On définit la matrice $H \in M_\ell(\mathbb{R}_+)$ par $H_{i,j} = r_j \mathscr{P}_{i,j}$. On suppose que la plus grande valeur propre λ de H est 1.

La proportion de sommets de chaque type converge vers r, avec des perturbations d'ordre $O(n^{\frac{2}{3}})$. S'il l'on se place dans le modèle $\tilde{G}^S(n, r, P^n)$, où le type de chaque sommet est choisi indépendamment selon la loi r, le théorème central limite entraîne (1.6) avec $\tilde{r} = 0$.

La matrice $\mathscr{P}$ représente les probabilités d'existence d'arête avant normalisation. Sans perte de généralité, on peut supposer les colonnes de $\mathscr{P}$ sont deux à deux différentes. Si deux colonnes sont identiques, deux types sont indifférentiables, et il est donc possible de les fusionner.

Le graphe aléatoire $G^S(N^n, P^n)$ sera noté G_n^S pour simplifier les notations.

1.1.3.e Lien entre un mouvement brownien de dimension ℓ et la taille des composantes de G_n^S

H est une matrice réelle positive irréductible, de plus grande valeur propre 1. Les sous-espaces propres à gauche et à droite associée à la valeur propre 1 sont donc de dimension 1, et il existe un vecteur propre à droite et un vecteur propre à gauche à coefficients strictement positifs. Soient x et y les vecteurs propres à droite et à gauche de H associé à la valeur propre 1 tels que :

- $\sum_i x_i = 1$.
- $\sum_i y_i = 1$.

On considère alors le mouvement brownien $W(s)$ de dimension ℓ et de matrice de covariance :

$$\begin{pmatrix} y_1 & 0 & 0 & \dots & \\ 0 & y_2 & 0 & \dots & \\ & & \dots & & \\ & & & \dots & \\ & & & 0 & y_\ell \end{pmatrix}$$

i.e. $V_{i,j} = \delta_{i,j} y_i$.

La dérive ρ est définie par :

$$\rho_i(s) = y_i s t + \frac{\tilde{r}_i y_i}{r_i} s - \frac{y_i^2}{2r_i} s^2.$$

Le mouvement brownien de dimension ℓ avec dérive parabolique $W^t(s)$ est défini par $W^t(s) = W(s) + \rho(s)$. Soit $\mathscr{W}^t$ le mouvement brownien avec dérive parabolique défini par :

$$\mathscr{W}^t(s) = \sum_i x_i W_i^t(s).$$

On note par $\mathscr{W}_+^t$ le mouvement brownien $\mathscr{W}^t$ réfléchi en 0, défini par

$$\mathscr{W}_+^t(s) = \mathscr{W}^t(s) - \inf_{s' \le s} \mathscr{W}^t(s').$$

On note par (γ_i) la suite décroissante des longueurs d'excursions de $\mathscr{W}_+^t$, qui est p.s. bien définie et dans $l^2_\searrow$.

Théorème 1.11. *Sous les conditions (1.5) et (1.6),*

$$(n^{-\frac{2}{3}} C_i(G^s(N^n, P^n)), i \ge 1) \xrightarrow[n\to\infty]{d} (\gamma_i, i \ge 1),$$

la convergence ayant lieu dans $l^2_\searrow$.

Remarque. En prenant un seul type, $r = 1$, $\tilde{r} = 0$ et $P = (1)$, on retrouve le résultat d'Aldous pour le graphe d'Erdős et Rényi [Ald97].

1.1.3.f Une généralisation multi-dimensionnelle du chemin de Lukasiewicz

Le chemin de Lukasiewicz ne faisant pas la différence entre les différents types de sommets, nous introduisons une variante ℓ-dimensionnelle du chemin de Lukasiewicz, plus adaptée à l'étude du modèle de Söderberg :

Soit $(Z_k^i, i \in \{1, \dots, \ell\}, k \ge 0)$ le processus défini par :

- Pour tout $i \in \{1, \dots, \ell\}$, $Z_0^i = 0$.

- Pour tout $i \in \{1, \dots, \ell\}$ et tout entier k non nul,

$$Z_k^i = Z_{k-1}^i + \Delta_k^i - H_{d_k,i}$$

où Δ_k^i est le nombre de sommets de type i appartenant à $V_w^{\mathcal{T}}$ à la k-ème itération de l'algorithme d'exploration et d_k est le type du sommet w_k.

Remarque.

- Le processus tient compte du types des sommets, et pas uniquement du nombre de sommets découverts.
- Chaque processus Z^i compte le nombre de sommets de type i découverts.
- Au lieu de soustraire 1 à chaque étape comme dans le chemin de Lukasiewicz, on soustrait $H_{d_k,i}$, ce qui permet d'obtenir une convergence vers un mouvement brownien. Ce choix découle de la limite locale décrite à la partie 1.1.3.c, puisque $H_{d_k,i}$ est une approximation de l'espérance de Δ_k^i.

Théorème 1.12.

$$(n^{-\frac{1}{3}} Z_{\lfloor sn^{\frac{2}{3}} \rfloor}, s \geq 0) \xrightarrow[n\to\infty]{d} (W^t(s), s \geq 0).$$

On définit $\mathscr{Z}$ par :

$$\mathscr{Z}_k = \sum_i x_i Z_k^i,$$

$\mathscr{Z}$ permet d'obtenir des informations sur la taille des composantes explorées par l'algorithme et sur $\mathcal{S}$, similaires à celles obtenues par le chemin de Lukasiewicz :

- Si l'algorithme termine l'exploration d'une composante à l'étape k, alors $-\mathscr{Z}$ atteint un nouveau record en k (mais, contrairement au chemin de Lukasiewicz, il peut exister d'autres records).
- Il existe trois constantes strictement positives K_1, K_2 et K_3, dépendant de H et x, telles que pour tout k,

$$K_1(\mathscr{Z}_k - \min_{k' \leq k} \mathscr{Z}_{k'}) \leq |S_k| \leq K_2(\mathscr{Z}_k - \min_{k' \leq k} \mathscr{Z}_{k'}) + K_3.$$

Remarque. L'existence de records *parasites* de $-\mathscr{Z}$ ne correspondant pas à la fin de l'exploration d'une composante pourrait poser problème pour la suite du raisonnement, puisque l'écart entre deux records consécutifs ne correspond plus forcément à la taille d'une composante, mais il est possible de prouver que l'effet de ces erreurs est négligeable devant $n^{\frac{2}{3}}$.

La convergence vers W implique que $(n^{-\frac{1}{3}}\mathscr{Z}_{\lfloor n^{\frac{2}{3}}s\rfloor}, s \geq 0) \xrightarrow[n\to\infty]{d} \mathscr{W}^t$. Nous pouvons ainsi relier les records de $-\mathscr{Z}$ avec les records de $-\mathscr{W}^t$, c'est-à-dire les points d'annulation de $\mathscr{W}^t_+$, et ainsi expliquer le lien entre la taille des composantes de G_n^S et la longueur des excursions de $\mathscr{W}^t_+$ donné dans le théorème 1.11.

1.1.4 Graphe aléatoire dynamique d'Erdős et Rényi avec contrainte de degré

1.1.4.a Le modèle

Cette partie est consacrée aux résultats obtenus dans le chapitre 3 de ma thèse, sur l'existence ou l'absence de composante géante dans le multigraphe dynamique $(G^k_{n,t})_{t\geq 0}$ où n et k sont deux entiers, désignant respectivement le nombre de sommets et le degré interdit. Ce problème m'a été suggéré par Itai Benjamini, que je remercie ici. On note par E l'ensemble des $\binom{n}{2}$ arêtes du graphe simple complet à n sommets. L'apparition des arêtes dans $(G^k_{n,t})_{t\geq 0}$ est décrite par un processus de Poisson ponctuel Π sur $\mathbb{R}_+ \times E$ d'intensité $\frac{1}{n}\mathbf{1_{t\geq 0}}dt \otimes \mu$ où μ est la mesure de comptage sur E.

Définition 1.12. Pour tout réel positif t, nous définissons le graphe $G^k_{n,t}$ de la manière suivante. On considère la suite $(t_i, e_i)_{1\leq i\leq l}$ des éléments de $\Pi([0,t]\times E)$, triés par ordre croissant de la première coordonnée. Le graphe $\mathscr{G}^i$ est défini par récurrence par :

- $\mathscr{G}^0$ est le graphe avec n sommets et aucune arête.
- Pour tout i, on part du graphe $\mathscr{G}^{i-1}$ et l'on augmente la multiplicité de e_i de 1. Si le degré de l'une des extrémités de e_i vaut k, alors toutes les arêtes adjacentes à cette extrémité sont supprimées (y compris e_i). Si les deux extrémités de e_i sont de degré k, les arêtes adjacentes aux deux extrémités sont supprimées. $\mathscr{G}^i$ est le graphe ainsi obtenu.
- $G^k_{n,t}$ est le graphe $\mathscr{G}^l$, obtenu une fois que tous les éléments de Π ont été utilisés.

Presque sûrement, pour tout temps t, $\Pi([0,t]\times E)$ est fini et les t_i sont tous différents, donc $(G^k_{n,t})_{t\geq 0}$ est p.s. bien défini. Il s'agit d'une chaîne de Markov à temps continu et à ensemble d'états finis (l'ensemble des graphes à n sommets de degré majoré par $k-1$). Si l'on ne supprime pas les arêtes, on obtient une version multigraphe du modèle d'Erdős-Rényi, qui sera notée $G^\infty_{n,t}$.

Il est parfois utile de considérer la suite de graphes $\mathscr{G}^i$, version à temps discret de ce modèle. Le modèle devient alors :

- Initialement le graphe possède n sommets mais pas d'arête.
- À chaque étape, une arête est choisie uniformément. Si le degré de l'une des extrémités atteint ou dépasse k, on supprime toutes les arêtes adjacentes à cette extrémité (y compris celle qui vient d'être ajoutée). Si les deux extrémités atteignent le degré k, on supprime les arêtes adjacentes aux deux extrémités.

1.1.4.b Difficultés du modèle

Plusieurs caractéristiques de ce modèle compliquent son étude :

- Le graphe n'est pas monotone : il y a non seulement adjonction, mais aussi suppression d'arêtes. En conséquence, la taille de la plus grande composante n'est pas monotone : l'existence d'une composante géante à un temps t ne garantit pas la présence d'une composante géante à un temps $t' > t$.

- Il n'est pas facile d'étudier le graphe localement. Pour connaître les arêtes adjacentes à un sommet donné dans $G^k_{n,t}$, il faut être capable de déterminer si les arêtes adjacentes ajoutées avant le temps t ont été supprimées depuis, ce qui dépend du degré de leur autre extrémité. Le degré de leur autre extrémité dépend des arêtes adjacentes supprimées et ainsi de suite. Cela implique que l'étude d'un voisinage d'un sommet peut nécessiter de regarder à une distance arbitraire de ce sommet.

1.1.4.c Résultats

Les principaux résultats de la partie 3 sont les suivants :

Théorème 1.13. *Pour $k \leq 3$, et toute suite de réels positifs t_n :*

$$C_1(G^k_{n,t_n}) = o_p(n).$$

Il n'y a donc pas de composante géante pour $k \leq 3$.

Théorème 1.14. *Lorsque $k \geq 5$, il existe t tel que*

$$C_1(G^k_{n,t}) = \Theta_p(n).$$

Il existe une composante géante pour $k \geq 5$, pour certains temps t.

Théorème 1.15. *Pour tout k et t fixé, il existe un arbre aléatoire T^k_t tel que pour tout entier l, $G^k_{n,t}$ enraciné en l sommets indépendants converge localement vers l copies indépendantes de T^k_t.*

De plus on peut caractériser l'existence d'une composante géante à l'aide de T^k_t :

Théorème 1.16. *Pour tout t et k :*

$$\frac{C_1(G^k_{n,t})}{n} \xrightarrow[n\to\infty]{p} \mathbb{P}(|T^k_t| = \infty).$$

1.1.4.d Étude de $k \leq 3$

Si $k \leq 3$, le degré maximal est inférieur ou égal à 2. Par conséquent les composantes sont des chemins ou des cycles. Soit A (resp. C) l'ensemble des composantes acycliques (resp. l'ensemble des cycles) de $G^k_{n,t}$. On définit la quantité

$$Z = \sum_{a \in A} |a|^2 + 2 \sum_{c \in C} |c|^2$$

qui va permettre de borner la taille des composantes de $G^k_{n,t}$. En étudiant la loi d'évolution de $G^k_{n,t}$, on arrive à montrer qu'il existe une constante K telle que pour tout t et n :

$$\mathbb{E}(Z) \leq Kn.$$

Cela implique que $C_1(G^k_{n,t})$ est borné stochastiquement par $\sqrt{n}$, et donc qu'il n'existe pas de composante géante.

1.1.4.e Étude de $k \geq 5$

Pour démontrer le théorème 1.14, nous minorons $G^k_{n,t}$ par un graphe $g^k_{n,t}$ plus simple à étudier et qui contient une composante géante pour t bien choisi. Le graphe $g^k_{n,t}$ est le graphe obtenu en partant de $G^\infty_{n,t}$ et en supprimant toutes les arêtes adjacentes à un sommet de degré au moins k dans $G^\infty_{n,t}$.

Lemme 1.17.

$$g^k_{n,t} \leq G^k_{n,t} \leq G^\infty_{n,t}$$

Par construction $g^k_{n,t}$ et $G^k_{n,t}$ sont des sous-graphes de $G^\infty_{n,t}$. Soit e une arête de $G^\infty_{n,t}$ qui n'est pas présente dans $G^k_{n,t}$. Cela signifie que l'une des extrémités de e atteint le degré k dans $G^k_{n,t'}$ pour $t' \leq t$, donc possède un degré supérieur ou égal à k dans $G^\infty_{n,t'}$ et donc dans $G^\infty_{n,t}$. Par conséquent, l'arête e n'est pas présente dans $g^k_{n,t}$.

L'intérêt du graphe $g^k_{n,t}$ est qu'il ne dépend que de l'ensemble des arêtes ajoutées, et pas de l'ordre d'ajout de ces arêtes. De plus, conditionné par le nombre de sommets de chaque degré, $g^k_{n,t}$ est un modèle de configuration, tel qu'introduit par Bender et Canfield [BC78, page 297] et Molloy et Reed [MR95, page 7] :

Définition 1.13. Soit (d_i) une suite finie d'entiers positifs, de somme n et telle que $\sum id_i$ est pair. Le modèle de configuration correspondant à $(d_i)_{i \geq 0}$ est le graphe aléatoire construit de la manière suivante :

- L'ensemble de sommets V est $\cup_i D_i$ avec D_i des ensembles disjoints tels que $|D_i| = d_i$.
- On considère l'ensemble $L = \cup_i D_i \times \{1, \ldots, i\}$. Pour chaque élément x de D_i, L contient donc $(x, 1)$, $(x, 2)$, $\ldots$, (x, i).

- On choisit un appariement aléatoire $E \subset L^2$ uniforme des éléments de L.
- L'ensemble des arêtes est E et la fonction d'incidence φ est définie par $\varphi(\{(x,i),(y,j)\}) = \{x,y\}$.

Le graphe aléatoire (V, E, φ) ainsi construit est appelé *le modèle de configuration associé à d*. Par construction ce graphe contient d_i sommets de degré i, pour tout i.

Pour démontrer l'existence de la composante géante, nous utilisons le résultat de Molloy et Reed [MR95, théorème 1] dont voici une version simplifiée suffisante pour notre usage :

Théorème 1.18 (Critère d'existence d'une composante géante pour le modèle de configuration)**.** *Soit* $((d_i(n))_{0\leq i\leq k-1}, n \in \mathbb{N})$ *k suites d'entiers positifs ou nuls tels que :*

- *pour tout n,* $\sum_i d_i(n) = n$ *;*
- *pour tout n,* $\sum_i i d_i(n)$ *est pair ;*
- *pour tout* $0 \leq i \leq k-1$, $\frac{d_i(n)}{n}$ *converge vers une limite, notée* λ_i*.*

On note par G^n *le modèle de configuration correspondant à* $(d_i(n))_{0\leq i\leq k-1}$*. On note par Q la quantité* $\sum_{i>0} i(i-2)\lambda_i$*. Si* $Q > 0$*, alors il existe une constante* $K > 0$ *telle que :*

$$C_1(G^n) \geq Kn \text{ a.p.s.}$$

Le résultat de Molloy et Reed est en réalité plus général, puisqu'il permet aussi de travailler avec des graphes dont le degré n'est pas uniformément borné, sous quelques restrictions techniques additionnelles. Dans notre cas, le degré des sommets est borné par $k-1$, donc ces hypothèses techniques sont automatiquement vérifiées et cette version simplifiée est suffisante. Pour appliquer ce théorème, il suffit donc de vérifier l'existence des λ_i et de calculer le signe de Q.

Lemme 1.19. *Pour tout* $i < k$, $\frac{|\{v \in g^k_{n,t} : deg(v) = i\}|}{n}$ *converge vers une limite* $p_{t,d}$ *et*

$$\sum_i i(i-2)p_{t,d} > 0 \Leftrightarrow e^{-t}\sum_{i=1}^{k-2}\frac{t^i}{(i-1)!} > 1.$$

Par construction de $g^k_{n,t}$, le degré d'un sommet v dans $g^k_{n,t}$ dépend uniquement de la boule de rayon 2 centré sur v dans $G^\infty_{n,t}$. Le graphe $G^\infty_{n,t}$ est un graphe d'Erdős-Rényi de paramètre $\frac{t}{n}$, donc d'après le théorème 1.2, $G^\infty_{n,t}$

enraciné en un sommet uniforme converge localement vers l'arbre de Galton-Watson T_t^{GW}, de loi de reproduction poissonnienne de paramètre t, ce qui permet de montrer l'existence de $p_{t,d}$ et de calculer sa valeur.

Pour tout $k \geq 5$, il existe $t_k^- < t_k^+$ tels que $\sum_{i=1}^{k-2} \frac{t^i}{(i-1)!} > 1$ si $t \in]t_k^-; t_k^+[$. En utilisant le théorème 1.18, cela implique le théorème 1.14 et donc l'existence d'une composante géante.

Pour $k = 4$, ou $k \geq 5$ et $t \geq t_k^+$, les deux arguments précédents ne permettent pas de montrer la présence ou l'absence d'une composante géante.

1.1.4.f Limite locale

$G_{n,t}^{\infty}$ est un graphe d'Erdős et Rényi de paramètre $\frac{t}{n}$, donc converge localement vers T_t^{GW}, l'arbre de Galton-Watson de loi de reproduction poissonnienne de paramètre t d'après le théorème 1.2. Néanmoins, cette convergence ne tient pas compte du temps d'ajout des arêtes, nécessaire pour connaître le graphe $G_{n,t}^k$. Pour cette raison, nous considérons le graphe $G_{n,t}^{\infty}$ dont les arêtes sont étiquetées par le temps d'ajout. En étudiant la limite locale, nous constatons qu'apparaît le *Poisson Weighted Infinite Tree* en dimension 1, défini par Aldous et Steele [AS04] :

Définition 1.14. Le *Poisson Weigthed Infinite Tree* (PWIT) en dimension d est un processus de branchement dont les arêtes sont étiquetées, construit récursivement de la manière suivante :

- à la racine est attachée une infinité d'arêtes, étiquetées par les valeurs d'un processus de Poisson d'intensité $f(x) = \frac{1}{d} x^{d-1} \mathbf{1}_{\mathbf{x \geq 0}}$.
- conditionnellement aux n premières générations, à chaque sommet de la n-ème génération on attache une infinité d'arêtes dont les étiquettes forment une copie d'indépendante d'un processus de Poisson d'intensité $f(x) = \frac{1}{d} x^{d-1} \mathbf{1}_{\mathbf{x \geq 0}}$.

Nous nous plaçons dans le cas $d = 1$, où l'intensité du Processus de Poisson est constante et égale à 1 et le PWIT sera noté dans ce cas par T_{∞}^{∞}. Considérons le sous-arbre du PWIT limité aux arêtes d'étiquettes inférieures ou égales à t :

Définition 1.15. On note par $PPP(t)$ le processus ponctuel de Poisson d'intensité 1 sur $[0, t]$.

T_t^{∞} est un processus de branchement dont les arêtes sont étiquetées défini de la manière suivante :

- on note par S_n les sommets de la n-ème génération de T_t^{∞} ;
- pour tout sommet v de T_t^{∞}, on note par Π_v l'ensemble des étiquettes entre v et ses enfants dans T_t^{∞} ;

- pour tout n, conditionnellement à $(\Pi_v)_{v\in\cup_{i<n}S_i}$, $(\Pi_v)_{v\in S_n}$ est une famille i.i.d. de copies de $PPP(t)$.

Remarque. En utilisant les propriétés des processus de Poisson, deux définitions équivalentes de T_t^∞ sont :

1. Considérons le graphe obtenu à partir du PWIT en dimension 1 et en supprimant toutes les arêtes d'étiquettes supérieures à t. La composante de la racine dans le graphe ainsi obtenu a même loi que T_t^∞. Cela justifie la notation T_∞^∞ adoptée ici pour le PWIT.

2. T_t^∞ est l'arbre obtenu en prenant l'arbre T^t et en étiquetant chaque arête indépendamment par une variable uniforme sur $[0,t]$.

Conditionnellement à $G_{n,t}^\infty$, le temps d'ajout de chaque arête de $G_{n,t}^\infty$ est uniforme sur $[0,t]$, indépendamment sur l'ensemble des arêtes. Comme le graphe d'Erdős-Rényi non étiqueté enraciné en un sommet uniforme converge localement vers T^t, et en utilisant la deuxième définition équivalent de T_t^∞, nous obtenons que le graphe $G_{n,t}^\infty$ enraciné en un sommet uniforme et dont les arêtes sont étiquetées par le temps d'ajout converge localement vers T_t^∞.

Même si $G_{n,t}^\infty$ converge localement, il n'est pas évident que $G_{n,t}^k$ converge localement, et que la limite puisse s'exprimer en fonction de T_t^∞. En effet, la version avec contrainte de degré d'un graphe n'est a priori bien définie que pour un graphe fini. On note par $\Omega_{<\infty}$ l'ensemble des graphes finis à arêtes étiquetées, dont les étiquettes sont toutes différentes. Nous utilisons la construction suivante pour définir $\tilde{T}_t^k$, la version avec contrainte de degré de T_t^∞ :

Définition 1.16. Soit e une arête de T_t^∞, l un entier positif. On considère l'ensemble $\Omega_{<\infty}(e,l)$ des éléments $\tilde{T}$ de $\Omega_{<\infty}$ avec une arête distinguée $\tilde{e}$ tels que les boules de rayon l $B^l(T_t^\infty, e)$ et $B^l(\tilde{T}_t^\infty, \tilde{e})$ soient isomorphes.

- Si $\tilde{e}$ est présente dans la version avec contrainte de degré de $\tilde{T}$ pour tous les graphes $\tilde{T} \in \Omega_{<\infty}(e,l)$, alors e appartient à la version avec contrainte de degré de T_t^∞.

- Si $\tilde{e}$ est supprimée dans la version avec contrainte de degré de $\tilde{T}$ pour tous les graphes $\tilde{T} \in \Omega_{<\infty}(e,l)$ alors e n'appartient pas à la version avec contrainte de degré de T_t^∞.

Si l'un de ces deux cas a lieu, on dit que la connaissance de la boule de rayon l est suffisante pour connaître si e est supprimée dans la version avec contrainte de degré de T_t^∞.

Lemme 1.20. *Presque sûrement pour tout temps t et toute arête e de T_t^∞, il existe un entier fini $l_{t,e}$ tel que la connaissance de la boule de rayon l est suffisante pour connaître si e est supprimée dans la version avec contrainte de degré de T_t^∞.*

Le lemme 1.20 implique donc que $\tilde{T}_t^k$ la version avec contrainte de degré de T_t^∞ est bien définie : $\tilde{T}_t^k$ est un sous-graphe de T_t^∞ avec le même ensemble de sommets, et il est presque sûrement possible de déterminer quelles arêtes de T_t^∞ appartiennent $\tilde{T}_t^k$ en regardant à une distance aléatoire mais finie. $\tilde{T}_t^k$ est un sous-graphe de T_t^∞, et est donc généralement une forêt. On note par T_t^k la composante connexe de T_t^k contenant et enracinée en la racine de T_t^∞.

Corollaire 1.21. *Pour tout entier l, le graphe étiqueté $G_{n,t}^k$ enraciné en l sommets aléatoires uniformes converge localement en distribution vers l copies indépendantes de T_t^k.*

$G_{n,t}^\infty$ enraciné en l sommets aléatoires uniformes $(v_1, \dots, v_l)$ converge localement en distribution vers l copies indépendantes de T_t^∞, que l'on note $T_1, \dots, T_l$. Grâce au théorème de représentation de Skorokhod, nous pouvons supposer que la convergence a lieu presque sûrement. Cela implique que pour tout entier d, pour n assez grand, la boule de rayon d centrée sur la l-ème racine est la même dans $G_{n,t}^\infty$ et dans T_l. Le lemme 1.20 implique qu'il est possible de définir T_i^k la version avec contrainte de degré de chaque T_i, et qu'il est possible de déterminer si une arête e de T_i appartient à T_i^k en considérant une boule de rayon aléatoire, mais fini presque sûrement. Par conséquent, $G_{n,t}^k$ enraciné en v converge localement vers $\cup_{1\leq i\leq l} T_i^k$, c'est-à-dire vers l copies indépendantes de T_t^k.

1.1.4.g Étude de la limite locale

T_t^∞ est un processus de branchement, mais qu'en est-il de T_t^k ? La partie 3.4.3 montre que, mis à part la première génération dont la loi est différente, T_t^k peut être décrit par un processus de branchement multitype, avec ensemble des types $[0, t]$, le type d'un sommet correspondant en réalité à la date d'ajout de l'arête le reliant à son père dans T_t^∞. Pour comprendre l'apparition de ce processus de branchement, nous allons raisonner sur divers graphes construits à partir de T_t^∞. Nous admettons ici que la version avec contrainte de degré de tous les graphes invoqués est bien définie.

Définition 1.17.

- Soit $(T_s^{\infty,u})_{0\leq s\leq t}$ le graphe dynamique obtenu en prenant le graphe dynamique $(T_s^\infty)_{0\leq s\leq t}$ et en y ajoutant au temps u un sommet v et une arête e entre v et la racine. Un exemple se trouve à la figure 1.2 page 36.
- Pour tout $0 \leq s \leq t$, soit $T_s^{k,u}$ la version de $T_s^{\infty,u}$ avec contrainte de degré.
- Soit $X_u \in [0, t] \cup \{\infty\}$ le premier instant s où l'arête e est supprimée de $(T_s^{k,u})_{0\leq s\leq t}$. Si l'arête e est présente dans $T_t^{k,u}$, on note $X_u = \infty$.

- Soit $T_t^{k+,u}$ le graphe ayant la loi de $T_t^{k,u}$ conditionné par $X_u = \infty$.

L'arête e est appelée *arête racine*.

On considère un arbre T avec les arêtes étiquetées, et une arête $e = (v_1, v_2)$ de ce graphe séparant T en deux arbres T_1 et T_2. On considère T_1^e (resp. T_2^e) égal à l'arbre T_1 (resp. T_2) auquel on ajoute v_2 (resp. v_1) et l'arête e, tel qu'illustré sur la figure 1.1. On note par u l'étiquette de e, T^k la version de T avec contrainte de degré, et $T_i^{k,e}$ la version de T_i^e avec contrainte de degré.

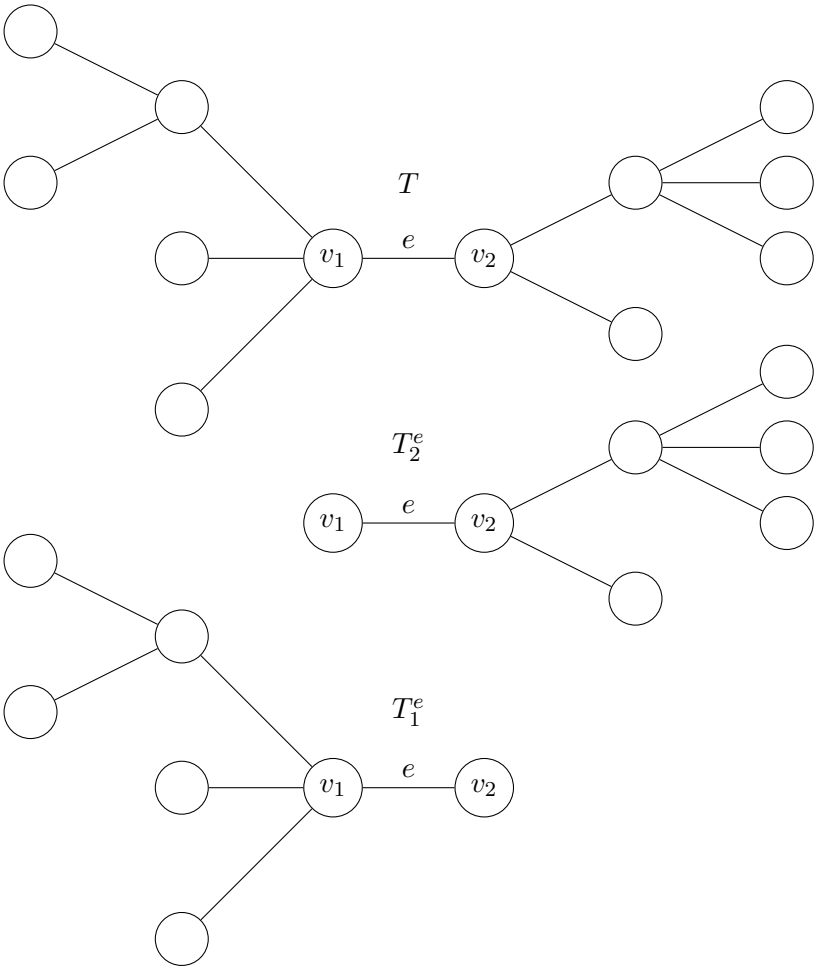

FIGURE 1.1 – Décomposition de T en T_i^e

Lemme 1.22. *L'arête e est présente dans T^k si et seulement si elle est présente dans $T_1^{k,e}$ et dans $T_2^{k,e}$.*

Démonstration. Pour tout $0 \leq s \leq t$, on considère le graphe T_s (resp. $T_{i,s}^e$) défini comme le sous-graphe de T (resp. T_i^e) restreint aux arêtes d'étiquettes

inférieures à s. Si l'on considère T comme un graphe dynamique croissant, T_s est l'état de T au temps s. On considère la variante de la version avec contrainte de degré associée à un arbre $\mathscr{T}$ et à une arête e de $\mathscr{T}$ définie de la manière suivante :

- Chaque arête de $\mathscr{T}$, d'étiquette s, est ajoutée au temps s.
- Si un sommet v atteint le degré k, toutes les arêtes adjacentes à v, à l'exception de e, sont supprimées.
- L'arête e n'est jamais supprimée.

Pour tout $0 \leq s \leq t$, on note par $\tilde{T}_s^k$, $\tilde{T}_{1,s}^{k,e}$ et $\tilde{T}_{2,s}^{k,e}$ cette variante associée respectivement à T_s, $T_{1,s}^e$ et $T_{2,s}^e$ et à l'arête e. Comme l'arête e n'est jamais supprimée, l'évolution d'un coté de l'arête e n'influe pas sur l'évolution de l'autre coté de l'arête e. Par conséquent, à tout moment s, le graphe $\tilde{T}_s^k$ est égal au recollement de $\tilde{T}_{1,s}^{k,e}$ et $\tilde{T}_{2,s}^{k,e}$ au niveau de l'arête e, par l'opération inverse de celle utilisée pour construire T_1^e et T_2^e.

De plus, par définition de $\tilde{T}_s^k$ et $\tilde{T}_{i,s}^{k,e}$, T_s^k (resp. $T_{i,s}^{k,e}$) est égal à $\tilde{T}_s^k$ (resp. $\tilde{T}_{i,s}^{k,e}$) tant que e n'est pas supprimée. Par conséquent, le premier instant s_e où l'une des extrémités de e atteint le degré k est le même dans $(T_s^k)_{0\leq s\leq t}$ et $(\tilde{T}_s^k)_{0\leq s\leq t}$. Si l'extrémité v_i atteint le degré k au temps s_e, v_i atteint le degré k dans $(\tilde{T}_{i,s}^{k,e})_{0\leq s\leq t}$ au temps s_e, et donc dans $(T_{i,s}^{k,e})_{0\leq s\leq t}$. Par conséquent, si l'arête e est supprimée dans T^k, elle est supprimée dans $T_1^{k,e}$ ou $T_2^{k,e}$. □

Pour chaque arête $e_s = (\emptyset, v_s)$ quittant la racine de $T_t^{\infty,u}$, étiquetée par s, et tout $0 \leq x \leq t$, on note par :

- $\tilde{T}_t^{\infty,s}$ le sous-arbre de $T_t^{\infty,u}$, enraciné en v_s, égal au sous-arbre de $T_t^{\infty,u}$ partant de v_s auquel on ajoute la racine $\emptyset$ de $T_t^{\infty,u}$ et l'arête e_s, tel qu'illustré par la figure 1.2 ;
- $\tilde{T}_x^{\infty,s}$ le sous-graphe de $\tilde{T}_t^{\infty,s}$ restreint aux arêtes d'étiquettes inférieures ou égales à x.
- $\tilde{T}_x^{k,s}$ la version de $\tilde{T}_x^{\infty,s}$ avec contrainte de degré ;
- $\tilde{X}_s \in [0,t] \cup \{\infty\}$ le premier instant où l'arête e_s est supprimée de $(\tilde{T}_x^{k,s})_{0\leq x\leq t}$.

Conditionnellement à l'ensemble S des étiquettes des arêtes quittant la racine de $T_t^{\infty,u}$, $(\tilde{T}_t^{\infty,s})_{s\in S}$ est une famille indépendante, et chaque $\tilde{T}_t^{\infty,s}$ a même loi que $T_t^{\infty,s}$. Par conséquent, conditionnellement à S et à $(\tilde{X}_s)_{s\in S}$, $(\tilde{T}_t^{\infty,s})_{s\in S}$ est une famille indépendante, et $\tilde{T}_t^{\infty,s}$ a même loi que $(T_t^{\infty,s})_{s\in S}$ conditionné par $X_s = \tilde{X}_s$. La connaissance de S et de $(\tilde{X}_s)_{s\in S}$ est suffisante pour connaître quelles arêtes adjacentes à la racine sont présentes dans $T_t^{k,u}$, via la procédure suivante :

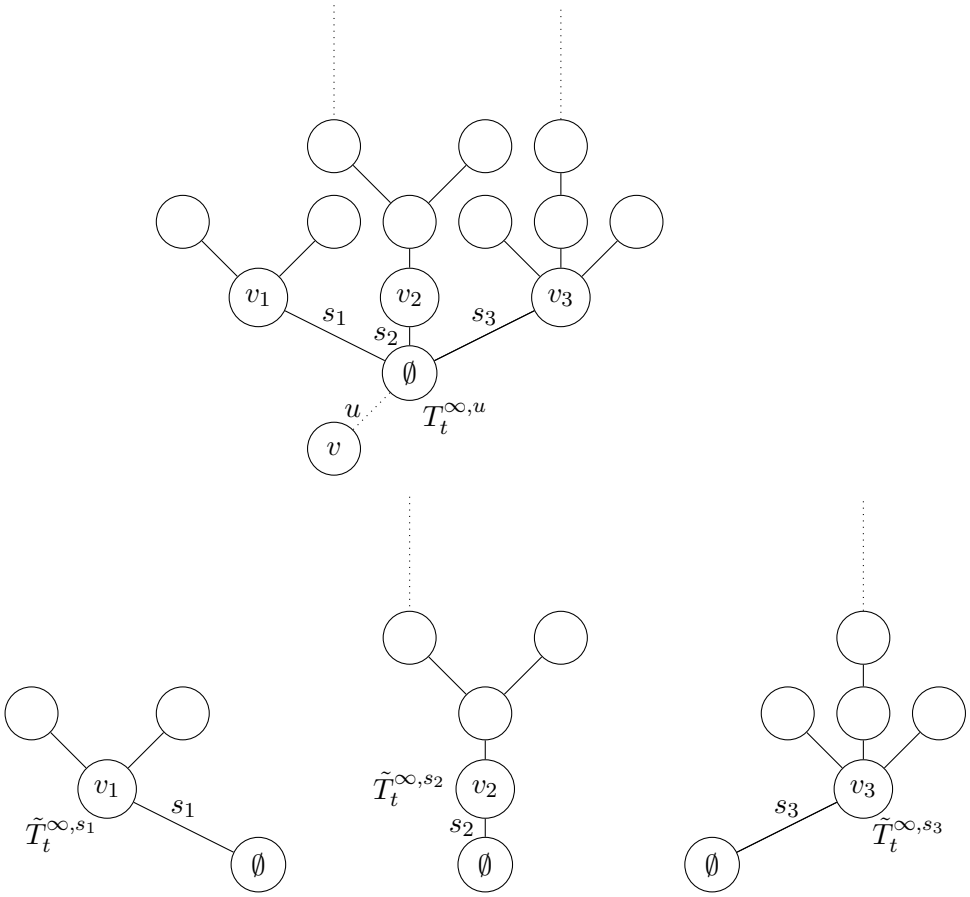

FIGURE 1.2 – Les arbres $T_t^{\infty,u}$ et $\tilde{T}_t^{\infty,s}$

- La connaissance de S permet de savoir à quel moment chaque arête adjacente à la racine est ajoutée.
- Si le degré de la racine atteint k (en comptant l'arête racine après le temps u), toutes les arêtes adjacentes à la racine sont supprimées.
- Si l'arête e_s est encore présente au temps $\tilde{X}_s$, elle est immédiatement supprimée à cause de son autre extrémité, d'après le lemme 1.22.

Par conséquent :

Théorème 1.23. *Conditionnellement à l'ensemble S' des étiquettes des arêtes quittant la racine de $T_t^{k,u}$, les arbres $\tilde{T}_t^{k,s}$ sont indépendants et de loi respective $T_t^{k+,s}$.*

Conditionnellement à l'ensemble S' des étiquettes des arêtes quittant la racine de $T_t^{k,u}$ et à l'événement « l'arête racine de $T^{k,u}$ est présente au temps t », les arbres $\tilde{T}_t^{k,s}$ sont indépendants et de loi respective $T_t^{k+,s}$.

Soit $S_t^{k+,u}$ l'ensemble des étiquettes des arêtes quittant la racine de $T_t^{k+,u}$. Le théorème 1.23 signifie que $T_t^{k+,\cdot}$ est un processus de branchement multitype, défini récursivement par :

- l'ensemble des types de enfants de la racine de $T_t^{k+,u}$ est $S_t^{k+,u}$;
- conditionnellement à $S_t^{k+,u}$, les sous-arbres commençant aux enfants de la racine de $T_t^{k+,u}$ sont indépendants, et de loi $T_t^{k+,s}$ avec s le type de l'enfant.

Par le même raisonnement, on montre que T_t^k est un processus de branchement multitype à deux étages, défini ainsi :

Théorème 1.24. *Soit S_t^k la loi de l'ensemble des étiquettes adjacentes à la racine dans T_t^k.*
T_t^k a même loi que l'arbre construit de la manière suivante :

- *l'ensemble des types de enfants de la racine de T_t^k est S_t^k ;*
- *conditionnellement à S_t^k, les sous-arbres commençant aux enfants de la racine de T_t^k sont indépendants, et de loi $T_t^{k+,s}$ avec s le type de l'enfant.*

La racine a donc une loi de progéniture particulière, mais la loi des sommets des générations suivantes est la même que pour $T_t^{k+,\cdot}$.

1.1.4.h Lien entre la limite locale et la composante géante

La partie 1.1.2.e explique le lien entre existence de la composante géante et limite locale. En particulier, si la limite locale est finie presque sûrement, alors il n'y a pas de composante géante. En revanche, le fait que la limite locale puisse être infinie n'implique pas l'existence d'une composante géante. Le lien se fait en démontrant le résultat suivant :

Lemme 1.25. *Soient v_1 et v_2 deux sommets uniformes indépendants de $G_{n,t}^k$. Alors*

$$\liminf_{n\to\infty} \mathbb{P}(v_1 \text{ et } v_2 \text{ sont dans la même composante}) \geq \mathbb{P}(|T_t^k| = \infty)^2,$$

Le lemme 1.25, avec le lemme 1.7, implique $\frac{C_1(G)}{n} \xrightarrow[n\to\infty]{p} \mathbb{P}(|T_t^k| = \infty)$.

Par définition, la limite locale ne donne des informations qu'à distance fixée de l'origine. Pour montrer le lemme 1.25, il est nécessaire d'avoir des informations sur une boule de rayon l centrée sur la racine, avec l qui augmente avec n. En étudiant plus finement le processus d'exploration il est possible d'obtenir successivement les résultats suivants :

- pour $t' = t - \epsilon$, avec probabilité proche de $\mathbb{P}(|T_{t'}^k| = \infty)^2$, les composantes de v_1 et v_2 dans $G_{n,t'}^k$ contiennent chacune au moins $n^{\frac{2}{3}}$ sommets ;
- avec probabilité proche de $\mathbb{P}(|T_t^k| = \infty)^2$, au moins $n^{\frac{3}{5}}$ de ces sommets sont dans la composante de v_1 (resp. v_2) dans $G_{n,t}^k$.
- avec probabilité proche de $\mathbb{P}(|T_t^k| = \infty)^2$, il existe une arête dans $G_{n,t}^k$ entre un sommet de la composante de v_1 dans $G_{n,t}^k$ et un sommet de la composante de v_2 dans $G_{n,t}^k$.

La dernière condition signifie que v_1 et v_2 sont dans la même composante. L'idée est donc de montrer que $G_{n,t'}^k$ est suffisament proche de $T_{t'}^k$ et donc que l'on peut découvrir un grand nombre de sommets dans les composantes de v_1 et v_2 dans $G_{n,t'}^k$, de ne regarder parmi ces sommets que ceux qui sont encore dans les composantes respectives de v_1 et v_2 dans $G_{n,t}^k$ et de montrer qu'il existe au moins une arête entre ces deux ensembles.

1.2 Hauteur d'arbres, grandes déviations et marches branchantes

Cette partie est consacrée à l'étude d'arbres aléatoires (arbre binaire de recherche, arbre de Lyndon) et à l'utilisation de grandes déviations et de marches branchantes dans ce cadre.

1.2.1 L'arbre binaire de recherche

Soit X un ensemble totalement ordonné. Un *arbre binaire de recherche* (dans la suite abrégé par ABR) étiqueté par des éléments de X est un arbre binaire dont chaque nœud interne est étiqueté par une *clé* appartenant à X, et tel que pour tout nœud interne, la clé du nœud soit supérieure à toutes les clés du sous-arbre gauche et inférieure à toutes les clés du sous-arbre droit partant de ce nœud. Les feuilles ne sont pas étiquetées.

Remarque. Il existe une autre définition d'un ABR, comme un arbre unaire-binaire où tous les sommets (y compris les feuilles) sont étiquetés, et tels que l'étiquette d'un sommet soit supérieure aux étiquettes présentes dans le sous-arbre gauche et inférieure aux étiquettes présentes dans le sous-arbre droit. Ces deux définitions sont en réalité équivalentes, ce qu'illustre la figure 1.2.1 :

- Pour passer d'un ABR selon la première définition à un ABR selon la seconde définition, il suffit de supprimer toutes les feuilles.
- Pour passer d'un ABR selon la deuxième définition à un ABR selon la première définition, il suffit d'ajouter des feuilles pour que tous les sommets aient deux fils.

L'ABR associé à une suite finie d'éléments distincts de X est défini de manière récursive. Soient $x_1, \dots, x_n$ n éléments distincts de X. L'ABR associé à $(x_1, \dots, x_n)$ est l'arbre binaire défini par :

- Si $n = 0$, l'arbre ne comprend qu'une racine, non étiquetée.
- Sinon la racine est étiquetée par x_1.
- Le sous-arbre gauche est l'ABR correspondant à la sous-suite de x_i contenant les termes strictement inférieurs à x_1.
- Le sous-arbre droit est l'ABR correspondant à la sous-suite de x_i contenant les termes strictement supérieurs à x_1.

L'ABR associé à une suite de n éléments deux à deux distincts est donc un arbre binaire avec n nœuds internes et $n + 1$ feuilles, dont les sommets internes sont étiquetés, de manière à ce que l'étiquette d'un sommet soit plus grande que les étiquettes présentes dans le sous-arbre gauche et plus petite que les étiquettes présentes dans le sous-arbre droit.

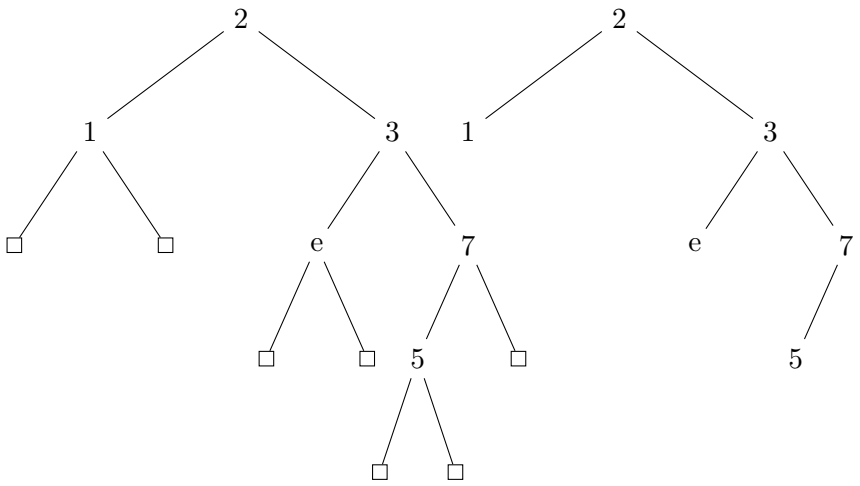

FIGURE 1.3 – ABR associé à $(2; 1; 3; 7; e, 5)$ selon les deux définitions

1.2.1.a Arbre binaire de recherche associé à une permutation uniforme

La forme de l'arbre associé à une suite finie ne dépend que de l'ordre entre les éléments, et pas de la valeur exacte de ceux-ci. Pour cette raison, il est souvent plus simple de remplacer la suite x par l'unique permutation $\sigma = (\sigma_1, \ldots, \sigma_n)$ dont les termes sont rangés dans le même ordre que ceux de x, donnée par

$$\sigma_i = |\{j : x_j \leq x_i\}|.$$

L'arbre binaire de recherche associé à une permutation aléatoire uniforme de $\{1, \ldots, n\}$ est noté par ABR_n. Il s'agit d'un arbre binaire aléatoire à n sommets, dont le comportement asymptotique est connu avec précision : par exemple sa hauteur $h(ABR_n)$ est proche de $\alpha \ln n - \beta \ln \ln n$, avec α la solution de $\alpha \ln(\frac{2e}{\alpha}) = 1$ plus grande que 1, et $\beta = \frac{3}{2 \ln(\frac{\alpha}{2})}$ [Pit84, Dev86, Drm01, Ree03].

1.2.2 Arbre de Yule

Soit λ un réel strictement positif. L'*arbre de Yule* $(Y_t)_{t \geq 0}$ de paramètre λ est la représentation d'un processus de branchement à temps continu définie de la manière suivante :

- Au temps $t = 0$, il y a une particule.
- Chaque particule meurt et donne naissance à deux sommets à taux λ. Chaque particule est représentée par une arête de longueur égale à sa durée de vie.

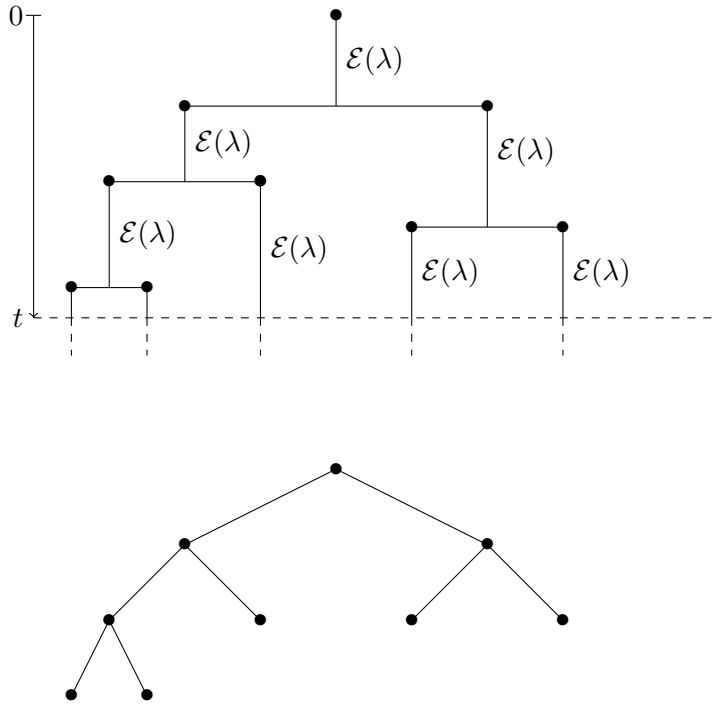

FIGURE 1.4 – Arbre de Yule et arbre binaire associé

La figure 1.4 présente un exemple d'arbre de Yule (tronqué au temps t).

On appelle *arbre binaire associé à* Y_t l'arbre combinatoire obtenu en oubliant la longueur des arêtes. Un exemple peut être trouvé figure 1.4.

L'arbre de Yule permet d'étudier facilement ABR_n, via le lemme suivant :

Lemme 1.26. *Soit* $\tau_n = \inf\{t : |Y_t| = n\}$ *le premier temps où l'arbre de Yule* Y *a* n *sommets. Alors*

1. *l'arbre binaire associé à* $Y_{\tau_{n+1}}$ *a la même loi que* ABR_n *;*
2. $Y_{\tau_{n+1}}$ *est indépendant de* τ_{n+1}.

Preuve du lemme. Pour toute permutation σ de $\{1, \dots, n\}$ et tout k entier entre 1 et $n+1$, on définit la permutation $f(\sigma, k)$ de $\{1, \dots, n+1\}$ par :

- Pour tout $i \in \{1, \dots, n\}$ tel que $\sigma(i) < k$, $f(\sigma, k)(i) = \sigma(i)$.
- Pour tout $i \in \{1, \dots, n\}$ tel que $\sigma(i) \geq k$, $f(\sigma, k)(i) = \sigma(i) + 1$.
- $f(\sigma, k)(n+1) = k$.

Cette construction revient à choisir l'image de $n+1$ (k), tout en gardant l'ordre relatif du reste de la permutation.

Exemple. Si $\sigma = (3, 1, 2, 4)$, alors $f(\sigma, 3) = (4, 1, 2, 5, 3)$.

On considère le processus aléatoire $(\sigma_n)_{n\geq 0}$ défini par récurrence de la manière suivante :

- σ_0 est l'unique permutation sur l'ensemble vide.
- Pour tout $n \geq 0$, $\sigma_{n+1} = f(\sigma_n, X_n)$, où X_n est un entier uniforme sur $\{1, \ldots, n+1\}$, indépendant de σ_n.

Lemme 1.27. *Pour tout $n \geq 0$, σ_n est une permutation uniforme parmi les permutations de $\{1, n\}$.*

Ce lemme se prouve par récurrence. Il est immédiat pour $n = 0$. On suppose qu'il est vrai pour $n \geq 0$. Soit σ^* une permutation de $\{1, \ldots, n+1\}$. Il existe une unique permutation ρ de $\{1, \ldots, n\}$ et un unique entier k entre 1 et $n+1$ tel que $\sigma^* = f(\rho, k)$, donc

$$\begin{aligned}\mathbb{P}(\sigma_{n+1} = \sigma^*) &= \mathbb{P}(\sigma_n = \rho)\mathbb{P}(X_n = k) \\ &= \frac{1}{n!}\frac{1}{n+1} \\ &= \frac{1}{(n+1)!}.\end{aligned}$$

Ce qui prouve le lemme pour $n+1$.

On considère maintenant le processus $(ABR_n)_{n\geq 1}$, où ABR_n est l'ABR associé à σ_n, pour tout entier n. Cela donne la construction suivante pour la suite des ABR_n :

- Initialement, ABR_0 ne contient qu'une feuille.
- Pour passer de ABR_n à ABR_{n+1}, on choisit un entier aléatoire X_n uniforme entre 1 et $n+1$. La X_n-ème feuille dans le contour de l'arbre est remplacée par un nœud interne ayant deux feuilles comme enfants.

Ce couplage permet de démontrer le lemme, par récurrence sur n, car $(Y_{\tau_{n+1}})_{n\geq 0}$ et $(ABR_n)_{n\geq 0}$ évoluent de la même manière : conditionnellement à l'arbre à n nœuds internes, une feuille est choisie au hasard uniformément, et remplacée par un nœud interne ayant deux feuilles comme enfant. L'indépendance entre $Y_{\tau_{n+1}}$ et τ_{n+1} découle du fait que la feuille choisie est indépendante de l'instant où elle est choisie. □

1.2.2.a Étude la hauteur de ABR_n via l'arbre de Yule

Le lien entre Y et ABR_n exhibé dans le lemme 1.26 permet d'étudier le profil de ABR_n, c'est-à-dire le nombre de feuilles à profondeur donnée.

On note par $U_k^{ABR_n}$ le nombre de feuilles à profondeur k dans ABR_n et par $U_k^{Y_t}$ le nombre de feuilles à profondeur k dans Y_t. Le but de cette partie

est d'illustrer sur un exemple simple la technique de couplage avec l'arbre de Yule, que nous utilisons pour l'analyse asymptotique de la hauteur de l'arbre de Lyndon. Nous allons pour cela montrer le résultat classique de Devroye[Dev86] :

$$\frac{h(ABR_n)}{\ln n} \xrightarrow[n\to\infty]{p} \alpha.$$

Cette preuve sera faite à l'aide du couplage arbre binaire de recherche-arbre de Yule de la section précédente, en s'inspirant des travaux de Pittel et Chauvin *et al.*. On pourra se reporter à [Pit94, CR04, CKMR05] pour l'analyse fine du profile de ABR_n à l'aide de cette technique de couplage.

Soit $\vec{x}$ une suite i.i.d. de Bernoulli($\frac{1}{2}$), indépendante de Y. On considère $\vec{x}$ comme une branche infinie de l'arbre binaire complet infini, de la manière suivante : en partant de la racine, si $\vec{x}_i = 0$, alors la branche infinie emprunte la branche menant au sous-arbre gauche à la profondeur i, et emprunte la branche menant au sous-arbre droit si $x_i = 1$.

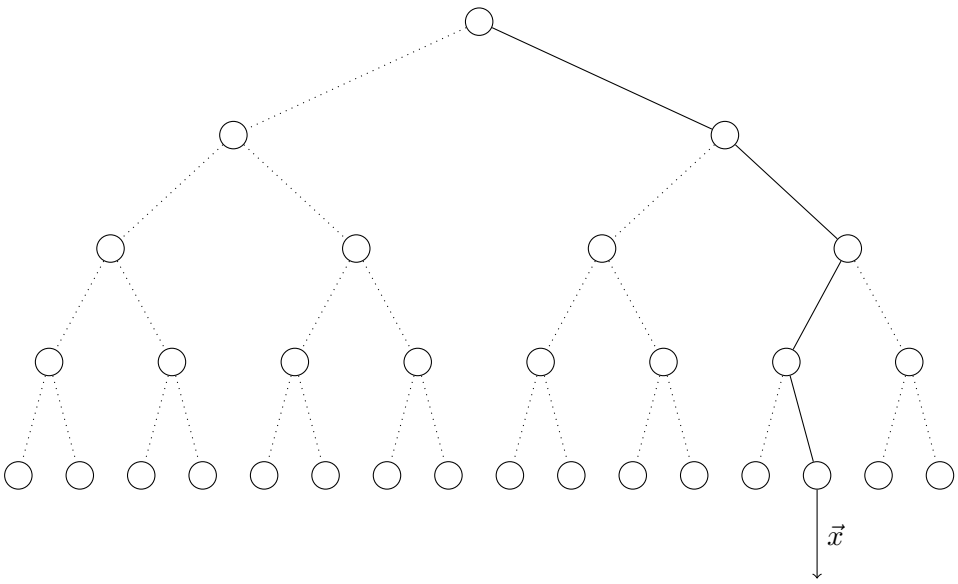

FIGURE 1.5 – Direction $\vec{x} = (1, 1, 0, 1, \dots)$ dans l'arbre binaire complet infini

Tout arbre binaire fini T peut être considéré comme un sous-arbre de l'arbre binaire complet infini. En utilisant ce plongement, il existe une unique feuille de T appartenant à la direction $\vec{x}$. Cette feuille sera appellée *la feuille de T dans la direction $\vec{x}$.*

L'étude de $U_k^{Y_t}$ nécessite d'étudier la loi de tout l'arbre. Le lemme suivant permet de relier $U_k^{Y_t}$ à l'étude d'une seule feuille.

Lemme 1.28. *Soit x la feuille de Y_t dans la direction $\vec{x}$, et h_x sa profondeur dans Y_t. Alors*

$$\mathbb{E}(U_k^{Y_t}) = 2^k \mathbb{P}(h_x = k)$$

Démonstration. On conditionne par Y_t. La probabilité que x soit une feuille donnée de profondeur k vaut 2^{-k}, puisque cela force le choix des k premiers termes de $\vec{x}$. Il y a $U_k^{Y_t}$ feuilles de profondeur k, donc $\mathbb{P}(h_x = k|Y_t) = 2^{-k}U_k^{Y_t}$. Le lemme est obtenu en prenant l'espérance. □

Dans un processus de Yule de paramètre 1, chaque individu meurt et donne naissance à deux sommets à taux 1. En conséquence, le long de la direction $\vec{x}$, les instants où un sommet se divise forment un processus de Poisson d'intensité 1 sur $[0, t]$. Comme la hauteur de x dans Y_t est égal au nombre de division de sommets le long de $\vec{x}$, on obtient :

$$\mathbb{P}(h_x = k) = \mathbb{P}(\mathrm{Poi}(t) = k)$$

puis

$$\mathbb{E}(U_k^{Y_t}) = 2^k e^{-t} \frac{t^k}{k!}.$$

Si l'on prend $\gamma > 0$ et k de la forme $k_t = \lfloor \gamma t \rfloor$, on obtient, par la formule de Stirling :

$$\begin{aligned}
\ln(\mathbb{E}(U_{k_t}^{Y_t})) &= k_t \ln 2 - t + k_t \ln(t) - \ln(k_t!) \\
&= \gamma t \ln 2 - t + \gamma t \ln(t) - \gamma t \ln(\gamma t) + \gamma t + o(t) \\
&= t(\gamma \ln 2 - 1 - \gamma \ln \gamma + \gamma) + o(t) \\
&= t(\gamma \ln(\frac{2e}{\gamma}) - 1) + o(t)
\end{aligned}$$

La fonction $\varphi : x \to x \ln(\frac{2e}{x}) - 1$ est strictement concave, tend vers -1 en 0 et vers $-\infty$ en $+\infty$, et est positive en 1. Il existe donc deux réels α^- et α tels que φ est positive sur $[\alpha^-; \alpha]$ et négative en dehors de cet intervalle. Ainsi le nombre moyen de feuilles à profondeur $\lfloor \gamma t \rfloor$ dans Y_t tend exponentiellement vite vers 0 lorsque $\gamma > \alpha$ et augmente exponentiellement lorsque $\alpha^- < \gamma < \alpha$. Cela permet de montrer que pour tout $\epsilon > 0$, avec grande probabilité, il n'existe aucune feuille à profondeur plus grande que $(\alpha + \epsilon)t$, et au moins une feuille à profondeur supérieure à $(\alpha - \epsilon)t$, et donc :

$$\frac{h(Y_t)}{t} \xrightarrow[t\to\infty]{p} \alpha \tag{1.8}$$

Le laps de temps $\tau_{n+1} - \tau_n$ est le temps qui s'écoule entre l'apparition de la n-ème feuille et l'apparition de la $n+1$ème feuille dans $(Y_t)_{t\geq 0}$. Chaque feuille se divise à taux 1 indépendamment, donc $\tau_{n+1} - \tau_n$ suit une loi exponentielle de paramètre n, indépendamment de τ_n. En conséquence, en notant E_i une

suite d'exponentielles indépendantes telle que E_i est de paramètre i, τ_{n+1} a même loi que $\sum_{i=1}^{n} E_i$.
$\mathbb{E}(E_i) = \frac{1}{n}$ et $Var(E_i) = \frac{1}{n^2}$, donc $\mathbb{E}(\tau_{n+1}) \simeq \ln n$ et $Var(\tau_{n+1}) \leq \frac{\pi^2}{6}$. Par conséquent :

$$\frac{\tau_{n+1}}{\ln n} \xrightarrow[n\to\infty]{p} 1 \tag{1.9}$$

En combinant (1.8) et (1.9), nous retrouvons le résultat de Devroye :

$$\frac{h(ABR_n)}{\ln n} \xrightarrow{p} \alpha$$

1.2.2.b Construction alternative de l'arbre de Yule

Soit t un réel fixé. Si l'on retourne le temps et que l'on considère $(Y_{t-s})_{s\geq 0}$, on observe un processus de coagulation dont la loi peut être décrite de la manière suivante :

- Soit $(X_i)_{i\geq 1}$ une suite i.i.d. de variables exponentielles de paramètre λ.
- Soit $i_t = \inf\{i : X_i \geq t\}$.
- Soit $X_0 = t$.
- Pour tout $1 \leq j < i_t$, $j^- = \sup\{i < j : X_i \geq X_j\}$. Par construction $j^- \geq 0$.
- Pour tout i, le sommet i est absorbé par le sommet j^- au temps X_i.

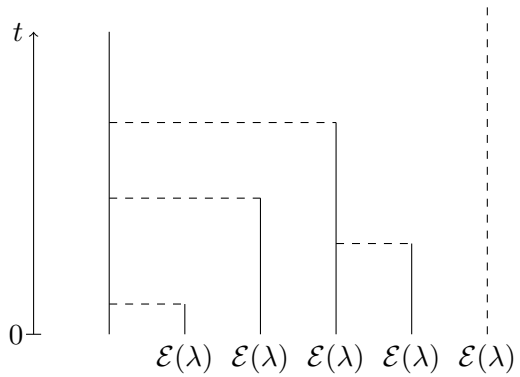

FIGURE 1.6 – Deuxième construction de l'arbre de Yule

Pour montrer l'égalité de loi, construisons la suite finie Z de la manière suivante :

- Soit a une variable exponentielle de paramètre 1.

- Si $a \geq t$, alors la suite Z est de longueur 0.
- Si $a < t$, soit $T = t - a$. Conditionnellement à T, considérons $(X_1^1, \ldots, X_{i_T^1 - 1}^1)$ et $(X_1^2, \ldots, X_{i_T^2 - 1}^2)$ deux copies indépendantes de $(X_1, \ldots, X_{i_T - 1})$.
- $Y = (X_1^1, \ldots, X_{i_T^1 - 1}^1, T, X_1^2, \ldots, X_{i_T^2 - 1}^2)$.

Lemme 1.29. *Z a même loi que $(X_1, \ldots, X_{i_t - 1})$.*

Preuve. Les deux suites sont finies presque sûrement, donc il est suffisant de prouver l'égalité de loi pour chaque longueur possible de suite. La probabilité que Y soit de longueur 0 vaut e^{-t} c'est-à-dire la probabilité que $i_t = 1$.

À longueur fixée $l \geq 1$, les lois des deux suites aléatoires sont absolument continues par rapport à la mesure de Lebesgue. La densité de la loi de Z est le produit de :

- $e^{-(t-T)}$ (choix de T),
- $\prod_{i=1}^{i_T^1 - 1} e^{-X_i^1}$ (densité pour les $i_T^1 - 1$ premières valeurs),
- e^{-T} (probabilité que la i_T^1-ème variable de X^1 dépasse T),
- $\prod_{i=1}^{i_T^2 - 1} e^{-X_i^2}$ (densité pour les $i_T^2 - 1$ dernières valeurs),
- e^{-T} (probabilité que la i_T^2-e variable de X^2 dépasse T).

Soit, après simplification :

$$e^{-t} \prod_{i=1}^{i_t - 1} Z_i$$

puisque $Z_{i_T^1} = T$ par construction. Cette quantité est également la densité de la suite $(X_i)_{i < i_t}$, ce qui prouve l'égalité en loi des deux suites.

□

Corollaire 1.30. *Le processus de coagulation décrit au-dessus a bien même loi qu'un arbre de Yule retourné.*

L'arbre créé par la suite Z a la même construction par récurrence qu'un arbre de Yule

- Attendre un temps exponentiel a.
- Si $a < t$, créer deux copies indépendantes de paramètre $t - a$ qui seront le fils gauche et le fils droit.
- Si $a > t$, l'arbre ne se sépare pas.

1.2.3 Grandes déviations pour les marches aléatoires

Soit X une suite i.i.d. de variables aléatoires réelles d'espérance nulle. On définit :

$$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

D'après la loi des grands nombres, $M_n \to 0$ p.s., donc $\mathbb{P}(M_n > x) \to 0$ pour tout $x > 0$. Grâce à la théorie des grandes déviations, il est possible de quantifier la vitesse de convergence :

Théorème 1.31. *La suite* $\frac{1}{n}\ln(\mathbb{P}(M_n > x))$ *converge et la limite vaut :*

$$\lim_{n\to\infty} \frac{1}{n}\ln(\mathbb{P}(M_n > x)) = -I(x)$$

La fonction $I(.)$ *est la transformée de Legendre-Fenchel de la fonction génératrice des cumulants de* X_1*, c'est-à-dire :*

$$I(x) = \sup_{\theta>0}[\theta x - \lambda(\theta)]$$

$$o\grave{u}\ \lambda(\theta) = \ln(\mathbb{E}(e^{\theta X_1}))$$

Il est possible de montrer l'inégalité $\limsup_n \frac{1}{n}\ln(\mathbb{P}(M_n > x)) \leq -\theta x + \lambda(\theta)$ pour tout $\theta > 0$ via une inégalité de Markov exponentielle :

$$\begin{aligned}
\mathbb{P}(M_n > x) &= \mathbb{P}(\sum_{i=1}^{n} X_i > nx) \\
&\leq \frac{\mathbb{E}(exp(\theta\sum_{i=1}^{n} X_i))}{e^{n\theta x}} \\
&= \frac{\mathbb{E}(\prod_{i=1}^{n} exp(\theta X_i))}{e^{n\theta x}} \\
&= \frac{\prod_{i=1}^{n}\mathbb{E}(exp(\theta X_i))}{e^{n\theta x}} \\
&= \frac{e^{n\lambda(\theta)}}{e^{n\theta x}} \\
&= exp(n(\lambda(\theta) - \theta x))
\end{aligned}$$

En optimisant en θ, on obtient $\limsup_n \frac{1}{n}\ln(\mathbb{P}(M_n > x)) \leq -I(x)$

La démonstration de l'autre inégalité, un peu plus longue, peut être trouvée dans [BZ79].

1.2.4 Marches branchantes

Une marche branchante décrit non seulement la généalogie des individus formant une population donnée, mais aussi leurs positions respectives. Dans les exemples que nous envisageons, les individus de la population vivent sur la droite réelle, et leur généalogie est codée par un arbre dont ils sont les nœuds. Les nœuds de l'arbre portent chacun, en étiquette, la position de l'individu correspondant sur la droite réelle. Dans la suite on parlera indifféremment d'individu ou de nœud, de position ou d'étiquette.

Définition 1.18 (Notation de Neveu)**.** Soit Z un arbre dont les sommets sont étiquetés par un nombre réel, correspondant à leur position. *La notation de Neveu* consiste à numéroter chaque sommet de Z par un mot de la manière suivante :

- La racine est numérotée par le mot vide.
- Si un sommet est numéroté par le mot fini u et a k enfants, alors ses enfants classés par position croissante sont numérotés par $u1, u2, \ldots, uk$, où ui désigne la concaténation du mot u avec l'entier i.

La profondeur d'un sommet u correspond à la longueur $|u|$ du mot le désignant. La position du sommet u est notée par x_u.

Définition 1.19. Soit $\mathscr{Z}$ un processus ponctuel aléatoire sur $\mathbb{R}$. La *marche branchante* Z de loi de reproduction $\mathscr{Z}$ est un processus de branchement Z dont les sommets sont étiquetés par un réel défini comme suit :

- La racine de Z est étiquetée par 0, *i.e.* la position de l'ancêtre est supposée être 0.
- Conditionnellement aux i premières générations de Z, la loi et la position des enfants de chaque individu de la i-ème génération est indépendante. La loi conditionnelle du nombre et de la position des enfants d'un individu situé à la position x est la loi de $(|\mathscr{Z}|, x + \mathscr{Z})$.

On note par Z_i le processus ponctuel comptant les étiquettes des sommets de la génération i, c'est-à-dire :

$$Z_i = \sum_{u \in Z, |u|=i} \partial_{x_u}$$

Dans certains ouvrages, la marche branchante est définie comme la suite de ces processus ponctuels, plutôt que comme un arbre aléatoire étiqueté.

Nous nous intéressons à l'évolution de Z_n. Dans [Kin75, Big76], Kingman puis Biggins étudient le comportement de la particule la plus à gauche de

chaque génération, $B^n = \inf Z_n$, quand n augmente. Pour cela, on considère la fonction $m : \mathbb{R}_+ \to \mathbb{R}_+$ définie par :

$$m(\theta) = \mathbb{E}(\sum_{x \in Z} e^{-\theta x}).$$

Lemme 1.32. *Pour tout entier $i \geq 1$,*

$$\mathbb{E}(\sum_{x \in Z_i} e^{-\theta x}) = m(\theta)^i$$

avec la convention que $(+\infty)^i = +\infty$.

Preuve. Cette égalité se montre par récurrence. Le cas $k = 1$ découle des définitions de Z_1 et m.

Conditionnellement à Z_i, chaque élément x de Z_i donne naissance à des enfants dont la loi de l'ensemble des positions est $x + Z$. Par conséquent :

$$\begin{aligned}
\mathbb{E}(\sum_{x \in Z_{i+1}} e^{-\theta x} | Z_i) &= \sum_{x \in Z_i} \mathbb{E}\left(\sum_{y \in Z} e^{-\theta(x+y)}\right) \\
&= \sum_{x \in Z_i} e^{-\theta x} \mathbb{E}\left(\sum_{y \in Z} e^{-\theta y}\right) \\
&= \sum_{x \in Z_i} e^{-\theta x} m(\theta) \\
\mathbb{E}(\sum_{x \in Z_{i+1}} e^{-\theta x}) &= m(\theta) \mathbb{E}(\sum_{x \in Z_i} e^{-\theta x}) \\
&= m(\theta)^{i+1}
\end{aligned}$$

□

On suppose $m(0) = \mathbb{E}(|Z|) > 1$, car dans ce cas la probabilité de non-extinction est strictement positive, et on note :

$$S = \cap_n \{Z_n \neq \emptyset\}.$$

S est l'événement de non-extinction.

On définit alors μ par :

$$\mu(x) = \inf_{\theta > 0} \{e^{x\theta} m(\theta)\}$$

La position de la particule la plus à gauche à chaque génération est liée à μ :

Théorème 1.33 ([Kin75, Big76])**.** $\lim_n \frac{B^n}{n} = \gamma$ *p.s. sur S, avec* $\gamma = \inf\{x : \mu(x) > 1\}$.

Comme pour les grandes déviations sur les marches aléatoires, une des deux inégalités se démontre par une inégalité de Markov exponentielle :

$$\begin{aligned}\mathbb{P}(B^n \leq n\alpha) &\leq e^{\alpha\theta n}\mathbb{E}(\sum_{x\in Z_n} e^{-\theta x}) \\ &= \left(e^{\alpha\theta} m(\theta)\right)^n\end{aligned}$$

En optimisant en θ, on obtient que

$$\mathbb{P}(B^n \leq n\alpha) < \mu(\alpha)^n.$$

Par conséquent, si $\alpha < \gamma$, alors par définition de γ, $\mu(\alpha) < 1$, et donc cette probabilité tend exponentiellement vite vers 0.

1.2.4.a Utilisation des marches branchantes pour retrouver la hauteur d'un arbre de Yule

Il est possible d'utiliser ce résultat pour démontrer la limite en probabilité de $\frac{h(Y_t)}{t}$ vue à la partie 1.2.2.a.

En effet, les temps d'apparition des sommets de profondeur i dans l'arbre de Yule $(Y_t)_{t\in\mathbb{R}}$ de paramètre 1 forment une marche branchante, de loi de reproduction $\mathscr{Z} = 2\partial_{\mathcal{E}(1)}$, avec $\mathcal{E}(1)$ une variable aléatoire exponentielle de paramètre 1. Notons $(Z_n^{Yule})_{n\geq 0}$ cette marche branchante. On a alors équivalence entre $h(Y_t) \geq k$ et $\inf Z_k^{Yule} \leq t$.

Comme Z^{Yule} ne peut pas s'éteindre, on obtient, d'après le théorème 1.2.4, que

$$\lim_n \frac{1}{n} \inf Z_n^{Yule} = \gamma \text{ p.s.}$$

puis

$$\lim_t \frac{h(Y_t)}{t} = \frac{1}{\gamma} \text{ p.s.}$$

Le calcul de $\gamma = \inf\{x : \mu(x) > 1\}$, avec les notations du théorème 1.2.4, permet de retrouver la même limite que dans l'équation (1.8).

1.2.5 Arbre de Lyndon

Soit Σ un alphabet fini ou dénombrable, muni d'un ordre total. L'ensemble des mots de longueur n est noté Σ^n et l'ensemble des mots finis ou infinis est noté Σ^*, et est muni de l'ordre lexigraphique défini de la manière suivante. Si w_1 et w_2 sont deux éléments de Σ^*, on dit que w_1 est inférieur à w_2 pour l'*ordre lexicographique* si l'on est dans l'un des deux cas suivants :

- w_1 est un préfixe strict de w_2.

- $w = uav_1'$ et $w_2 = ubv_2$ où u est le plus long préfixe commun à w_1 et w_2, a et b deux éléments de Σ tels que $a < b$ et v_1 et v_2 deux mots sur l'alphabet Σ.

L'ordre lexicographique est un ordre total sur Σ^*.

Définition 1.20. On dit qu'un mot fini w est un *mot de Lyndon* s'il est plus petit que tous ses suffixes propres non vides.

Pour tout entier n on note par $\mathcal{L}_n$ l'ensemble des mots de Lyndon de longueur n et par $\mathcal{L} = \cup_n \mathcal{L}_n$ l'ensemble des mots de Lyndon.

Exemple. Le mot 001 est un mot de Lyndon, car il est plus petit que 1 et 01, ses deux suffixes propres non vides.

Remarque. Un mot de Lyndon est apériodique. En effet, pour tout mot u et entier $k \geq 2$, $u < u^k$.

Définition 1.21. Soit w un mot fini. On dit que w' est une *rotation* de w s'il existe u et v tels que $w = uv$ et $w' = vu$. On dit que la rotation est *non triviale* si u et v sont non vides.

On appelle *collier* de w l'ensemble des rotations de w.

Remarque. La relation « *être une rotation de* » est une relation d'équivalence.

Exemple. Le collier de 1001 est $\{1001, 0011, 0110, 1100\}$.

Proposition 1.34. *Soit w un mot fini ; w est un mot de Lyndon si et seulement si w est strictement inférieur à toutes ses rotations non triviales.*

Démonstration. Si w est Lyndon, et s'écrit $w = uv$ avec u et v non vides, alors $w < v < vu$.

Supposons que w est strictement inférieur à toutes ses rotations non triviales. Supposons qu'il existe deux mots u et v non vides tels que $w = uv$ et $v < w$. Par hypothèse $w < vu$. Nous avons donc $v < w < vu$. Cela signifie que v est un préfixe de w, donc qu'il existe t tel que $w = vt$. Nous avons alors successivement :

$$\begin{aligned} vt = w &< vu \\ t &< u \\ tv &< uv = w \quad \text{car } t \text{ et } u \text{ sont de même longueur} \end{aligned}$$

tv étant une rotation non triviale de w, cette dernière inégalité est par hypothèse impossible. □

Corollaire 1.35. *Si w est un mot apériodique, il existe un unique mot de Lyndon dans le collier de w (le mot minimal parmi les rotations de w). Si w est périodique, aucun mot du collier de w n'est de Lyndon.*

Corollaire 1.36. *Pour tout n,*

$$|\mathcal{L}_n| = \frac{1}{n}\sum_{d|n} \mu(\frac{n}{d})|\Sigma|^d$$

où μ est la fonction de Möbius.

Soit A_n l'ensemble des mots de longueur n apériodiques. Alors le corollaire 1.35 implique l'égalité $|\mathcal{L}_n| = \frac{1}{n}|A_n|$, puisqu'il y a n mots apériodiques par collier apériodique, dont un mot de Lyndon. En décomposant l'ensemble des mots de longueur n selon la longueur de leur plus petite période, on obtient :

$$|\Sigma|^n = \sum_{d|n} |A_d|$$

et donc $A_n = \sum_{d|n} \mu(\frac{n}{d})|\Sigma|^d$.

1.2.5.a Factorisation standard

Théorème 1.37. *Soit w un mot de Lyndon. Si v est le plus petit suffixe strict de w, et u un mot tel que $w = uv$, alors u et v sont des mots de Lyndon. Le couple (u, v) est appelé la* factorisation standard *de w.*

Démonstration. Soient a et b deux mots non vides tels que $v = ab$. Alors b est un suffixe strict de w, donc $v < b$. Par conséquent, v est un mot de Lyndon.

Montrons maintenant par l'absurde que u est un mot de Lyndon. Les mots a, b, c et d seront tous non vides. On suppose que u n'est pas un mot de Lyndon. Il existe donc deux mots a et b non vides tels que $u = ab$ et $ba \leq u$. Nous avons alors à la fois $b < ba \leq u$ et $bv > w = uv$ (car bv est un suffixe strict de w, qui est un mot de Lyndon), donc b est un préfixe strict de u (car u est strictement plus long que b). Il existe c non vide tel que $u = bc$.

Nous avons alors $ba \leq u = bc$, donc $a \leq c$. De plus $abv = w = bcv < bva$ (car bva est une rotation non triviale du mot de Lyndon w) donc $cv < va$. De plus $v < cv$ (car v est le plus petit suffixe strict de w). Donc v est un préfixe strict de cv, et il existe d non vide tel que $cv = vd$. Cela implique que $bvd = w < bva$, donc $d < a$, et en conséquence $dbv < abv = w$, ce qui est impossible car dbv est une rotation de w.

□

Une autre caractérisation de la factorisation standard est la suivante :

Proposition 1.38. *Soit w un mot de Lyndon de longueur au moins 2. Il existe au moins une factorisation de w de la forme $w = uv$ avec u et v deux mots de Lyndon non vides. La factorisation standard est la factorisation de ce type avec v de longueur maximale.*

Démonstration. L'existence d'une telle factorisation est une conséquence du théorème précédent. Il reste donc à prouver que la factorisation avec v de longueur maximale est bien la factorisation standard. Cela est une conséquence de l'observation suivante : si v' est un suffixe de w plus long que le suffixe strict minimal v, alors v est un suffixe de v', et donc v' n'est pas un mot de Lyndon. □

1.2.5.b Arbre de Lyndon

La factorisation standard permet de décomposer n'importe quel mot de Lyndon de longueur au moins 2 en deux mots de Lyndon plus courts. Si l'on itère cette décomposition, on obtient un arbre de Lyndon :

Définition 1.22. *L'arbre de Lyndon* $T(w)$ associé à un mot de Lyndon w est l'arbre binaire planaire enraciné étiqueté défini par :

- Si w est de longueur 1, $T(w)$ est l'arbre n'ayant qu'un seul sommet étiqueté par w.
- Si w est de longueur au moins 2, soit (u, v) la factorisation standard de w. Alors $T(w)$ est l'arbre dont la racine est étiquetée par w, le sous-arbre gauche est $T(u)$ et le sous-arbre droit est $T(v)$.

Si w est de longueur n, $T(w)$ est un arbre à n feuilles et $n-1$ sommets internes. Les feuilles sont étiquetées par les lettres de w, et l'étiquette de chaque nœud interne est la concaténation des étiquettes des deux fils.

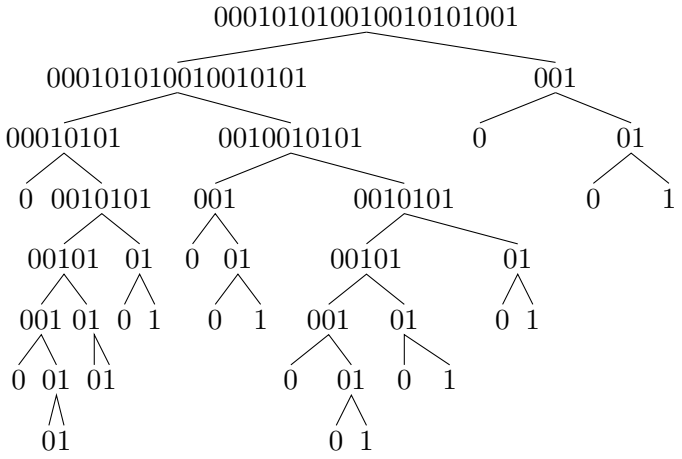

FIGURE 1.7 – Arbre de Lyndon associé à 00010101001001010100.

1.2.5.c Arbre de Lyndon uniforme

Des combinatoriciens, parmi lesquels Jean Perrin, ou encore Bassino et al. [BCN05], ont soulevé la question du comportement typique des grands arbres

de Lyndon , et en particulier de leur hauteur. Nous donnons ici une première réponse. On se place dans le cadre d'un alphabet Σ de taille 2 que l'on suppose égal à $\{0,1\}$. Soit L_n un mot de Lyndon de longueur n choisi uniformément parmi $|\mathcal{L}_n|$, et $T(L_n)$ son arbre de Lyndon associé. La hauteur d'un arbre T est noté par $h(T)$.

Notons $(A(n,k))_{n,k}$ la famille des nombres eulériens, c'est-à-dire le nombre de permutations σ de $\{1,\dots,n\}$ ayant exactement k descentes (k positions i où $\sigma(i) \geq \sigma(i+1)$). Définissons :

$$\Xi(\theta) = \lim_n \frac{1}{n} \ln(A(n, \lfloor \theta n \rfloor)/n!), \tag{1.10}$$

Une expression de Ξ peut-être trouvée dans [GK94, p. 299]. Définissons également :

$$\Psi(\lambda,\mu,\nu) = \ln\left(\frac{(1+\mu)^{1+\mu}}{\mu^\mu}\ \frac{(e\lambda\ln 2)^\nu \ln 2}{\nu^\nu 2^\lambda}\right) +\ \Xi(\lambda-\mu), \tag{1.11}$$

Alors $\Delta^\star$ est défini comme suit :

$$\Delta^\star = \sup_{\lambda,\mu,\nu>0} \left((1+\nu+\mu)\ln 2 + \Psi(\lambda,\mu,\nu)\right)/(\lambda(\ln 2)^2) \tag{1.12}$$

ce qui permet d'estimer que $\Delta^\star = 5.092 \pm 0.5.10^{-4}$.

Théorème 1.39.

$$\frac{h(T(L_n))}{\ln n} \xrightarrow[n\to\infty]{p} \Delta^\star.$$

Le théorème 1.39 est le résultat principal du chapitre 4 de ce manuscrit.

1.2.6 Transformations successives du problème

1.2.6.a Utilisation d'une source i.i.d.

Pour étudier les propriétés de $T(L_n)$, il est utile de comprendre les propriétés de L_n. D'après le corollaire 1.35, chaque collier apériodique de longueur n contient n mots apériodiques dont 1 mot de Lyndon. Par conséquent, L_n a la même loi que la rotation minimale d'un mot apériodique choisi uniformément parmi les mots mots apériodiques de longueur n. Pour éviter d'avoir à vérifier l'apériodicité, il est possible d'utiliser l'artifice suivant :

- Soit W_∞ un mot infini aléatoire, obtenu comme le développement binaire d'un réel U uniforme sur $[0,1]$.
- Soit W_n le mot W_∞ tronqué après n lettres.
- Si W_n est apériodique, alors notons L_n la rotation minimale de w. Si W_n est périodique, choisir $L_n = 0^{n-1}1$.

Si l'on conditionne par le fait que W_n est apériodique, L_n est uniforme sur $\mathcal{L}_n$. Néanmoins, la loi de L_n diffère de la loi uniforme sur $\mathcal{L}_n$ à cause de la probabilité que W_n soit périodique. Comme cette probabilité tend exponentiellement vite vers 0, il est équivalent de prouver le théorème 1.39 avec L_n et avec un mot uniforme sur $\mathcal{L}_n$. Pour cette raison, dans la suite, on considérera que L_n est construit via cet artifice.

1.2.6.b Transformation exponentielle

Pour obtenir L_n, il est nécessaire de trouver la rotation minimale de W_n. S'il existe une unique série de 0 consécutifs de longueur maximale, L_n est la rotation commençant par cette série. En revanche, s'il existe deux ou plusieurs séries de 0 consécutifs de même longueur maximale, il est nécessaire de considérer les suffixes les suivant. Comme la probabilité qu'au moins deux séries aient la même longueur maximale ne tend pas vers 0, il est plus simple de considérer la construction suivante, donnant un mot de Lyndon de longueur aléatoire : soit W^ℓ le mot formé par le `1`, suivi du W_∞ tronqué à la position τ_ℓ du ℓ-ème 0 de la première série de ℓ 0 consécutifs dans W_∞.

Dans ce cas, $\mathscr{W}^\ell$ est obligatoirement apériodique. Dénotons par L^ℓ la rotation minimale de W^ℓ, c'est-à-dire :

$$W^\ell = \mathtt{1}\underbrace{\mathtt{010110\ldots1}\overbrace{\mathtt{000000}}^{\ell 0s}}_{\text{préfixe de } W_\infty} \quad \text{et} \quad L^\ell = \overbrace{\mathtt{000000}}^{\ell 0s}\mathtt{1}\underbrace{\mathtt{010110\ldots1}}_{\text{préfixe de } W_\infty}.$$

La longueur de L^ℓ croît exponentiellement avec ℓ : $\mathbb{E}(|L^\ell|) = 2^{\ell+1} - 1$. En conséquence, on peut s'attendre à ce que la hauteur de $\mathfrak{L}(L^\ell)$ croisse linéairement avec ℓ :

Théorème 1.40. $\frac{h(\mathfrak{L}(L^\ell))}{\ell} \xrightarrow[\ell\to\infty]{\mathbb{P}} \Delta^\star \ln 2$.

On en déduit le théorème 1.39 à l'aide d'une factorisation de L_n en facteurs de type L^ℓ.

1.2.7 Couplage avec un arbre de Yule

1.2.7.a Réduction à un squelette, partie haute de l'arbre

Soit k un entier. Pour tout mot w de Lyndon, $\mathfrak{L}^k(w)$ est défini par récurrence par :

- Si w a au plus un seul facteur 0^k (et donc en particulier pas de facteur 0^{k+1}), alors $\mathfrak{L}^k(w)$ est un seul nœud, étiqueté par w.
- Sinon, soit (u, v) la factorisation standard de w. Alors la racine de $\mathfrak{L}^k(w)$ est étiquetée par w, le fils gauche est $\mathfrak{L}^k(u)$ et le fils droit est $\mathfrak{L}^k(v)$.

Dans la suite, on considèrera un squelette $\mathfrak{T}_\ell := \mathfrak{L}^{a_\ell}(L^\ell)$ de l'arbre de Lyndon $\mathfrak{L}(L^\ell)$, approximable de manière précise par un ABR, pourvu que le seuil a_ℓ croisse lentement vers l'infini.

Une feuille v de cet arbre est soit étiquetée par 0 et est dans ce cas appelée *aiguille* (*needle*), soit par un mot de préfixe $0^{a_\ell}1$, et est alors appelée *limbe* (*blade*). Pour retrouver l'arbre de Lyndon complet à partir du squelette, il suffit de remplacer chacun des limbes de $\mathfrak{T}_\ell$ par l'arbre de Lyndon associé à l'étiquette de la feuille, l'arbre de Lyndon associé à un limbe sera appelé *arbuste*. On numérote les limbes de $\mathfrak{T}_\ell$ dans l'ordre de parcours en profondeur, de 1 à N_ℓ le nombre de limbes.

À tout limbe v de $\mathfrak{T}_\ell$, associons le couple (n_v, t_v) tel que $0^{n_v}1t_v$ est la concaténation des étiquettes des feuilles situées strictement après le limbe $v-1$ dans le parcours en profondeur de $\mathfrak{T}_\ell$. La construction du squelette peut être décrite à partir de la suite (n_v, t_v) de la manière suivante.

1.2.7.b Arbre de Lyndon d'une suite

Étant donné un ensemble ordonné $(\mathcal{R}, \prec)$, on peut munir $\mathbb{N} \times \mathcal{R}$ d'un ordre lexicographique, également noté $\prec$, et défini de la manière suivante : $(n,t) \prec (m,u)$ si $n > m$ ou si $n = m$ et $t \prec u$. Soit $\epsilon^{(1)}$ la suite d'entiers définie par $\epsilon_i^{(1)} = \delta_{i,1}$ pour tout i.

Soit $B = (l_i, r_i)_{1 \leq i \leq N}$ une suite finie d'éléments de $\mathbb{N} \times \mathcal{R}$ telle que les r_i sont tous distincts et les l_i tous supérieurs ou égaux à k. On suppose également que l'élément minimal de B pour $\prec$ est le premier élément (l_1, r_1).

Définition 1.23. L'arbre de Lyndon $\mathfrak{L}^k(B)$ est défini par récurrence par :

1. Si $N = 1$ et $l_1 = k$, alors $\mathfrak{L}^k(B)$ est réduit à une seule feuille, étiquetée par (k, t_1) qui sera un limbe.
2. Sinon, on considère la suite finie $B' = (r - \epsilon^{(1)}, l)$, et on note par i_0 l'indice de son élément minimal, pour $\prec$.
   (a) Si $i_0 = 1$, alors $\mathfrak{L}^k(B)$ est l'arbre binaire avec une aiguille (étiquetée par 0) comme fils gauche et $\mathfrak{L}^k(B')$ comme fils droit.
   (b) Si $i_0 \geq 2$, $\mathfrak{L}^k(B)$ a $\mathfrak{L}^k((l_i, r_i)_{1 \leq i < i_0})$ comme fils gauche et $\mathfrak{L}^k((l_{i+i_0}, r_{i+i_0})_{0 \leq i \leq N - i_0})$ comme fils droit.

Alors $\mathfrak{L}^{a_\ell}((n_v, t_v))$ et $\mathfrak{T}_\ell$ sont identiques, aux étiquettes près, puisque mis à part les aiguilles, les sommets sont étiquetés par (l,t) dans $\mathfrak{L}^{a_\ell}((n_v, t_v))$ et par $0^l t$ dans $\mathfrak{T}_\ell$.

1.2.7.c Modification des suffixes

Les suffixes t_v ne sont pas indépendants, puisque que t_v est un suffixe de $t_{v'}$ si $v > v'$. Néanmoins, l'arbre de Lyndon n'utilisant que l'ordre sur les suffixes,

et non pas leurs valeurs précises, il est possible de modifier légérement les suffixes t_v sans modifier significativement l'arbre de Lyndon. Soit $(\zeta_i)_{i\in\mathbb{N}}$ une suite i.i.d. de mots infinis uniformes, et soit s_v la concaténation du préfixe de longueur a_ℓ de t_v avec ζ_v. Si t_{N_ℓ} a moins de a_ℓ lettres, t_{N_ℓ} est complétée par des 0 jusqu'à avoir la bonne longueur. On considère

$$\mathfrak{S}_\ell = \mathfrak{L}^{a_\ell}((n_v, s_v)_{1\leq v\leq N_\ell}).$$

Les arbres $\mathfrak{S}_\ell$ et $\mathfrak{T}_\ell$ différent peu, uniquement quand les a_l premières lettres ne sont pas suffisantes pour distinguer deux suffixes. La proposition 4 du chapitre 3 permet de quantifier cette proximité.

Par ailleurs, la loi de $((n_v, s_v)_{1\leq v\leq N_\ell})$ possède une description simple :

Proposition 1.41. *La suite $\tilde{S} = (0^{n_v-a_l}1s_v)_{v\geq 2}$ suivie de $0^{n_1-a_\ell}1s_1$ est une suite i.i.d. de copies de W_∞ arrêtée après la premier mot w commençant par $\ell - a_\ell$ 0, à une légère modification près : on enlève si nécessaire des 0 initiaux du mot w jusqu'à ce que w commence par $0^{l-a_l}1$.*

Chaque mot binaire infini est le développement dyadique d'un réel dans $[0, 1]$. La suite $\tilde{S}$ peut donc être vue comme une suite de réels uniformes sur $[0, 1]$, arrêtée au premier réel inférieur à $2^{a_\ell-\ell}$, ce réel étant éventuellement multiplié par une puissance de 2 jusqu'à appartenir à $[2^{-\ell+a_\ell-1}, 2^{-\ell+a_\ell}]$. En posant $X_i = -\log_2 \tilde{S}_i$, on obtient alors une suite i.i.d. $X = (X_i)_{i\geq 1}$ de variables aléatoires exponentielles, arrêtée après le premier terme supérieur à $\ell - a_\ell$, ce dernier terme étant décalé d'un entier pour appartenir à $[\ell - a_\ell, \ell - a_\ell + 1[$.

1.2.8 Étude via l'arbre de Yule

À l'aide de la suite X, on peut construire, comme on l'a vu à la section 1.2.2.b, un arbre de Yule, noté $\mathfrak{Y}_{\ell-a_\ell}$. Les arbres $\mathfrak{S}_\ell$ et $\mathfrak{Y}_{\ell-a_\ell}$ ont des structures très voisines, du fait que la transformation $(n_v, s_v) \to X_v$ est croissante. En effet, en comparant la construction par récurrence de $\mathfrak{S}_\ell$ avec la définition de $\mathfrak{Y}_{\ell-a_\ell}$ on constate que les branchements de type 2.b de la définition 1.23 de $\mathfrak{S}_\ell$ correspondent aux branchements de $\mathfrak{Y}_{\ell-a_\ell}$. Ainsi $\mathfrak{Y}_{\ell-a_\ell}$, vu comme un arbre plan sans tenir compte de la longueur des arêtes, peut-être obtenu à partir de $\mathfrak{S}_\ell$ en supprimant toutes les aiguilles et en contractant le sommet père des aiguilles tel qu'illustré sur la figure 1.8. La profondeur d'une feuille v dans $\mathfrak{S}_\ell$ correspondant donc au nombre de nœuds internes entre la racine et v. La profondeur correspond donc à la somme du nombre de nœuds internes entre la racine et v dans $\mathfrak{Y}_{\ell-a_\ell}$ et du nombre d'aiguilles supprimées sur le chemin de la racine à v. D'après la construction de l'arbre de Lyndon associé à une suite B, une aiguille est présente à la racine lorsque l'élément minimal de $B - \epsilon^{(1)}$ est le premier terme, ce qui correspond sur la suite X à ce que le premier terme de la suite X soit supérieur d'au moins 1 à tous les autres termes de la suite.

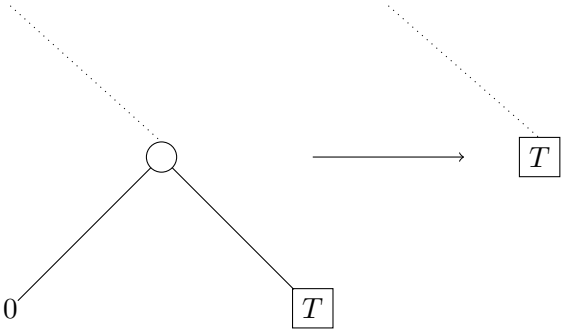

FIGURE 1.8 – Exemple de contraction d'une aiguille : le sous-arbre T est déplacé vers le père de sa racine

Pour calculer la profondeur d'un limbe v dans $\mathfrak{S}_\ell$, introduisons les notations suivantes :

- Considérons l'ensemble des nœuds internes sur le chemin de la racine à v dans $\mathfrak{Y}_{\ell-a_\ell}$, vu comme un sous-ensemble de $[0, \ell - a_\ell]$ selon le temps du processus de Yule. Notons par Π_v un processus ponctuel marqué, correspondant à l'ensemble de ces nœuds internes, étiqueté par 0 (resp. 1) s'il s'agit d'un fils gauche (resp. droit). On note par $\Pi_v^{(0)}$ (resp. $\Pi_v^{(1)}$) la restriction de Π_v aux points étiquetés par 0 (resp. 1).
- Pour tout processus ponctuel $\pi = \{\xi_1 < \xi_2 < \cdots < \xi_{k-1} < \xi_k\}$ à valeur dans un intervalle $[0, m]$, on note par $G(\pi)$ la quantité

$$G(\pi) = \sup_{i=1}^{k+1} \lfloor \xi_{i+1} - \xi_i \rfloor$$

où par convention $\xi_0 = 0$ et $\xi_{k+1} = m$.

Alors le nombre d'aiguilles entre la racine et v appartient à $\{G(\Pi_v^{(1)}), G(\Pi_v^{(1)}) + 1\}$, l'incertitude de 1 venant d'une condition au bord, qui ne pose pas de problème pour connaître la hauteur asymptotique. Par conséquent la profondeur d'un limbe v vaut

$$|\Pi_v| + G(\Pi_v^{(1)}) + \epsilon$$

avec $\epsilon \in \{0, 1\}$.

À l'exception du terme d'erreur ϵ, d'effet négligeable sur le comportement asymptotique, la profondeur d'une feuille peut donc être calculée comme une fonctionnelle relativement simple de l'arbre de Yule.

1.2.8.a Passage d'un arbre à un chemin par une formule *many to one*

On cherche à estimer le nombre de limbes ayant une profondeur donnée dans $\mathfrak{S}_\ell$. Cette hauteur est la somme de trois termes : $|\Pi_v^{(0)}| + |\Pi_v^{(1)}| + G(\Pi_v^{(1)})$.

Pour tous m, n et A, on dit qu'une feuille v de $\mathfrak{Y}_\ell$ est de type (m,n,A) si $|\Pi_v^{(0)}| = m$, $|\Pi_v^{(1)}| = n$ et $\Pi_v^{(1)} \in A$, on note par $\pi_{\ell,m,n,A}(\mathfrak{Y}_\ell)$ le nombre de limbes v de type (m,n,A) et par $\pi_{\ell,m,n,A}$ l'espérance du nombre de limbes v de type (m,n,A) dans $\mathfrak{Y}_\ell$.

On choisit $\vec{x}$ un mot infini à droite uniforme, indépendant de $\mathfrak{Y}_\ell$, vu comme une direction infinie dans l'arbre binaire complet infini, telle que fait dans la figure 1.5 page 43.

Cette direction infinie, dans $\mathfrak{Y}_\ell$ passe par une unique feuille que l'on note par $v_{\vec{x}}$, et on note par $E_{\ell,m,n,A}$ l'événement que $v_{\vec{x}}$ soit de type (m,n,A). Alors

$$\mathbb{P}(E_{\ell,m,n,A}|\mathfrak{Y}_\ell) = 2^{-m-n}\pi_{\ell,m,n,A}(\mathfrak{Y}_\ell)$$

En effet, la probabilité que $\vec{x}$ passe par une feuille donnée de type (m,n,A) correspond à la probabilité que $\vec{x}$ commence par un préfixe de longueur $m+n$ donné. En passant à l'espérance, on obtient :

$$\mathbb{P}(E_{\ell,m,n,A}) = 2^{-m-n}\pi_{\ell,m,n,A}$$

Par ailleurs, par définition d'un arbre de Yule, les marques de $\Pi_{v_{\vec{x}}}$ forment un processus de Poisson d'intensité $\ln 2$, chaque point ayant une marque 0 ou 1 indépendamment avec probabilité $\frac{1}{2}$. Par les proprités classiques des processus de Poisson, $\Pi^{(0)}_{v_{\vec{x}}}$ et $\Pi^{(1)}_{v_{\vec{x}}}$ sont deux processus de Poisson indépendants d'intensité $\frac{\ln 2}{2}$ sur $[0,\ell]$, ce qui nous permet d'obtenir la formule suivante :

$$\mathbb{P}(E_{\ell,m,n,A}) = \frac{(\frac{\ln 2}{2})^m}{m!}\frac{(\frac{\ln 2}{2})^n}{n!}\mathbb{U}_{\ell,n}(A)$$

où $\mathbb{U}_{\ell,n}$ est la loi d'un processus de Poisson sur $[0,\ell]$ d'intensité $\frac{\ln 2}{2}$, et conditionné à avoir n points : $\mathbb{U}_{\ell,n}$ est donc la loi uniforme sur le simplexe $\{0 < \xi_1 < \xi_2 < \cdots < \xi_n < \ell\}$. Nous avons ainsi prouvé la formule suivante :

Proposition 1.42 (Formule many to one)**.**

$$\pi_{\ell,m,n,A} = \frac{(\ell \ln 2)^{m+n}2^{-\ell}}{m!n!}\ \mathbb{U}_{n,\ell}(A)$$

Cette formule permet de passer d'une fonctionnelle de l'arbre de Yule tout entier à une fonctionnelle de processus de Poisson, plus facile à étudier.

1.2.8.b Utilisation des grandes déviations

On cherche à estimer le comportement de π lorsque ℓ tend vers l'infini. Par des techniques de grandes déviations, on montre qu'il existe une fonction Ψ telle que :

$$\Psi(\lambda,\mu,\nu) \ = \ \lim_n n^{-1}\ \ln\left(\pi_{\ell,m,n,g}\right)$$

lorsque $\frac{1}{n}(\ell,m,g)$ converge vers (λ,μ,ν).

Cela signifie que le nombre de limbes de type (m,n,g) dans $\mathfrak{S}_\ell$ est de l'ordre de $e^{n\Psi(\lambda,\mu,\nu)} = e^{\ell\Psi(\lambda,\mu,\nu)/\lambda}$.

1.2.8.c Greffe des arbustes

Pour obtenir l'arbre de Lyndon complet $\mathfrak{L}_\ell$ à partir de $\mathfrak{T}_\ell$, chaque limbe v est remplacé par un arbuste, correspondant à l'arbre de Lyndon associé à v. Dans la section 3.4 du chapitre 3, on prouve que la hauteur maximale dans un groupe de k arbustes se comporte comme le maximum de k géométriques indépendantes de paramètre $\frac{1}{2}$, et vaut ainsi approximativement $\log_2 k$. En conséquence, la profondeur maximale d'une feuille appartenant à un arbuste correspondant à un limbe de type (m, n, g) est proche de

$$\begin{aligned} \ell\frac{1+\mu+\nu}{\lambda} + \log_2(e^{\ell\Psi(\lambda,\mu,\nu)/\lambda}) &= \ell(\frac{1+\mu+\nu}{\lambda} + \frac{\Psi(\lambda,\mu,\nu)}{\lambda \ln 2}) \\ &= \ell\Delta(\lambda,\mu,\nu). \end{aligned}$$

En optimisant en λ, μ et ν, on obtient que la hauteur de $\mathfrak{L}_\ell$ est environ $\Delta^\bullet \ell$, avec

$$\Delta^\bullet = \sup_{\lambda,\mu,\nu>0} \Delta(\lambda,\mu,\nu).$$

1.2.8.d Retour à un mot de Lyndon de longueur fixée

Nous avons jusque là considéré $\mathfrak{L}(L^\ell)$, l'arbre de Lyndon associé à un mot de longueur aléatoire. Même si la hauteur de l'arbre de Lyndon L_n n'est pas une suite croissante, il est possible, en comparant $T(L_n)$ avec des arbres $T(L^\ell)$ pour des ℓ bien choisis de l'ordre de $\log_2 n$, d'encadrer la hauteur de L_n pour des n fixés, et d'ainsi déduire le théorème 1.39 du théorème 1.40.

1.3 Index des notations (français)

Ci-dessous se trouvent l'ensemble des notations utilisées dans ce manuscrit, avec un rapide rappel de leur sens et la partie où elles sont introduites.

1.3.1 Graphes aléatoires et limites locales

Ci-dessous se trouve une liste des notations utilisées dans la partie 1.1.

- $C(v, G)$: composante du sommet v dans le graphe G (partie 1.1.2.d).
- $C_i(G)$: taille de la i-ème plus grande composante du graphe G (partie 1.1.2.b).
- d_{loc} : distance locale sur les graphes enracinés (définition 1.8).
- d_k : type du sommet w_k (partie 1.1.3.f).
- $G(n, p)$: second modèle de graphe aléatoire d'Erdős et Rényi, ou modèle de Gilbert, à n sommets, où chaque arête est présente indépendamment avec probabilité p (définition 1.6).

- $\bar{G}(n,m)$: premier modèle de graphe aléatoire d'Erdős et Rényi, choisi uniformément parmi les graphes simples à n sommets et m arêtes (définition 1.5).
- $G^S(N,P)$: modèle de graphe aléatoire de Söderberg avec nombre de sommets de chaque type N et une matrice de probabilité P (définition 1.11).
- $\tilde{G}^S(n,r,P)$: modèle de graphe aléatoire de Söderberg avec n sommets i.i.d. de loi r et une matrice de probabilité P (définition 1.10).
- G_n^S : modèle de graphe aléatoire de Söderberg au seuil critique (partie 1.1.3.d).
- G_n^t : graphe $G(n, \frac{1}{n} + tn^{-\frac{4}{3}})$ (partie 1.1.2.g).
- H : matrice de terme $H_{i,j} = r_j \mathscr{P}_{i,j}$, dans le modèle de Söderberg (partie 1.1.3.b).
- L_k : chemin de Lukasiewicz (définition 1.9).
- N_i^n : nombre de sommets de G^n appartenant à des composantes de taille au moins i (partie 1.1.2.e).
- $\mathscr{P}$: matrice de probabilité d'existence d'arête, utilisée dans le modèle de Söderberg (partie 1.1.3.b).
- $\mathscr{P}_2(V)$: ensemble des parties à deux éléments de V.
- Poi(c) : variable aléatoire de Poisson de paramètre c.
- $PPP(t)$: processus ponctuel de Poisson d'intensité 1 sur $[0;t]$.
- r : proportion de sommets de chaque type, dans le modèle de Söderberg (définition 1.10 et partie 1.1.3.d).
- $\mathcal{S}$: ensemble des sommets *en réserve* de l'algorithme d'exploration défini à la partie 1.1.2.d.
- $\mathcal{T}$: ensemble des sommets non encore découverts par l'algorithme d'exploration défini à la partie 1.1.2.d.
- T_c^{GW} : arbre de Galton-Watson de loi de reproduction Poi(c).
- T_y^∞ : processus de branchement dont les arêtes sont étiquetés et de loi de reproduction $PPP(t)$ (définition 1.15).
- $\mathcal{U}$: ensemble des sommets *déjà utilisés* par l'algorithme d'exploration défini à la partie 1.1.2.d.
- $V_w^{\mathcal{S}}$: voisins de w appartenant à $\mathcal{S}$, lors de l'instruction 2' de l'algorithme d'exploration (partie 1.1.2.d).
- $V_w^{\mathcal{T}}$: voisins de w appartenant à $\mathcal{T}$, lors de l'instruction 2 de l'algorithme d'exploration (partie 1.1.2.d).
- w_k : sommet utilisé lors de la k-ème itération de l'algorithme d'exploration (la partie 1.1.2.d).
- W^t : mouvement brownien avec dérive parabolique de dimension ℓ

(partie 1.1.3.e).

- $\mathscr{W}^t$: mouvement brownien avec dérive parabolique de dimension 1 (partie 1.1.3.e).
- $\mathscr{W}_+^t$: mouvement brownien réfléchi en 0 (partie 1.1.3.e).
- x : vecteur propre à droite de H de valeur propre 1 (partie 1.1.3.e).
- y : vecteur propre à gauche de H de valeur propre 1 (partie 1.1.3.e).
- Z_k : généralisation du chemin de Lukasiewicz(partie 1.1.3.f).
- Δ_k : nombre de sommets découverts par l'algorithme d'exploration après k itérations (définition 1.9).
- Δ_k^i : nombre de sommets de type i découverts par l'algorithme d'exploration après k itérations(partie 1.1.3.f).
- Π : famille de processus de Poisson indiquant les temps d'apparition d'arêtes (éventuellement multiples) dans le second graphe d'Erdős et Rényi et ses variantes.
- λ : plus grande valeur propre de la matrice H (partie 1.1.3.b).
- ρ : dérive du mouvement brownien W^t (partie 1.1.3.e).
- $\Omega_{<\infty}$: ensemble des graphes finis à arêtes étiquetées, dont les étiquettes sont toutes différentes (partie 1.1.4.f).

1.3.2 Hauteur d'arbres, grandes déviations et marches branchantes

Ci-dessous se trouve une liste des notations utilisées dans la partie 1.2.

- ABR_n : arbre binaire de recherche correspondant à une permutation uniforme de $\{1, \dots, n\}$ (partie 1.2.1.a).
- L_n : mot de Lyndon uniforme sur $\mathcal{L}_n$ (partie 1.2.5.c).
- $\mathcal{L}$: ensemble des mots de Lyndon (définition 1.20).
- $\mathcal{L}_n$: ensemble des mots de Lyndon de longueur n (définition 1.20).
- $\mathfrak{L}^k(\cdot)$: partie haute de l'arbre de Lyndon $T(\cdot)$ (partie 1.2.7.a), ou partie haute de l'arbre de Lyndon d'une suite (définition 1.23).
- M_n : marche aléatoire (partie 1.2.4).
- $m(\theta)$: fonctionnelle utilisée dans l'étude des marches branchantes (partie 1.2.4).
- $\mathfrak{S}_\ell$: partie haute de l'arbre de Lyndon $\mathfrak{L}^{a_\ell}((n_v, s_v)_{1\le v\le N_\ell})$, associé à la suite modifiée des suffixes (partie 1.2.7.c).
- $T(w)$: arbre de Lyndon associé au mot de Lyndon w (définition 1.22).
- $\mathfrak{T}_\ell$: squelette $\mathfrak{L}^{a_\ell}(L^\ell)$ (partie 1.2.7.a).
- $U_k^{ABR_n}$: nombre de feuilles à profondeur k dans ABR_n (partie 1.2.2.a).
- $\vec{x}$: direction infinie dans l'arbre binaire complet (figure 1.5 page 43).

- Y_t : arbre de Yule au temps t (partie 1.2.2).
- $\mathfrak{Y}$: arbre de Yule construit à partir de la suite des suffixes modifiés de L_n (partie 1.2.8).
- $\mu(\theta)$: fonctionnelle utilisée dans l'étude des marches branchantes, définies à la partie 1.2.4.
- Σ : alphabet fini ou dénombrable.
- Σ^* : ensembles des mots finis ou infinis sur l'alphabet Σ.
- τ_n : premier instant où l'arbre de Yule Y_t a n sommets (lemme 1.26).
- $\prec$: relation d'ordre (partie 1.2.7.b).

1.4 Index (English)

In this part are the notations used the introduction chapter, with a short summary of their meaning and the part where they appear.

1.4.1 Graphes aléatoires et limites locales

Here is the list of notations used in part 1.1.

- $C(v, G)$: component of the vertex v in the graph G (part 1.1.2.d).
- $C_i(G)$: size of the i-th largest connected component of the graph G (part 1.1.2.b).
- d_{loc}: local distance on rooted graphs (*définition* 1.8).
- d_k: type of the vertex w_k (part 1.1.3.f).
- $G(n, p)$: second Erdős-Rényi random graph model, or Gilbert model, with n vertices, where each edge is present independently with probability p (*définition* 1.6).
- $\bar{G}(n, m)$: second Erdős-Rényi random graph model, chosen uniformly at random among simple graphs with n vertices and m edges (*définition* 1.5).
- $G^S(N, P)$: Söderberg random graph model, where N denote the number of vertices of each type, and P the probability matrix (*définition* 1.11).
- $\tilde{G}^S(n, r, P)$: Söderberg random graph model with n i.i.d. vertices of law r and a probability matrix P (*définition* 1.10).
- G_n^S: Söderberg random graph at the critical threshold (part 1.1.3.d).
- G_n^t: graph $G(n, \frac{1}{n} + tn^{-\frac{4}{3}})$ (part 1.1.2.g).
- H: matrix with entries $H_{i,j} = r_j \mathscr{P}_{i,j}$, in the Söderberg model (part 1.1.3.b).
- L_k: Lukasiewicz path (*définition* 1.9).

- N_i^n: number of vertices in components of size larger than or equal to i in G^n (part 1.1.2.e).
- $\mathscr{P}$: matrix driving the probability of existence of edges, used in the Söderberg model (part 1.1.3.b).
- $\mathscr{P}_2(V)$: family of sets of size 2 over V.
- Poi(c): Poisson random variable of parameter c.
- $PPP(t)$: Poisson point process of intensity 1 on $[0, t]$.
- r: proportion of each vertex type, in the Söderberg model (*définition* 1.10 and part 1.1.3.d).
- $\mathcal{S}$: set of vertices *in the stack* in the exploration algorithm defined at part 1.1.2.d.
- $\mathcal{T}$: set of undiscovered vertices in the exploration algorithm defined at part 1.1.2.d.
- T_c^{GW}: Galton-Watson tree with offspring law Poi(c).
- T_y^{∞}: branching process with labelled edges and offspring law $PPP(t)$ (*définition* 1.15).
- $\mathcal{U}$: set of *already used* vertices in the exploration algorithm defined at part 1.1.2.d.
- $V_w^{\mathcal{S}}$: neighbors of w belonging to $\mathcal{S}$, as in instruction 2' of the exploration algorithm (part 1.1.2.d).
- $V_w^{\mathcal{T}}$: neighbors of w belonging to $\mathcal{T}$, as in instruction 2 of the exploration algorithm (part 1.1.2.d).
- w_k: vertex used at the k-th iteration of the exploration algorithm (part 1.1.2.d).
- W^t: ℓ-dimensional Brownian motion with parabolic drift (part 1.1.3.e).
- $\mathscr{W}^t$: 1-dimensional Brownian motion with parabolic drift (part 1.1.3.e).
- $\mathscr{W}_+^t$: reflecting Brownian motion (part 1.1.3.e).
- x: right-eigenvector of H with eigenvalue 1 (part 1.1.3.e).
- y: left-eigenvector of H with eigenvalue 1 (part 1.1.3.e).
- Z_k: generalization of Lukasiewiczpath (part 1.1.3.f).
- Δ_k: number of discovered vertices after k iterations of the exploration algorithm (*définition* 1.9).
- Δ_k^i: number of discovered type i vertices after k iterations of the exploration algorithm (part 1.1.3.f).
- Π: Poisson point process family marking the times of additions of each edge in the second Erdős-Rényi random graph and its variants.
- λ: largest eigenvalue of H (part 1.1.3.b).
- ρ: drift of the Brownian motion W^t (part 1.1.3.e).

- $\Omega_{<\infty}$: set of finite graphs with labelled edges, such that all the labels are different (part 1.1.4.f).

1.4.2 Hauteur d'arbres, grandes déviations et marches branchantes

Here is the list of notations used in part 1.2.

- ABR_n: binary search tree associated to a random uniform permutation of $\{1, \dots, n\}$ (part 1.2.1.a).
- L_n: Lyndon word chosen uniformly at random on $\mathcal{L}_n$ (part 1.2.5.c).
- $\mathcal{L}$: set of the Lyndon words (*définition* 1.20).
- $\mathcal{L}_n$: set of the Lyndon words of length n (*définition* 1.20).
- $\mathfrak{L}^k(\cdot)$: top part of the Lyndon tree $T(\cdot)$ (part 1.2.7.a), or top part of the Lyndon tree of a sequence (*définition* 1.23).
- M_n: random walk (part 1.2.4).
- $m(\theta)$: function used in the study of branching random walks (part 1.2.4).
- $\mathfrak{S}_\ell$: top part of the Lyndon tree $\mathfrak{L}^{a_\ell}((n_v, s_v)_{1 \leq v \leq N_\ell})$ of the sequence of modified suffixes (part 1.2.7.c).
- $T(w)$: Lyndon tree of the Lyndon word w (*définition* 1.22).
- $\mathfrak{T}_\ell$: skeleton $\mathfrak{L}^{a_\ell}(L^\ell)$ (part 1.2.7.a).
- $U_k^{ABR_n}$: number of leaves at depth k in ABR_n (part 1.2.2.a).
- $\vec{x}$: infinite direction in the complete binary tree (figure 1.5 page 43).
- Y_t: Yule tree at time t (part 1.2.2).
- $\mathfrak{Y}$: Yule t built from the sequence of modified suffixes of L_n (part 1.2.8).
- $\mu(\theta)$: function used in the study of branching random walks (part 1.2.4).
- Σ: finite or countably infinite alphabet.
- Σ^*: set of finite of infinite words on the alphabet Σ.
- τ_n: first time where the Yule tree Y_t has n vertices (*lemme* 1.26).
- $\prec$: partial order (part 1.2.7.b).

Chapter 2

The size of the largest connected component at the critical window in an inhomogeneous graph

Contents

Consider the inhomogeneous random graph model studied by Söderberg [SÖ2] with ℓ types of vertices and look at the sizes of the largest components, when the parameters are close to the critical phase. The sequence of size of components, once rescaled by a factor $n^{\frac{2}{3}}$, converges in distribution

to the law of the excursions lengths of a reflecting Brownian motion with a drift similar to the one used by Aldous in [Ald97].

2.1 Introduction

In 1960, Erdős and Rényi [ER60] introduced and studied a simple model of random graph with n vertices, in which each edge is present with probability $\frac{c}{n}$ independently. When n tends to infinity, there is a threshold at $c = 1$ with a giant component of size $\Theta(N)$ appearing when c is larger than 1. Then Aldous [Ald97] studied the behavior at the critical window, *i.e.* when the probability of a given edge is close to $\frac{1}{n}$.

The Erdős-Rényi model is homogeneous (in the sense that all the vertices are roughly the same), and therefore fails to capture the inhomogeneity of many real-life networks. Consequently, several models have been proposed to overcome this restriction (see for example [BA99, NR06, SÖ2]). The existence of a threshold has been shown in various models ([ER60, BJR07, Tur06]). In particular, Bollobás, Janson and Riordan [BJR07] defined a general model of inhomogeneous random graph, where a type is assigned to each vertex, and the probability of an edge is a function (called the *kernel*) of the types of the endpoints of the edge, and showed that there is a phase transition in this general model.

Various results have been shown on the behavior of this graph at the critical window when the kernel is of rank 1 [Tur13, BvdHvL10], with an infinite number of types (see for example [Tur13] for a more precise description of the rank 1-model). Joseph [Jos14] obtained results in the configuration model and the graph with a given degree sequence.

The aim of this article is to study the behavior around the critical point in a model with a finite number of types, but with no additional restriction on the kernel.

2.1.1 The inhomogeneous random graph

Söderberg [SÖ2] introduced a model of inhomogeneous random graph model with a finite number of types. This random graph has with n vertices labelled from 1 to n, which can be of ℓ different types.

The first parameter is a vector r in the unit $(n-1)$-simplex that represents the proportion of each type of vertices. There exists two slightly different models: in the first one, the type of each vertex form a family of i.i.d. r.v. with common distribution r, whereas in the second one the number N_i^n of vertices of type i is chosen non-random and close to nr_i, and a partition of $[\![1; n]\!]$ in ℓ subsets is chosen uniformly with N_1^n vertices in the first subset, N_2^n in the second, and so on. For the sake of simplicity, we shall focus on the second model, with the possibility of a small deviation in proportion (up to an order $n^{-\frac{1}{3}}$).

More precisely, let $\tilde{r}$ be a vector in $\mathbb{R}^\ell$ such that $\sum_{i=1}^{\ell} \tilde{r}_i = 0$, and $(N_i^n)_{i\in[\![1,\ell]\!],n\in\mathbb{N}}$ be a family of non-negative integers such that:

- $\forall n \in \mathbb{N}, \sum_{i=1}^{\ell} N_i^n = n$ (2.1)
- $\forall i \in [\![1;\ell]\!], N_i^n = r_i n + \tilde{r} n^{\frac{2}{3}} + o(n^{\frac{2}{3}})$ (2.2)

It should be noted that these conditions are almost surely true with $\tilde{r} = 0$ if the first model is used, in which the type of each vertex is chosen independently. Therefore any result on the second model, with a fixed number of vertices of each type, can be extended to the first model, by conditioning on the number of vertices of each type.

Once a type is assigned to each vertex, the edges are drawn independently and with a probability that depends on the types of the endpoints of the edge. More precisely, let P^n be a $\ell \times \ell$ symmetric matrix with coefficients in $[0;1]$: if i_1 (resp. i_2) denotes the type of the vertex v_1 (resp. v_2), the edge (v_1, v_2) is present with probability $P^n_{i_1,i_2}$. The resulting random graph is denoted by $\mathscr{G}_n$ (or $\mathscr{G}$ when no confusion is possible). Söderberg showed in [SÖ2] that if $P^n = \frac{1}{n}\mathscr{P}$, the behavior of $\mathscr{G}_n$ depends mostly on the matrix H with general term:

$$H_{i,j} = r_j \mathscr{P}_{i,j}.$$

- If the largest eigenvalue of H is smaller than or equal to 1, then the largest component of $\mathscr{G}$ has size $o(n)$.
- If the largest eigenvalue of H is strictly larger than 1, then the largest component of $\mathscr{G}$ has size $\Theta(n)$.

In the following, we consider a matrix $\mathscr{P}$ such that the largest eigenvalue of H is 1, and we take

$$P^n = (\frac{1}{n} + tn^{-\frac{4}{3}})\mathscr{P}, \text{ with } t \text{ a real number.}$$

In this article, the following hypotheses will be assumed:

- $\forall i, r_i > 0$,
- $\mathscr{P}$ is irreducible,
- The columns of $\mathscr{P}$ are all different.

The first two hypotheses allow to be sure that the graph is not in fact two disjoint independent graphs, and the last hypothesis that two different types are really different.

Lemma 2.1. *All the eigenvalues of H of modulus 1 are simple.*

Proof. As a consequence of the hypotheses, for all (i, j) there exists an integer $k > 0$ such that the coefficient (i, j) of H^k is positive. Thus the non-negative matrix H is irreducible, and, the largest eigenvalue of H in absolute value is real (and therefore equal to 1) and is simple, by the Perron-Frobenius theorem. Nevertheless, H can have other eigenvalues of absolute value 1, who will then be simple and roots of unity[1]. □

Let x and y be the right-eigenvector and the left-eigenvector associated to the eigenvalue 1, such that:

- $\sum_{i=1}^{\ell} x_i = 1$
- $\sum_{i=1}^{\ell} y_i = 1$

The Perron-Frobenius theorem entails that such vectors x and y exist and that all the entries of x and y are positive. The largest entry of $\mathscr{P}$ will be denoted by $\mathscr{P}^+$, and x^+ and x^- will denote the largest and the smallest entry of x. It should be noticed that $x^- > 0$.

2.1.2 A Brownian motion with parabolic drift

Let $W(s)$ denote a ℓ-dimensional Brownian motion (without drift), with diagonal covariance matrix V:

$$\begin{pmatrix} y_1 & 0 & & \dots & 0 \\ 0 & y_2 & 0 & \dots & 0 \\ \dots & & \dots & & \\ & & & \dots & 0 \\ 0 & & & 0 & y_\ell \end{pmatrix}$$

i.e. $V_{i,j} = \delta_{i,j} y_i$. The ℓ-dimensional drift ρ is defined by:

$$\rho_i(s) = y_i s t + \frac{\tilde{r}_i y_i}{r_i} s - \frac{y_i^2}{2 r_i} s^2.$$

The Brownian motion with drift $W^t(s)$ is defined by $W^t(s) = W(s) + \rho(s)$. Theorem 2.2 will link this Brownian motion to the size of the largest components of $\mathscr{G}$.

[1] It will be the case if the graph is bipartite.

2.1.3 The relation between the size of the connected components of the graph and the Brownian motion

Let $\mathscr{W}^t$ and W_+^tbe defined as:

$$\mathscr{W}^t = \sum_{i=1}^{\ell} x_i W^{t,i};$$

$$\mathscr{W}_+^t(s) = \mathscr{W}^t(s) - \min_{u \leq s} \mathscr{W}^t(u).$$

$\mathscr{W}^t$ is a 1-dimensional Brownian motion with parabolic drift and $\mathscr{W}_+^t$ is a reflecting Brownian motion with parabolic drift.

Let $C_n(1) \geq C_n(2) \geq \cdots \geq C_n(j) \ldots$ be the sorted sequence of components' size of $\mathscr{G}_n$. $(C_n(j), j \geq 1)$ can be seen as an infinite non-increasing sequence by adding zeros to the end.

Let $\gamma(1) \geq \gamma(2) \ldots$ be the ordered lengths of the excursions of $\mathscr{W}_+^t$ (which is well defined, see Aldous [Ald97, Lemma 25]. The excursions above 0 of $\mathscr{W}_+^t$ are the excursions of $\mathscr{W}$ above its current minimum.

The main theorem of this article is the following:

Theorem 2.2. $(n^{-\frac{2}{3}} C_n(j))_{j \in \mathbb{N}} \xrightarrow[n \to \infty]{d} (\gamma(j))_{j \in \mathbb{N}}$. *The convergences holds for the l^2-topology.*

2.2 Preliminary results on the size of the components

2.2.1 The exploration process

In this part will be described a classical deterministic process that explores the connected components by constructing a rooted spanning tree on each connected component. Recall that the vertices are labelled from 1 to n.

The process uses a *FIFO* (*First In, First Out*) stack $\mathcal{S}$ of vertices. Initially, the stack contains only the vertex 1. In order to keep track of the vertices discovered along the way, the algorithm uses a dynamic partition of $[\![1; n]\!]$ in three sets:

- The set $\mathcal{U}$, initially empty, contains the vertices that have been in $\mathcal{S}$ at some point, but no longer are.
- The set $\mathcal{S}$ contains the vertices of the stack ($\mathcal{S}$ denotes both the stack, and the set of its vertices). At the beginning of the algorithm, $\mathcal{S} = \{1\}$.
- The set $\mathcal{T}$ contains the vertices that have never been in $\mathcal{S}$. At the beginning of the algorithm, $T = [\![2; n]\!]$.

The vertices in $\mathcal{U}$ will be called *used*, the vertices in $\mathcal{S}$ will be called *in the stack* and the vertices in $S \cup U$ will be called *discovered*.

As long as the stack $\mathcal{S}$ is not empty, an iteration of the process will consist of the following two instructions:

Instruction 1: Take the first vertex v out of the stack $\mathcal{S}$ and put it in $\mathcal{U}$.

Instruction 2: Ignoring the edges between v and vertices in $S \cup U$, remove any neighbor of v that belongs to $\mathcal{T}$, and add it, according to the FIFO rule, to the stack $\mathcal{S}$.

Remark.

- This process explores the connected component of 1, by iteratively looking at its neighbors, then at its neighbors' neighbors, and so on.
- Not looking at the edges between v and a vertex previously discovered does not prevent the algorithm from exploring the whole component. It only discards edges that would create cycle. Thus the algorithm generates a spanning tree.
- By the *FIFO* property, the process explores the component in a breadth-first order.
- At iteration 2, one often needs to add several vertices to the stack, and by the FIFO rules, they are added at the end of the stack. By convention, vertices that are added at the same time will be added in increasing label order (*i.e.* the largest label at the end of the stack).

This algorithm runs until the whole connected component of the vertex 1 is explored, and at that moment the stack $\mathcal{S}$ becomes empty. In order to continue the exploration of the graph, an extra instruction needs to be added:

Instruction 1: Take the first vertex v out of the stack $\mathcal{S}$ and put it in $\mathcal{U}$.

Instruction 2: Ignoring the edges between v and vertices in $S \cup U$, remove any neighbor of v that belongs to $\mathcal{T}$, and add it, according to the FIFO rule, to the stack $\mathcal{S}$.

Instruction 3: If the stack $\mathcal{S}$ is empty, take the vertex of smallest label in $\mathcal{T}$, remove it from $\mathcal{T}$ and add it to the stack $\mathcal{S}$.

This modification allows this exploration process to run until the whole graph is explored, generating a spanning forest in the process. The connected components are explored one by one, and in the order of the label of their smallest vertex. See figure 2.1 for an example of the exploration order.

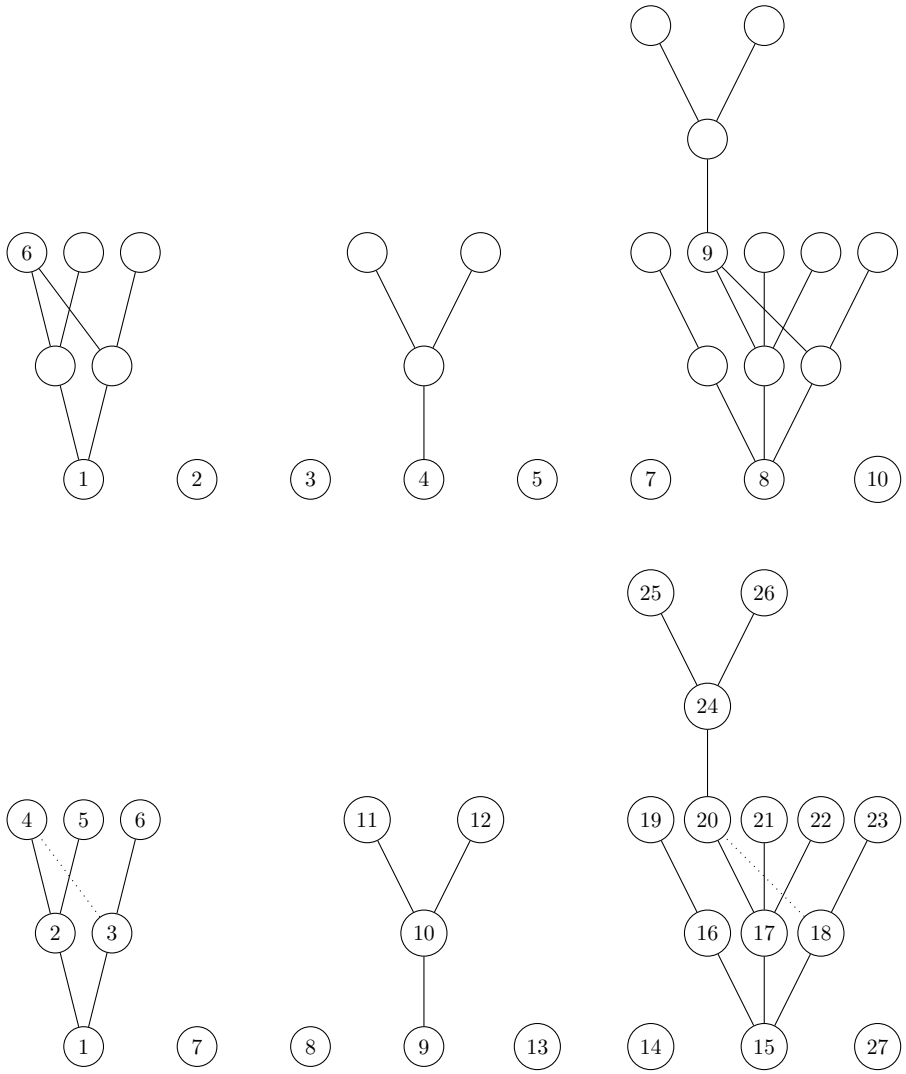

Figure 2.1: The first ten vertices in their component (top) and the order of exploration of the algorithm (bottom)

The set $\mathcal{U}$ (resp. $\mathcal{S}$, $\mathcal{T}$) after k iterations of the algorithm will be denoted by $\mathcal{U}_k$ (resp. $\mathcal{S}_k$, $\mathcal{T}_k$). The vertex added to $\mathcal{U}$ at the kth iteration will be denoted by v_k and its type by d_k. The set $\mathcal{U}$ (resp. $\mathcal{S}$, $\mathcal{T}$) after the first $k-1$ iterations of the algorithm, and the first two instructions of the kth iteration of the algorithm will be denoted by $\tilde{\mathcal{U}}_k$ (resp. $\tilde{\mathcal{S}}_k$, $\tilde{\mathcal{T}}_k$).

Let us state a few facts about this algorithm:

- At each iteration, a vertex is added to $\mathcal{U}$. Thus after k iterations, $\mathcal{U}_k$ has k vertices.

- The only difference between $\mathcal{U}_k, \mathcal{S}_k, \mathcal{T}_k$ and $\tilde{\mathcal{U}}_k, \tilde{\mathcal{S}}_k, \tilde{\mathcal{T}}_k$ comes from instruction 3. Therefore for every k, $\mathcal{U}_k = \tilde{\mathcal{U}}_k$. For this reason, the notation $\tilde{\mathcal{U}}$ will not be used.

- The instruction 3 only occurs when the process ends the exploration of a connected component. Then $\mathcal{S}_k$ and $\mathcal{T}_k$ differ from $\tilde{\mathcal{S}}_k$ and $\tilde{\mathcal{T}}_k$ only if the process ends the exploration of a connected component. In that case $\mathcal{S}_k = \tilde{\mathcal{S}}_k \cup \{v_{k+1}\}$ and $\mathcal{T}_k = \tilde{\mathcal{T}}_k \setminus \{v_{k+1}\}$.

The integer U_k^i (resp. S_k^i, T_k^i, $\tilde{S}_k^i$,$\tilde{T}_k^i$)) will denote the number of vertices of type i in $\mathcal{U}_k$ (resp. $\mathcal{S}_k$, $\mathcal{T}_k$, $\tilde{\mathcal{S}}_k$, $\tilde{\mathcal{T}}_k$).

2.2.2 A dynamic view

In order to study this algorithm dynamically, two filtrations $(\mathscr{F}_k)_{k\geq 1}$ and $(\tilde{\mathscr{F}}_k)_{k\geq 1}$ are needed. $\mathscr{F}_k$ (resp. $\tilde{\mathscr{F}}_k$) denotes the σ-algebras generated by two types of events: on one hand the presence or absence of the edges with at least one end in $\mathcal{U}_k$, on the other hand, the type of each vertex of $\mathcal{U}_k \cup \mathcal{S}_k$ (resp. $\mathcal{U}_k \cup \tilde{\mathcal{S}}_k$). It is then clear that, for every k:

$$\mathscr{F}_{k-1} \subset \tilde{\mathscr{F}}_k \subset \mathscr{F}_k$$

These σ-algebras have the following interpretations: $\mathscr{F}_k$ represents the information revealed by the first k iterations of the exploration process and $\tilde{\mathscr{F}}_k$ the information revealed by the first $k-1$ iterations and the first two instructions of the kth iteration.

Consider the random partition π_k (resp. $\tilde{\pi}_k$) of $\mathcal{T}_k$ (resp. $\tilde{\mathcal{T}}_k$) induced by the vertices' type: conditionally on $\mathscr{F}_k$ (resp. $\tilde{\mathscr{F}}_k$), π_k (resp. $\tilde{\pi}_k$) is uniform among partitions with T_k^i (resp. $\tilde{T}_k^i$) vertices of type i, which can be seen as some kind of self-similarity property. Moreover, conditionally on $\mathscr{F}_k$ and π_k (or $\tilde{\mathscr{F}}_k$ and $\tilde{\pi}_k$), the edges with no endpoints in $\mathcal{U}_k$ are present independently and with probability $P_{i,j}$ if the types of their endpoints are i and j.

2.2.3 A probabilistic construction

The law of the set of sizes of components is invariant under any permutation of vertices' labels. Thus the following recursive construction, that ignores labels, is more convenient for our purposes: consider $\mathbf{S}$ a finite word, with letters in the alphabet $[\![1;\ell]\!]$. The word $\mathbf{S}$ can be seen as the sequence of the types of the elements of $\mathcal{S}$ in the previous algorithm. Consider also U a vector in $\mathbb{N}^\ell$. $(U^i)_i$ will play the same role as $(U_k^i)_i$ in the previous algorithm. Similarly, S^i denotes the number of times the letter i appears in the word $\mathbf{S}$. Initially $U = 0$ and $\mathbf{S}$ contains only one letter, chosen randomly, equal to i with probability $\frac{N_i^n}{n}$. Let us iterate the following algorithm n times:

Instruction 1: Remove the first letter of the word $\mathbf{S}$, let say it is a "d", and increase U^d by one.

Instruction 2: Add to the end of $\mathbf{S}$ a random number of letters, with δ^i times the letter i. The law of $(\delta^i)_{i\in[\![1\,\ell]\!]}$, given the previous steps, is as follows:

- δ^i is a binomial random variable with parameters $(N_i^n - U^i - S^i, P_{d,i})$;
- the family $(\delta^i)_{1\leq i\leq \ell}$ is independent;
- the $\sum_{1\leq i\leq \ell}\delta^i$ letters are shuffled uniformly and then added at the end of the word $\mathbf{S}$.

Instruction 3: If $\mathbf{S}$ is the empty word, add to $\mathbf{S}$ a random letter, equal to i with a probability proportional to $N_i^n - U^i$.

Remark. The distributions in instructions 2 and 3 come from the properties of self-similarity described in Section 2.2.2.

In the following, U_k (resp. $\mathbf{S}_k$) denotes the state of U (resp. $\mathbf{S}$) after k iterations, and $\tilde{\mathbf{S}}_k$ denotes the state of $\mathbf{S}$ after $k-1$ iterations and the first two instructions of the kth iteration and d_k the value of d in the instruction 1 of the kth iteration of the algorithm.

The latter process, with the word $\mathbf{S}$, is necessary for rigorous proofs (in particular to be able to condition on the first k step of the exploration process), but the arguments will sometimes be expressed in the framework of the former process, on $\mathscr{G}_n$ for clarity reasons and to avoid confusing the reader used to considerations on graphs. The equivalences below allow the reader to express it in the framework of words if needed:

- The letters in $\mathbf{S}$ correspond to the types of the vertices contained in the stack $\mathcal{S}$.
- Adding a letter i to the word $\mathbf{S}$ is equivalent to discovering a vertex of type i.
- Using the instruction 3 is equivalent to ending the exploration of a connected component.

2.2.4 A variant of the Lukasiewicz path

Definition 2.1. Let δ_k^i denote the number of vertices of type i added to $\mathcal{S}$ in the kth iteration. Let $(Z_k)_{k\geq 1}$ be the discrete process with values in $\mathbb{R}^\ell$ defined by induction:

$$\begin{aligned} Z_0 &= 0 \qquad \text{and} \\ Z_k^i &= Z_{k-1}^i - H_{d_k,i} + \delta_k^i. \end{aligned}$$

For $\ell = 1$, we recover the Erdős-Rényi case: $H_{1,1} = 1$ and Z is the Lukasiewicz path, as defined in [LG10], also called *breadth-first walk* in Aldous [Ald97]. Thus Z can be seen as a ℓ-dimensional generalisation of the Lukasiewicz path.

If L_k is the usual Lukasiewicz path associated to a forest, $-\min_{j\leq k} L_j$ is the number of components explored exhaustively (*i.e.* how many times the instruction 3 is used) before the kth iteration of the algorithm, and $|\mathcal{S}_k|$ is equal to $L_k - \min_{j\leq k} L_j + 1$. These results comes from the following observations:

- L is increased by 1 each time a letter/vertex is added to $\mathcal{S}$ at instruction 2.
- L is decreased by 1 each time a letter/vertex is removed from $\mathcal{S}$ at instruction 1.
- Therefore the only vertices that affect the value of L are on one hand the vertices in the stack (that have increased the value of L but not already decreased it) counted with value $+1$ each and on the other hand the vertices added at the instruction 3, *i.e.* the first vertex of each component (that have not increased the value of L when discovered but still have decreased it when used) counted with value -1 each.

As with the Lukasiewicz path, estimations of the number of components, and of the number of vertices in the stack can also be derived from Z, by considering a suitable combination:

$$\mathscr{Z} = \sum_{i=1}^{\ell} x_i Z^i.$$

Proposition 2.3.

a) Each time the process ends the exploration of a component, say at iteration k, $\mathscr{Z}$ reaches a new strict downward record.

b) The number Λ_k of components explored exhaustively before iteration k satisfies:

$$-\frac{1}{x^+}\min_{j\leq k}\mathscr{Z}_j - \frac{x^+ - x^-}{x^+} \leq \Lambda_k \leq -\frac{1}{x^-}\min_{j\leq k}\mathscr{Z}_j.$$

c) The cardinal of $\mathcal{S}$ after k iterations satisfies

$$\frac{1}{x^+}(\mathscr{Z}_k - min_{j\leq k}\mathscr{Z}_j) + \frac{x^-}{x^+} \leq |\mathcal{S}_k| \leq \frac{1}{x^-}(\mathscr{Z}_k - min_{j\leq k}\mathscr{Z}_j) + \frac{x^+}{x^-}.$$

Remark.

- Unlike the Lukasiewicz path, there can be other downward records during the exploration of a given component.
- If $\ell = 1$, $x^- = x^+ = 1$ and we recover some known facts about the Lukasiewicz path.

Proof. The variation of $\mathscr{Z}$ between k and $k+1$ satisfies:

$$\mathscr{Z}_{k+1} = \mathscr{Z}_k - \underbrace{\sum_{j=1}^{\ell} H_{d_k,j} x_j}_{x_{d_k}} + \sum_{j=1}^{\ell} x_j \delta_k^j.$$

That means that instead of increasing (resp. decreasing) by 1 when a vertex is discovered (resp. used), $\mathscr{Z}$ increases (resp. decreases) by x_i, when a vertex of type i is discovered in instruction 2 (resp. used in instruction 1).

More precisely, for all $k \in \mathbb{N}$, let τ_k denote the step when the exploration of the kth component ends with the use of the instruction 3 (with $\tau_0 = 0$). Then the following lemma holds:

Lemma 2.4. *For each $k \geq 0$, the following formula holds:*

$$\mathscr{Z}_k = \sum_{i=1}^{\ell} x_i S_k^i - \sum_{j \geq 0, \tau_j \leq k} x_{d_{1+\tau_j}}.$$

Proof. This formula can be proven by induction:

- For $k = 0$, $\mathscr{Z}_k = 0$, for the only element of $\mathcal{S}$ is $d_1 = d_{1+\tau_0}$.
- Assume that the formula holds for $k-1$.

  During instruction 1, one d_k is removed from $\mathcal{S}$.

  During instruction 2, for each j, the letter j is added δ_k^j times j to $\mathcal{S}$. Therefore the variation of the first sum in the right-hand side during the first two instructions is $-x_i + \sum_{j=1}^{\ell} x_j \delta_k^j$, which is by definition the variation of the left-hand side between $k-1$ and k. Therefore:

  $$\mathscr{Z}_k = \sum_{i=1}^{\ell} x_i \tilde{S}_k^i - \sum_{j \geq 0, \tau_j \leq k-1} x_{d_{1+\tau_j}}.$$

  To end the proof of the lemma, we have to show that substituting S to $\tilde{S}$ in the first sum and k to $k-1$ in the second does not change the value of the right-hand side expression.

Case 1: $\tilde{\mathcal{S}}$ is not empty. Then $\mathcal{S}_k = \tilde{\mathcal{S}}_k$ and the second sum is not modified.

Case 2: $\tilde{\mathcal{S}}$ is empty. Then $\mathcal{S} = \{d_{k+1}\}$ and $k = \tau_j$, for some j, then the term $x_{d_{k+1}}$ is added in the two sums. Therefore the value of the right-hand side is not changed.

□

Proposition 1.9 can be derived from Lemma 2.4:

$$\mathscr{Z}_{\tau_m} = x_{d_{1+\tau_m}} - \sum_{j=0}^{m} x_{d_{1+\tau_m}} = \sum_{j=0}^{m-1} x_{d_{1+\tau_j}}, \text{ because } S_{\tau_m} = \{d_{1+\tau_m}\}.$$

For all integer j in $[\tau_m, \tau_{+1})$,

$$\mathscr{Z}_k \geq x^- - \sum_{j=0}^{m} x_{d_{\tau_j+1}}.$$

The x^- comes from the fact that, due to instruction 3, S_k is never empty, and thus the first sum is at least x^-.

Therefore, this entails a relation between the current minimum and the first sum of Lemma 2.4:

$$\forall m, k \text{ s.t. } \tau_m \leq k < \tau_{m+1}, -\sum_{j=0}^{m} x_{d_{1+\tau_j}} + x^- \leq \min_{j\leq k} \mathscr{Z}_j \leq -\sum_{j=0}^{m-1} x_{d_{1+\tau_j}}. \tag{2.3}$$

As $\Lambda_k = |\{m, \tau_m \leq k\}| - 1$ (the -1 is due to τ_0), this inequalities gives, by bounding x_j by x^- and x^+:

$$-\frac{1}{x^+}\min_{j\leq k}\mathscr{Z}_j - \frac{x^+ - x^-}{x^+} \leq \frac{1}{x^+}\sum_{j=0}^{m-1} x_{d_{1+\tau_j}} \leq \Lambda_k \leq \frac{1}{x^-}\sum_{j=0}^{m-1} x_{d_{1+\tau_j}} \leq -\frac{1}{x^-}\min_{j\leq k}\mathscr{Z}_j.$$

By using (2.3) and Lemma 2.4, the following inequalities hold for k such that $\tau_m \leq k < \tau_{m+1}$:

$$\begin{aligned}
-\min_{j\leq k}\mathscr{Z}_j + x^- \leq & \quad \sum_{j=0}^{m} x_{d_{1+\tau_j}} & \leq -\min_{j\leq k}\mathscr{Z}_j + x^+ \\
\mathscr{Z}_k - \min_{j\leq k}\mathscr{Z}_j + x^- \leq & \quad \mathscr{Z}_k + \sum_{j=0}^{m} x_{d_{1+\tau_j}} & \leq \mathscr{Z}_k - \min_{j\leq k}\mathscr{Z}_j + x^+ \\
\mathscr{Z}_k - \min_{j\leq k}\mathscr{Z}_j + x^- \leq & \quad \sum_{i=i}^{\ell} x_i S_k^i & \leq \mathscr{Z}_k - \min_{j\leq k}\mathscr{Z}_j + x^+ \\
\frac{1}{x^+}\left(\mathscr{Z}_k - \min_{j\leq k}\mathscr{Z}_j + x^-\right) \leq & \quad |S_k| & \leq \frac{1}{x^-}\left(\mathscr{Z}_k - \min_{j\leq k}\mathscr{Z}_j + x^+\right)
\end{aligned}$$

□

We now define a continuous-time version of Z and of the algorithm. The instruction 1 and 3 remain the same, and the instruction 2 is modified in the following way:

Instruction 2: Add to $\mathcal{S}$ (at the end of the stack, as it is a *FIFO* stack) at random times a random number of elements of $[\![1, \ell]\!]$. δ^i denotes the number of times the element i is added. The law of $(\delta^i)_{i \in [\![1\, \ell]\!]}$, given the previous steps, is:

- δ^i is a binomial random variable with parameters $(N^i - U^i - S^i, P_{d,i})$;
- The family $(\delta^i)_{1 \leq i \leq \ell}$ is independent;

For all i in $[\![1, \ell]\!]$, let Δ_k^i denote the multiset containing δ^i times i, and let Δ_k denote the disjoint union of the sets Δ_i^k for i in $[\![1, \ell]\!]$, *i.e.* the multiset containing δ^i times i for each i in $[\![1, \ell]\!]$. To determine the times when the elements of Δ_k are added to $\mathbf{S}$, let $(Y_\zeta)_{\zeta \in \Delta_k}$ be a family of i.i.d. random variables of uniform law on $[0, 1]$. An element ζ of Δ_k is added at the end of $\mathbf{S}$ at time $k + Y_\zeta$.

Remark. This construction implies that the elements of Δ_k are added in a random uniform order.

This construction can also be extended to have a continuous version of Z: If s is a real number in $[k, k+1]$, Z_n^i is defined by

$$Z_n^i(s) = Z_k^i - (s-k)H_{d_k,i} + \sum_{\zeta \in \Delta_k^i} \mathbf{1}_{\mathbf{Y_\zeta \leq s-k}}.$$

This definition coincides with Definition 2.1 at integer times. Let $\tilde{Z}_n$ be the rescaled version of Z, defined by:

$$\tilde{Z}_n(s) = n^{-\frac{1}{3}} Z_n(n^{\frac{2}{3}} s).$$

This rescaled version is a key argument for proving Theorem 2.2, due to the following property

Theorem 2.5. $\tilde{Z}_n \xrightarrow[n \to \infty]{d} W^t$

Recall that $\mathscr{W}^t(s) = \sum_{i=1}^{\ell} x_i W_i^t(s)$. If $\tilde{\mathscr{Z}_n}(s)$ denotes $n^{-\frac{1}{3}} \mathscr{Z}_n(n^{\frac{2}{3}} s) = \sum_{i=1}^{\ell} x_i \tilde{Z}_n^i(s)$, Theorem 2.5 implies that $\tilde{\mathscr{Z}_n} \to \mathscr{W}^t$ allowing to see the link between the size of the connected components and the lengths of the excursions of $\mathscr{W}_+^t$.

For all non-integer times s, let say in $(k, k+1)$ for some integer k, let $\mathscr{F}_s$ be defined as the σ-algebra generated by:

- the σ-algebra $\mathscr{F}_{\lfloor s \rfloor}$,
- the sets $\{Y_\zeta : Y_\zeta \leq s - k; \zeta \in \Delta_k^i\}$, for $i \in [\![1, \ell]\!]$.

This allows us to interpolate the σ-algebra $\mathscr{F}_k$ at non-integer times.

2.3 Proofs

2.3.1 Theorem 2.5

By general theory, we may write:

$$Z_n(s) = M_n(s) + A_n(s)$$

$$\forall (i,j) \in [\![1;\ell]\!]^2,\ M_n^i(s)M_n^j(s) = Q_n^{i,j}(s) + B_n^{i,j}(s)$$

such that:

- $M_n(.)$ and $Q_n(.)$ are martingales with respect to $(\mathscr{F}_s)_{s\geq 0}$,
- $A_n(.)$ is a continuous, bounded variation process,
- $B_n(.)$ is a continuous matrix with positive-definite increment,
- $M_n(0) = A_n(0) = Q_n(0) = B_n(0) = 0$.

For all positive real numbers s_0 and all $i \in [\![1;\ell]\!]$, the following limits will be proven:

$$n^{-\frac{1}{3}} \sup_{s\leq n^{\frac{2}{3}}s_0} \left| A_n^i(s) - y_i s t n^{-\frac{1}{3}} + \frac{(y_i s)^2}{2nr_i} - n^{-\frac{1}{3}}\frac{\tilde{r}_i y_i}{r_i} s \right| \xrightarrow[n\to\infty]{\text{p}} 0 \quad (2.4)$$

$$n^{-\frac{2}{3}} B_n(n^{\frac{2}{3}}s_0) \xrightarrow[n\to\infty]{\text{p}} s_0 V \quad (2.5)$$

$$n^{-\frac{2}{3}}\mathbb{E} \sup_{s\leq n^{\frac{2}{3}}s_0} |M_n(s) - M_n(s^-)|^2 \xrightarrow[n\to\infty]{\text{p}} 0 \quad (2.6)$$

The rescaled processes are defined in a consistent way by $\tilde{A}_n(s) = n^{-\frac{1}{3}}A_n(n^{\frac{2}{3}}s)$, $\tilde{M}_n(s) = n^{-\frac{1}{3}}M_n(n^{\frac{2}{3}}s)$, $\tilde{Q}_n(s) = n^{-\frac{2}{3}}Q_n(n^{\frac{2}{3}}s)$ and $\tilde{B}_n(s) = n^{-\frac{2}{3}}B_n(n^{\frac{2}{3}}s)$. The limits become:

$$\sup_{s\leq s_0} \left|\tilde{A}_n(s) - \rho(s)\right| \xrightarrow[n\to\infty]{\text{p}} 0$$

$$\tilde{B}_n(s_0) \xrightarrow[n\to\infty]{\text{p}} s_0 V$$

$$\mathbb{E}\sup_{s\leq s_0} \left|\tilde{M}_n(s) - \tilde{M}_n(s-)\right|^2 \xrightarrow[n\to\infty]{\text{p}} 0$$

This implies that $\tilde{Z}_n = \tilde{M}_n + \tilde{A}_n \xrightarrow{d} W + \rho$, uniformly on every compact, by standard arguments, see [EK86, Theorem 7.1.4(b)].

From now on, s_0 is a fixed positive real number. $Z_n^i(s)$ is a process with jumps $+1$ every time an item "i" is added to the stack $\mathcal{S}$ (*i.e.* a vertex of type i is discovered) and a drift $-H_{d_{\lfloor s\rfloor},i}$, therefore $\sup_{s\leq s_0}|M_n(s)-M_n(s^-)|^2 \leq 1$ and the limit (2.6) holds.

In the following, $(\eta_i)_{i\in[\![1,\ell]\!]}$ will denote the random piecewise constant function defined by:

$$\begin{aligned}\eta : \mathbb{R}_+ &\to \mathbb{R}^\ell \\ s &\to (H_{d_{\lfloor s\rfloor},i})_{1\le i\le \ell}.\end{aligned}$$

For each $i \in [\![1;\ell]\!]$, $Z_n^i(u)$ is a random process, with piecewise constant downward drift $-\eta_i(u)$ and jumps of height $+1$ whenever a vertex of type i is discovered. For $i \in [\![1;\ell]\!]$, and $u \ge 0$, let $a^i(u)$ denote the rate of jumps at time u, *i.e.* $a^i(u)du$ is the probability of a jump of Z^i between u and $u+du$, conditionally on $\mathscr{F}_u$. Similarly, for any $i,j \in [\![1;\ell]\!]$, let $b^{i,j}(u)$ be the rate of simultaneous jumps of Z_i and Z_j, *i.e.* $b^{i,j}(u)du$ is the probability that Z^i and Z^j both jumps between u and $u+du$, conditionally on $\mathscr{F}_u$. For any $i \neq j$, as the jumps (*i.e.* the time of addition of discovered vertices) are almost surely different, $b_n^{i,j} = 0$ a.s. For any i, $b_n^{i,i}(u) = a_n^i(u)$, by definition.

The processes $A_n^i(s)$ and $B_n^{i,j}(s)$ can be computed from the drift and rates of jumps via the following formulas:

$$A_n^i(s) = \int_0^s \left(a_n^i(u) - \eta_i(u)\right) du \tag{2.7}$$

$$B_n^{i,j}(s) = \int_0^s b_n^{i,j}(u)du. \tag{2.8}$$

Therefore B_n is a diagonal matrix. Let us estimate $a_n^i(u)$ on $[k, k+1[$.

Let $\nu^i(u)$ denote the number of vertices of type i already discovered at time u. The jumps correspond to discovered edges between v_k and vertices of type i. For an eligible vertex $\tilde{v}$ (*i.e.* a not already discovered vertex of type i) the edge $(v_k, \tilde{v})$ is discovered between u and $u+du$ if $\tilde{v}$ is a neighbour of v_k (which happens with probability $P_{d_k,i}$) and the associated uniform random variable $Y_{\tilde{v}}$ is in $[u, u+du]$. A simple computation shows that if an event occurs with probability p, and, conditionally on its occurrence, happens uniformly on $[0,1]$, then

$$\mathbb{P}(\text{event occurs during } [s, s+ds] | \text{does not occur before s}) = \frac{p}{1-sp}ds.$$

At time s, there are $N_i^n - \nu^i(s)$ eligible vertices of type i.

Thus, for any positive s:

$$a_n^i(s) = (N_i^n - \nu_n^i(s))\frac{P_{d_{\lfloor s\rfloor},i}}{1-(s-\lfloor s\rfloor)P_{d_{\lfloor s\rfloor},i}}.$$

To simplify the computation, we want to remove the denominator (which is close to 1) and substitute $r_i n + \tilde{r}_i n^{\frac{2}{3}}$ to N_i^n. Let $\bar{a}_n^i(s)$ be defined by:

$$\bar{a}_n^i(s) = (r_i n + \tilde{r}_i n^{\frac{2}{3}} - \nu_n^i(s))P_{d_{\lfloor s\rfloor},i}.$$

$$
\begin{aligned}
|a_n^i(s) - \bar{a}_n^i(s)| &\leq |a_n^i(s) - (N_i^n - \nu_n^i(s))P_{d_{\lfloor s \rfloor},i}| + |(N_i^n - \nu_n^i(s))P_{d_{\lfloor s \rfloor},i} - \bar{a}_n^i(s)| \\
&= (N_i^n - \nu_n^i(s))P_{d_{\lfloor s \rfloor},i}\frac{(s - \lfloor s \rfloor)P_{d_{\lfloor s \rfloor},i}}{1 - (s - \lfloor s \rfloor)P_{d_{\lfloor s \rfloor},i}} + P_{d_{\lfloor s \rfloor},i}|N_i^n - (r_i n + \tilde{r}_i n^{\frac{2}{3}})| \\
&\leq n(\frac{1}{n} + tn^{-\frac{4}{3}})^2\frac{\mathscr{P}^{+2}}{1 - \mathscr{P}^{+}(\frac{1}{n} + tn^{-\frac{4}{3}})} + \mathscr{P}^{+}(\frac{1}{n} + tn^{-\frac{4}{3}})\underbrace{\left|N_i^n - (r_i n + \tilde{r}_i n^{\frac{2}{3}})\right|}_{o(n^{\frac{2}{3}}),\text{ by (2.2)}} \\
&=: f^i(n)
\end{aligned}
$$

It should be noted that $f^i(n)$ does not depend on s, and that $f^i(n) = o(n^{-\frac{1}{3}})$. Let $f(n) = \max_{i \in [\![1;\ell]\!]} f^i(n)$, so $f(n)$ is a bound uniform in s and i: for all $s \geq 0$ and $i \in [\![1;\ell]\!]$

$$
|a_n^i(s) - \bar{a}_n^i(s)| \leq f(n). \tag{2.9}
$$

To continue, an estimation of $\nu_n(s)$ is needed. The following lemma shows that $\nu_n(s)$ is close to sy:

Lemma 2.6. *There exists a function g such that $g(n) \xrightarrow[n\to\infty]{} 0$ and, with probability tending to 1 when n tends to ∞:*

- $\forall i,\ \max_{k \leq s_0 n^{\frac{2}{3}}} \left|U_k^i - y_i k\right| \leq g(n)n^{\frac{2}{3}}$ *(2.10)*
- $\forall i,\ \max_{s \leq s_0 n^{\frac{2}{3}}} |\nu_n^i(s) - y_i s| \leq g(n)n^{\frac{2}{3}}$ *(2.11)*
- $\forall i,\ \max_{k \leq s_0 n^{\frac{2}{3}}} \left|\sum_{m=0}^{k-1}(k + \frac{1}{2})\mathbf{1}_{\mathbf{d_k=i}} - y_i\frac{k^2}{2}\right| \leq g(n)n^{\frac{4}{3}}$ *(2.12)*

Lemma 2.6 claims that, up to small error terms, one can assume that the types of the discovered vertices are distributed according to the eigenvector y. It should be noticed that this repartition is different from the repartition in the initial vertex set, close to r. The proof of Lemma 2.6 can be found in Section 2.3.3. We will also assume that $g(n) \geq \max(n^{-\frac{1}{3}}, n^{\frac{1}{3}}f(n))$. In the following computations, K will be a generic constant that does not depend on s or n and which value can change along the computation (it can depend on s_0).

In the next computation, we assume that the inequalities in Lemma 2.6 hold. Starting from the definition of $\bar{a}_n^i$ and using Lemma 2.6, inequality

(2.9) and $\eta_i(s) = r_i \mathscr{P}_{d_{\lfloor s \rfloor},i}$, the following inequalities hold:

$$\begin{aligned}
\bar{a}_n^i(s) - (r_i n + \tilde{r}_i n^{\frac{2}{3}} - \nu_n^i(s))\mathscr{P}_{d_{\lfloor s \rfloor},i}(\frac{1}{n} + tn^{-\frac{4}{3}}) &= 0 \\
|\bar{a}_n^i(s) - r_i \mathscr{P}_{d_{\lfloor s \rfloor},i} - n^{-\frac{1}{3}} r_i \mathscr{P}_{d_{\lfloor s \rfloor},i} t - n^{-\frac{1}{3}} \tilde{r}_i \mathscr{P}_{d_{\lfloor s \rfloor},i} + \frac{y_i s}{n} \mathscr{P}_{d_{\lfloor s \rfloor},i}| &\leq K\left(g(n)n^{-\frac{1}{3}} + n^{-\frac{2}{3}}\right) \\
|a_n^i(s) - \eta_i(s) - n^{-\frac{1}{3}}(t + \frac{\tilde{r}_i}{r_i})\eta_i(s) + \frac{y_i s}{n r_i}\eta_i(s)| &\leq K(g(n)n^{-\frac{1}{3}} + f(n)) \\
&\leq K(g(n)n^{-\frac{1}{3}})
\end{aligned}$$

Integrating over s:

$$\begin{aligned}
\left| A_n^i(s) - n^{-\frac{1}{3}}(t + \frac{\tilde{r}_i}{r_i}) \sum_{j=1}^{\ell} U_{\lfloor s \rfloor}^j H_{j,i} + \frac{1}{n}\frac{y_i}{r_i} \sum_{j=1}^{\ell} \sum_{k \leq s} (k + \frac{1}{2}) H_{j,i} \mathbf{1}_{\mathbf{j=d_k}} \right| &\leq K g(n) n^{-\frac{1}{3}} s \\
\left| A_n^i(s) - n^{-\frac{1}{3}}(t + \frac{\tilde{r}_i}{r_i}) s \underbrace{\sum_{j=1}^{\ell} y_j H_{j,i}}_{y_i} + \frac{1}{n}\frac{y_i}{r_i} \underbrace{\sum_{j=1}^{\ell} H_{j,i} y_j}_{y_i} \frac{s^2}{2} \right| &\leq K g(n) n^{\frac{1}{3}} \\
\left| A_n^i(s) - n^{-\frac{1}{3}}(t + \frac{\tilde{r}_i}{r_i}) y_i s + \frac{1}{n}\frac{y_i^2 s^2}{2 r_i} \right| &\leq K g(n) n^{\frac{1}{3}}
\end{aligned}$$

As $g(n) = o(1)$, and these inequalities hold if the inequalities of Lemma 2.6 hold, *i.e.* with probability tending to 1, the limit (2.4) is proven.

We now need to prove the limit (2.5). As B_n and V are diagonal, the limit is trivial for the non-diagonal coefficients. Let i be an integer in $[\![1; l]\!]$. By previous computations,

$$B_n^{i,i}(s) = A_n^i(s) + \int_0^s \eta_i(u) du.$$

$$\begin{aligned}
\left| \int_0^s \eta_i(u) du - \sum_{j=1}^{\ell} \sum_{k \leq \lfloor s \rfloor} H_{j,i} \mathbf{1}_{\mathbf{d_k = j}} \right| &\leq K \\
\left| \int_0^s \eta_i(u) du - \sum_{j=1}^{\ell} H_{j,i} U_{\lfloor s \rfloor}^j \right| &\leq K \\
\left| \int_0^s \eta_i(u) du - \sum_{j=1}^{\ell} H_{j,i} y_j s \right| &\leq K g(n) n^{\frac{2}{3}} \\
\left| \int_0^s \eta_i(u) du - y_i s \right| &\leq K g(n) n^{\frac{2}{3}}
\end{aligned}$$

As $\max_{s\leq s_0 n^{\frac{2}{3}}} |A_n(s)| = n^{\frac{1}{3}} \max_{s\leq s_0} \tilde{A}_n(s) = o_p(n^{\frac{2}{3}})$ according to limit (2.4), this proves the limit (2.5).

The limits (2.4), (2.5) and (2.6) have been proven ending the proof of the convergence of $\tilde{Z}_n$ and of Theorem 2.5.

2.3.2 Theorem 2.2

Let now see how Theorem 2.2 can be obtained from Theorem 2.5. By the Skorokhod representation theorem, it can assumed that W and all the $\mathscr{G}_n$ are on the same probability space, with $\tilde{Z}_n \to W$ uniformly on every compact. Then $\tilde{\mathscr{Z}} \xrightarrow[n\to\infty]{} \mathscr{W}$ uniformly on every compact. The properties of $\tilde{\mathscr{Z}}$ allow us to use this deterministic lemma, proposed by Aldous, in [Ald97, Lemma 7]

Lemma 2.7. *Let $f : \mathbb{R}^+ \to \mathbb{R}$ be a continuous function, f_n be a sequence of functions, and $(\tau_{n,i}, (n,i) \in \mathbb{N}^2)$ be a collection of non-negative real numbers. Let $\mathscr{E}$ be the set of non-empty intervals $e = (l, r)$ such that:*

$$f(r) = f(l) = \min_{s\leq l} \text{ and } f(s) > f(l) \text{ for } l < s < r$$

We assume that the following hypotheses hold:

1. *For all intervals (l_1, r_1), (l_2, r_2) in $\mathscr{E}$ with $l_1 < l_2$, we have $f(l_1) > f(l_2)$*
2. *The complement of $\bigcup_{e\in\mathscr{E}} e$ has Lebesgue measure zero.*
3. *$f_n \to f$ uniformly on every compact.*
4. *$\forall n, 0 = \tau_{n,0} < \tau_{n,1} < \dots$ and $\tau_{n,i} \xrightarrow[i\to+\infty]{} +\infty$.*
5. *$\forall i, n, f_n(\tau_{n,i}) = \min_{u\leq\tau_{n,i}} f_n(u)$.*
6. *$\forall s_0 \in \mathbb{R}^+, \max_{i:\tau_{n,i}\leq s_0} (f_n(\tau_{n,i}) - f_n(\tau_{n,i+1})) \xrightarrow[n\to+\infty]{} 0$.*

Let $\Xi = \{(l, r-l) : (l,r) \in \mathscr{E}\}$ and $\Xi^n = \{(\tau_{n,i}, \tau_{n,i+1} - \tau_{n,i})\}$. Then $\Xi^n \xrightarrow[n\to+\infty]{} \Xi$ for the vague convergence on counting measures on $(0,+\infty)\times(0,+\infty)$.

For all integer n and i, let $\gamma(n,i)$ the iteration of the algorithm when the exploration of the ith component ends ($\gamma(n,0) = 0$). With I denoting the number of components in $\mathscr{G}_n$, let $\gamma(n,i) = n + i - I$ for $i \geq I$. We can apply Lemma 2.7 with $f = \mathscr{W}$, $f_n = \tilde{\mathscr{Z}}_n$, and $\tau_{n,i} = n^{-\frac{2}{3}}\gamma(n,i)$:

- the Brownian motion $\mathscr{W}$ satisfies almost surely the hypotheses 1 and 2;

- hypothesis 3 is Theorem 2.5;
- hypotheses 4, 5 and 6 are direct consequences of the construction of $\mathscr{Z}$ and the Lukasiewicz path-like properties of $\mathscr{Z}$ proven in Proposition 2.3.

Let $C_{n,i}$ denote the size of the ith component of $\mathscr{G}$ (in the order of exploration, *i.e.* not sorted), then $\{(n^{-\frac{2}{3}}\gamma(n,i), n^{-\frac{2}{3}}C_{n,i})\} \xrightarrow[n\to+\infty]{} \{(l(\gamma), |\gamma|), \gamma \text{ an excursion of } \mathscr{W}_+^t\}$, with $l(\gamma)$ denoting the starting point of the excursion γ. This convergence implies a restricted version of Theorem 2.2, where one only considers the components of $\mathscr{G}_n$ whose exploration ended before $s_0 n^{\frac{2}{3}}$ iterations of the algorithm and the excursion of $\mathscr{W}_+^t$ ended before s_0. To fully prove Theorem 2.2, we are going to prove that no component of size $\Omega(n^{\frac{2}{3}})$ is missed when s_0 tends to ∞.

Let $T(z)$ and $T_n(z)$ be defined by:

$$T(z) = min\{s : \mathscr{W}^t(s) \leq -x^+ z - x^+\}$$

$$T_n(z) = min\{s : \tilde{\mathscr{Z}}(s) \leq -x^+ z - x^+\}$$

Proposition 2.3 entails that by iteration $n^{\frac{2}{3}}T_n(z)$ the algorithm has ended the exploration of at least $\lfloor \frac{n^{\frac{1}{3}}(x^+z+x^+)}{x^+} \rfloor - 1 = \lfloor zn^{\frac{1}{3}} \rfloor$ components, and therefore has discovered all the vertices whose label is smaller than $\lfloor zn^{\frac{1}{3}} \rfloor$. By Theorem 2.5, $T_n(z) \xrightarrow[n\to+\infty]{(d)} T(z)$.

Let $(C_n^z(j))_j$ be defined as the sorted (in decreasing order) sequence of the sizes of the components of $\mathscr{G}_n$ explored exhaustively before iteration $T_n(z)$ and let $(\gamma^z(j))_j$ be the sorted sequence of lengths of excursions of $\mathscr{W}_+^t$ before $T(z)$. As $T(z)$ is a.s. finite, the restricted version of Theorem 2.2 entails:

$$\forall z > 0, (n^{-\frac{2}{3}}C_n^z(j))_j \xrightarrow[n\to\infty]{} (\gamma^z(j))_j.$$

with convergence on every compact (and in fact, as implied later by the proof, also with respect to the l^2 topology).

To prove the unrestricted Theorem 2.2, we use the following lemma showing that no large component is missed when z tends to ∞:

Lemma 2.8. *Let $p(n,z,\delta)$ be the probability that $\mathscr{G}_n$ contains a component of size greater than $\delta n^{\frac{2}{3}}$ which does not contain any vertex i in $[\![1; \lfloor zn^{\frac{1}{3}} \rfloor]\!]$. Then*

$$\forall \delta > 0, \lim_{z\to\infty} \limsup_n p(n,z,\delta) = 0.$$

Proof. Let $q(\beta, n)$ denote the expected number of components of size greater than β in $\mathscr{G}_n$. Conditionally on the unlabelled graph, the labels of the

vertices are random, therefore the probability that none of the vertices of a given component of size larger than $\delta n^{\frac{2}{3}}$ has a label smaller than $zn^{\frac{1}{3}}$ is smaller than $\left(1-\frac{zn^{\frac{1}{3}}}{n}\right)^{\delta n^{\frac{2}{3}}} \xrightarrow[n\to\infty]{} exp(-\delta z)$. By summing over all the components, one can obtain:

$$\limsup_n p(n,z,\delta) \leq exp(-\delta z) \limsup_n q(\delta n^{\frac{2}{3}}, n)$$

Showing that $\limsup_n q(\delta n^{\frac{2}{3}}, n) < \infty$ for all $\delta > 0$ is sufficient to prove Lemma 2.8, and is a consequence of the following lemma:

Lemma 2.9. *Let w_k^n denote the probability that the component of the vertex 1 has a size larger than k. Then there exists a constant A (depending on H and t) such that*

$$\forall n,\, k,\, w_k^n \leq A(k^{-\frac{1}{2}} + n^{-\frac{1}{3}})$$

The proof of Lemma 2.9 is in Section 2.3.4.

As the law of $\mathscr{G}_n$ is invariant by permutation over the vertices, Lemma 2.9 remains true if 1 is substituted by any other vertex. By summing over the vertices, the expected number of vertices in component whose size is larger than k is smaller than nw_k^n, and consequently:

$$q(k,n) \leq \frac{nw_k^n}{k} \leq \frac{nA(k^{-\frac{1}{2}} + n^{-\frac{1}{3}})}{k} \tag{2.13}$$

Therefore $q(\delta n^{\frac{2}{3}}, n) \leq \frac{A(\delta^{-\frac{1}{2}}+1)}{\delta} < \infty$. □

Lemma 2.8 entails that the probability of missing a component of size at least $\delta n^{\frac{2}{3}}$ by stopping the exploration process at iteration $T_n(z)$ goes to 0 when z goes to infinity uniformly on n. Therefore $n^{-\frac{2}{3}}(C_n(j) - C_n^z(j))_j$ goes to 0, on every compact, uniformly on n, when z tends to infinity. As $(\gamma(j)) - \gamma^z(j))_j$ tends to 0 on every compact (as the set of excursion of $\mathscr{W}_+^t$ can a.s be sorted in decreasing order, by [Ald97, Lemma 25]), we have proven that

$$n^{-\frac{2}{3}}(C_n(j))_j) \xrightarrow[n\to\infty]{d} (\gamma(j))_j$$

uniformly on every compact.

Theorem 2.2 is true if one considers the convergence with respect to the product topology. To conclude with the l^2-topology, let us temporarily assume the following lemma, whose proof can be found shortly afterwards:

Lemma 2.10.

$$\lim_{z\to 0} \limsup_{n\to+\infty} \mathbb{E}\left(\sum_{i:C_n(i)\leq zn^{\frac{2}{3}}} (n^{-\frac{2}{3}}C_n(i))^2\right) = 0$$

For any integer m, the l^2-distance between $n^{-\frac{2}{3}}(C_n(j))_j$ and $(\gamma(j))_j$ can be bounded by the following sum:

$$\sum_{j=1}^{\infty}(n^{-\frac{2}{3}}C_n(j)-\gamma(j))^2 \leq \underbrace{\sum_{j\leq m}(n^{-\frac{2}{3}}C_n(j)-\gamma(j))^2}_{S_1^m} + 2\underbrace{\sum_{j>m}(n^{-\frac{2}{3}}C_n(j))^2}_{S_2^m} + 2\underbrace{\sum_{j>m}\gamma(j)^2}_{S_3^m}$$

Let $\epsilon > 0$ be a positive real number.

$(\gamma(j))_j$ is a.s. l^2 [Ald97, Lemma 25], thus there exists an integer m_3 such that with probability larger than $1-\epsilon$, S_3^m is smaller than ϵ for every $m \geq m_3$.

By Lemma 2.10, there exist a $z > 0$ such that, for n large enough,

$$\mathbb{E}\left(\sum_{i:C_n(i)\leq zn^{\frac{2}{3}}}(n^{-\frac{2}{3}}C_n(i))^2\right) \leq \epsilon^2.$$

Therefore, by Markov inequality, with probability larger than $1-\epsilon$,

$$\underbrace{\sum_{i:C_n(i)\leq zn^{\frac{2}{3}}}(n^{-\frac{2}{3}}C_n(i))^2}_{S_4} \leq \epsilon.$$

The expected number of components of size larger than $zn^{\frac{2}{3}}$ is $q(zn^{\frac{2}{3}}, n)$, and is finite according to the proof of Lemma 2.8. Therefore, by Markov inequality, there exists a $m_2 \geq m_3$ such that with probability larger than $1-\epsilon$, there are at most m_2 components of size larger than $zn^{\frac{2}{3}}$. In that case, $S_2^{m_2} \leq S_4$.

$S_1^{m_2}$ is the l^2-distance between γ and C_n restricted to their m_2 first components. The convergence with respect to the product topology has already been proven, thus $S_1^{m_2} \xrightarrow[n\to\infty]{} 0$ a.s. All these considerations prove that the convergence also holds for the l^2-topology.

The proof of Lemma 2.10 comes from the following computation, using

inequality (2.13):

$$\begin{aligned}
\mathbb{E}\Big(\sum_{i:C_{n,i}\le\lfloor zn^{\frac{2}{3}}\rfloor}(n^{-\frac{2}{3}}C_{n,i})^2\Big) &= n^{-\frac{4}{3}}\sum_{i:C_{n,i}\le\lfloor zn^{\frac{2}{3}}\rfloor}\mathbb{E}(C_{n,i}^2)\\
&= n^{-\frac{4}{3}}\sum_{k=1}^{\lfloor zn^{\frac{2}{3}}\rfloor}k^2\mathbb{E}(card\{i:C_{n,i}=k\})\\
&= n^{-\frac{4}{3}}\sum_{k=1}^{\lfloor zn^{\frac{2}{3}}\rfloor}(2k-1)\underbrace{\mathbb{E}(card\{i:C_{n,i}\ge k\})}_{q(k,n)}\\
&\le Kn^{-\frac{1}{3}}\sum_{k=1}^{\lfloor zn^{\frac{2}{3}}\rfloor}\underbrace{\frac{2k-1}{k}}_{\le 2}(k^{-\frac{1}{2}}+n^{-\frac{1}{3}})\\
&\le 2Kn^{-\frac{1}{3}}\left(\sum_{k=1}^{\lfloor zn^{\frac{2}{3}}\rfloor}k^{-\frac{1}{2}}+\sum_{k=1}^{\lfloor zn^{\frac{2}{3}}\rfloor}n^{-\frac{1}{3}}\right)\\
&\xrightarrow[n\to+\infty]{} 2K(2\sqrt{z}+z)\\
&\xrightarrow[z\to 0]{} 0
\end{aligned}$$

2.3.3 Lemma 2.6

Let us recall Lemma 2.6:

Lemma 2.6. *There exists a function g such that $g(n)\xrightarrow[n\to\infty]{}0$ and, with probability tending to 1 when n tends to ∞:*

- $\forall i,\ \max_{k\le s_0 n^{\frac{2}{3}}}\left|U_k^i-y_i k\right|\le g(n)n^{\frac{2}{3}}$ *(2.10)*
- $\forall i,\ \max_{s\le s_0 n^{\frac{2}{3}}}|\nu_n^i(s)-y_i s|\le g(n)n^{\frac{2}{3}}$ *(2.11)*
- $\forall i,\ \max_{k\le s_0 n^{\frac{2}{3}}}\left|\sum_{m=0}^{k-1}(k+\frac{1}{2})\mathbf{1}_{\mathbf{d_k=i}}-y_i\frac{k^2}{2}\right|\le g(n)n^{\frac{4}{3}}$ *(2.12)*

As $\nu^i(s)$ in non-decreasing, it is sufficient to prove (2.11) for s integer. The difference between (2.10) and (2.11) is the stack $\mathcal{S}$: $v_n^i(k)-U_k^i$ is the number of vertices of type i in the stack $\mathcal{S}_k$. Proposition 2.3 allow us to use $\mathscr{Z}$ to obtain bounds on the sets created by the exploration process. $\mathscr{Z}$ can be bound via the following lemma:

Lemma 2.11. *For a fixed s_0, $\max_{s\leq s_0 n^{\frac{2}{3}}} |\mathscr{Z}_s|$ is stochastically bounded by $n^{\frac{1}{3}}$.*

Proof of Lemma 2.11. The same decomposition is used for $\mathscr{Z}$ as for Z, with $\mathscr{Z} = \mathscr{M} + \mathscr{A}$ and $\mathscr{M}^2 = \mathscr{Q} + \mathscr{B}$, where $\mathscr{M}$ and $\mathscr{Q}$ are martingales, $\mathscr{A}$ a continuous, bounded variation process and $\mathscr{B}$ an increasing continuous function. Then, by linearity, $\mathscr{A} = \sum_{i=1}^{\ell} x_i A^i$ and $\mathscr{B} = \sum_{i=1}^{\ell} x_i^2 B^{i,i}$ (the non-diagonal terms are null). Let X be a constant and define the stopping time $T_{X,n}$ by $\min(s_0 n^{\frac{2}{3}}, \min\{s : |\mathscr{Z}_s| \geq X n^{\frac{1}{3}}\})$. In this computation, K will be a generic constant, whose value may vary, and does not depend on X or n (but can depend on s_0 and the parameters of the model).

Recall that $f(n)$ is a bound from above of the difference between $\bar{a}_n^i(s)$ and $a_n^i(s)$ and that $f(n) = o(n^{-\frac{1}{3}})$. Then, by the optional sampling theorem and equation (2.8):

$$\begin{aligned}
\mathbb{E}\mathscr{M}^2(T_{X,n}) &= \mathbb{E}\mathscr{B}(T_{X,n}) \\
&= \mathbb{E}\left(\int_0^{T_{X,n}} \sum_{i=1}^{\ell} x_i^2 a_n^i(s) ds\right) \\
&\leq \mathbb{E}\left(\int_0^{T_{X,n}} \sum_{i=1}^{\ell} x_i^2 (\bar{a}_n^i(s) + f(n)) ds\right) \\
&\leq \sum_{i=1}^{\ell} x_i^2 \int_0^{s_0 n^{\frac{2}{3}}} n\mathscr{P}^+(\frac{1}{n} + t n^{-\frac{4}{3}}) + f(n) ds \\
&\leq K n^{\frac{2}{3}}
\end{aligned}$$

$$\mathbb{E}(|A(T_{X,n})|) \leq \mathbb{E}\left(\int_0^{T_{X,n}} \sum_{i=1}^{\ell} x_i |a_n^i(s) - \eta_i(s)| ds\right)$$

$$\begin{aligned}
|a_n^i(s) - \eta_i(s)| &\leq |\bar{a}_n^i(s) - \eta_i(s)| + f(n) \\
&= |(r_i n + \tilde{r}_i n^{\frac{2}{3}} - \nu^i(s)) P_{d_{\lfloor s \rfloor}, i} - H_{d_{\lfloor s \rfloor}, i}| + f(n) \\
&\leq K n^{-\frac{1}{3}} + K \frac{\nu_n^i(s)}{n}
\end{aligned}$$

Moreover, for $s \leq T_{X,n}$, we can bound from above $\nu^i(s)$, by the number of vertices already discovered of any type, which is the sum of the number of vertices already used (*i.e.* $\lceil s \rceil$) and of the number of vertices in the stack.

By Proposition 2.3, the number of vertices in the stack is smaller than:

$$\frac{1}{x^-}(\max_{s\leq T_{X,n}} \mathscr{Z} - \min_{s\leq T_{X,n}} \mathscr{Z}) + \frac{x^+}{x^-} \leq \frac{2}{x^-} X n^{\frac{1}{3}} + K.$$

Then, by integrating, as $T_{X,n} \leq s_0 n^{\frac{2}{3}}$:

$$\mathbb{E}(|A(T_{X,n})|) \leq K(n^{\frac{1}{3}} + X).$$

This inequality implies:

$$\begin{aligned}
\mathbb{P}(\sup_{s \leq s_0 n^{\frac{2}{3}}} |\mathscr{Z}_n(s)| \geq X n^{\frac{1}{3}}) &= \mathbb{P}(|\mathscr{Z}_n(T_{X,n})| \geq X n^{\frac{1}{3}}) \\
&\leq \frac{\mathbb{E}|\mathscr{Z}_n(T_{X,n})|}{X n^{\frac{1}{3}}} \\
&\leq \frac{\mathbb{E}|\mathscr{M}_n(T_{X,n})| + \mathbb{E}|\mathscr{A}_n(T_{X,n})|}{X n^{\frac{1}{3}}} \\
&\leq \frac{\sqrt{\mathbb{E}(\mathscr{M}_n^2(T_{X,n}))} + K n^{\frac{1}{3}} + KX}{X n^{\frac{1}{3}}} \\
&\leq \frac{K}{X} + K n^{-\frac{1}{3}}
\end{aligned}$$

This bound ends the proof of Lemma 2.11. □

By Proposition 2.3, Lemma 2.11 has two useful implications. First it implies that the number of vertices in the stack, is stochastically bounded by $n^{\frac{1}{3}}$ and therefore the inequalities (2.10) and (2.11) of Lemma 2.6 are equivalent (the difference between the two being the vertices in the stack). Lemma 2.11 also entails that the number of components whose exploration started before iteration $\lfloor s_0 n^{\frac{2}{3}} \rfloor$ is stochastically bounded by $n^{\frac{1}{3}}$.

The next step of the proof of Lemma 2.6 is to estimate U_k^i. By definition of the process, there are k vertices used in the first k iterations, but the repartition of the types in these k vertices is random. We are going to show that the proportion of used vertices of type i in the first k iterations is close to y_i.

The instruction 2 of the kth iteration of the algorithm adds a random number of elements at the end of the stack $\mathcal{S}$. The number of times the element j is added is a binomial of parameter $N_j - \nu^j$ and $P^n_{d_k,j}$. Let $h^k_{i,j}$ denotes $\sum_{k' \leq k} (N_j - \nu^j(k')) \mathbf{1}_{\mathbf{d_{k'}=i}}$ (this can be seen as the number of Bernoulli random variables used by such binomials in the first k iterations of the algorithm). The exploration process can be coupled with ℓ^2 sequences of independent Bernoulli random variables, $\mathscr{B}^{i,j}_k$, for $(i,j) \in [\![1;\ell]\!]^2$, with $\mathbb{P}(\mathscr{B}^{i,j}_k = 1) = P_{i,j}$ such that the number of times the element i is added to the stack $\mathcal{S}$ in the kth iteration is the number of successes in the sequence $\mathscr{B}^{d_k,i}$ between positions $h^{k-1}_{d_k,i} + 1$ and $h^k_{d_k,i}$.

The following lemma says that the number of successes in the binomials is close to its expectation:

Lemma 2.12.

$$\forall (i,j) \in [\![1;\ell]\!]^2, \lim_{X\to+\infty} \left(\limsup_{n\to\infty} \mathbb{P}\left(\max_{k\leq s_0 n^{\frac{5}{3}}} \left| \sum_{m=1}^{k} \mathscr{B}_m^{i,j} - kP_{i,j} \right| \geq Xn^{\frac{1}{3}} \right) \right) = 0$$

Proof. We want to bound the gap between the number of successes in the Bernoulli variables and its expected value. $R_k = \sum_{m=1}^{k} \mathscr{B}_m^{i,j} - kP_{i,j}$ is a martingale. For a positive real number X, let define the stopping time T_X by $T_X = \min\{s_0 n^{\frac{5}{3}}, \min\{k : |R_k| \geq Xn^{\frac{1}{3}}\}\}$.

By the optional sampling theorem, $\mathbb{E}(R_{T_X}) = 0$, and

$\mathrm{Var}(R_{T_X}) = \mathbb{E}(T_X)P_{i,j}(1-P_{i,j}) \leq s_0 n^{\frac{5}{3}} P_{i,j} \leq Kn^{\frac{2}{3}}$ with K a constant independent of X. Then, by the Bienaymé-Chebyshev inequality:

$$\mathbb{P}(|R_{T_X}| \geq Xn^{\frac{1}{3}}) \leq \frac{Kn^{\frac{2}{3}}}{X^2 n^{\frac{2}{3}}} = \frac{K}{X^2} \xrightarrow[X\to\infty]{} 0.$$

As $|R_{T_X}| < Xn^{\frac{1}{3}}$ is equivalent to $\max_{k\leq s_0 n^{\frac{5}{3}}} |R_k| < Xn^{\frac{1}{3}}$, this concludes the proof of Lemma 2.12. □

From now on, the following conditions are assumed to hold:

(C1) There are less than $n^{\frac{1}{2}}$ components whose exploration started before the kth iteration.

(C2) $\forall (i,j) \in [\![1;\ell]\!]^2, \max_{k\leq s_0 n^{\frac{5}{3}}} \left| \sum_{m=1}^{k} \mathscr{B}_m^{i,j} - kP_{i,j} \right| \leq n^{\frac{1}{2}}$.

(C3) At any time before $s_0 n^{\frac{2}{3}}$, there are less than $n^{\frac{1}{2}}$ vertices in the stack.

Lemmata 2.11 and 2.12 imply that these three quantities are stochastically bounded by $n^{\frac{1}{3}}$, thus the three properties are asymptotically almost surely true. The end of the proof will consist in showing that these three conditions imply deterministically the inequalities of Lemma 2.6.

Lemma 2.13. *There exists a constant K (depending only on H) such that for all $u \in \mathbb{R}^\ell$, $||(u_i - y_i \sum_{j=1}^{\ell} u_j)_i|| \leq K||uH - u||$*

Lemma 2.13 holds for any norms, as all the norms are equivalent (the value of K will of course depend on the norms used).

Proof. Let φ_1 and φ_2 be two endomorphisms of $\mathbb{R}^\ell$ defined by:

$$\varphi_1(u) = uH - u$$

$$\varphi_2(u) = (u_i - y_i \sum_{j=1}^{\ell} u_j)_i$$

As, by Lemma 2.1, 1 is a simple eigenvalue of H, with left-eigenvector y, there exists F stable by H such that $\mathbb{R}^\ell = \mathbb{R}y \bigoplus F$. $\mathrm{Ker}(\Phi_2) = \mathbb{R}y$, therefore on the finite-dimensional F, $\varphi_{2|F} : F \to Im\Phi_2$ is invertible, and its inverse is continue. Therefore there exists a constant K such that for all u in $Im\Phi_2$, $||\varphi_1 \circ {\varphi_{2|F}}^{-1}(u)|| \leq K||u||$. Applying this inequality to $\varphi_2(u)$ gives for all $u \in F$:

$$||\varphi_1(u)|| \leq K||\varphi_2(u)||. \tag{2.14}$$

Let $u \in R^\ell$. There exists $(u_F, u_y) \in F \times \mathbb{R}y$. $\varphi_1(u) = \varphi_1(u_F)$ and $\varphi_2(u) = \varphi_2(u_F)$, therefore inequality (2.14) holds for every $u \in \mathbb{R}^\ell$.

□

Let $k \leq s_0 n^{\frac{2}{3}}$. To end the proof, we are going to prove that U_k is close to $U_k H$, andn therefore, by Lemma 2.13, close to ky. Let $\bar{\mathcal{U}}_k$ denote the set of children of the first k vertices and let $\bar{U}_k$ denote the vector of the numbers of vertices of $\bar{\mathcal{U}}_k$ of each type. The sets $\mathcal{U}_k$ and $\bar{\mathcal{U}}_k$ are close: the set $\mathcal{U}_k \setminus \bar{\mathcal{U}}_k$ contains the first vertex of each component seen before iteration k, and $\bar{\mathcal{U}}_k \setminus \mathcal{U}_k$ is the stack. By the conditions (C1) and (C3), there are less than $n^{\frac{1}{2}}$ vertices in each of these two relative complements. By definition of $h^k_{i,j}$, $\bar{U}^j_k = \sum_{i=1}^{\ell} \sum_{m=1}^{h^k_{i,j}} \mathscr{B}^{i,j}_m$.

Recall that $h^k_{i,j} = \sum_{k' \leq k} (N_j - \nu^j(k')) \mathbf{1}_{\mathbf{d_{k'}=i}}$. Therefore, $h^k_{i,j} \leq nk \leq s_0 n^{\frac{5}{3}}$, and by using the condition (C2)

$$|\bar{U}^j_k - \sum_{i=1}^{\ell} h^k_{i,j} P_{i,j}| \leq \ell n^{\frac{1}{2}}. \tag{2.15}$$

Moreover, as $\nu^j(k') \leq s_0 n^{\frac{2}{3}} + n^{\frac{1}{2}}$ (the first term bound from above the number of used vertices, and the second the number of vertices in the stack), then

$$\begin{aligned} (N_j - s_0 n^{\frac{2}{3}} - n^{\frac{1}{2}}) \sum_{k' \leq k} \mathbf{1}_{\mathbf{d_{k'}=i}} \leq \quad & h^k_{i,j} \quad \leq N_j \sum_{k' \leq k} \mathbf{1}_{\mathbf{d_{k'}=i}} \\ (N_j - s_0 n^{\frac{2}{3}} - n^{\frac{1}{2}}) U^i_k \leq \quad & h^k_{i,j} \quad \leq N_j U^i_k \\ |h^k_{i,j} - N_j U^i_k| \leq \quad & K n^{\frac{4}{3}} \end{aligned} \tag{2.16}$$

Inequalities (2.15) and (2.16) allow us to conclude:

$$||U_k H - U_k|| \leq ||U_k H - (\sum_{i=1}^{\ell} N_j U^i_k P_{i,j})_j|| + ||(\sum_{i=1}^{\ell} N_j U^i_k P_{i,j})_j - \bar{U}_k|| + ||\bar{U}_k - U_k||$$

As $U_k \leq s_0 n^{\frac{2}{3}}$ and $|r_j n - N_j| = O(n^{\frac{2}{3}})$:

$$
\begin{aligned}
|(U_kH)_j - \sum_{i=1}^{\ell} N_j U_k^i P_{i,j}| &\leq \sum_{i=1}^{\ell} |H_{i,j}U_k^i - N_j U_k^i \frac{1}{n}(1+tn^{-\frac{1}{3}})\mathscr{P}_{i,j}| \\
&\leq \sum_{i=1}^{\ell} |H_{i,j}U_k^i - r_j n U_k^i \frac{1}{n}(1+tn^{-\frac{1}{3}})\mathscr{P}_{i,j}| \\
&\quad + |r_j n - N_j|(U_k^i \frac{1}{n}|1+tn^{-\frac{1}{3}}|\mathscr{P}_{i,j}) \\
&= \sum_{i=1}^{\ell} H_{i,j}U_k^i |t| n^{-\frac{1}{3}} + |r_j n - N_j|(U_k^i \frac{1}{n}|1+tn^{-\frac{1}{3}}|\mathscr{P}_{i,j}) \\
&= O(n^{\frac{1}{3}}) \\
&= o(n^{\frac{2}{3}})
\end{aligned}
$$

$$
\begin{aligned}
|\sum_{i=1}^{\ell} N_j U_k^i P_{i,j} - \bar{U}_k^j| &\leq |\bar{U}_k^j - \sum_{i=1}^{\ell} h_{i,j}^k P_{i,j}| + \sum_{i=1}^{\ell} |N_j U_k^i P_{i,j} - h_{i,j}^k P_{i,j}| \\
&\leq \ell n^{\frac{1}{2}} + \sum_{i=1}^{\ell} P_{i,j}|N_j U_k^i - h_{i,j}^k| \\
&\leq \ell n^{\frac{1}{2}} + \sum_{i=1}^{\ell} P_{i,j} K n^{\frac{4}{3}} \\
&= O(n^{\frac{1}{2}}) \\
&= o(n^{\frac{2}{3}}) \\
|\bar{U}_k - U_k| &\leq 2n^{\frac{1}{2}}
\end{aligned}
$$

Thus $||U_kH - U_k|| = o(n^{\frac{2}{3}})$. This proves that, assuming conditions (C1), (C2) and (C3) hold, there exists $g(n) = o(n^{\frac{2}{3}})$ such that, for all $i \in [\![1;\ell]\!]$, and $k \leq s_0 n^{\frac{2}{3}}$, $|U_k^i - y_i k| \leq g(n)$ ($g(n)$ is uniform in $k \leq s_0 n^{\frac{2}{3}}$).

We now need to prove the inequality (2.12) of Lemma 2.6. We assume that the three conditions (C1), (C2) and (C3) still hold.

Let D be a function such that $D(n)n^{\frac{2}{3}} \xrightarrow[n\to\infty]{} \infty$ and $(g(n)+1)D(n) \xrightarrow[n\to\infty]{} 0$. The error terms will be denoted by θ_i, which will be proven subsequently to be $o(n^{\frac{4}{3}})$. The θ_i will be out of the summations.

Let $k_0 \leq s_0 n^{\frac{2}{3}}$ be a fixed integer.

$$\begin{aligned}
\sum_{k=1}^{k_0} k\mathbf{1}_{\mathbf{a_k=i}} &= \sum_{j=0}^{k_0 D(n)-1} \sum_{\frac{j}{D(n)}<k\leq\frac{(j+1)}{D(n)}} k\mathbf{1}_{\mathbf{a_k=i}} + \theta_0 \\
&= \sum_{j=0}^{k_0 D(n)-1} \sum_{\frac{j}{D(n)}<k\leq\frac{(j+1)}{D(n)}} \frac{j}{D(n)}\mathbf{1}_{\mathbf{a_k=i}} + \theta_0 + \theta_1 \\
&= \sum_{j=0}^{k_0 D(n)-1} \frac{j}{D(n)}\left(U^i_{\lfloor\frac{j+1}{D(n)}\rfloor} - U^i_{\lfloor\frac{j}{D(n)}\rfloor}\right) + \theta_0 + \theta_1 \\
&= \sum_{j=0}^{k_0 D(n)-1} \frac{j}{D(n)}\frac{y_i}{D(n)} + \theta_0 + \theta_1 + \theta_2 \\
&= \frac{y_i\lfloor k_0 D(n)\rfloor\lfloor k_0 D(n)-1\rfloor}{2D(n)^2} + \theta_0 + \theta_1 + \theta_2 \\
&= y_i\frac{k_0^2}{2} + \theta_0 + \theta_1 + \theta_2 + \theta_3
\end{aligned}$$

We now need to show that the θ_i can be bounded by a function $\tilde{g}(n) = o(n^{\frac{4}{3}})$, uniformly on $k_0 \leq s_0 n^{\frac{2}{3}}$.

θ_0 is the error due to the removal of some terms of the summation during the renumbering:

$$\begin{aligned}
\theta_0 &= \sum_{k\leq k_0} k\mathbf{1}_{\mathbf{a_k=i}} - \sum_{k=0}^{\frac{\lfloor k_0 D(n)\rfloor}{D(n)}} k\mathbf{1}_{\mathbf{a_k=i}} \\
\theta_0 &= \sum_{\lfloor\frac{\lfloor k_0 D(n)\rfloor}{D(n)}\rfloor+1}^{k=k_0} k\mathbf{1}_{\mathbf{a_k=i}} \\
|\theta_0| &\leq \frac{1}{D(n)} s_0 n^{\frac{2}{3}} \\
&= o(n^{\frac{4}{3}})
\end{aligned}$$

θ_1 is the error due to approximating k by $\frac{j}{D(n)}$:

$$\begin{aligned}
\theta_1 &= \sum_{j=0}^{k_0 D(n)-1} \sum_{\frac{j}{D(n)}<k\leq\frac{(j+1)}{D(n)}} (k-\frac{j}{D(n)})\mathbf{1_{a_k=i}} \\
|\theta_1| &\leq \sum_{j=0}^{k_0 D(n)-1} \sum_{\frac{j}{D(n)}<k\leq\frac{(j+1)}{D(n)}} (k-\frac{j}{D(n)}) \\
&\leq \sum_{j=0}^{k_0 D(n)-1} \sum_{\frac{j}{D(n)}<k\leq\frac{(j+1)}{D(n)}} \frac{1}{D(n)} \\
&\leq \frac{k_0}{D(n)} \\
&\leq \frac{s_0 n^{\frac{4}{3}}}{n^{\frac{2}{3}} D(n)} \\
&= o(n^{\frac{4}{3}})
\end{aligned}$$

θ_2 is the error due to approximating the number of vertices of type i between $\frac{j}{D(n)}$ and $\frac{j+1}{D(n)}$ by $\frac{y_i}{D(n)}$:

$$\begin{aligned}
\theta_2 &= \sum_{j=0}^{k_0 D(n)-1} \frac{j}{D(n)} \left(U^i_{\lfloor \frac{j+1}{D(n)} \rfloor} - y_i \frac{j+1}{D(n)} + y_i \frac{j}{D(n)} - U^i_{\lfloor \frac{j}{D(n)} \rfloor} \right) \\
|\theta_2| &\leq \sum_{j=0}^{k_0 D(n)-1} \frac{j}{D(n)} \left(\left| U^i_{\lfloor \frac{j+1}{D(n)} \rfloor} - y_i \left\lfloor \frac{j+1}{D(n)} \right\rfloor \right| + \left| U^i_{\lfloor \frac{j}{D(n)} \rfloor} - y_i \left\lfloor \frac{j}{D(n)} \right\rfloor \right| \right) + 2y_i \frac{j}{D(n)} \\
&\leq \sum_{j=0}^{k_0 D(n)-1} \frac{j}{D(n)} 2g(n) + 2y_i \frac{j}{D(n)} \\
&\leq (k_0 D(n))^2 \frac{(g(n)+y_i)}{D(n)} \\
&\leq s_0^2 n^{\frac{4}{3}} D(n)(g(n)+y_i) \\
&= o(n^{\frac{4}{3}}) \\
|\theta_3| &\leq \frac{3}{2} \frac{k_0}{D(n)} \\
&= o(n^{\frac{4}{3}})
\end{aligned}$$

It should be noted that upper bounds of the $|\theta_i|$ are uniform on $k_0 \leq s_0 \frac{2}{3}$, which ends the proof.

2.3.4 Lemma 2.9

Let us first recall Lemma 2.9

Lemma 2.9. *Let w_k^n denote the probability that the component of the vertex 1 has a size larger than k. Then there exists a constant A (depending on H and t) such that*

$$\forall n,\, k,\, w_k^n \leq A(k^{-\frac{1}{2}} + n^{-\frac{1}{3}})$$

We need to bound from above the probability w_k^n that the component of the vertex 1 in $\mathscr{G}_n$ has more than k vertices, which is also the probability that the tree created by the algorithm described in the part 2.2.3 has more than k vertices.

First, assume that $k \leq n^{\frac{2}{3}}$. Recall that the law of the number of children of type j of a vertex of type i is a binomial of parameters $(N_i^n - U_i - S_i, P_{i,j})$. As $N_i^n \leq r_i n + (\tilde{r}_i + 1)n^{\frac{2}{3}}$ for n large enough, by a coupling argument, w_k^n can be bounded from above by the probability that there is at least k nodes in a multitype Galton-Watson tree with the following law for the children: for a vertex of type i, the numbers of children of each type are independent, and the number of children of type j follow a binomial law of parameter $r_i n + (\tilde{r}_i + 1)n^{\frac{2}{3}}$ and $P_{i,j}^n$. Le Cam's theorem entails that this Galton-Watson tree can be coupled with a Galton-Watson tree with Poisson law of parameters $(r_i n + (\tilde{r}_i + 1)n^{\frac{2}{3}})P_{i,j}^n)_{i,j}$ in such a way that the probability that they differ in the first k vertices is smaller than

$$k\ell \max_i \left(r_i n + (\tilde{r}_i + 1)n^{\frac{2}{3}}\right) \max_{i,j} {P_{i,j}^n}^2 = O(n^{-\frac{1}{3}}).$$

This means that it is sufficient to prove Lemma 2.9 on this Galton-Watson tree instead of the component of 1 (up to a change in the constant A).

As this parameter is smaller than $H_{i,j} + n^{-\frac{1}{3}}\left((\tilde{r}_i + 1)\mathscr{P}_{i,j} + tH_{i,j} + 1\right)$ for n large enough, this last Galton-Watson tree can be bound from above by a Galton-Watson tree with Poisson law of parameters $H_{i,j} + n^{-\frac{1}{3}}\tilde{H}_{i,j}$ with $\tilde{H}_{i,j} = ((\tilde{r}_i + 1)\mathscr{P}_{i,j} + tH_{i,j} + 1)$.

The event that this Galton-Watson tree has more than k vertices is included in the union of two events: either the tree has more than $\sqrt{k}$ generations (*i.e.* has a depth larger than $\sqrt{k}$), or there are more than k vertices in the first $\sqrt{k}$ generations.

Let us bound from above the probability of the second event. If e_i denotes the row-vector of length ℓ, with $\ell - 1$ "0" and a "1" in the ith position, the expected number of vertices of type j after m generations of this Galton-Watson tree starting with a vertex of type i is:

$$e_i(H + n^{-\frac{1}{3}}\tilde{H})^m e_j^T$$

Lemma 2.14. *There exists a sub-multiplicative norm $N(.)$ on $\mathcal{M}_n(\mathbb{C})$ such that $N(H) = 1$.*

Proof. If P is an invertible matrix, and $\tilde{N}$ a sub-multiplicative norm, then $N : X \to \tilde{N}(P^{-1}XP)$ is a sub-multiplicative norm. Therefore it is sufficient to prove Lemma 2.14 with a matrix similar to H. Let H_1 be the Jordan normal form of H. As by Lemma 2.1, the eigenvalues of modulus 1 are simple, the blocks are either a simple complex number of modulus 1, or a block with λ on the diagonal, $|\lambda| < 1$, and 1 on the extra-diagonal. For any positive s, let $P = Diag(s, s^2, \ldots, s^\ell)$. $P^{-1}\tilde{H}P$ is the matrix H where every extra-diagonal 1 has be substituted by a s. With s small enough, $\tilde{N}(P^{-1}\tilde{H}P) = 1$ where $\tilde{N}(A) = \max_{1\leq j\leq \ell} \sum_{i=1}^{\ell} |A_{i,j}|$ is the induced norm with respect to the 1-norm.

□

This implies that $N((H + \frac{\tilde{H}}{n^{\frac{1}{3}}})^m) \leq (1 + \frac{N(\tilde{H})}{n^{\frac{1}{3}}})^m \leq exp(mn^{-\frac{1}{3}}N(\tilde{H}))$. Therefore there exists a constant A such that $e_i(H + n^{-\frac{1}{3}}\tilde{H})^m e_j^T \leq A$ for all i, j, n and for all $m \leq n^{\frac{1}{3}}$. Thus the expected number of vertices in the $\sqrt{k}$ first generations is smaller than $A\sqrt{k}$. By Markov inequality, the probability that there are more than k vertices in these $\sqrt{k}$ generations is smaller than $Ak^{-\frac{1}{2}}$.

Let us now deal with the probability of the first event. The probability of survival after k generations can be easily computed with the generating function (see for example [Har63] for a theory of multitype Galton-Watson trees, and branching processes in general). In this case, the generating function $f_n : \mathbb{R}^\ell \to \mathbb{R}^\ell$ is defined by:

$$\begin{aligned} f_n^{(i)}(s_1, \ldots s_\ell) &= \prod_{j=1}^{\ell} exp\left((H_{i,j} + n^{-\frac{1}{3}}\tilde{H}_{i,j}^n)(s_j - 1)\right) \\ &= exp(\sum_{j=1}^{\ell}(H_{i,j} + n^{-\frac{1}{3}}\tilde{H}_{i,j}^n)(s_j - 1)). \end{aligned}$$

f_n^k will denote the kth functional power of f_n. Then the probability of survival after k generations starting from a single vertex of type i is $1 - f_n^{k(i)}(0, 0, \ldots, 0)$.

By property of the exponential function, there is a constant $K > 0$ such that $1 - exp(z) \leq -z - Kz^2$ for all z smaller in absolute value than $\sum_{i,j} H_{i,j} + |\tilde{H}_{i,j}^n|$. Then:

$$1 - f_n^{(i)}(s) \leq -\sum_{j=1}^{\ell}(H_{i,j} + n^{-\frac{1}{3}}\tilde{H}_{i,j}^n)(s_j - 1) - K\left((H_{i,j} + n^{-\frac{1}{3}}\tilde{H}_{i,j}^n)(s_j - 1)\right)^2. \tag{2.17}$$

Let u_n^m be defined by $u_n^m = \sum_{i=1}^{\ell} y_i(1 - f_n^{m(i)}(0))$. By applying inequality (2.17) to $s = f_n^m(0)$ and summing over i, we obtain, for n large enough:

$$u_n^{m+1} \leq (1 + K'n^{-\frac{1}{3}})u_n^m - \frac{K}{2}\sum_{1=i}^{\ell} y_i\Big(\sum_{j=1}^{\ell} H_{i,j}(1 - f_n^{m(j)}(0))\Big)^2. \quad (2.18)$$

As $s \to \|s\|_H = \sqrt{\frac{K}{2}\sum_i y_i\left(\sum_j H_{i,j}s_j\right)^2}$ and $s \to \|s\|_y = \sum y_i|s_i|$ are strictly positive on $[0;\infty)^\ell \setminus \{0\}$ and are positive homogeneous of degree 1, there exists a constant K_2 such that, for all s in $[0,\infty)^\ell$, $\|s\|_H \geq K_2\|s\|_y$. Inequality (2.18) implies $u_n^{m+1} \leq (1 + K'n^{-\frac{1}{3}})u_n^m - (K_2u_n^m)^2$.

Let us define $\tilde{v} = K'n^{-\frac{1}{3}}$ and $v_m = K_2^2u^m - \tilde{v}$. By multiplying each side by K_2^2, this inequality becomes:

$$\begin{aligned} v_{m+1} + \tilde{v} &\leq (1 + \tilde{v})(v_m + \tilde{v}) - (v_m + \tilde{v})^2 \\ v_{m+1} &\leq v_m(1 - v_m - \tilde{v}) \end{aligned} \quad (2.19)$$

v_m is a decreasing sequence (it is an increasing function of u^m, a positive linear combination of probabilities of survival for m generations), therefore if $v_m \leq 0$ for some $m \leq \sqrt{k}$, $v_{\sqrt{k}} \leq 0 \leq \frac{1}{\sqrt{k}}$.

Assume that v_m is positive for all $m \leq \sqrt{k}$. Inequality (2.19) implies:

$$v_{m+1} \leq v_m(1 - v_m)$$

By using the inequality $\frac{1}{x}(1 - \frac{1}{x}) \leq \frac{1}{x+1}$, by induction, for all $m \leq \sqrt{k}$, $v_m \leq \frac{1}{m+1}$, thus $v_{\sqrt{k}} \leq k^{-\frac{1}{2}}$. This implies that $u_n^{\sqrt{k}} \leq K'(n^{-\frac{1}{3}} + k^{-\frac{1}{2}})$. As $1 - f^{m(i)}(0) \leq \frac{u_n^m}{y_i}$, this conclude the proof of Lemma 2.9 for $k \leq n^{\frac{2}{3}}$.

If k is larger than $n^{\frac{2}{3}}$, the probability that there are more than k vertices is smaller than the probability that there are more than $n^{\frac{2}{3}}$ vertices, which is smaller than $A((n^{\frac{2}{3}})^{-\frac{1}{2}} + n^{-\frac{1}{3}}) = 2An^{-\frac{1}{3}} \leq 2A(n^{-\frac{1}{3}} + k^{-\frac{1}{2}})$.

Chapter 3

Erdős-Rényi random graph process with forbidden degree

Contents

As suggested by Itai Benjamini, we introduced a variant of the Erdős-Rényi random graph process with a forbidden degree k, in which every edge adjacent to a vertex v is removed when the degree of v reaches k (but the removed edges may very well be added again later). We study the existence of a giant component, depending on the forbidden degree k and the time parameter t. We prove that for $k > 4$ a giant component appears at some point, while for $k < 4$, a giant component never occurs. The main tool of our study is the local limit of the random graph process: it provides useful information about the cases $k > 4$, but also the threshold case k=4.

3.1 Introduction

3.1.1 The model

We consider a sequence of multigraph-valued stochastic processes $G^k_n = (G^k_{n,t})_{t\geq 0}$, in which n and k, two positive integers, stand for the number of vertices and for the forbidden degree, respectively. The vertices are labelled from 1 to n, and E denotes the set of their $\binom{n}{2}$ eventual edges.

The multiplicity of the edges depends on a Poisson point process Π with intensity $\frac{1}{n}\mathbf{1}_{\mathbf{t}\geq\mathbf{0}}dt \otimes \mu$ on $\mathbb{R}_+ \times E$, where μ is the counting measure on E. Π will be seen as a marked point process, with elements in $\mathbb{R}_+$, marked by an edge in E. For any $e \in E$, Π_e denotes the point process on $\mathbb{R}_+$ of the elements of Π marked by e, and is the set of times when the multiplicity of the edge e is increased by 1. More precisely, initially, $G^k_{n,0}$ does not contain any edge. At each time $t \in \Pi_e$, the multiplicity of the edge e is increased by 1, and, for any endpoint w of e that reaches degree k at this step, every edge incident to w, including e, is removed (when both endpoints reach degree k simultaneously, the destruction of edges takes place on both sides)[1]. A given edge can be added several times to the process, therefore an edge can be removed, and added again at a later time; multiple edges can also occur. A vertex of $G^k_{n,t}$ is said *saturated* if its degree is $k-1$, the maximum possible degree.

At some point, a discrete time version of the continuous time process G^k_n will be needed: the unmarked version of Π is a Poisson point process with intensity $\frac{n-1}{2}$ on $\mathbb{R}_+$, so it is a.s. possible to order the points of Π in increasing order. For any integer i, let $\tau_i = \inf\{t \geq 0 : \Pi([0,t] \times E) \geq i\}$. By the memoryless property of the Poisson point process, the times τ_i are independent of the discrete process $(G^k_{n,\tau_i})_i$. This discrete process can be described as follows:

- At step $i = 0$, there is no edge.
- At each step $i \geq 1$, choose an edge uniformly at random among the $\binom{n}{2}$ elements of E and increase its multiplicity by one.
- For each endpoint reaching degree k, remove every edge incident to this endpoint.

The multigraph-valued stochastic process in which the edges appear according to $(\Pi_e)_{e\in E}$, but are never erased, is denoted $G^\infty_n = (G^\infty_{n,t})_{t\geq 0}$. We shall call it the *Erdős-Rényi multigraph process*, or the *Erdős-Rényi multigraph*. For any graph G, $C_{\max}(G)$ will denote the size (the number of vertices) of the largest connected component of G.

Possible generalisations Let G be a locally finite simple graph. For every edge e, let Π_e be a locally finite subset of $\mathbb{R}_+$. Define the forbidden degree version of G as the multigraph process with the same set of vertices as G evolving in the following way:

- Initially no edge is present.

[1] Technically, the endpoints will never reach degree k, going directly from degree $k-1$ to 0, but we will describe this situation as reaching degree k and immediately going to degree 0.

- For every point t of Π_e, increase the multiplicity of e by one at time t.

- If a vertex reaches the degree k, remove every adjacent edges.

This model is not always well-defined when the graph is infinite. A sufficient condition will be described in Section 3.4.2.

It is possible to allow the forbidden degree to depend on the vertex by using the following generalisation:

- Let $(k_v)_{v\in V}$ be a sequence of integers, indexed by the set of vertices. The integer k_v will be called *the forbidden degree of* v.

- The edges are added as previously. Whenever a vertex v reaches degree k_v, remove every edge adjacent to v.

This generalization allows to interpolate between two forbidden degrees, by setting the proportion of vertices of each forbidden degree, *e.g.* by having a proportion λ of the vertices with forbidden degree $k+1$ and a proportion $1-\lambda$ with forbidden degree $k+1$.

Heuristically, we expect that the resulting random graph process is stochastically increasing with the forbidden degree, even if it is not deterministically increasing, as explained in part 3.1.3.

3.1.2 Context

In [ER60], Erdős and Rényi obtained a striking result: the largest component of the Erdős-Rényi graph with tn edges and n vertices has two radically different behaviors depending on t. If $t \leq \frac{1}{2}$, $C_{\max}(G^{\infty}_{n,m}) = o(n)$ a.a.s whereas as soon as $t > \frac{1}{2}$ $C_{\max}(G) = \Theta(n)$. For this reason we will study the evolution of $C_{\max}(G^{k}_{n,t})$, depending on (k,t,n).

Another model with a degree constraint is the graph process with degree restriction [RW92]. There also exists several models with local constraints, *e.g.* the triangle-free process[Boh09] and H-free processes[ESW95, BR00, OT01] where H is a given subgraph. These models differ with the forbidden degree on many aspects, but in our eye the essential difference is that the graph processes are increasing in the former models, while in the forbidden degree model the graph process is not, as edges are routinely removed. Another model with edge removal is the Drossel-Schwab forest-fire model[DS92], where full components are removed with a rate proportional to their size.

3.1.3 Issues we have to deal with

Several points need to be considered when studying this process:

1. *Chronology.* For a fixed t, the Erdős-Rényi multigraph $G_{n,t}^{\infty}$ does not depend on the chronology of the apparition of edges, but only on the set of added edges, while due the forbidden degree constraint, chronology suddenly matters for $G_{n,t}^{k}$.

2. *Monotony.* G_n^k is not an increasing process, and the existence of a giant component at a given time does not imply the existence of a giant component at a later time. Actually, having more edges at a given time can lead to a smaller graph at a later time.

3. *Locality.* The degree of a given vertex depends on which adjacent edges have been removed. In turn, this depends on the degrees of the other endpoints of these edges, and these degrees depend on the neighbors of these endpoints, and so on... As a consequence, the existence of a local limit is questionable, much more than for the Erdős-Rényi random multigraph.

3.1.4 Results

The main results obtained in this article are the following:

Theorem 3.1. *For every $k \leq 3$ and for every sequence of non negative real numbers $(t_n)_{n\geq 0}$,*

$$C_{\max}(G_{n,t_n}^{k}) = o(n).$$

Theorem 3.1 means that there is no giant component for $k \leq 3$.

Theorem 3.2. *If $k \geq 5$, there exists an interval I such that for any time $t \in I$, $C_{\max}(G_{n,t}^{k}) = \Theta(n)$.*

According to Theorem 3.2, there exists a giant component at some point if $k \geq 5$, but this does not entail that a giant component still exists at a later time.

Theorem 3.3. *Let v be a random vertex of $G_{n,t}$, chosen uniformly among the vertices of $G_{n,t}$ independently of $(G_{n,t}^{\infty})_{t\geq 0}$. Then for any $k \in \mathbb{N} \cup \{\infty\}$, $(G_{n,t}^{k}, v)$ converges in distribution, for the local limit topology, to T_t^k, when n tends to ∞.*

T_t^{∞} is a Galton-Watson tree with Poissonian offspring distribution. T_t^k is the forbidden degree version of T_t^{∞}, and turns out to be a two-stages multitype branching process. The convergence is with respect to the local topology, as introduced by Benjamini and Schramm [BS01].

Theorem 3.4. *For any $t \geq 0$ and $k \in \mathbb{N}$,*

$$\frac{C_{\max}(G_{n,t}^{k})}{n} \xrightarrow[n\to\infty]{p} \mathbb{P}(|T_t^k| = \infty).$$

Expectedly, the behavior of the local limit predicts somehow the existence of the giant component, as the supercriticality of the local limit T_t^k is equivalent to the existence of a giant component.

Section 3.2 will be about Theorem 3.1, Theorem 3.2 will be discussed in Section 3.3, the existence and properties of the local limit in Section 3.4 and the Theorem 3.4 will be discussed in Section 3.5.

3.2 There is no giant component if $k \leq 3$

If $k \leq 2$, no component has more than 2 vertices. Thus, this section dealing only with the case $k = 3$, $G_{n,t}^3$ is denoted $G_{n,t}$. The largest allowed degree being 2, the connected components are either paths or cycles, which limits their growth. Theorem 3.1 follows from the next proposition.

Proposition 3.5. *There exists a constant A such that for any $\ell \in \mathbb{N}$, and for $n \geq 10$,*

$$\mathbb{E}\left[C_{\max}(G_{n,\tau_\ell})^2\right] \leq An.$$

Proof of Theorem 3.1. Let $|a|$ denote the cardinality of a finite set a, and set $N(t) = |\Pi \cap [0,t] \times E|$, resp. $N_e(t) = |\Pi_e \cap [0,t]|$. Since the two families $(\tau_i)_{i\in\mathbb{N}}$ and $(G_{n,\tau_i})_{i\in\mathbb{N}}$ are independent, we obtain:

$$\begin{aligned}
\mathbb{E}\left[C_{\max}(G_{n,t_n})^2\right] &\leq \sum_i \mathbb{E}\left[C_{\max}(G_{n,t_n})^2 \,|N(t_n) = i\right] \mathbb{P}(N(t_n) = i) \\
&\leq \sum_i \mathbb{E}\left[C_{\max}(G_{n,\tau_i})^2\right] \mathbb{P}(N(t_n) = i) \\
&\leq An.
\end{aligned}$$

□

Proof of Proposition 3.5. For any n and ℓ, let B_ℓ (resp. A_ℓ, C_ℓ) denote the set of connected components of G_{n,τ_ℓ} (resp. the set of acyclic connected components of G_{n,τ_ℓ}, the set of cycles of G_{n,τ_ℓ}). Consider Z_ℓ defined by:

$$Z_\ell = \sum_{a\in A_\ell} |a|^2 + 2\sum_{b\in C_\ell} |b|^2. \tag{3.1}$$

The reason for counting cycles twice will be clear later. Proposition 3.5 follows at once from the next Proposition.

Proposition 3.6. *There exists a positive constant A such that for all integers n and ℓ, $\mathbb{E}(Z_\ell) \leq An$.*

Proof. Set

$$u_\ell = \mathbb{E}\left[\frac{Z_\ell}{n}\right].$$

Let $\mathscr{F}_\ell$ be the σ-algebra generated by $(G_{n,t})_{t\le\tau_\ell}$, and let Δ denote the variation of Z_ℓ:

$$\Delta_{\ell+1} = Z_{\ell+1} - Z_\ell.$$

We shall prove that, for suitable constants α and β, both larger than 1,

$$\mathbb{E}(\Delta_{\ell+1}|\mathscr{F}_\ell) \le \alpha + \beta\tfrac{Z_\ell}{n} - \tfrac{1}{4}\tfrac{Z_\ell^2}{n^2}. \tag{3.2}$$

As a consequence,

$$n(u_{\ell+1} - u_\ell) \quad \le \quad \alpha + \beta u_\ell - \tfrac{1}{4}u_\ell^2,$$

Note that $u_0 = 1$ and let r denote the positive root of $\alpha + \beta X - \frac{1}{4}X^2$. Now,

- if $r \le u_\ell \le r + \alpha + \beta^2$, then (3.2) entails that $u_{\ell+1} \le u_\ell$;
- if $0 \le u_\ell \le r$, then (3.2) entails that $u_{\ell+1} - u_\ell \le \frac{1}{n}(\alpha + \beta^2)$.

Thus $A = r + \alpha + \beta^2$ is a suitable choice since $u_0 = 1 \le A$. □

Now, due to (3.1),

$$Z_\ell \ge C_{\max}(G_{n,\tau_\ell})^2,$$

which concludes the proof of Proposition 3.5, assuming relation (3.2). □

Proof of (3.2). Consider a graph process $\Gamma = (\Gamma_\ell)_{\ell\ge0}$ that starts from n vertices and no edges. At each step $\ell \ge 1$ a first random vertex v_ℓ is picked uniformly and a second random vertex w_ℓ different from v_ℓ is picked uniformly too. Then the multiplicity of the edge $e_\ell = \{v_\ell, w_\ell\}$ is increased by 1, provided that the forbidden degree rule allows it. If either v_ℓ or w_ℓ has degree 2 in $\Gamma_{\ell-1}$ then its edges are erased and e_ℓ is not added. Since Γ has the same distribution as $(G_{n,\tau_\ell})_{\ell\ge0}$, we shall consider, in what follows, that the process (Z_ℓ) is a functional of Γ, rather than a functional of $(G_{n,\tau_\ell})_{\ell\ge0}$.

Let us decompose the variation Δ_ℓ according to the connected components of v_ℓ and of w_ℓ. For $c, d \in B_\ell$, set

$$X_{c,d} = \Delta_\ell \mathbf{1}_{\mathbf{v}_\ell\in\mathbf{c}}\mathbf{1}_{\mathbf{w}_\ell\in\mathbf{d}},$$

so that

$$\Delta_\ell = \sum_{(c,d)\in B_\ell^2} X_{c,d}.$$

Since A_ℓ, B_ℓ, C_ℓ are measurable with respect to $\mathscr{F}_\ell$,

$$\mathbb{E}\left[\Delta_\ell|\mathscr{F}_\ell\right] = \sum_{(c,d)\in B_\ell^2} \mathbb{E}\left[X_{c,d}|\mathscr{F}_\ell\right].$$

Let us list the cases in which Z_ℓ increases:

1. **The edge e_ℓ makes c a cycle.** This entails that $c = d \in A_\ell$ and that v_ℓ and w_ℓ are the two endpoints of c, which happens with probability at most $\frac{2}{n(n-1)}$. In this case

$$\Delta_\ell = |c|^2.$$

2. **The edge e_ℓ merges c and d.** The components c and d merge only if $c, d \in A_\ell$, $c \neq d$ and v_ℓ and w_ℓ are endpoints of c and d. There are at most 2 endpoints per component, so, given c and d, this happens with probability at most $\frac{4}{n(n-1)}$. In this case

$$\Delta_\ell = (|c| + |d|)^2 - |c|^2 - |d|^2 = 2|c||d|.$$

Now let us list two of the cases in which Z_ℓ decreases:

3. **c is a cycle.** If $c \neq d$, the cycle c is split into a path of length $|c| - 1$ on one hand and an isolated vertex on the other hand. If $c = d$ the cycle can even be split in smaller parts, but in both cases $X_{c,d} \leq -|c|^2$ with a probability $\frac{|c|}{n}$.

4. **c has endpoints but v_ℓ is not one of them.** Let us say that v_ℓ is the mth vertex of c. If $c \neq d$, the path c is split into three paths, of length $m - 1$, $|c| - m$ and 1. If $c = d$, one of the three previous paths can be split again due to the effect of w_ℓ. Thus, for any m with $1 < m < |c| - 1$, with probability $\frac{1}{n}$,

$$X_{c,d} \leq (m-1)^2 + (|c| - m)^2 + 1 - |c|^2 \leq 2m(m - |c|).$$

These 4 cases cover the possible contributions of (c, v_ℓ) to Δ_ℓ. The two first cases, in which the two sides of e_ℓ play symmetric rôles, also cover the positive contributions of the other side of e_ℓ, (d, w_ℓ), to Δ_ℓ. As we aim to provide an upper bound for Δ_ℓ, we do not need to discuss the negative contributions of the other side, and considering the 4 cases, we obtain:

$$\sum_{(c,d) \in B_\ell^2} \mathbb{E}\left[X_{c,d} | \mathscr{F}_\ell\right] \leq \sum_{1 \leq i \leq 4} D_i,$$

in which

$$\begin{aligned} D_1 &= \tfrac{2}{n(n-1)} \sum_{c \in A_\ell} |c|^2 \leq \frac{2Z_\ell}{n(n-1)}, \\ D_2 &= \tfrac{8}{n(n-1)} \sum_{c,d \in A_\ell} |c||d| \leq \frac{8n}{n-1}, \\ D_3 &= -\tfrac{1}{n} \sum_{c \in C_\ell} |c|^3 \end{aligned}$$

and

$$\begin{aligned} D_4 &= \tfrac{1}{n}\sum_{c\in A_\ell}\sum_{m=2}^{|c|-1} 2m(m-|c|) \\ &= \tfrac{1}{3n}\sum_{c\in A_\ell}(-|c|^3+7|c|-6)\mathbf{1}_{|\mathbf{c}|\geq \mathbf{2}} \\ &\leq \tfrac{7}{3}-\tfrac{1}{3n}\sum_{c\in A_\ell}|c|^3. \end{aligned}$$

Thus, for $n \geq 9$,

$$\mathbb{E}\left[\Delta_\ell|\mathscr{F}_\ell\right] \leq 12+\frac{3Z_\ell}{n^2}-\frac{1}{4n}\left(\sum_{c\in A_\ell}|c|^3+\sum_{c\in C_\ell}4|c|^3\right).$$

Let $(a_c)_{c\in B_\ell}$ and $(b_c)_{c\in B_\ell}$ denote two sequences defined as follows:

- if $c \in A_\ell$, $a_c = \sqrt{|c|}$ and $b_c = |c|^{\frac{3}{2}}$;
- if $c \in C_\ell$, $a_c = \sqrt{|c|}$ and $b_c = 2|c|^{\frac{3}{2}}$.

Then, by the Cauchy-Schwarz inequality, we have:

$$(Z_\ell)^2 = (\sum_{c\in B_\ell} a_c b_c)^2 \leq \sum_{c\in B_\ell} a_c^2 \sum_{c\in B_\ell} b_c^2 = n(\sum_{c\in A_\ell}|c|^3+\sum_{c\in C_\ell}4|c|^3).$$

As a consequence, (3.2) holds true for $(\alpha,\beta)=(12,3)$. □

3.3 At some point, there is a giant component if $k \geq 5$

In this part, rather than considering the process with forbidden degree, we shall consider a lower bound, i.e. a new process in which all the edges incident to a vertex to which at least k edges have been added are removed. For $k \geq 5$, this new process is supercritical, allowing us to prove Theorem 3.2.

3.3.1 Approximation by a simpler model

Consider the graph $g^k_{n,t}$ obtained when one erases all the edges of $G^\infty_{n,t}$ that are incident to a vertex with degree k or more. As opposed to $G^k_{n,t}$, $g^k_{n,t}$ depends on Π only through $(N_e(t))_{e\in E}$, thus it does not depend on the chronology. Let T_k denote the transformation that maps $(N_e(t))_{e\in E}$ to $g^k_{n,t}$:

$$g^k_{n,t} = T_k(N_\cdot(t)).$$

Lemma 3.7. *For any n and t,*

$$g^k_{n,t} \leq G^k_{n,t} \leq G^\infty_{n,t}.$$

If $G(e)$ denotes the multiplicity of the edge e in G, the order we consider in this section is the product order for $(G(e))_{e\in E}$.

Proof. The second inequality is clear. For the first one, note that if $0 \leq G^k_{n,t}(e) < G^\infty_{n,t}(e)$, then one of the endpoints of e reaches the threshold k in G^∞_n at some point s before t, and, as a consequence, $g^k_{n,t}(e) = 0$. Else none of the endpoints' degrees of e reach the threshold k in $G^\infty_{n,t}$, and as a consequence $g^k_{n,t}(e) = G^k_{n,t}(e) = G^\infty_{n,t}(e)$. □

The graph $g^k_{n,t}$, conditionally on its degree sequence, is distributed as a configuration model conditioned on being loopless, as will be proven in Corollary 3.9. The configuration model, introduced by Bender and Canfield [BC78, page 297] is defined as follow:

Definition 3.1. Let $(c_v)_{1\leq v\leq n}$ be a finite sequence of non-negative integers, such that $\sum_v c_v$ is even. The configuration model associated to $(c_v)_{1\leq v\leq n}$ is the random graph constructed as follow:

- The set of vertices V is $\{1, \dots, n\}$.
- Let L be the multiset containing c_v copies of each vertex v of V.
- A random uniform pairing E of the elements of L is chosen.
- The set of edges is defined by the pairing, with each pair (x, y) corresponding to an edge between x and y.

By construction, the vertex v has degree c_v in the resulting graph.

Remark. The configuration model is sometimes defined from the sequence $(d_j)_{j\geq 0}$ where $d_j = \#\{v : c_v = j\}$ denotes the number of vertices of degree j, instead of the degree sequence c.

Let $c = (c_i)_{1\leq i\leq n}$ be a sequence of integers smaller than k such that $\sum c_i$ is even and let $\mathscr{G}_c$ denote the set of loopless multigraphs with degree sequence c. An element of $\mathscr{G}_c$ is described by the sequence $(g_e)_{e\in E}$ of its edges' multiplicity.

Lemma 3.8. *For g_1, $g_2 \in \mathscr{G}_c$, let m^j_i denote the number of edges with multiplicity i in g_j. Then for any $t \geq 0$,*

$$\mathbb{P}(g^k_{n,t} = g_1) \prod_{i\geq 2} (i!)^{m^1_i} = \mathbb{P}(g^k_{n,t} = g_2) \prod_{i\geq 2} (i!)^{m^2_i}.$$

Corollary 3.9. *The graph $g^k_{n,t}$, conditionally given its degree sequence, is a configuration model conditioned to be loopless.*

In the configuration model, the number of configurations corresponding to a given multigraph is $\frac{\prod_v c_v!}{\prod_i (i!)^{m_i}}$, thus Lemma 3.8 implies Corollary 3.9.

Proof of Lemma 3.8. Set $B = \{v \in V : c_v > 0\}$. Let $g \in \mathscr{G}_c$. Let $s(e)_{e \in E}$ be a multigraph on the set of vertices $\{1, \dots, n\}$.

Lemma 3.10. *For all $e \in E$, let $r(e) = s(e) - g(e)$. $T_k(s) = g$ if and only if the following three conditions hold:*

1. *For all e in E, $r(e) \geq 0$.*
2. *For all $v \in B$, the degree of v in the graph r is strictly smaller than $k - c_v$.*
3. *For all $e \in E$ such that $r(e) > 0$, one of the endpoints of e is not in B and has degree at least k in r.*

Proof. Let us first assume that $T_k(s) = g$. The graph g is obtained from s by removing every edge adjacent to a vertex of degree at least k in s. Therefore g is a subgraph of s, r denotes the removed edges, and Condition 1 holds. If the degree of v is at least k in s, then every edge adjacent to v is removed when building the graph $T_k(s)$. Therefore, if v is in B (*i.e.* if the degree of v is positive in g), the degree of v in s is strictly smaller than k. As v has degree c_v in g, v has degree strictly smaller than $k - c_v$ in r, proving Condition 2. If $r(e) > 0$, this means that the edge e is removed when building $T_k(s)$, *i.e.* that one of its endpoint v has degree at least k in s. As by Condition 2, v cannot be in B, this means that $c_v = 0$ and therefore that v has degree at least k in r, proving Condition 3.

Conversely, let us assume that r satisfies conditions 1-3. Then by Condition 1, g is a subgraph of $s = g + r$. The endpoints of the edges e of g have positive degree in g and therefore they belong to B. By Condition 2, their degree in s is strictly smaller than k, and therefore the edges of g are not removed when building $T_k(s)$. The edges of r are removed, as by Condition 3, one of their endpoints has degree at least k in r and therefore degree at least k in $s = g + r$.

□

Let $\mathcal{E}_c$ denote the set of graphs satisfying Conditions 1-3 of Lemma 3.10. It should be noted that $\mathcal{E}_c$ only depends on c, not on g. It should also be noted that for any $r \in \mathcal{E}_c$ and $g \in \mathscr{G}_c$, the set of edges of r and g are disjoint, as the endpoints of edges of g are in B, whereas at least one endpoint of the edges of r is not in B, by Condition 3.

Since $N_e(t)$ is a Poisson random variable with parameter $\lambda = \frac{t}{n}$, for any graph s:

$$\begin{aligned}\mathbb{P}(G_{n,t}^{\infty} = s) &= \prod_{e \in E} \mathbb{P}(N_e(t) = s(e)) \\ &= e^{-(n-1)t/2} \lambda^{\sum_e s(e)} \prod_{e \in E} \frac{1}{s(e)!}\end{aligned}$$

Therefore, for $g \in \mathscr{G}_c$,

$$\begin{aligned}\mathbb{P}(g_{n,t}^{k} = g) &= \sum_{s : T_k(s) = g} \mathbb{P}(G_{n,t}^{\infty} = s) \\ &= \sum_{r \in \mathcal{E}_c} \mathbb{P}(G_{n,t}^{\infty} = g + r) \\ &= e^{-(n-1)t/2} \sum_{r \in \mathcal{E}_c} \lambda^{\sum_e g(e)+r(e)} \prod_{e \in E} \frac{1}{(g(e)+r(e))!} \\ &= e^{-(n-1)t/2} \frac{\lambda^{\sum_e g(e)}}{\prod_{e \in E} g(e)!} \sum_{r \in \mathcal{E}_c} \lambda^{\sum_e r(e)} \prod_{e \in E} \frac{1}{r(e)!}.\end{aligned}$$

Note that for any edge e, $r(e)g(e) = 0$, hence $(g(e) + r(e))! = g(e)!r(e)!$ holds. Thus

$$\begin{aligned}\mathbb{P}(g_{n,t}^{k} = g) \prod_{i \geq 1} (i!)^{m_i} &= \mathbb{P}(g_{n,t}^{k} = g) \prod_{e \in E} g(e)! \\ &= e^{-(n-1)t/2} \lambda^{\sum_e g(e)} \Psi \\ &= e^{-(n-1)t/2} \lambda^{\sum_{i=1}^{n} c_i/2} \Psi,\end{aligned}$$

in which Ψ, defined by

$$\Psi = \sum_{r \in \mathcal{E}_c} \lambda^{\sum_e r(e)} \prod_{e \in E} \frac{1}{r(e)!}$$

depends only on c, as the set $\mathcal{E}_c$ does. □

In the rest of the section, t is fixed, so $G_{n,t}^{\infty}$ and $g_{n,t}^{k}$ are abridged in G_n^{∞} and g_n^k for readability. In [MR95], Molloy and Reed first study the configuration model associated to a degree sequence, and then extend the results to the configuration model conditioned on being simple (this conditional model is called random graph with given degree sequence). In our case, we will first study the configuration model associated to the degree sequence of g_n^k, using [MR95, Theorem 1] and then extend the result to g_n^k by mirroring the proof used for [MR95, Lemma 2]. For a vertex v, the degree of v in G_n^{∞} (resp. in g_n^k) is denoted $C_n(v)$ (resp. $c_n(v)$). Set

$$d_n(i) = \#\{v : c_n(v) = i\}.$$

Let G_n^{conf} be the configuration model associated to the degree sequence c_n. We shall see that the sequence d_n satisfies the assumptions of [MR95, Theorem 1], i.e. for each i, there exists a constant $p_{t,i}$ such that

$$\lim_n \frac{d_n}{n}(i) = p_{t,i}.$$

As the degree in the graph is bounded by $k-1$, the sequence d_n is sparse and well-behaved (with the vocabulary of [MR95]).

Let $\mathfrak{PGW}(t)$ denote a Galton-Watson tree whose offspring is Poisson distributed with parameter t. By [DM10, Proposition 2.6], for any given vertex v, the rooted graph (G_n^∞, v) converges in distribution to $\mathfrak{PGW}(t)$ when n tends to infinity. Since $c_n(v)$ depends only on the vertices of the ball with radius 2 in $(G_{n,t}^\infty, v)$ and on the edges between them, the asymptotic distribution of $c_n(v)$ can be read on the first two levels of $\mathfrak{PGW}(t)$. Two cases arise:

- if $C_n(v) \geq k$, the degree of v in G_n^∞ exceeds the threshold k, and $c_n(v) = 0$;
- if $C_n(v) < k$, an edge $\{v, w\}$ incident to v in G_n^∞ is erased in g_n^k if and only if $C_n(w)$ exceeds the threshold k.

Owing to [DM10, Proposition 2.6], $(C_n(v), C_n(w))$ weakly converges to $(C(v), C(w))$ such that $(C(v), C(w) - 1)$ are i.i.d. and Poisson distributed with parameter t, for both its offspring and its father v contribute to the degree of w in $\mathfrak{PGW}(t)$. Thus the degree of a neighbor of v is less than $k-1$ with an asymptotic probability:

$$\pi_k(t) = e^{-t} \sum_{i=0}^{k-2} \frac{t^i}{i!},$$

and the asymptotic distribution $c(v)$ of $c_n(v)$ is obtained by the following algorithm:

1. draw a Poisson random variable $C(v)$ with parameter t;
2. build a Poisson random variable Y with parameter $t\pi_k(t)$ by a thinning of $C(v)$ with parameter $\pi_k(t)$, so that the conditional distribution of Y given $C(v)$ is binomial with parameters $C(v)$ and $\pi_k(t)$;
3. set $c(v) = Y \mathbf{1}_{\mathbf{C(v) \leq k-1}}$.

Set

$$p_{t,i} = \mathbb{P}(c(v) = i) = \lim_n \mathbb{P}(c_n(v) = i).$$

Lemma 3.11. *For $0 \leq i < k$, $d_n(i)/n \xrightarrow[n\to\infty]{(P)} p_{t,i}$.*

Proof. Since $\mathbb{P}(c_n(1)=i) \xrightarrow[n\to\infty]{} p_{t,i}$, and the vertices play symmetric rôles,

$$\mathbb{E}\left[d_n(i)\right] = \sum_{i=1}^{n} \mathbb{P}(c_n(v)=i) = n\mathbb{P}(c_n(1)=i),$$

and the limit holds for expectations. Let us show that it holds in probability.

$$\begin{aligned}
\mathbb{E}\left[d_n(i)^2\right] &= \sum_{v,w\in[\![n]\!]} \mathbb{P}(c_n(v)=i=c_n(w)) \\
&= \sum_{v} \mathbb{P}(c_n(v)=i) + \sum_{v\neq w} \mathbb{P}(c_n(v)=i=c_n(w)) \\
&= n\mathbb{P}(c_n(1)=i) + n(n-1)\mathbb{P}(c_n(1)=i=c_n(2))
\end{aligned}$$

As the birooted local limit of $G_{n,t}^{\infty}$ is a couple of independent $\mathfrak{PGW}(t)$, $\mathbb{P}(c_n(1)=c_n(2)=i)$ converges to $p_{t,i}^2$, and:

$$\mathbb{E}\left[d_n(i)^2\right] = n^2 p_{t,i}^2 + o(n^2).$$

Thus $Var(d_n(i)) = o(n^2)$, leading to the desired result. □

Let

$$Q_t = \sum_i i(i-2)p_{t,i}.$$

According to [MR95, Theorem 1], if $Q_t > 0$, there exists a constant α

$$\mathbb{P}(C_{\max}(G_n^{conf}) \geq \alpha n | c_n) \xrightarrow[n\to\infty]{(P)} 1.$$

Moreover, the law of G_n^{conf} conditionally on c_n and on G_n^{conf} being loopless is equal to the law of g_t^k conditionally on c_n. By the main result of [McK85], there exists a constant $\beta > 0$ such that $\mathbb{P}(G_n^{conf}$ is loopless$|c_n) > \beta$ a.a.s, and therefore, if $Q_t > 0$:

$$\mathbb{P}(C_{\max}(g_n^k) \geq \alpha n | c_n) \xrightarrow[n\to\infty]{(P)} 1.$$

To conclude, we need to to compute the sign of

$$Q_t = \mathbb{E}\left[c(v)(c(v)-2)\right] = \mathbb{E}\left[c(v)(c(v)-1)\right] - \mathbb{E}\left[c(v)\right].$$

According to the points 2 and 3 of the description of the law of $c(v)$ above, $\mathbb{E}\left[c(v)|C(v)\right] = \pi_k(t)C(v)\mathbf{1}_{\mathbf{C(v)<k}}$ and $\mathbb{E}\left[c(v)\right] = \pi_k(t)\mathbb{E}\left[C(v)\mathbf{1}_{\mathbf{C(v)<k}}\right]$. Similarly, $\mathbb{E}\left[c(v)(c(v)-1)\right] = \pi_k(t)^2\mathbb{E}\left[C(v)(C(v)-1)\mathbf{1}_{\mathbf{C(v)<k}}\right]$. Moreover, we

have:

$$\begin{aligned}
\mathbb{E}(C(v)\mathbf{1}_{\mathbf{C(v)<k}}) &= e^{-t}\sum_{i=1}^{k-1} i\frac{t^i}{i!} \\
&= e^{-t}\sum_{i=1}^{k-1} \frac{t^i}{(i-1)!} \\
&= e^{-t}\sum_{i=0}^{k-2} \frac{t^{i+1}}{i!} \\
&= t\pi_k(t)
\end{aligned}$$

Thus the equation for the existence of a giant component become successively:

$$\begin{aligned}
\pi_k(t)^2\mathbb{E}(C(v)(C(v)-1)\mathbf{1}_{\mathbf{C(v)<k}}) &> \pi_k(t)\mathbb{E}(C(v)\mathbf{1}_{\mathbf{C(v)<k}}) \\
\pi_k(t)^2\mathbb{E}(C(v)(C(v)-1)\mathbf{1}_{\mathbf{C(v)<k}}) &> t\pi_k(t)^2 \\
e^{-t}\sum_{i=2}^{k-1} i(i-1)\frac{t^i}{i!} &> t \\
te^{-t}\sum_{i=0}^{k-3} \frac{t^i}{i!} &> 1
\end{aligned}$$

- If $k \geq 5$, there exists a t such that the condition is satisfied. (e.g. $t = 2$).
- If $k \leq 4$, the condition does not hold for any t.

This ends the proof of Theorem 3.2.

Remark. This part does not prove the existence (nor the non-existence) of a giant component for $k = 4$, and only proves that there is a giant component for some bounded interval of t (and not, as it could be expected, for every t larger than some t_0) for $k \geq 5$.

3.4 The local limit

The aim of this part is to prove the existence of a local limit, and to study the link between this local limit and the existence of a giant component.

3.4.1 Local topology

For the purpose of this study, the objects used are multigraphs with labelled edges, where the edges are labelled by the time of addition. A root of a graph G is either a vertex or an edge of G. We consider the local topology, as introduced by Benjamini and Schramm[BS01].

Definition 3.2. Given a graph G, an edge $e = (v_1, v_2)$ of G, a vertex v of G, and a non-negative integer l, $B_l(G, v)$ denotes the set of vertices of the ball of radius l centered at v in G, and $B_l(G, e)$ denotes $B_l(G, v_1) \cup B_l(G, v_2)$. The notation $B_l(G, v)$ (resp. $B_l(G, e)$) will also denote the subgraph of G induced by the set of vertices $B_l(G, v)$ (resp. $B_l(G, e)$), and $S_l(G, v)$ (resp. $S_l(G, e)$) will denote the sphere of radius l centered at v (resp. e) *i.e.* the set of vertices $B_l(G, v) \setminus B_{l-1}(G, v)$ (resp. $B_l(G, e) \setminus B_{l-1}(G, e)$), where by convention $B_{-1} = \emptyset$. The ball (resp. sphere) of radius l will be called l-ball (resp. l-sphere).

Definition 3.3. For an integer l, a l-rooted multigraph (G, r) with labelled edges is the data of a multigraph G with labelled edges equipped with a l-tuple $r = (r_1, \ldots, r_l)$ of roots of G: the set of roots is ordered, and a given root can appear several times.

For an integer j and an l-rooted graph (G, r), $B_j(G, (r))$ denotes the l-rooted graph $\cup_{1 \leq i \leq l} B_j(G, r_i)$.

Definition 3.4 (Isomorphism and distance)**.**

- Two l-rooted multigraphs (G, r) and $(\tilde{G}, \tilde{r})$ are said to be isomorphic if there is a graph isomorphism Φ from G to $\tilde{G}$ preserving the edges' labels such that for all i, $\Phi(r_i) = \tilde{r}_i$.

- For any integer j, (G, r) and $(\tilde{G}, \tilde{r})$ are said to be j-isomorphic if their j-balls are isomorphic.

- The pseudo-distance between two l-rooted graphs G and $\tilde{G}$ is defined as 2^{-j}, with j the largest integer such that G and $\tilde{G}$ are j-isomorphic (as l-rooted labelled multigraphs). The convergence according to this distance is called the local convergence (this is a straightforward extension of the definition of the local convergence, according to Benjamini *et al*).

If (G, r) and $(\tilde{G}, \tilde{r})$ are isomorphic, then they are j-isomorphic for any j. The converse holds if there is at least one root in each connected component of each graph, but can be false otherwise.

Definition 3.5. Let T_t^∞ be the labelled random graph defined by:

- The shape of T_t^∞ is a Galton-Watson tree with Poisson offspring with mean t.

- Conditionally on the shape of T_t^∞, the edges' labels are uniform on $[0, t]$ and independent.

Lemma 3.12. *For any integer l, the* unlabelled *graph $G_{n,t}^\infty$, rooted at l vertices chosen independently uniformly at random, converges toward*

l independent copies of a Galton-Watson tree with Poisson offspring with mean t, while the labelled *graph $G_{n,t}^{\infty}$, rooted at l vertices chosen independently uniformly at random, converges toward l independent copies of T_t^{∞}.*

The one-rooted version of the first part of Lemma 3.12 is a well-known result, for various versions of the Erdős-Rényi graph. One instance can be found in [DM10, Proposition 2.6], for $G_{n,t}^{\infty}$ with no multiple edges. For completeness, a proof of this multi-roots, multigraph version of the result in [DM10] can be found below:

Proof. For l and j some positive integers, consider (G, r), $r = (r_i)_{1\leq i}$, a l-rooted finite graph with radius not larger than j (such that $\cup_{1\leq i\leq l} B_j(G, r_i) = G$). Let us compute $p_n(G)$, the probability that $B_j(G_{n,t}^{\infty}, v)$ is isomorphic to (G, r), in which $(v_i)_{1\leq i\leq l}$ is a sequence of l vertices chosen independently at random. We need a few notations:

- V_G is the vertex set of G, and $N_V = |V_G|$;
- E_G is the edge set of G and $N_E = |E_G|$;
- for any edge $e \in E_G$, let m_e denote the multiplicity of e in G.
- N_V^b is the number of vertices of G that are in $B_{j-1}(G, r)$;
- $(w_i)_{1\leq i\leq N_V}$ is a specific ordering of V_G, such that $(w_i)_{1\leq i\leq N_V^b}$ are the elements of $B_{j-1}(G, r)$, and $(w_i)_{N_V^b+1\leq i\leq N_V}$ are the remaining vertices of G.

There are at most l components of $G = \cup_{1\leq i\leq l} B_j(G, r_i)$, therefore $l + N_E \geq N_V$. If $l + N_E = N_V$, then G is a forest of l simple trees.

Lemma 3.13. *If $l + N_E > N_V$, then $p_n(G) \to 0$.*

Proof of Lemma 3.13. Let Φ be an injection from V_G to $\{1, \ldots, n\}$. If Φ induces an isomorphism between (G, r) and $B_j(G_{n,t}^{\infty}, v)$, then the following conditions holds:

1. For every $1 \leq i \leq l$, $v_i = \Phi(r_i)$.
2. For every edge (w_i, w_j) of G, there is an edge, with the same multiplicity, between $\Phi(w_i)$ et $\Phi(w_j)$, in $G_{n,t}^{\infty}$.

These conditions are only necessary, not sufficient. For any given Φ, the probability of the first condition is $\frac{1}{n^l}$. The probability of the second condition, being of the following form:

$$\prod_{e\in E_G} \frac{t^{m_e}}{n^{m^e} m^e!}$$

is smaller than $(\frac{t}{n})^{N_E}$. Therefore, by independence between $G_{n,t}^{\infty}$ and v, the probability that Φ induces an isomorphism between (G,r) and $G(_{n,t}^{\infty},v)$ is smaller than $\frac{t^{N_E}}{n^{l+N_E}}$. By union bounds, as there are less than n^{N_V} injections between $(w_i)_{1\leq i\leq N_V}$ to $\{1,\dots,n\}$, we obtain:

$$p_n(G)\leq t^{N_E}n^{N_V-l-N_E}$$

proving Lemma 3.13. □

A Galton-Watson is usually described as a genealogy tree, *i.e.* the planer embedding of a rooted tree, or "*plane tree*": a plane tree is a rooted tree with an order relation between the children of the same node (see [FS09][Annex A.9]). An isomorphism of plane trees (resp. of a sequence of plane trees) is an isomorphism preserving the root (resp. the sequence of roots, with order) and the children's order. Assuming that $B_j(G_{n,t}^{\infty},v)$ is a forest with l components, let $B_j^*(G_{n,t}^{\infty},v)$ be the plane representation of $B_j(G_{n,t}^{\infty},v)$, where the children are ordered according to their original label, in $\{1,\dots,n\}$. The rooted graph $B_j(G_{n,t}^{\infty},v)$ and G are isomorphic as rooted graphs, if and only if the plane representation of $B_j(G_{n,t}^{\infty},v)$ is isomorphic to one of the plane representations of (G,r), as a plane tree or forest. Therefore

$$\mathbb{P}(B_j(G_{n,t}^{\infty},v)\sim(G,r))=\sum_{(G^*,r)}\mathbb{P}(B_j^*(G_{n,t}^{\infty},v)\underset{plane}{\sim}(G^*,r))$$

where the summation is taken over the plane representations of (G,r).

Let us compute the number of injections Φ from V_G to $\{1,\dots,n\}$ preserving the order among the children of each node of G. First we choose the N_V elements of $\Phi(V_G)$, without ordering (there are $\binom{n}{N_V}$ possible choices). From there, choosing Φ is equivalent to choosing a plane forest isomorphic to G^* on N_V vertices. The number of such plane forests is:

$$\frac{N_V!}{\prod_{i=1}^{N_V^b}d_{w_i}^{out}!}$$

where $d_{w_i}^{out}$ is the outdegree of w_i, *i.e.* the number of children of w_i (cf. [Spe97][p.2], [CM01][p.5] or [CF03][p.12]). Therefore the number of injections Φ from V_G to $\{1,\dots,n\}$ preserving the order among the children of each node of G is

$$\frac{n!}{(n-N_V)!\prod_{i=1}^{N_V^b}d_{w_i}^{out}!}. \tag{3.3}$$

It should be noted that

$$N_E=\sum_{i=1}^{N_V^b}d_{w_i}^{out} \tag{3.4}$$

Such an injection Φ induces an isomorphism of plane forests between (G^*, r) and $(B_j^*(G_{n,t}^\infty, v)$ if and only if:

1. For each i, the random uniform root v_i is equal to $\Phi(r_i)$;
2. For any $1 \leq i \leq N_V^b$ and $1 \leq j \leq N_V$, there is the same number of edges (either 0 or 1) between w_i and w_j in G and between $\Phi(w_i)$ and $\Phi(w_j)$ in $G_{n,t}^\infty$;
3. For any $i \leq N_V^b$ and vertex $j \in \{1, \ldots, n\} \setminus \Phi(V_G)$, there is no edge between $\Phi(w_i)$ and j in $G_{n,t}^\infty$.

For a given Φ, the probability of the first property is $\frac{1}{n^l}$. As the multiplicity of each edge of $G_{n,t}^\infty$ is a Poisson random variable of parameter $\frac{t}{n}$, the probability of the last two properties is:

$$\left(e^{-\frac{t}{n}}\frac{(\frac{t}{n})^1}{1!}\right)^{N_E}\left(e^{-\frac{t}{n}}\frac{(\frac{t}{n})^0}{0!}\right)^{(n-1)N_V^b-N_E} = e^{-t\frac{n-1}{n}N_V^b}\left(\frac{t}{n}\right)^{N_E}. \tag{3.5}$$

Combining (3.3), (3.4), (3.5), the fact that the probability of the first property is $\frac{1}{n^l}$ and $N_v = l + N_e$, we obtain:

$$\begin{aligned}
\mathbb{P}\left(B_j^*(G_{n,t}^\infty, v) \underset{plane}{\sim} (G^*, r)\right) &= \frac{n!}{(n-N_V)!\prod_{i=1}^{N_V^b} d_{w_i}^{out}!} e^{-t\frac{n-1}{n}N_V^b}\left(\frac{t}{n}\right)^{N_E} \\
&\xrightarrow[n\to\infty]{} \frac{e^{-tN_V^b}t^{N_E}}{\prod_{i=1}^{N_V^b} d_{w_i}^{out}!} \\
&= \prod_{i=1}^{N_V^b}\left(e^{-t}\frac{t^{d_{w_i}^{out}}}{d_{w_i}^{out}!}\right) \\
&= \mathbb{P}(GW_j^l \underset{plane}{\sim} (G^*, r))
\end{aligned}$$

where GW_j^l denotes l independent copies of a Galton-Watson tree with Poisson offspring with mean t. By summing over all the plane representations of G, one obtains that:

$$\mathbb{P}\left(B_j(G_{n,t}^\infty, v) \sim (G, r)\right) \to \mathbb{P}\left(GW_j^l \sim (G, r)\right)$$

for all G and j, ending the proof of the first part of Lemma 3.12.

By the properties of Poisson point processes, conditionally given its unlabelled version, the labels of the edges of $G_{n,t}^\infty$ are independent and uniform on $[0, t]$. Therefore, the labelled graph $G_{n,t}^\infty$ l-rooted at l independent uniform random roots converges weakly toward l independent copies of T_t^∞.

□

By the standard properties of Poisson processes, T_t^∞ can be described as a branching tree such that the law of the set of labels of outgoing edges is a Poisson point process of intensity 1 on $[0, t]$.

Let T_∞^∞ denote the *Poisson-Weighted Infinite Tree* in dimension 1, or *PWIT*, as defined by Aldous and Steele [AS04]. The distribution of T_t^∞ can also be described as follows:

- Remove every edge of T_∞^∞ whose label is larger than t.
- Let T_t^∞ be the connected component of the root in the resulting subgraph.

It should be noted that the notion of convergence used here is not exactly the same as the one used in [AS04], as the latter allows the edges' labels to converge toward their limit, whereas the former requires the edges' labels to be eventually constant.

3.4.2 The forbidden degree version of infinite graphs.

3.4.2.a Neighborhood with boundaries

Definition 3.6. Let Ω be the set of locally finite graphs with labelled edges such that no two edges have the same label. Let $\Omega_{<\infty}$ be the set of finite graphs in Ω. For any non-negative t and graph G , let G_t be the subgraph of G restricted to its edges with label smaller than t.

For any graph $G \in \Omega_{<\infty}$, the forbidden degree version $G^k = \Phi(G)$ of G is defined by adding the edges of G in their labelling order and by removing the edges adjacent to any vertex reaching degree k. If $G \in \Omega \setminus \Omega_{<\infty}$, the labels are not necessarily well-ordered, so some care is needed to extend Φ to a larger class of graphs of Ω. This is the aim of this section.

Definition 3.7. Let e be an edge of G. (G_B, B, S) is called a *neighborhood with boundary* of e in G if:

- S and B are subset of the set of vertices of G.
- G_B is the induced subgraph of G restricted to B.
- The edge e is in G_B.
- S is a subset of B.
- S contains the boundary of B in G (*i.e.* the sets of vertices of B with a neighbor not in B in the graph G)

Remark. The last condition means that elements of $B \setminus S$ in G are never endpoints of edges between B and $G \setminus G_B$, while vertices in S may be so. Sometimes we shorten (G_B, B, S) in (G_B, S) as B can be retrieved from G_B.

Example. For any graph G, integer l, and vertex or edge x of G, $(B_l(G,x), S_l(G,x))$ is a neighborhood with boundary.

Definition 3.8. Let e be an edge of G, $\mathcal{V} = (G_B, B, S)$ a neighborhood with boundary of e in G. Let us consider the set $\mathcal{S}_\mathcal{V}$ of finite graphs $\tilde{G}$ with an edge $\tilde{e} \in \tilde{G}$ and a neighborhood with boundary $(\tilde{G}_{\tilde{B}}, \tilde{B}, \tilde{S})$ of $\tilde{e}$ in $\tilde{G}$ such that (G_B, B, S, e) is isomorphic to $(\tilde{G}_{\tilde{B}}, \tilde{B}, \tilde{S}, \tilde{e})$. If either,

1. $\tilde{e}$ is present in $\Phi(\tilde{G})$ for every $\tilde{G} \in \mathcal{S}_\mathcal{V}$,
2. or $\tilde{e}$ is removed in $\Phi(\tilde{G})$ for every $\tilde{G} \in \mathcal{S}_\mathcal{V}$,

then we say that the knowledge of $\mathcal{V}$ is sufficient to know whether e is removed in the forbidden degree version of G.

The set $\mathcal{S}_\mathcal{V}$ depends both on (G_B, B) and on S: the subgraph of $\tilde{G}$ induced by $\tilde{B}$ is isomorphic to G_B, but one can obtain $\tilde{G}$ only by growing new edges on the vertices of $\tilde{S}$. That is, only the vertices of $\tilde{S}$ may have a different degree in $\tilde{G}_{\tilde{B}}$ and in $\tilde{G}$. Thus S matters in the definition above, in a somehow hidden way. If finally we are able to extend Φ to some infinite graph G, we expect, obviously, that the status of e in $\Phi(G)$ (present or absent) is the same as the status of its counterpart $\tilde{e}$ in $\Phi(\tilde{G})$.

3.4.2.b Locality of the forbidden version

For each edge e of G, let $l(e)$ denote the smallest integer l such that, for all t, the knowledge of $(B_l(G,e), S_l(G,e))$ is sufficient to ascertain whether e is removed in the forbidden degree version of G_t. If there is no such l, let $l(e) = \infty$.

For any vertex v of G, let $l(v)$ be the smallest integer such that the knowledge of $(B_l(G,v), S_l(G,v))$ is sufficient to know which edges adjacent to v are removed in the forbidden degree version of G_t for all t.

In this section, we shall prove that for any $T \geq 0$, a.s. $l(v) < \infty$ (resp. $l(e) < \infty$) for any vertex v (resp. any edge e) of T_T^∞.

Definition 3.9. Let $\Omega^+ \subset \Omega$ be the set of labelled graphs G such that for all edges e of G, $l(e) < \infty$.

For any graph $G \in \Omega^+$, $\Phi(G)$ is defined as follows:

- the set of vertices of $\Phi(G)$ is V;
- the set of edges of $\Phi(G)$ is a subset of E;
- for each $e \in E$, e is an edge of $\Phi(G)$ if and only if e is present in $\Phi(B_{l(e)}(G,e))$.

The main result of this part is the following lemma:

Lemma 3.14. *Almost surely, for all $t \geq 0$, $T_t^\infty \in \Omega^+$.*

Lemma 3.14 will be proven in part 3.4.2.c. By Lemma 3.14, $\Phi(T_t^\infty)$ is well defined, and, as a subgraph of the tree T_t^∞, $\Phi(T_t^\infty)$ is a forest. Let T_t^k be the connected component of the root $\emptyset$ in $\Phi(T_t^\infty)$.

Corollary 3.15. *For any i, $G_{n,t}^k$ rooted at m independent uniform vertices locally converges in distribution toward m independent copies of $(T_t^k, \emptyset)$.*

Proof. By Lemma 3.12, $G_{n,t}^\infty$ rooted at m random uniform vertices $(v_i)_{1\leq i\leq m}$ locally converges in distribution toward $\mathscr{T} := \cup_{1\leq i\leq m}(T_i, r_i)_{1\leq i\leq m}$, where (T_i, r_i) are m independent copies of $(T_t^\infty, \emptyset)$.

Lemma 3.16. *For any positive integer m, let Ω_m^+ be the set of m-rooted elements of Ω^+. Then the application*

$$\begin{array}{rcl} \Omega_m^+ & \to & \Omega_m^+ \\ (G, (r_i)_{1\leq i\leq m}) & \to & (\Phi(G), (r_i)_{1\leq i\leq m}) \end{array}$$

is continuous for the local topology.

Lemma 3.16 entails that $(G_{n,t}^k, v)$ locally converges toward $(\Phi(\mathscr{T}), r)$, as $G_{n,t}^k = \Phi(G_{n,t}^\infty)$. As $\Phi(\mathscr{T}) = \cup_i(\Phi(T_i))$, and T_t^k is the connected component of the root in $\Phi(T_t^\infty)$, $(G_{n,t}^k, v)$ locally converges toward m independent copies of T_t^k. □

Proof of Lemma 3.16. Let $(G^n, (r_i^n)_{1\leq i\leq m})$ be a sequence of n-rooted graphs converging toward $(H, (s_i)_{1\leq i\leq m}) \in \Omega_m^+$ for the local topology and let j denote a positive integer. As $H \in \Omega^+$, $\cup_i B_j(H, s_i)$ is finite, and $l(e)$ is finite for every edge e of H. Therefore, $L = \max_{e\in\cup_i B_j(H,s_i)} l(e) < \infty$. By definition of the local convergence, for n large enough, there exists an isomorphism Ψ_n between $\cup_i B_{j+L+1}(G^n, r_i^n)$ and $\cup_i B_{j+L+1}(H, s_i)$. By definition of L and $l(e)$, every edge e of $\cup_i B_j(G^n, r_i^n)$ is present in $\Phi(G^n)$ if and only if $\Psi_n(e)$ is present in $\Phi(H)$. Therefore Ψ_n induces an isomorphism between $\cup_i B_j(\Phi(G^n), r_i^n)$ and $\cup_i B_j(\Phi(H), s_i)$. As this construction works for any integer j, $(\Phi(G^n), (r_i^n)_{1\leq i\leq m})$ locally converges toward $(\Phi(H), (s_i)_{1\leq i\leq m})$. □

3.4.2.c Propagation paths

Definition 3.10. For any element G of Ω, a *propagation path* of length $l \in \mathbb{N}$ is the data of a self-avoiding path $(v_0, e_1, v_1, e_2, \ldots, e_l, v_l)$ of length l and a sequence of edges $(\tilde{e}_1, \tilde{e}_2, \ldots, \tilde{e}_{l-1})$ of length $l-1$ in G such that

- for all $1 \leq i \leq l-1$, $\tilde{e}_i$ is adjacent to v_i;
- for all $i \neq i'$, $\tilde{e}_i \neq e_{i'}$;
- the labels of the sequence $(\tilde{e}_i)$ are decreasing.

For any vertex v of G, let $l'(v) \in \mathbb{N} \cup \{\infty\}$ be the supremum of the lengths of the propagation paths starting at v, *i.e.* such that $v = v_0$.

Lemma 3.17. *Let $G \in \Omega$. For every vertex v in G:*

$$l(v) \leq l'(v) + 1.$$

Lemma 3.17 has a useful consequence:

Corollary 3.18. *Let Ω^- be the set of graphs $G \in \Omega$ such that for all vertex $v \in G$, $l'(v) < \infty$. Then*

$$\Omega^- \subset \Omega^+.$$

Lemma 3.14 is a consequence of Corollary 3.18 and the following lemma:

Lemma 3.19. *For all t, a.s. $T_t^\infty \in \Omega^-$.*

The rest of this subsection is devoted to the proof of Lemmata 3.17 and 3.19.

Proof of Lemma 3.17. Let G be a graph in Ω, e^* be an edge of G and l a positive integer. In order to simplify the notations, B will denote $B_l(G, e^*)$ and S will denote $S_l(G, e^*)$.

Definition 3.11. An edge e of B will be said **certain at time t** if, for all $t' \leq t$, the knowledge of (B, S) allows one to determine if e is removed in the forbidden degree version of $G_{t'}$ and **uncertain at time** t otherwise.

The edges in $G \setminus B$ are said uncertain at any time.

By definition, $l(e^*) > l$ if and only if e^* is uncertain at some time t. G is a locally finite graph, therefore B is a finite graph. Let J denote the finite set of labels of B. The application $t \to B_t$ is a right-continuous piecewise constant function jumping only at elements of J.

For every edge $e \in B$, let y_e denote the infimum of the times where e is uncertain, and t_e denote the label of e.

Lemma 3.20. *For every edge $e \in B \setminus S$, there exists:*

- *an edge e' adjacent to one endpoint v of e such that $t_{e'} = y_e$;*
- *an edge $e'' \neq e'$ adjacent to v such that $y_{e''} < y_e$.*

Proof. Let e be an edge in $B \setminus S$. y_e corresponds to a jump of $t \to B_t$, therefore there exists an edge e' such that $t_{e'} = y_e$. As e is certain at time y_e^- and uncertain at time y_e, then,, $\tilde{e}$ is removed at time y_e in some graph $\tilde{G}$, and it is not in some other graph $\tilde{G}$ (according to the notations of Definition 3.8). Let e' denote the only edge with label y_e (e' is unique by definition of Ω). Only the endpoints of e' can reach the forbidden degree

at time y_e, and only edges adjacent to an endpoint of e' can be removed at time y_e thus e' and e share a common endpoint v. Moreover, in order for the removal of e to be uncertain, the degree of v must be uncertain before time $y_e = t_{e'}$ in the forbidden degree version of G.[2], *i.e.* one of the edge adjacent to v is uncertain before time y_e. Let e'' denote such an edge. As an edge cannot be uncertain before its label, $e'' \neq e'$. □

Lemma 3.21 (Necessary condition for the uncertainty to spread)**.** *Given an edge $e \in B$ such that $y_e < \infty$, there exists a self-avoiding path $e_1, v_1, e_2, \dots e_p, v_p$ in B with $e_1 = e$ and a sequence of edges $(\tilde{e}_i)_{1 \leq i \leq p-1}$ in B such that:*

1. *for every $i \in \{1, p-1\}$, $\tilde{e}_i$ is adjacent to v_i;*
2. *for every $i \neq j$, $\tilde{e}_i \neq e_j$;*
3. *$t_{\tilde{e}_1} > t_{\tilde{e}_2} \cdots > t_{\tilde{e}_{p-1}}$;*
4. *the vertex v_p is in S.*

Proof. Let Ξ denote the set of integers p such that there exists a path $e = e_1, v_1, e_2, \dots e_p, v_p$ in B and a sequence of edges $(\tilde{e}_i)_{1 \leq i \leq p-1}$ such that:

1. for every $i \in \{1, p-1\}$, $\tilde{e}_i$ is adjacent to v_i;

2'. for every i, $\tilde{e}_i \neq e_{i+1}$;

3a. $y_{e_1} > y_{e_2} \cdots > y_{e_{p-1}}$;

3b. for every $i \in \{1, p-1\}$, $t_{\tilde{e}_i} = y_{e_i}$.

Such a path trivially exists for $p = 1$. The edges e_i in the path are distinct by condition 3a, and the graph B is finite, thus Ξ is bounded from above. Let $p_{\max} = \max \Xi$. By Lemma 3.20, if $e_{p_{\max}} \in B \setminus S$, the path can be extended one step further, by taking $e = e_p$, $\tilde{e}_p = e'$ and $e_{p+1} = e''$, and therefore p is not maximal. This entails that for any maximal path, either $v_{p_{\max}}$ or $v_{p_{\max}-1}$ is in S, and therefore there exists a path satisfying conditions 1, 2', 3a, 3b and 4 (and condition 3, as a consequence of 3a and 3b). The shortest path satisfying conditions 1, 2', 3a, 3b and 4 is self-avoiding (otherwise it would not be the shortest). If this path does not satisfy condition 2, while satisfying condition 2', there exists (i, j) such that $j \notin \{i, i+1\}$ and $\tilde{e}_i = e_j$. In that case, e_j is incident to v_i, and the path is not self-avoiding. □

As a consequence of Lemma 3.21, if $l(e^*) > l$, then e^* is uncertain at some time t, and there exists a propagation path of length l starting at an endpoint of e^*, proving Lemma 3.17.

□

[2] it is possible that e and e' share both their endpoints, e.g. if $e = e'$. In that case, the degree of at least one of their endpoints must be uncertain. Let v denote such an endpoint.

Proof of Lemma 3.19. We will now prove that for all $t \geq 0$, $T_t^\infty \in \Omega^-$ a.s. If T_t^∞ contains only the root, $T_t^\infty \in \Omega^-$. With positive probability, the root has at least one child. Let $\mathscr{T}$ be the tree T_t^∞ conditioned on having at least one edge, and let v be one children of the root of $\mathscr{T}$ chosen uniformly at random. The law of $\mathscr{T}$ rerooted at v is absolutely continuous with respect to the law of $\mathscr{T}$. Therefore, it is sufficient to prove that almost surely $l'(\emptyset) < \infty$ in $\mathscr{T}$, and therefore in T_t^∞.

Definition 3.12. A possible propagation path of length l is the data of:

- a self-avoiding path $\emptyset = v_0, e_1, v_1, e_2, v_2, \dots e_l, v_l$ in B_l from $\emptyset$ to a vertex v_l of S_l.
- a sequence of edges $(\tilde{e}_i)_{1 \leq i \leq l-1}$ in B_l such that:
  - for every $i \in \{1, l-1\}$, $\tilde{e}_i$ is adjacent to v_i;
  - for every $i \neq i'$, $\tilde{e}_i \neq e_{i'}$.

Let H_l be the set of such possible propagation paths.

A possible propagation path is a propagation path if the labels of $\tilde{e}$ are decreasing. Let $p(l)$ denote the probability that there exists a propagation path of length l starting at the root of T_t^∞. The set H_l only depends on the unlabelled version of the tree T_t^∞. Conditionally on the unlabelled version of the tree T_t^∞, the labels of the edges are i.i.d. uniform on $[0, t]$, thus the conditional probability that a given possible propagation path is actually a propagation path is $\frac{1}{(l-1)!}$. Therefore, by union bound, $p(l) \leq \frac{\mathbb{E}(|H_l|)}{(l-1)!}$.

The following lemma gives an explicit formula for the expected size of H_l:

Lemma 3.22. *Let X_t be a Poisson random variable of parameter t. Then:*

$$\mathbb{E}(|H_l|) = \mathbb{E}(X_t)\mathbb{E}(X_t^2)^{l-1} = t(t+t^2)^{l-1}$$

Proof. The proof is done by induction on l.

H_1 is the set of self-avoiding paths of length 1 starting from the root, equal to the degree of the root, distributed as X_t.

The set H_{l+1} can be constructed from H_l in the following way:

- Take an element of H_l and denote its endpoint by v_l.
- Take a child v_{l+1} of v_l and extend the self-avoiding path to v_{l+1}.
- Choose w a neighbor of v_l different from v_{l+1} (either a child of v_l, or v_{l-1}). Let $\tilde{e}_l$ be the edge between v_l and w.

This construction gives every element of H_{l+1} once. Therefore an element of H_l gives d^2 elements of H_{l+1} with d the number of children of v_l. By the branching property of T_t^∞, this implies that $\mathbb{E}(|H_{l+1}|) = \mathbb{E}(X_t^2)\mathbb{E}(|H_l|)$. □

By a quick computation:

$$p(l) \leq \frac{\mathbb{E}(|H_l|)}{(l-1)!} = \frac{t(t+t^2)^{l-1}}{(l-1)!} \xrightarrow[l\to\infty]{} 0$$

ending the proof of Lemma 3.19.

□

Lemma 3.14 allows us to study the local limit, and therefore the convergence of l balls of radius i, with i and l fixed. In Section 3.5, more precise results will be needed, in which i and l can depend on n. For this reason, the following lemma will be useful:

Lemma 3.23. *There exist two sequences $b_n = o(\ln n)$ and $c_n = o(1)$ such that, asymptotically almost surely:*

1. *For every v in $G^\infty_{n,t}$, $l(v) \leq b_n$.*
2. *For every v in $G^\infty_{n,t}$, the ball of radius b_n centered at v in $G^\infty_{n,t}$ contains at most x_n vertices, where $x_n = n^{c_n}$.*

Definition 3.13 (Good event characteristic functions)**.** To avoid dealing with problematic events of vanishing probability, we will introduce several characteristic functions, denoted $GE_1, GE_2, \ldots$ (GE stands for *"Good Event"*) such that GE_i is a Bernoulli variable of parameter tending to 1. Let GE_1 be the characteristic function of the events of Lemma 3.23.

Proof. By Lemma 3.17, proving that a.a.s. there is no propagation path of length b_n in $G^\infty_{n,t}$ implies the first part of Lemma 3.14. The total number of edges in $G^\infty_{n,t}$ is a Poisson variable of parameter $\binom{n}{2}\frac{t}{n} = \frac{t(n-1)}{2}$ and is smaller than tn with high probability. As a consequence, $G^\infty_{n,t}$ is a.a.s. a subgraph of $G^\infty_{tn} = G^\infty_{n,\tau_{\lfloor nt \rfloor}}$, the graph $G^\infty_{n,\cdot}$ stopped the first time $\tau_{\lfloor tn \rfloor}$ there are $\lfloor tn \rfloor$ edges in $G^\infty_{n,\cdot}$, thus it is sufficient to prove the absence of propagation path of length b_n in G^∞_{nt} instead of $G^\infty_{n,t}$, and for this proof we shall rely on combinatorial arguments.

To detect a propagation path, we do not need the label of each edge, only its rank among the $\lfloor nt \rfloor$ edges of G^∞_{nt}. It is convenient to see G^∞_{nt} as the result of a random allocation of $\lfloor nt \rfloor$ balls in $\frac{n(n-1)}{2}$ urns (the edge $r \leq \lfloor nt \rfloor$ being incident to 2 random vertices $\{v(r), \tilde{v}(r)\}$).

Remark (Alternative description of propagation paths)*.* A propagation path of length l can be described by a sequence of vertices $(v_0, v_1, \ldots, v_l)$ without repetition together with a l-uple without repetition $(r_1, \ldots, r_l)$ of integers smaller than tn and a strictly decreasing $l-1$-uple $(\tilde{r}_1, \ldots, \tilde{r}_{l-1})$ of integers smaller than tn such that:

(C1) For every $i \in \{1, \ldots, l\}$, $\{v(r_i), \tilde{v}(r_i)\} = \{v_{i-1}, v_i\}$.

(C2) For every $i \in \{1, \dots, l-1\}$, $v_i \in \{v(\tilde{r}_i), \tilde{v}(\tilde{r}_i)\}$.

(C3) If $j \neq j'$, then $\tilde{r}_j \neq r_{j'}$.

This description is equivalent to the description in Definition 3.10, where the edge e_i (resp. $\tilde{e}_i$) in Definition 3.10 corresponds to the edge r_i (resp. $\tilde{r}_i$) in this remark.

The collision set of a propagation path is the set of integers i such that $r_i = \tilde{r}_i$. The collision set is a subset of $\{1, \dots, l-1\}$.

Definition 3.14. A potential propagation path with collision set S is the data of a l-uple without repetition $(r_1, \dots, r_l)$ of integers smaller than tn and a strictly decreasing $l-1$-uple $(\tilde{r}_1, \dots, \tilde{r}_{l-1})$ of integers smaller than tn such that:

- $r_i = \tilde{r}_i \Leftrightarrow i \in S$.
- If $i \neq j$, then $\tilde{r}_i \neq r_j$.

A potential propagation path is only two sequences of edges (denoted by their corresponding integer), without any constraint on their endpoints: the sequence of edges described by $(r_i)_{1 \leq i \leq l}$ does not need to constitute a path nor does the edges r_i and $\tilde{r}_i$ need to share a common endpoint. A potential propagation path will be a propagation path if and only if there exists a sequence of vertices $(v_0, v_1, \dots, v_l)$ without repetition such that conditions (C1) and (C2) are satisfied.

The number of potential propagation paths with collision set S is:

$$\binom{\lfloor tn \rfloor}{l-1} \frac{(\lfloor tn \rfloor - l)!}{(\lfloor tn \rfloor - 2l + |S|)!} \leq \frac{(tn)^{2l-1-|S|}}{(l-1)!}$$

This formula is obtained by choosing first $(\tilde{r}_i)_{1 \leq i \leq l-1}$, and then choosing the $l - |S|$ elements of $(r_i)_{1 \leq i \leq l}$ that are still unknown.

Given two vertices v_1 and v_2 and an integer r, $\mathbb{P}(\{v(r), \tilde{v}(r)\} = \{v_1, v_2\}) = \binom{n}{2}^{-1}$, and $\mathbb{P}(v_1 \in \{v(r), \tilde{v}(r)\}) = \frac{2}{n}$. Therefore, given a potential propagation path $((r_i)_i, (\tilde{r}_i)_i)$ and a sequence of vertices $(v_0, v_1, \dots, v_l)$ without repetition, the probability that conditions (C1) and (C2) hold is:

$$\binom{n}{2}^{-l} \left(\frac{2}{n}\right)^{l-1-|S|} = (n-1)^{-l} \left(\frac{2}{n}\right)^{2l-1-|S|}.$$

There are less than $n(n-1)^l$ possible choices for $(v_i)_{0 \leq i \leq n}$, therefore, by union bound, the probability p_l that there exists a propagation path of

length b_n in G_{tn}^{∞} is bounded by:

$$\begin{aligned} p_l &\leq \sum_{S\subset\{1,\dots,l-1\}} \frac{(tn)^{2l-|S|-1}}{(l-1)!} n(n-1)^l (n-1)^{-l} \left(\frac{2}{n}\right)^{2l-1-|S|} \\ &= \sum_{S\subset\{1,\dots,l-1\}} n\frac{(2t)^{2l-|S|-1}}{(l-1)!} \\ &= \frac{n}{(l-1)!}\sum_{i=1}^{l-1}\binom{l-1}{i}(2t)^{2l-i-1} \\ &= \frac{n(2t)^l(2t+1)^{l-1}}{(l-1)!} \end{aligned}$$

Taking $l = \frac{\ln n}{\sqrt{\ln\ln n}} =: b_n$, and using Stirling formula, one obtains that $p_l \to 0$.

We now need to bound the volume of the ball of radius b_n in $G_{n,t}^{\infty}$. For any $w \in S_l(G_{n,t}^{\infty}, 1)$, let $d^*(w)$ be the number of edges that connects w to some vertex of $G_{n,t}^{\infty} \setminus B_l(G_{n,t}^{\infty}, 1)$. Conditionally on $B_l(G_{n,t}^{\infty}, 1)$, $(d_w^*)_{w\in S_l(G_{n,t}^{\infty},1)}$ is an i.i.d. family of Poisson variables of parameter $\left(n - |B_l(G_{n,t}^{\infty}, 1)|\right)\frac{t}{n} \leq t$. Therefore the volume of the ball of radius b_n in $G_{n,t}^{\infty}$ is stochastically dominated by the volume of the first b_n generations of T_t^{∞}. We assume that $t > 1$.

Let Z_l be the number of vertices in the lth generation of T_t^{∞}. Conditionally on Z_i, Z_{i+1} is the sum of Z_i independent Poisson variables of parameter t, *i.e.* a Poisson variable of parameter tZ_i. Using large deviations theory (or direct computation), there exists a positive constant α such that:

$$\forall x \geq 1,\ \mathbb{P}(\mathrm{Poi}(x) \geq 2x) \leq exp(-\alpha x) \tag{3.6}$$

For all integers i, let $p_i = \mathbb{P}\left(Z_{i+1} > (2t)^{i+1}\ln^2 n \text{ and } Z_i \leq (2t)^i \ln^2 n\right)$.

$$\begin{aligned} p_i &= \mathbb{E}\left(\mathbb{P}\left(Z_{i+1} > (2t)^{i+1}\ln^2 n | Z_i\right)\mathbf{1}_{\mathbf{Z_i \leq (2t)^i \ln^2 n}}\right) \\ &= \mathbb{E}\left(\mathbb{P}\left(Poi(tZ_i) > (2t)^{i+1}\ln^2 n | Z_i\right)\mathbf{1}_{\mathbf{Z_i \leq (2t)^i \ln^2 n}}\right) \\ &\leq \mathbb{E}\left(\mathbb{P}\left(Poi(t(2t)^i\ln^2 n) > (2t)^{i+1}\ln^2 n | Z_i\right)\mathbf{1}_{\mathbf{Z_i \leq (2t)^i \ln^2 n}}\right) \\ &\leq \mathbb{P}\left(Poi(t(2t)^i\ln^2 n) > (2t)^{i+1}\ln^2 n\right) \\ &\leq exp(-\alpha t(2t)^i \ln^2 n) \\ &\leq exp(-\alpha \ln^2 n) \\ &= o(\frac{1}{n^2}) \end{aligned}$$

Therefore $P_n := \sum_{i=0}^{b_n} p_i = o(\frac{1}{n})$. With probability larger than $1 - P_n$, for all $i \leq b_n$, $Z_i \leq (2t)^i \ln^2 n$, and therefore $\sum_{i=0}^{b_n} Z_i \leq \frac{(2t)^{b_n+1}-1}{2t-1}\ln^2 n =: n^{c_n}$,

where $c_n = o(1)$. Therefore, $B_{b_n}(G^\infty_{n,t}, 1)$ contains less than n^{c_n} vertices with probability larger than $1 - P_n$. By union bounds, all the balls of radius b_n in $G^\infty_{n,t}$ contain less than n^{c_n} vertices with probability larger than $1 - nP_n \to 1$. □

3.4.3 The local limit is a branching process

Definition 3.15. For any $G \in \Omega^+$ and $t \geq 0$, let $\Phi_t(G) = \Phi(G_t)$, where G_t still denotes the subgraph of G restricted to edges with labels smaller than or equal to t.

Given a vertex v of T^∞_∞, different from the root, let w denote the parent of v and $T^{\infty,v}_\infty$ denote the connected component of v in $T^\infty_\infty \setminus \{v, w\}$ *i.e.* the subtree of T^∞_∞ starting at v, and let $\tilde{T}^{\infty,v}_\infty = (T^{\infty,v}_\infty \cup \{v, w\})$ and $\tilde{T}^{\infty,v}_t = (T^{\infty,v}_\infty)_t$. See figure 3.1 for an example.

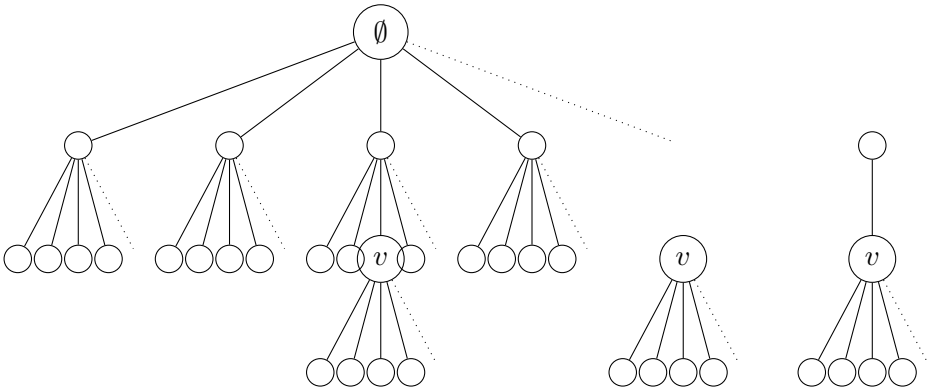

Figure 3.1: Example of T^∞_∞, $T^{\infty,v}_\infty$ and $\tilde{T}^{\infty,v}_\infty$.

Lemma 3.24. *For a graph $G \in \Omega^+$, an edge $e \in G$ and a non-negative t, we say that e is removed in $\Phi_t(G)$ if $e \in G_t$ and $e \notin \Phi_t(G)$.*

Let t be a non-negative number and v vertex of T^∞_∞. If $T^\infty_t \in \Omega^-$, then, at least one of the following propositions hold:

- $\Phi(T^\infty_t) \cap \tilde{T}^{\infty,v}_\infty = \Phi(\tilde{T}^{\infty,v}_t)$;
- *the edge between v and its parent w is removed in $\Phi_t(T^\infty_\infty)$.*

An immediate corollary of Lemma 3.24 is:

Corollary 3.25. *With the notation of Lemma 3.24, if (v, w) is removed in $\Phi(\tilde{T}^{\infty,v}_t)$, then it is removed in $\Phi(T^\infty_t)$.*

In other words, until (v, w) is removed in $\Phi_t(T^\infty_\infty)$, the knowledge of $\tilde{T}^{\infty,v}_\infty$ is sufficient to know the evolution of the subtree starting at v in the forbidden degree version of T^∞_∞.

Proof. The graphs $\Phi(T_t^\infty) \cap \tilde{T}_t^{\infty,v}$ and $\Phi_t(\tilde{T}_t^{\infty,v})$ have the same vertex set (the vertex set of $\tilde{T}_t^{\infty,v}$), therefore any difference comes from the set of edges. Given two graphs G_1 and G_2 with the same set of vertices and an edge e, we are going to say that e *separates* G_1 and G_2 if e is present in one of the graphs but not the other. Let us assume that $\Phi(T_t^\infty) \cap \tilde{T}_t^{\infty,v} \neq \Phi(\tilde{T}_t^{\infty,v})$, and let e be an edge separating $\Phi(T_t^\infty) \cap \tilde{T}_t^{\infty,v}$ and $\Phi(\tilde{T}_t^{\infty,v})$.

The graph T_t^∞ is in Ω^- a.s., so is $\tilde{T}_t^{\infty,v}$, and therefore these two graphs are in Ω^+. By definition of Ω^+, there exists an integer l such that the knowledge of the balls of radius l centered at e and (v,w) in T_t^∞ and $\tilde{T}_t^{\infty,v}$ are sufficient to know whether e and (v,w) are present in $\Phi(T_t^\infty)$ and $\Phi(\tilde{T}_t^{\infty,v})$. Let $B^a = B_l(T_t^\infty, e) \cup B_l(T_t^\infty, (v,w))$ and $B^b = B_l(\tilde{T}_t^{\infty,v}, e) \cup B_l(\tilde{T}_t^{\infty,v}, (v,w))$. By Definition 3.8, e and (v,w) are present in $\Phi(T_t^\infty)$ (resp. $\Phi(\tilde{T}_t^{\infty,v})$) if and only if they are present in $\Phi(B^a)$ (resp. $\Phi(B^b)$). It should be noted that $B^b = B^a \cap \tilde{T}_t^{\infty,v}$. B^a is a finite graph, therefore the set of labels of B^a (and B^b) is finite. Let s be the first time such that $\Phi_s(B^a) \cap \tilde{T}_t^{\infty,v} \neq \Phi_s(B^b)$ holds, and let e' be an edge of $T_t^{\infty,v}$ that separates $\Phi_s(B^a) \cap \tilde{T}_t^{\infty,v}$ and $\Phi_s(B^b)$, *i.e.* e' is removed at time s in either $\Phi_\cdot(B^a)$ or $\Phi_\cdot(B^b)$. Let x be an endpoint of e' that reaches the forbidden degree at time s in one of the graphs, but not the other. Therefore the degree of x is different in $\Phi(B^a)$ and in $\Phi(B^b)$ strictly before time s. This implies that an edge e'' adjacent to x is present in either $\Phi_{s^-}(B^a)$ or $\Phi_{s^-}(B^a)$ but not in the other graph. If $x \neq w$, every edge adjacent to x in either B^a or B^b is in $\tilde{T}_t^{\infty,v}$, and e'' contradicts the definition of s and e'. Therefore $x = w$ and $e' = (v,w)$ (as (v,w) is the only edge of $\tilde{T}_t^{\infty,v}$ adjacent to w). As the degree of w in B^b is 1, w cannot reach the forbidden degree in $\Phi(B^b)$. Therefore w reaches the forbidden degree in $\Phi(B^a)$ and (v,w) is removed from $\Phi(B^a)$ *i.e.* (v,w) is removed in $\Phi(T_t^\infty)$. □

Let us add to the set of rooted graphs an element $\mathscr{X}$, used to denote a graph that is no longer *relevant*: let $\tilde{T}_t^{k,v}$ (resp. $T_t^{k,v}$) be defined as equal to $\Phi(\tilde{T}_t^{\infty,v})$ (resp. $\tilde{T}_t^{k,v} \cap T_t^{\infty,v}$) if the edge between v and its parent has not been removed from $\Phi(\tilde{T}_t^{\infty,v})$ and equal to $\mathscr{X}$ otherwise. Let $\tau_1, \dots, \tau_l$ denote the labels of the edges adjacent to the root in T_t^∞, ordered in increasing order, and let $v_1, \dots, v_l$ denote the other endpoints of these edges. For any $s \leq t$, let $\tilde{S}_s$ denote the set of integers such that $\{v_i : i \in \tilde{S}_s\}$ is the set of neighbors of the root in $\Phi_s(T_t^\infty)$ *i.e.* $\tilde{S}_s$ records which edge is present in the forbidden degree version of T_t^∞ at time s. For any $i \leq l$, let $\rho_i = \inf\{s : T_s^{k,v_i} = \mathscr{X}\}$. Let $\Gamma_\tau = \{\tau_i, 1 \leq i \leq l\}$ and $\Gamma_\rho = \{\rho_i, 1 \leq i \leq l\}$.

Lemma 3.26. *Almost surely, for every $1 \leq i < j \leq l$, $\tau_i \neq \tau_j$, $\rho_i \neq \tau_j$, and either $\rho_i \neq \rho_j$ or $\rho_i = \rho_j = \infty$.*

Proof. By definition, τ_i is the label of $(\emptyset, v_i)$ and ρ_i is the first time $(\emptyset, v_i)$ is removed in $\Phi(\tilde{T}_t^{\infty,v_i})$, *i.e.* is either infinite or equal to the label of an edge adjacent to v_i. As $T^\infty \in \Omega$, no two edges have the same label. □

Lemmata 3.24 and 3.26 allow us to consider the following dynamic set $(S_s)_{0\leq s\leq t}$:

- at time 0, S_0 is empty;
- $s \to S_s$ is piecewise constant, and its set of jumps is included in $\Gamma_\rho \cup \Gamma_\tau$;
- for every $\tau_i \notin \Gamma_\rho$, if $|S_{\tau_i^-}| = k-1$, then $S_{\tau_i} = \emptyset$, otherwise $S_{\tau_i} = S_{\tau_i^-} \cup \{i\}$;
- for every finite $\rho_i \notin \Gamma_\tau$, $S_{\rho_i} = S_{\rho_i^-} \setminus \{i\}$;
- for every i such that $\tau_i = \rho_i$, if $|S_{\tau_i^-}| = k-1$, then $S_{\tau_i} = \emptyset$, otherwise $S_{\tau_i} = S_{\tau_i^-}$.

In other words, the element i is added at time τ_i, removed at time ρ_i, and S becomes empty whenever it reaches size k, even fleetingly[3].

Corollary 3.27. *For any time s, $S_s = \tilde{S}_s$ and the connected component of the root is the same in $\Phi(T_s^\infty)$ and in $\cup_{i\in S_s}\Phi(\tilde{T}_s^{\infty,v_i})$.*

Proof. Let first assume that for some s, $S_s \neq \tilde{S}_s$. The set S can only change at the times $(\tau_i)_{1\leq i\leq l}$ and $(\rho_i)_{1\leq i\leq l}$; and the set $\tilde{S}$ can only change at times equal to a label of an edge in $B_2(T_t^\infty, \emptyset)$. These two sets of times are finite, therefore there exists a smallest s such that $S_s \neq \tilde{S}_s$. Let $i \in S_s \Delta \tilde{S}_s$, in which Δ denotes the symmetric difference. The element i is added to S and $\tilde{S}$ at time τ_i, therefore the difference eventually comes from the removal times: when i is in S and $\tilde{S}$ at time s^- and is removed from one, but not the other at time s. An element can be removed from S or $\tilde{S}$ at time s for the following reasons:

1. S (resp. $\tilde{S}$) reaches size k at time s. Therefore $|S_{s^-}| = k-1$ (resp. $|\tilde{S}_{s^-}| = k-1$) and an element is added to S (resp. $\tilde{S}$) at time s. As $S_{s^-} = \tilde{S}_{s^-}$ by definition of s, and the times of addition are identical for S and $\tilde{S}$, S reaches size k at time s if and only if $\tilde{S}$ reaches size k at time s.

2. $\rho_i = s$, that is $(\emptyset, v_i)$ is removed from $\Phi(\tilde{T}_\cdot^{\infty,v})$ at time s. By Corollary 3.25, $(\emptyset, v_i)$ is also removed from $\Phi(T_s^\infty)$, and i is removed from $\tilde{S}$.

3. $(\emptyset, v_i)$ is removed from $\Phi(T_\cdot^\infty)$ at time s because the vertex v_i reaches degree k in $\Phi(T_\cdot^\infty)$, *i.e.* v_i has degree $k-1$ in $\Phi(T_{s^-}^\infty)$, and an edge is added to v_i. $i \in S_{s^-}$, therefore $(\emptyset, v_i)$ is present in $\Phi(\tilde{T}_{s^-}^{\infty,v_i})$. By Lemma 3.24, $\Phi(T_{s^-}^\infty) \cap \tilde{T}_\infty^{\infty,v_i} = \Phi(\tilde{T}_{s^-}^{\infty,v_i})$, and therefore v_i has degree

[3] Actually, S never reaches size k, going directly from size $k-1$ to 0, but we informally say that S reaches size k for an instant, as described in the footnote 1 page 101

$k-1$ in $\Phi(\tilde{T}_{s^-}^{\infty,v_i})$, and reaches degree k at time s. Therefore $\rho_i = s$, and i is removed from S at time s.

Therefore, for all s, $S_s = \tilde{S}_s$. Moreover, for any $i \in S_s$, $(\emptyset, v_i) \in \Phi(\tilde{T}_s^{\infty,v_i})$, therefore, by Lemma 3.24, $\Phi(T_s^\infty) \cap \tilde{T}_\infty^{\infty,v_i} = \Phi(\tilde{T}_s^{\infty,v_i})$, ending the proof of Corollary 3.27.

□

Definition 3.16. Given a rooted graph $(G, \emptyset)$ in Ω and a non-negative number y, we introduce the following notations:

- $\Theta^y(G, \emptyset)$ denotes the graph G rooted at $\emptyset$ with an extra vertex w and an edge between w and $\emptyset$, labelled by y;
- if $\Theta^y(G, \emptyset) \in \Omega^+$, then $\tilde{\Phi}^y(G, \emptyset)$ and $\Phi^y(G, \emptyset)$ are defined as follows: if $(w, \emptyset)$ is removed from $\Phi(\Theta^y(G, \emptyset))$, then let $\tilde{\Phi}^y(G, \emptyset) = \mathscr{X}$ and $\Phi^y(G, \emptyset) = \mathscr{X}$; otherwise, let $\tilde{\Phi}^y(G, \emptyset) = \Phi(\Theta^y(G, \emptyset))$ and $\Phi^y(G, \emptyset) = \tilde{\Phi}^y(G, \emptyset) \cap G$;
- let $T_t^{\infty,y} = \Theta^y(T_t^\infty)$, $\tilde{T}_t^{k,y} = \tilde{\Phi}^y(T_t^\infty)$ and $T_t^{k,y} = \Phi^y(T_t^\infty)$.

Let $m(t, y) = \mathbb{P}(T_t^{k,y} \neq \mathscr{X})$. The tree T_t^∞ contains no edge with probability $e^{-t} > 0$ and in that case $T_t^{k,y} \neq \mathscr{X}$, therefore $m(t, y) > e^{-t} > 0$. Let $T_t^{k+,y}$ (resp. $\tilde{T}_t^{k+,y}$) be the random tree $T_t^{k,y}$ (resp. $\tilde{T}_t^{k,y}$) conditioned on not being equal to $\mathscr{X}$.

Lemma 3.28. *Conditionally on $(\tau_i)_{1 \leq i \leq l}$ and $(S_s)_{0 \leq s \leq t}$:*

- *the graphs $(T_t^{k,v_i})_{i \in S_t}$ (resp. $(\tilde{T}_t^{k,v_i})_{i \in S_t}$) are independent:*
- *for each $i \in S_t$, T_t^{k,v_i} (resp. $\tilde{T}_t^{k,v_i}$) has the same law as T_t^{k+,τ_i} (resp. $\tilde{T}_t^{k+,\tau_i}$).*

Proof. Let $\mathscr{F}$ be the σ-algebra generated by the sequences $(\tau_i)_{1 \leq i \leq l}$ and $(\rho_i)_{1 \leq i \leq l}$, and $\mathscr{F}_1$ be the σ-algebra generated by $(\tau_i)_{1 \leq i \leq l}$ and $(S_s)_{0 \leq s \leq t}$. As $\mathscr{F}_1 \subset \mathscr{F}$, it is sufficient to prove Lemma 3.28 with $\mathscr{F}$ instead of $\mathscr{F}_1$. By the branching property, conditionally on $(\tau_i)_{1 \leq i \leq l}$, the trees $(\tilde{T}_t^{\infty,v_i})_{1 \leq i \leq l}$ are independent, and for each i, $\tilde{T}_t^{\infty,v_i}$ is a copy of $\tilde{T}_t^{\infty,\tau_i}$. Each ρ_i only depends on $\tilde{T}_t^{\infty,v_i}$, and S_t is $\mathscr{F}$-measurable. Therefore, conditionally on $\mathscr{F}$: the trees $(\tilde{T}_t^{\infty,v_i})_{1 \leq i \leq l}$ are independent and for each i, $\tilde{T}_t^{\infty,v_i}$ has the same law as $\tilde{T}_t^{\infty,\tau_i}$ conditionally on $\inf\{s : \tilde{T}_s^{k,\tau_i} = \mathscr{X}\} = \rho_i$. It follows that conditionally on $\mathscr{F}$: the trees $(\tilde{T}_t^{k,v_i})_{i \in S_t}$ are independent and for each i in S_t, $\tilde{T}_t^{k,v_i}$ has the same law as T_t^{k+,τ_i}.

As $T_t^{k,v_i} = \tilde{T}_t^{k,v_i} \cap T_t^{\infty,v_i}$ and $T_t^{k,\tau_i} = \tilde{T}_t^{k,\tau_i} \cap T_t^{\infty,\tau_i}$, Lemma 3.28 with $\tilde{T}_t^{k,v_i}$ imply Lemma 3.28 with T_t^{k,v_i}. □

Lemma 3.28 implies the following theorem:

Theorem 3.29. *Let $B(y)$ be the law of the set of the labels of the edges adjacent to the root of $T_t^{k+,y}$. Let BP be the multitype branching process with offspring law $B(\cdot)$. Then $T_t^{k+,y}$ has same law as BP starting with a root of label y.*

Proof. $T_t^{k,y}$ is either equal to T_t^k or to $\mathscr{X}$. With the notations of the proof of Lemma 3.28, the event $T_t^{k,y} = \mathscr{X}$ is $\mathscr{F}_1$-measurable, therefore, conditionally on $T_t^{k,y} \neq \mathscr{X}$ and the sequence $(\tau_1, \ldots, \tau_l)$ of labels of the edges adjacent to the root, the subtrees starting at the children of the root are independent copies of T_t^{k+,τ_i}. □

Lemma 3.28 also implies that T_t^k is a two-stages branching process:

Theorem 3.30. *Conditionally on the set $(\tau_i)_{i\in S_t}$ of labels of edges adjacent to the root in T_t^k, the subtrees starting at the root's children are independent, and copies of T_t^{k+,τ_i}.*

3.4.3.a Properties of the branching process

Let $\mu_{t,y}$ (resp. $\mu_{t,y}^+$, ν_t) be the law of the set of labels of edges adjacent to the root in $T_t^{k,y}$ (resp. $T_t^{k+,y}$, T_t^k). This measure can be decomposed according to the degree of the root in these trees:

$$\begin{aligned}
\mu_{t,y} &= \sum_{i=-1}^{k-2} \mu_{t,y}^i; \\
\mu_{t,y}^+ &= \sum_{i=0}^{k-2} \mu_{t,y}^{i+}; \\
\nu_t &= \sum_{i=0}^{k-1} \nu_t^i.
\end{aligned}$$

in which $\mu_{t,y}^i$ (resp. $\mu_{t,y}^{i+}$, ν_t^i) only puts mass on sets of cardinality i and $\mu_{t,y}^{-1}$ is equal to $(1 - m(t,y))\delta_{\mathscr{X}}$.

Definition 3.17. For X be a finite subset of $[0,t]$, let us define the random rooted tree $T_t^{\infty,X}$ as follows:

- let P be a Poisson point process of intensity 1 on $[0,t]$;
- conditionally on P, let $(T^z)_{z\in P\cup X}$ be an i.i.d family of copies of T_t^∞;
- then, adding a root $\emptyset$ to the forest $(T^z)_{z\in P\cup X}$, and, for each $z \in P\cup X$, adding an edge with label z between $\emptyset$ and the root of T^z, one obtains $T_t^{\infty,X}$. $T^{\infty,X}$ is rooted at $\emptyset$.

Let E_t^X denote the event *"the only edges adjacent to the root in $\Phi(T_t^{\infty,X})$ are the edges labelled by X"*. If $y \notin X$, let $E_t^{y,X}$ denote the event *"$\Phi^y(T_t^{\infty,X}) \neq \mathscr{X}$ and the only edges adjacent to the root in $\Phi^y(T_t^{\infty,X})$ are the edges labelled by X"*.

Informally, $T_t^{\infty,X}$ is the conditional distribution of the tree T_t^∞, given that the root is incident to edges with labels in X. This informal explanation will be justified rigorously with the help of Campbell formulas in the following pages. If $y \notin X$, a.s., $\Theta^y(T_t^{\infty,X}) \in \Omega^+$ and $\Phi^y(T_t^{\infty,X})$ is well-defined.

Definition 3.18. Let $0 \leq j \leq i \leq k-2$ be two integers. Q_i^j is the set $\{(t, y, x_1, \ldots, x_i) : 0 \leq x_1 \leq \ldots x_j \leq y \leq x_{j+1} \leq \cdots \leq x_i \leq t\}$, and Q_i is the simplex $\{(t, x_1, \ldots, x_i) : 0 \leq x_1 \leq \cdots \leq x_i \leq t\}$.

Lemma 3.31. *For any integer i and j, $(t, y, x_1, \ldots, x_i) \to \mathbb{P}(E_t^{y,\{x_1,\ldots,x_i\}})$ (resp. $(t, x_1, \ldots, x_i) \to \mathbb{P}(E_t^{\{x_1,\ldots,x_i\}})$) is continuous on $\mathring{Q}_i^j$ (resp. $\mathring{Q}_i$), and both are larger than $exp(-(i+1)t)$.*

Lemma 3.32. *$(y, t) \to m(t, y)$ is continuous on $\{0 \leq y < t\}$ and is larger than e^{-t}. For any integer i and $y < t$, $\mu_{t,y}^i$, $\mu_{t,y}^{i+}$ and ν_t^i are absolutely continuous with respect to the Lebesgue measure on $[0, t]^i$, with respective densities:*

$$\begin{aligned} \frac{\partial \mu_{t,y}^i}{\partial x} = f_i(t, y, x) &= \mathbb{P}(E_t^{y,\{x_1,\ldots,x_i\}}) \\ \frac{\partial \mu_{t,y}^{i+}}{\partial x} = g_i(t, y, x) &= \mathbb{P}(E_t^{y,\{x_1,\ldots,x_i\}})/m(t, y). \\ \frac{\partial \nu_t^i}{\partial x} = h_i(t, x) &= \mathbb{P}(E_t^{\{x_1,\ldots,x_i\}}) \end{aligned}$$

The functions f_i, g_i and h_i are not defined on the null set where two or more coordinates are equal. They are by definition symmetric in the variables $(x_j)_{1 \leq j \leq i}$. By Lemma 3.31, f_i and g_i are continuous on every set $\mathring{Q}_i^j$, and h_i is continuous on Q_i, and f_i, g_i and h_i are larger than $exp(-(i+1)t)$.

The continuity in Lemma 3.31 will be proven by a coupling argument:

- Let P^{coupl} be a Poisson point process of intensity 1 on $(0, \infty)$.
- Let $(T^{z,coupl})_{z \in P^{coupl}}$ be a family of independent copies of the process T_∞^∞.
- Let $(T^{j,coupl})_{j \leq i}$ be i independent copies of T_∞^∞.

Let $T^{\infty,\{x_1,\ldots,x_i\}}$ be constructed in the following way:

- P is the restriction of P^{coupl} to $(0, t)$.

- For any $z \in P$, $T^z = T_t^{z,coupl}$.
- For any x_j, $T^{x_j} = T_t^{j,coupl}$.

This construction allows to couple every $T_t^{\infty,\{x_1,\dots,x_i\}}$ by using the same randomness. Let T be a positive real number, and let $(t, y, x_1 \dots, x_l)$ and $(\tilde{t}, \tilde{y}, \tilde{x}_1 \dots, \tilde{x}_l)$ be two elements of $\mathring{Q}_i^j$ such that $T \geq t$ and $T \geq \tilde{t}$. Using the notation $[a, b]$ to describe the non-empty interval $[\min(a, b), \max(a, b)]$, let $I = [\tilde{t}, t] \cup [y, \tilde{y}] \cup_i [x_i, \tilde{x}_i]$. Let l be a positive integer. Let $X = \{x_1 \dots, x_i\}$ and $\tilde{X} = \{\tilde{x}_1 \dots, \tilde{x}_i\}$

Lemma 3.33. *If the following properties hold,*

1. $P^{coupl} \cap I = \emptyset$;
2. *For every $z \in P^{coupl} \cap [0, T]$, no edge in the first l generation of $T_T^{z,coupl}$ has a label in I.*
3. *For every j, no edge in the first l generation of $T_T^{j,coupl}$ has a label in I.*
4. *For every $z \in P^{coupl} \cap [0, T]$, there is no propagation path of length l starting from the root of $T_T^{z,coupl}$.*
5. *For every j, there is no propagation path of length l starting from the root of $T_T^{j,coupl}$.*

then $E_t^{X,y}$ holds if and only if $E_{\tilde{t}}^{\tilde{X},\tilde{y}}$ holds, and E_t^X holds if and only if $E_{\tilde{t}}^{\tilde{X}}$ holds.

Proof. The knowledge of $B^l(\Theta^y(T_t^X), \emptyset)$ (resp. $B^l(\Theta^{\tilde{y}}(T_{\tilde{t}}^{\tilde{X}}), \emptyset)$) is sufficient to know which edges are adjacent to the root in $\Phi(\Theta^y(T_t^X))$ (resp. $\Phi(\Theta^{\tilde{y}}(T_{\tilde{t}}^{\tilde{X}}))$), by properties 4 and 5. By properties 1, 2 and 3, the unlabelled versions of $B^l(\Theta^y(T_t^X), \emptyset)$ and $B^l(\Theta^{\tilde{y}}(T_{\tilde{t}}^{\tilde{X}}), \emptyset)$ are equal. For any graph $G \in \Omega^+$, the set of removed edges in $\Phi(G)$ only depends on the unlabelled graph G and the respective order of the edges' labels, not on the actual labels. By hypothesis, the elements of (t, y, x) and $(\tilde{t}, \tilde{y}, \tilde{x})$ are in the same respective order. Therefore, by properties 2 and 3, the respective order of the labels is the same in $B^l(\Theta^y(T_t^X), \emptyset)$ and $B^l(\Theta^{\tilde{y}}(T_{\tilde{t}}^{\tilde{X}}), \emptyset)$, and therefore $E_t^{x,y}$ holds if and only if $E_{\tilde{t}}^{\tilde{x},\tilde{y}}$ holds.

The proof for E_t^X is identical upon removing Θ^y. □

Proof of Lemma 3.31. For a given T, the probability of the properties 4 and 5 tends to 1 when l tends to ∞. For a given T and l, the properties of properties 1, 2 and 3 tends to 1 when $(\tilde{t}, \tilde{y}, \tilde{x}_1, \dots \tilde{x}_l)$ tends to $(t, y, x_1, \dots x_l)$, proving the continuity of $\mathbb{P}(E_t^{y,X})$ and $\mathbb{P}(E_t^X)$.

If $P = \emptyset$ and every subtree starting from a children is empty, the events $E_t^{y,X}$ and E_t^X hold. $|P|$ is a Poisson variable of parameter t, and T_t^∞ is empty

with probability $exp(-t)$, therefore $\mathbb{P}(E_t^{y,X}) =$ (resp. $\mathbb{P}(E_t^X)$) is larger than $exp(-(i+1)t)$ for all set X of size $i \leq k-2$ and any $y \notin X$ (resp. for all set X of size $i \leq k-1$). □

Proof of Lemma 3.32. For any finite set $X \subset [0,t]$ and $y \in [0,t] \setminus X$, we consider the following construction:

- Let $(T_t^z)_{z \in X}$ be an i.i.d. family of copies of T_t^∞.
- Let $T_t^{\infty,(X)}$ is the tree obtained by taking a vertex $\emptyset$, every tree $(T_t^z)_{z \in X}$ and adding an edge, labelled by z, between $\emptyset$ and the root of T_t^z. $T_t^{\infty,(X)}$ is rooted at $\emptyset$.
- Let $Z_X^y \subset X$ (resp. Z_X) be equal to the random set of labels of the edges adjacent to the root of $\Phi^y(T_t^{\infty,(X)})$ (resp. $\Phi(T_t^{\infty,(X)})$), if defined.

Remark. The difference between $T_t^{\infty,X}$ and $T_t^{\infty,(X)}$ is that the latter have only edges labelled by elements of X adjacent to the root, whereas the former have edges labelled by X and additional edges, according to the Poisson point process P.

By the branching property, conditionally on P, the set of edges adjacent to the root of T_t^∞, T_t^∞ has same distribution as $T_t^{\infty,(P)}$, and therefore the set of edges adjacent to $\Phi^y(T_t^\infty)$ has same distribution as Z_P^y.

Let λ be a bounded positive continuous function $[0,t]^i \to [0,\infty)$.

$$\int_{[0,t]^i} \lambda(x_1,\dots,x_i)\mu_{y,t}^i(d(x_1\dots x_i)) = \int_{[0,t]^i} \lambda(x_1,\dots,x_i) f(t,y,x_1\dots,x_l)dx_1\dots dx_i.$$

Let A denote the left-hand side:

$$\begin{aligned}
A &= \mathbb{E}\left(\sum_{x_1,\dots,x_i \in P}^{\neq} \lambda(x_1,\dots,x_i)\mathbf{1}_{\mathbf{Z_P^y=\{x_1,\dots,x_i\}}}\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(\sum_{x_1,\dots,x_i \in P}^{\neq} \lambda(x_1,\dots,x_i)\mathbf{1}_{\mathbf{Z_P^y=\{x_1,\dots,x_i\}}}\middle| P\right)\right) \\
&= \mathbb{E}\left(\sum_{x_1,\dots,x_i \in P}^{\neq} \lambda(x_1,\dots,x_i)\mathbb{P}(Z_P^y = \{x_1,\dots,x_i\})\right) \\
&= \mathbb{E}\left(\sum_{x_1,\dots,x_i \in P}^{\neq} \gamma(x_1,\dots,x_i,P)\right)
\end{aligned}$$

in which $\gamma(x_1,\dots,x_i,Y) = \lambda(x_1,\dots,x_i)\mathbb{P}(Z_Y^y = \{x_1,\dots,x_i\})$ for every set Y and real numbers $x_1,\dots,x_i$. By the reduced Campbell formula, [BB09,

formula (9.12)]:

$$\mathbb{E}\left(\sum_{x_1,\dots,x_i\in P}^{\neq}\gamma(x_1,\dots,x_i,P)\right)=\int_{[0,t]^i}\mathbb{E}(\gamma(x_1,\dots,x_l,P_{x_1,\dots,x_i}))M^{(i)}(dx_1\dots dx_n).$$

in which, for a Poisson point process of intensity Λ equal to the Lebesgue measure, $M^{(i)}=\Lambda^i$, by [BB09, Proposition 9.1.3] and $P_{x_1,\dots,x_i}$ has same law as $P\cup\{x_1,\dots,x_i\}$ by [BB09, Corollary 9.2.5]. Therefore:

$$\begin{aligned}A &= \int_{[0,t]^i}\mathbb{E}(\gamma(x_1,\dots,x_l,P_{x_1,\dots,x_i}))dx_1\dots dx_i\\ &= \int_{[0,t]^i}\lambda(x_1,\dots,x_i)\mathbb{P}\left(Z^y_{\{x_1,\dots,x_i\}\cup P}=\{x_1,\dots,x_i\}\right)dx_1\dots dx_i\\ &= \int_{[0,t]^i}\lambda(x_1,\dots,x_i)\mathbb{P}(E_t^{y,X})dx_1\dots dx_i\end{aligned}$$

Therefore $\mu^i_{y,t}$ is absolutely continuous with respect to the Lebesgue measure with density $(x_1,\dots,x_i)\to\mathbb{P}(E_t^{y,\{x_1\dots,x_i\}})$. This density is a probability, and is therefore bounded by 1, and is continuous by Lemma 3.31. Therefore $(t,y)\to m(t,y)=\sum_{i=0}^{k-2}\mu^i_{y,t}(Q_i)$ is continuous.

As the random tree $T_t^{k+,y}$ is equal to the tree $T_t^{k,y}$ conditioned on not being equal to $\mathscr{X}$, $\mu^{i+}_{y,t}$ is absolutely continuous with respect to the Lebesgue measure with density $(x_1,\dots,x_i)\to\mathbb{P}(E_t^{y,X})/m(t,y)$.

The proof for ν^i_t is identical to the proof for $\mu^i_{y,t}$ upon replacing Φ^y by Φ, $E_t^{y,X}$ by E_t^X and Z_X^y by Z_X.

□

3.4.3.b Link between $\mu_{.,t}$ and ν_t

Lemma 3.34. *For every integer $i\geq 1$ and iuple $(x_1,\dots x_i)\in(0,t)^i$, $h_i(t,x_1\dots,x_i)=m(t,x_1)f_{i-1}(t,x_1,\dots,x_l)$.*

Let $X=\{x_1,\dots,x_i\}$ and $\tilde{X}=\{x_2,\dots,x_i\}$. By Lemma 3.32, the density $h_i(t,x_1\dots,x_i)$ is equal to $\mathbb{P}(E_t^X)$, *i.e.* the probability that the only edges adjacent to the root in $\Phi(T_t^{\infty,X})$ are the edges labelled by elements of X.

Let v be the other endpoint of the edge labelled by x_1, and T_t^{∞,x_1} the subtree of $T_t^{\infty,X}$ starting at v and rooted at v. L et $T_t^{\infty,X,\setminus x_1}$ be the tree $T_t^{\infty,X}\setminus T_t^{\infty,x_1}$. The figure 3.4.3.b illustrates these trees.

By definition of $T_t^{\infty,X}$, T_t^{∞,x_1} and $T_t^{\infty,X,\setminus x_1}$ are independent, and have same distribution as respectively T_t^{∞} and $T_t^{\infty,\tilde{X}}$.

E_t^X implies that the edge labelled by x_1 is present in $\Phi(T_t^{\infty,X})$, therefore, by Lemma 3.24, E_t^X is equivalent to:

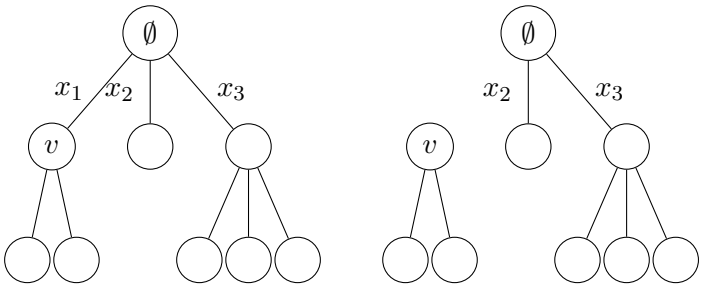

Figure 3.2: Example of $T_t^{\infty,X}$, T_t^{∞,x_1} and $T_t^{\infty,X,\setminus x_1}$, from left to right.

- the edge labelled by x_1 is present in $\Phi(\Theta^{x_1}(T_t^{\infty,x_1}))$, *i.e.* $\Phi^{x_1}(T_t^{\infty,x_1}) \neq \mathscr{X}$, and
- the edges adjacent to the root in $\Phi(\Theta^{x_1}(T^{\infty,X,\setminus x_1}))$ are the edges labelled by $x_1, \dots, x_i$.

This two events are independent and of probability $m(t, x_1)$ and $\mathbb{P}(E_{x_2}^{\{x_2,\dots,x_i\}})$, therefore $\mathbb{P}(E_t^X) = m(t, x_1)\mathbb{P}(E_t^{x_1,\tilde{X}})$, proving Lemma 3.34.

This formula allows to compute the density of any ν_t^i for $i \geq 1$ from the density of $\mu_{t,y}^i$. As the probability that the root of T_t^k is isolated is $1 - \sum_{i=1}^{k-1} \nu^i(Q_i)$, it is sufficient to study the measure μ to know ν. This computation also implies that $m(t, x_1)f_{i-1}(t, x_1, x_2 \dots, x_i)$ is symmetric in the x_i.

3.5 The equivalence between supercriticality of the local limit and the existence of the giant component

Let a_t^k be the probability that the tree T_t^k is infinite. The goal of this section is to prove Theorem 3.4.

Theorem 3.4. *For all $t \geq 0$, $\frac{C_{\max}(G_{n,t}^k)}{n}$ converges in probability to a_t^k.*

For readability, $C_{\max}\left(G_{n,t}^k\right)$ is abridged in $C_{\max}$.

Remark. Theorem 3.4 is meaningful both in the subcritical case and in the supercritical case.

If T_t^k is critical or subcritical, $a_t^k = 0$, and therefore the largest component's size is $o_p(n)$.

If T_t^k is supercritical, $a_t^k > 0$, so there is a giant component of size equivalent to $a_t^k n$.

This result is analogous to the result on the size of the largest component in the Erdős-Rényi graph, where a_t is the probability of survival of a Galton-Watson whose offspring is Poisson distributed with parameter t.

The local limit results imply that a_t^k is close to the expected proportion of vertices in large components. To prove Theorem 3.4, one needs to prove that a_t^k is a.s. close to the actual proportion of vertices in large components (not only in expectation), and that almost every vertices in large components are in the same component.

3.5.1 The subcritical or critical case

For any integer i, let N^i be the number of vertices of $G_{n,t}^k$ in components of size at least i.

Lemma 3.35. *For all $\epsilon > 0$, $\mathbb{P}(\frac{N^{\sqrt{n}}}{n} \geq a_t^k + \epsilon) \xrightarrow[n\to\infty]{} 0$.*

Lemma 3.35 allows to bound $C_{\max}$ from above:

Corollary 3.36. *For all $\epsilon > 0$, $\mathbb{P}(\frac{C_{\max}}{n} \geq a_t^k + \epsilon) \xrightarrow[n\to\infty]{} 0$.*

For all i, if $C_{\max} \geq i$, then $N^i \geq C_{\max}$. Therefore for all integer i, $C_{\max} \leq \max(N^i, i) \leq i + N^i$. Using $i = \sqrt{n}$ gives Corollary 3.36.

Corollary 3.36 is sufficient to prove Theorem 3.4 in the subcritical or critical case, as in that case $a_t^k = 0$ and $\frac{C_{\max}}{n}$ is a positive random variable.

Proof of Lemma 3.35. Let v_1 and v_2 be two independent uniform random vertices of $G_{n,t}^k$, and i a non-negative integer. Let $C(v_1)$ (resp. $C(v_2)$) denote the component of v_1 (resp. v_2) in $G_{n,t}^k$:

$$\begin{aligned}
\mathbb{P}(|C(v_1)| \geq i | G_{n,t}^k) &= \frac{N^i}{n} \\
\mathbb{P}(|C(v_1)| \geq i) &= \mathbb{E}\left(\frac{N^i}{n}\right) \\
\mathbb{P}(|C(v_1)| \geq i \text{ and } |C(v_2)| \geq i | G_{n,t}^k) &= \left(\frac{N^i}{n}\right)^2 \\
\mathbb{P}(|C(v_1)| \geq i \text{ and } |C(v_2)| \geq i) &= \mathbb{E}\left(\frac{N^i}{n}\right)^2.
\end{aligned}$$

Being in a component of size at least i is a local event, therefore by the birooted local convergence proven in Lemma 3.12:

$$\begin{aligned}
\mathbb{P}(|C(v_1)| \geq i) &\xrightarrow[n\to\infty]{} \mathbb{P}(|T_t^k| \geq i) \\
\mathbb{P}(|C(v_1)| \geq i \text{ and } |C(v_2)| \geq i) &\xrightarrow[n\to\infty]{} \mathbb{P}(|T_t^k| \geq i)^2.
\end{aligned}$$

Therefore, for all i,

$$\begin{aligned}
\mathbb{E}\left(\frac{N^i}{n}\right) &\xrightarrow[n\to\infty]{} \mathbb{P}(|T_t^k| \geq i) \\
\mathrm{Var}\left(\frac{N^i}{n}\right) &\xrightarrow[n\to\infty]{} 0 \\
\text{and therefore } \frac{N^i}{n} &\xrightarrow[n\to\infty]{p} \mathbb{P}(|T_t^k| \geq i).
\end{aligned}$$

N^i is non-increasing in i, and $\mathbb{P}(|T_t^k| \geq i) \xrightarrow[i\to\infty]{} a_t^k$, therefore for all $\epsilon > 0$, $\mathbb{P}\left(\frac{N\sqrt{n}}{n} \geq a_t^k + \epsilon\right) \xrightarrow[n\to\infty]{} 0$. □

3.5.2 The supercritical case

If $a_t^k > 0$, a lower bound is needed. The following lemma gives such a lower bound.

Lemma 3.37. *Let v_1 and v_2 be two independent uniform vertices of G_n^k. Then $\liminf_n \mathbb{P}(v_1$ and v_2 are in the same component$) \geq (a_t^k)^2$.*

Let us first see why Lemma 3.37 implies Theorem 3.4. Let $(C_i)_{i\geq 1}$ be the sequence of the sizes of the component in $G_{n,t}^k$ (in any order). Conditionally on $G_{n,t}^k$, the probability that v_1 and v_2 are in the same component is $\frac{1}{n^2}\sum C_i^2$.

Let $\epsilon > 0$. Lemmata 3.35 and 3.37 imply that for n large enough:

$$\begin{aligned}
(a_t^k)^2 - \epsilon &\leq \frac{1}{n^2}\mathbb{E}(\sum_i C_i^2) \\
&= \frac{1}{n^2}\mathbb{E}\left(\sum_{i:C_i\geq\sqrt{n}} C_i^2 + \sum_{i:C_i<\sqrt{n}} C_i^2\right) \\
&\leq \frac{1}{n^2}\mathbb{E}\left(\sum_{i:C_i\geq\sqrt{n}} C_i^2 + \sqrt{n}\sum_{i:C_i<\sqrt{n}} C_i\right) \\
&\leq \frac{1}{n^2}\mathbb{E}\left(\sum_{i:C_i\geq\sqrt{n}} C_i^2 + \sqrt{n}\cdot n\right) \\
&\leq \mathbb{E}\left(\frac{C_{\max}}{n}\frac{\sum_{i:C_i\geq\sqrt{n}} C_i}{n}\right) + n^{-\frac{1}{2}} \\
&= \mathbb{E}\left(\frac{C_{\max}}{n}\frac{N\sqrt{n}}{n}\right) + n^{-\frac{1}{2}} \\
(a_t^k)^2 - \epsilon &\leq \mathbb{E}\left(\frac{C_{\max}}{n}\right)(a_t^k + \epsilon) + \epsilon \qquad (3.7)
\end{aligned}$$

The last inequality holding for n large enough because $\frac{C_{\max}}{n} \leq 1$ a.s., $\frac{N\sqrt{n}}{n} \leq 1$ a.s. and $\mathbb{P}(\frac{N\sqrt{n}}{n} \leq a_t^k + \epsilon) \to 1$.

By taking $\epsilon \to 0$, (3.7) implies that $\liminf \mathbb{E}(\frac{C_{\max}}{n}) \geq a_t^k$. As $\frac{C_{\max}}{n}$ is smaller than 1 a.s. and smaller than $a_t^k + \epsilon$ with high probability by Corollary 3.36, this implies that $\frac{C_{\max}}{n} \xrightarrow{p} a_t^k = \mathbb{P}(|T_t^k| = \infty)$.

The proof of Lemma 3.37 will follow these steps:

1. With probability close to $(a_t^k)^2$, v_1 and v_2 are in components of $G_{n,t'}^k$ of size larger than a threshold (that depends on the number of vertices) at time $t' = t - \epsilon$;

2. In that case, with probability close to one, a path between w_1 and w_2 will exist at time t in G_t^k.

In order to do the first step, and use the local limit, we need to construct a graph that approximates $T_{t'}^k$ and $G_{n,t'}^k$ simultaneously.

3.5.2.a A few standard results on multitype branching processes

This part will summarize the results on multitype branching processes that will be used, using the results and notations of [Har63]. Let T be a multitype branching process, with type set $\mathbb{X} = [0, \tau]$ for some positive τ. For every integer i, let Z_i be the random set equal to the labels of the elements of ith generation of T. Let $\mathbb{P}_x$ (resp. $\mathbb{E}_x$) denote the probability (resp. expectation) when the root of T has type x. For all $A \subset \mathbb{X}$, let $M(x, A) = \mathbb{E}_x(|A \cup Z_1|)$. We will assume that the branching processes considered always satisfy the following two conditions:

(C1) For all $x \in \mathbb{X}$, $|Z_1| \leq k$ $\mathbb{P}_x$-a.s.

(C2) There exists two positive real numbers a and b such that for all $x \in \mathbb{X}$, $M(x, \cdot)$ is absolutely continuous with respect to the Lebesgue measure, and its density $m(x, y)$ is a uniformly positive bounded function, $0 < a \leq m(x, y) \leq b < \infty$.

Conditions (C1) and (C2) will be direct consequences of Lemmata 3.31 and 3.32 for all branching processes we will consider. (C1) and (C2) implies that T satisfies technical conditions 10.1 and 13.1 with the notations of [Har63].

By [Har63, Theorem 10.1], the operator M has a real positive eigenvalue ρ, larger than any other eigenvalue. This eigenvalue ρ will be called the spectral radius associated to T. Let $q(x) = \mathbb{P}_x(|T| < \infty)$ denote the extinction probability function of T. By [Har63, Theorem 12.1, Theorem 14.1 and following remarks]:

Lemma 3.38.

1. *If $\rho \leq 1$, then for all $x \in \mathbb{X}$, $q(x) = 1$*

2. *If $\rho > 1$:*

   - *For all $x \in \mathbb{X}$, $q(x) < 1$.*
   - *For all $x \in \mathbb{X}$, conditionally on $|T| = \infty$, $\frac{|Z_i|}{\rho^i}$ converges P_x-a.s. toward a random non-zero variable W when i tends to ∞.*

For all non-negative function s, let $\varphi_x(s)$ be defined by:

$$\varphi_x(s) = \mathbb{E}_x \left(\prod_{y \in Z_1} s(y) \right).$$

Remark. With the notation of [Har63], this is actually $\Phi(-\ln s)$, but this version is more suited for our use.

By [Har63, Theorem 15.1]:

Lemma 3.39. *If $\rho > 1$,*

- *q is the only uniformly positive and uniformly less than 1 function satisfying $q(x) = \varphi_x(q)$ for all $x \in \mathbb{X}$.*
- *If q^0 is a positive and uniformly less than 1 function, and if we define the sequence of functions $(q^i)_{i \geq 1}$ by $q^{i+1}(x) = \varphi_x(q^i)$, then q^i converges everywhere toward q.*

Corollary 3.40.

- *If q^0 is a positive and uniformly less than 1 function such that for all x, $q^0(x) \geq \varphi_x(q^0)$, then for all $x \in \mathbb{X}$, $q(x) \leq q^0(x)$.*
- *If q^0 is a positive and uniformly less than 1 function such that for all x, $q^0(x) \leq \varphi_x(q^0)$, then for all $x \in \mathbb{X}$, $q(x) \geq q^0(x)$.*

Proof. We define $(q^i)_{i \geq 0}$ as in Lemma 3.39. Then by Lemma 3.39, for all x, $q^i(x)$ converges toward $q(x)$. For the first part of Corollary 3.40, $q^1(x) \leq q^0(x)$. As $\varphi_x(s)$ is an increasing function of the positive function s, the sequence $(q^i)_{i \geq 0}$ is a non-increasing sequence and is therefore larger than its limit. Similarly, in the second part, the sequence q^i is non-decreasing, and therefore smaller than its limit. □

Let T' be a two stages-branching process, with a different law for the root, and same law as T for the remaining of the tree. Let Z'_i be the random set of labels of the ith generation of T'. Recall that q and ρ denote the extinction probability function and the spectral radius of T.

Lemma 3.41.

- $\mathbb{P}(|T'| < \infty) = \mathbb{E}\left(\prod_{y \in Z'_1} q(y)\right)$.
- *If $\rho > 1$, conditionally on $|T'| = \infty$, $\frac{[Z'_i]}{\rho^i}$ converges toward a random non-zero variable W.*

For such a two-stages branching process T', ρ will be called the spectral radius associated to T'.

Proof. Conditionally on the first generation Z_1, the subtrees starting at elements of Z_1 are independent, so the probability of extinction is $\prod_{y \in Z'_1} q(y)$. The first part of Lemma 3.41 is obtained by taking the expectation. The second part is obtained by using Lemma 3.38 on the surviving subtrees. □

For all t, let ρ_t be the spectral radius associated to $T_t^{k+,\cdot}$ or T_t^k.

Lemma 3.42. *The spectral radius ρ_t is a upper semi-continuous function of t.*

Lemma 3.42 will be proven in part 3.5.6.

3.5.3 Approximating $G_{n,t}^k$ by an idealised graph

Let t be such that $a_t^k > 0$, *i.e.* $\rho_t > 1$. Let $\frac{1}{2} > \epsilon_1 > 0$ and $t' = t - \epsilon_1$. Let $\epsilon_2 > 0$. Let $K_n = \frac{(1-\epsilon_2)\ln n}{\ln(\rho_{t'})}$. Using Lemma 3.42, we can choose ϵ_1 small enough so $\rho_{t'} > 1$.

The graph G^{mod} will be constructed dynamically while exploring $G_{n,t'}^{\infty}$. G^{mod} will denote the growing graph process and $G_{\text{end}}^{\text{mod}}$ the graph G^{mod} at the end of its construction. At any point, G^{mod} will satisfy the following properties:

- G^{mod} is either empty, one planar tree with labelled edges or two planar trees with labelled edges, such that no two edges have the same label.
- In each planar tree, the children of a given vertex are ordered, for the planar tree order, in increasing order of the labels of the outgoing edges.

The graph $G_{\text{end}}^{\text{mod}}$ will have the same law as two independent copies of $T_{t'}^{\infty}$. Informally, $G_{\text{end}}^{\text{mod}}$ will be built in such a way that the components of the roots in $\Phi(G_{\text{end}}^{\text{mod}})$ are close to the components of v_1 and v_2 in $G_{n,t'}^k = \Phi(G_{n,t'}^{\infty})$. In order to achieve this, we define an exploration process, that will only look at the parts of the graph $G_{n,t'}^{\infty}$ that are useful to construct the component of v_1 and v_2 in $G_{n,t'}^k$. At the beginning of its exploration, $G_{n,t'}^{\infty}$ (seen from v_1 and v_2) looks like two independent copies of $T_{t'}^{\infty}$. But, as the exploration continues, these two graphs differ: $G_{n,t'}^{\infty}$ can have cycles or multiple edges,

and the numbers of neighbors of the vertices decreases (in law), as a growing fraction of $G^{\infty}_{n,t'}$ is known. The following construction explains how to couple $G^{\infty}_{n,t'}$ with its idealized branching version $G^{\text{mod}}_{\text{end}}$ in such a way that the differences between $\Phi(G^{\text{mod}}_{\text{end}})$ and $G^{k}_{n,t'}$ are small.

Let V^G denote the set of vertices of $G^{\infty}_{n,t}$. In order to construct G^{mod}, the process will use $G^{\infty}_{n,t'}$ and some extra randomness. Initially, G^{mod} is empty, and will be created dynamically along with the exploration of $G^{\infty}_{n,t'}$. The process will add vertices to G^{mod}, from V^G and extra vertices. The extra vertices will be called *dummy vertices*.

As $G^{\text{mod}}_{\text{end}}$ is two planar trees, we can define the following notions:

Definition 3.19. At any point, G^{mod} is a rooted planar tree or two rooted planar trees. For any vertex w of G^{mod} and integer i, *the i-children* of w are defined as the vertices at distance i of w in the subtree starting at w.

The set of vertices of $G^{\text{mod}}_{\text{end}}$ is ordered with the breadth-first order, denoted by $<$. For any two vertices w and w' of $G^{\text{mod}}_{\text{end}}$, $w < w'$ if either:

- w is in the first component of $G^{\text{mod}}_{\text{end}}$ and w' is in the second component.
- w and w' are in same component, and the distance between w and the root is strictly smaller than the distance between w' and the root.
- w and w' are in the same component, in the same generation, and the father of w is strictly smaller than the father of w' for the breadth-first order.
- w and w' are two children of the same vertex w'', and the label of the edge between w and w'' is smaller than the label of the edge between w' and w''.

The construction of $G^{\text{mod}}_{\text{end}}$ will use the following increasing sets:

- $A^G \subset V^G$ will denote the set of vertices of $G^{\infty}_{n,t'}$ that have already been discovered.
- $B^G \subset A^G$ will denote the set of the vertices whose neighbors are known. These vertices will be called *used*.
- $A^C \subset A^G$ will denote the set of *corrupted* vertices. They indicate where G^{mod} and $G^{\infty}_{n,t'}$ are different.
- $B^C \subset A^C$ will denote the set of corrupted vertices whose children in $G^{\text{mod}}_{\text{end}}$ have already been constructed.
- A^{Dum} will denote the set of dummy vertices. At any point, $A^{Dum} \cap V^G = \emptyset$.

- $B^{Dum} \subset A^{Dum}$ will denote the subset of dummy vertices whose children in $G^{\text{mod}}_{\text{end}}$ have already been constructed.
- $A = A^G \cup A^{Dum}$ is the set of the vertices of G^{mod}. The sets A^G and A^{Dum} are disjoint.
- $B = B^G \cup B^C \cup B^{Dum}$ is the set of the vertices of G^{mod} whose children in G^{mod} have already been constructed. The sets $B^G \cup B^C$ and B^{Dum} are disjoint, but B^G and B^C might not be disjoint.

During the process, all these sets will only increase.

$G^{\text{mod}-}$ is the isomorphism class of the graph G^{mod}, with respect to the isomorphism of graphs with labelled edges and unlabelled vertices. This allows for example to consider $G^{\text{mod}-}$ without knowing if a given vertex is a dummy vertex. As the labels of the edges of G^{mod} are all different, the only automorphism of G^{mod}, seen as a graph with labelled edges and unlabelled vertices, is the identity, and therefore a unique way to map $G^{\text{mod}-}$ back to G^{mod}, allowing us to consider the vertex of G^{mod} associated to a given vertex of $G^{\text{mod}-}$.

Let $\mathscr{F}$ denote the increasing σ-algebra, generated by A^G, B^G, A^C, B^C, A^{Dum}, B^{Dum}, the set of labelled edges with at least one endpoint in B in the graph $G^{\infty}_{n,t'}$ and the set of labelled edges with at least one endpoint in B in $G^{\text{mod}}_{\text{end}}$. $\mathscr{F}$ represents the current knowledge obtained with the exploration of $G^{\infty}_{n,t'}$ and the construction of G^{mod}.

The following tools will be used in the construction of G^{mod}.

3.5.3.a Split a vertex/edge:

There is no multiple edge in $T^{\infty}_{t'}$, whereas there can be multiple edges in $G^{\infty}_{n,t'}$. To solve this issue, any multiple edge will be substituted upon discovery by the appropriate number of simple edges. If a multiple edge e of multiplicity l is discovered between w and w' while looking at the neighbors of w, add w' to G^{mod} and only one edge between w and w', labelled by the smallest label of e. For any label y among the $l-1$ other labels, considered in increasing order, add a dummy vertex w_y to G^{mod} and A^{Dum}, and add an edge between w and w_y labelled by y in G^{mod}. An example can be found in figure 3.3. This operations allows to avoid adding multiple edges to G^{mod} without altering the degree of w (it does alter the degree of w').

3.5.3.b Probing a vertex w:

This operation will be done for a vertex $w \in A^G \setminus B$. The goal is to approximate the set of labelled edges adjacent to w in $G^{\infty}_{n,t'}$ by a Poisson point process of intensity 1 on $[0, t']$.

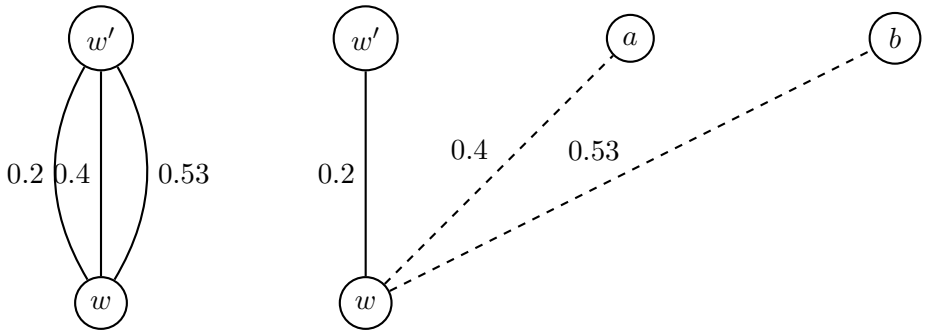

Figure 3.3: The triple edge between w and w' in $G^{\infty}_{n,t'}$ is split in G^{mod} by adding two dummy vertices a and b.

- Let μ_w^{dummy} be an independent Poisson point process of intensity $\frac{|A^G|}{n}$ on $[0, t']$. For each point s of μ_w, in increasing order, add a dummy vertex w' and an edge labelled by s between w and w' to G^{mod}. Add w' to A^{Dum}. If there is at least one such edge, add w to A^C and B^C.
- For every vertex $w' \in V^G \setminus A^G$, look at the edge between w' and w:
  - If there is a single edge between w and w', add this edge and w' to G^{mod}. Add w' to A^G.
  - If there is a multiple edge between w and w', split it. Add w' to A^G and A^C. Add w to A^C and B^C.
  - If there is no edge between w and w', do not do anything.
- For every vertex w' in $A^G \setminus B^G$, do not add the edges between w and w' to G^{mod}. If there is at least one such edge, add w to A^C and B^C and add w' to A^C.
- Add w to B^G.

3.5.3.c Fake-probing a vertex w:

1. If w is a vertex of $A^{Dum} \setminus B^{Dum}$, fake-probing w is the following operations:
   - Add w to B^{Dum}.
   - Let Y_w be an independent Poisson point process of intensity 1 on $[0, t']$. For each point y of Y_w, add a dummy vertex w_y to G^{mod} and A^{Dum}, and add an edge labelled by y between w and w_y to G^{mod}.
2. If w is a vertex of $A^G \setminus B$, fake-probing w is the following operations:
   - Add w to A^C and B^C.

- Let Y_w be an independent Poisson point process of intensity 1 on $[0, t']$. For each point y of Y_w, add a dummy vertex w_y to G^{mod} and A^{Dum}, and add an edge labelled by y between w and w_y to G^{mod}.

For any vertex w, and any set, σ-algebra or graph X, let X_w denote the value of X before the probing or the fake-probing of w. Similarly, let X_{w+} denote the value of X after the probing or fake-probing of w.

The following lemma summarizes the properties that will hold through the construction of $G^{\mathrm{mod}}_{\mathrm{end}}$:

Lemma 3.43.

1. *For every vertex $w' \in G^{mod}_w$, $w' \neq w$, the set of edges adjacent to w' in G^{mod} is not modified during the probing or fake-probing of w.*

If the following properties hold before the probing or the fake-probing of a vertex w in $A_w \setminus B_w$, then the following properties will also hold after the probing or fake-probing of w.

2. *The vertex set of G^{mod} is A.*
3. *For every vertex $w' \in B^G \setminus B^C$, the set of edges adjacent to w' is the same in G^{mod} and $G^{\infty}_{n,t'}$.*
4. *For every vertex $w' \in A \setminus B$, w' has no child in G^{mod}.*
5. *If $w' \in A^G \setminus A^C$, the set of edges between w' and elements of B^G is the same in G^{mod} and in $G^{\infty}_{n,t'}$.*
6. *If $w' \in V^G \setminus A$ then there is no edge between w' and any vertex of B^G in $G^{\infty}_{n,t'}$.*

We assume that the choice of w and of whether to do a probing or a fake-probing are $\mathscr{F}$-measurable. If the following property holds before the probing or fake-probing of w, then it holds after:

7. *Let $E^B \subset E$ be the set of edges with at least one endpoint in B^G. For any $e \in E^B$, the set of labels of e in $G^{\infty}_{n,t'}$ is $\mathscr{F}$-measurable. Conditionally on $\mathscr{F}$, the set of labels of edges in $E \setminus E^B$ is an i.i.d. family of Poisson point processes of intensity $\frac{1}{n}$ on $[0, t']$.*

If the properties 2-7 hold before the probing of w and if one of these two properties holds before the probing of a vertex, they still holds after:

8a. G^{mod} is a tree rooted at $\tilde{v}_1$.

8b. G^{mod} is two trees rooted at $\tilde{v}_1$ and $\tilde{v}_2$.

If the properties 2-7 hold before the probing of w, then

9. *conditionally on $\mathscr{F}_w$, the law of set of all labels of edges between w and the children of w in G^{mod}, including dummy vertices, is a Poisson point process of intensity 1 on $[0, t']$.*

Proof. The properties 1-6 and 8a/8b are direct consequences of the description of the probing and fake-probing of a vertex.

If $w \in A^G \setminus B^G$ is probed, the only vertex added to B^G during the probing of w is w, and the edges of $G^{\infty}_{n,t'}$ looked at are the edges adjacent to w. If w is fake-probed, no vertex is added to B^G and no edge is looked at during the fake-probing of w. Therefore property 7 holds after the probing if it holds before the probing.

If $w \in A$ is probed, by property 7, conditionally on $\mathscr{F}_w$, the set of labels of edges between w and elements of $V^G \setminus |A^G|$ is a Poisson point process of intensity $\frac{n-|A^G|}{n}$. Adding an independent Poisson point process of intensity $\frac{|A^G|}{n}$ gives the wanted distribution for the property 9. If w is fake-probed, property 9 is a direct consequence of the definition of the fake-probing. □

Each time several vertices are probed (or fake-probed) at the same time by the algorithm, they are probed in breadth-first order.

3.5.3.d The initialisation of a component:

This tool is used at the beginning of the construction of each of the two components of $G^{\text{mod}}_{\text{end}}$. Let $j \in \{1, 2\}$. The aim of this part is to construct the ball of radius b_n centered at $\tilde{v}_j$ in G^{mod}.

The description assume that, before the initialisation, the properties 2-7 of Lemma 3.43 hold and that either G^{mod} is empty (if $j = 1$) or 8a holds (if $j = 2$).

For every i, j, $S_{j,i}$ is the sphere of radius i centered at $\tilde{v}_j$ in $G^{\text{mod}}_{\text{end}}$. Let $\mathscr{B}_{1,i} = \cup_{i' \leq i} S_{1,i}$ and $\mathscr{B}_{2,i} = \cup_{i' \leq i} S_{2,i} \cup G^{\text{mod}}_{\text{end},1}$. For any $w \in G^{\text{mod}}$, generation(w) is the ordered pair (j, i) such that $w \in S_{j,i}$.

Definition 3.20 (Initialisation failed)**.** If one of the following events happen, the initialisation is said to have failed:

- A previous initialisation has failed.
- Before the initialisation, v_j was already in A^G.
- At any point of the construction of a $S_{j,i}$, a vertex is added to A^C.

If $v_j \in A^G$ or if a previous initialisation has failed, then add to G^{mod} and A^{Dum} a dummy vertex $\tilde{v}_j$, as the root of the second component. Otherwise add v_j to A^G and to G^{mod} and let $\tilde{v}_j = v_j$.

Let $0 \leq i < b_n$. We assume that $S_{j,i}$ has already been built. If the initialisation has not yet failed, probe every vertex w of $S_{j,i}$. If the initialisation has failed, fake-probe every vertex w of $S_{j,i}$. In both cases, $S_{j,i+1}$ is the set of children of elements of $S_{j,i}$ in G^{mod}, *i.e.* the sphere of radius $i+1$ in G^{mod} centered at $\tilde{v}_j$.

Starting with $S_{j,0} = \{\tilde{v}_j\}$, the sets $S_{j,i}$ are built until S_{j,b_n} has been built. For any set, σ-algebra of graph X, let $X_{j,i}$ denote the value of X after the construction of $S_{j,i}$.

Lemma 3.44.

- *After the construction of $S_{j,i}$, if the initialisation has not yet failed, $S_{j,i} \subset A^G \setminus (B^G \cup A^C)$.*
- *After the construction of $S_{j,i}$, for every $i' < i$, $S_{j,i'} \subset B$.*

These properties are direct consequences of the algorithm: if at any point a vertex is added to A^C, the initialisation is said to have failed. The first property allows us to probe the elements of $S_{j,i}$ when the initialisation has not yet failed.

3.5.3.e Subsequent construction of the component of v_j in G^{mod}:

If the initialisation has not failed, the balls of radius b_n centered at v_j are the same in G^{mod} and in $G^{\infty}_{n,t'}$, by property 3 of Lemma 3.43. Therefore, if $GE_1 = 1$, it is possible to determine the neighbors of v_j in $G^k_{n,t'}$ (by Lemma 3.23). Let $D_{j,0} = \{v_j\}$, $E_{j,0} = \{v_j\}$ and $D_{j,1}$ be the set of neighbors of v_j in $G^k_{n,t'}$. If the initialisation has failed or if it is not possible to determine the neighbors of v_j in $G^k_{n,t'}$, let $D_{j,1} = D_{j,0} = E_{j,0} = \emptyset$.

After the initialisation, the component of $\tilde{v}_j$ in G^{mod} is a planar tree rooted at $\tilde{v}_j$ of depth at most b_n, the vertices of the $b_n - 1$ first generations belonging to B and the vertices of the b_nth generation belonging to $A \setminus B$.

3.5.3.f One iteration of the main part of the algorithm:

We now describe the iteration i, constructing the sets $E_{j,i}$, $D_{j,i+1}$. For all j, i, $E_{j,i} \subset D_{j,i} \subset S_{j,i}$. For any element w of $D_{j,i}$, if we are able to determine the children of w in the forbidden degree version of G^{mod}, w is added to $E_{j,i}$ and its children are added to $D_{j,i+1}$. More precisely, assuming $D_{j,i}$ and $\mathscr{B}_{j,b_n+i-1}$ are known, the iteration i is the following operations:

- Initially, $E_{j,i}$ and $D_{j,i+1}$ are empty.
- For each $w \in S_{j,i+b_n-1}$, in the breadth-first order, if w is a $b_n - 1$-children of an element of $D_{j,i}$, probe w. Otherwise, fake-probe w.

- Consider every w in $D_{j,i}$ such that no (b_n-1)-child or b_n-child of w is in A^C. If the knowledge of G^{mod} is sufficient to determine the set N of neighbors of w in $G^k_{n,t'}$ add w to $E_{j,i}$, and add every element of $N \setminus p(w)$ to $D_{j,i+1}$ where $p(w)$ is the parent of w in G^{mod}.

Remark. If $GE_2 = 1$, the knowledge of the ball of radius b_n centered at w in $G^\infty_{n,t'}$ is sufficient to know the set N. Lemma 3.45 will prove that the balls of radius b_n centered at w in $G^{\text{mod}}_{\text{end}}$ and $G^\infty_{n,t'}$ are identical and therefore that the knowledge of G^{mod} is sufficient to know the set N if $GE_2 = 1$. Lemma 3.45 will also prove that for every b_n1-children of elements of $D_{j,i}$ is in $A^G_{j,i+b_n}$, and therefore can be probed.

An iteration is the execution of the algorithm above, for a given value of i. For any set, σ-algebra or graph X, let $X_{j,i+b_n}$ denote the value of X after the ith iteration in the construction of the component of $\tilde{v}_j$, *i.e.* the construction of $E_{j,i}$ and $D_{j,i+1}$.

3.5.3.g Finalisation of a component:

After K_n iterations of this algorithm are done, and $\mathscr{B}_{j,K_n+b_n}$ is constructed, graft to every vertex of S_{j,b_n+K_n} an independent copy of $T^\infty_{t'}$. Add every added vertex to A^{Dum} and B^{Dum}. Add every vertex of S_{j,b_n+K_n} to A^C and B^C.

The value of the sets and σ-algebra X after the finalisation of the jth component will be denoted by $X_{j,\infty}$. If the properties 2-7 and either 8a or 8b hold before the finalisation of a component, they hold after the finalisation of a component, as no edge of $G^k_{n,t'}$ is looked at, and the copies are independent.

3.5.3.h Construction of $G^{\text{mod}}_{\text{end}}$:

The construction of $G^{\text{mod}}_{\text{end}}$ is done by starting with $A^G = B^G = A^C = B^C = A^{Dum} = B^{Dum} = \emptyset$ and doing the following operations:

1. First initialisation, for the component of $\tilde{v}_1$.
2. Do the iteration i of the algorithm for i from 1 to K_n, to construct $\mathscr{B}_{1,b_n+i}$ and the sets $E_{1,i}$ and $D_{1,i+1}$ for every $i \leq K_n$.
3. Finalise the construction of the first component.
4. Second initialisation, for the component of $\tilde{v}_2$.
5. Do the iteration i of the algorithm for i from 1 to K_n, to construct $\mathscr{B}_{2,i+b_n}$ and the sets $E_{2,i}$ and $D_{2,i+1}$ for every $i \leq K_n$.
6. Finalise the construction of the second component.

For any graph, σ-algebra of set X, X_{end} denotes the value of X after the complete algorithm.

3.5.3.i First properties of $G^{\mathbf{mod}}$

Lemma 3.45. *For any $i < b_n$, $j \in \{1,2\}$, after the construction of $S_{j,i}$ in the initialisation:*

1. *the properties 2-7 of Lemma 3.43 hold;*
2. *if $j = 1$, 8a holds; if $j = 2$, 8b holds.*

For every $i \leq K_n$ and $j \in \{1,2\}$, after the iteration i constructing $E_{j,i}$ and $D_{j,i+1}$:

3. *for every vertex w of $E_{j,i}$, the ball of radius b_n centered at w is the same in G^{mod} and $G^{\infty}_{n,t'}$;*
4. *every vertex of $D_{j,i+1}$ is a child of a vertex of $E_{j,i}$ in G^{mod};*
5. *the set of vertices of $G^{mod}_{j,i+b_n}$ is $\mathscr{B}_{j,i+b_n}$;*
6. *$A_{j,i+b_n} = \mathscr{B}_{j,i+b_n}$;*
7. *$B_{j,i+b_n} = \mathscr{B}_{j,i+b_n-1}$;*
8. *the properties 2-7 of Lemma 3.43 hold;*
9. *if $j = 1$, 8a holds; if $j = 2$, 8b holds.*

After the construction of G^{mod}_{end}:

10. *G^{mod-}_{end} has same law as two independent copies of $T^{\infty}_{t'}$. Moreover the second component of G^{mod-}_{end} is independent of $\mathscr{F}_{1,\infty}$.*

Proof. The properties 1, 2, 8 and 9 are consequences of Lemma 3.43 and the fact that the properties 2-7 of Lemma 3.43 are preserved by the initialisation or the finalisation of a component. The properties 4, 5, 6 and 7 are consequences of the description of the algorithm, as every vertex of $S_{j,i+b_n-2}$ is either probed or fake-probed for the construction of $D_{j,i}$.

The property 3 is proven by induction. As noted in the subsection 3.5.3.e, property 3 holds for $i = 0$ and $j \in \{1,2\}$. Let us assume that the property 3 holds for $i \geq 0$ and $j \in \{1,2\}$. Let w be a vertex of $E_{j,i+1} \subset D_{j,i+1}$. By property 4, w is a children of a vertex w' of $E_{j,i}$. As the ball of radius b_n centered at w' is the same in G^{mod} and $G^{\infty}_{n,t'}$, the only possible difference between the balls of radius b_n centered at w is the neighbors of the $b_n - 1$-children of w. On one hand, if none of the $b_n - 1$-children or the b_n-children of w are added to A^C, then every ball of radius 1 centered at such children is the same in G^{mod} and in $G^{\infty}_{n,t'}$ and therefore the ball of radius b_n centered in w is the same in G^{mod} and $G^{\infty}_{n,t'}$. One the other hand, if one such vertex is added to A^C, w is not added to $E_{j,i}$.

To prove property 10, it is sufficient to show that for any i and j, conditionally on $\mathscr{B}_{j,i}$, the law of the set of labels of the edges outgoing of the vertices of $S_{j,i}$ is an i.i.d. family of Poisson point processes of intensity 1 on $[0, t']$. If $i < b_n$, conditionally on $\mathscr{F}_{j,i}$, if the initialisation have failed at an earlier stage, every fake-probing gives the correct law. If the initialisation has not yet failed, the algorithm is going to probe each vertex of $S_{j,i}$. By Lemma 3.43, the conditional law of the set of exiting edges is also a Poisson point process of intensity 1 on $[0, t']$ (some of them might point to dummy vertices). As $\mathscr{B}_{j,i}$ is measurable with respect to $\mathscr{F}_{j,i}$, the conditional law is the same conditioned on $\mathscr{B}_{j,i}$.

For all $i \in \{b_n, \dots, K_n + b_n - 1\}$, $\mathscr{B}_{j,i}$ is measurable with respect to $\mathscr{F}_{j,i}$. Each vertex of the ball of radius i is either probed or fake-probed. Therefore, by property 7 of Lemma 3.43, conditionally on $\mathscr{F}_{j,i}$, or on $\mathscr{B}_{j,i}$, the set of labels of edges between each vertex of the ball of radius i and their children is an i.i.d. family of Poisson point processes of intensity 1 on $[0, t']$.

If $i \geq K_n + b_n$, the law of the subsequent generations comes from the grafting of independent copies of $T_{t'}^{\infty}$, and gives the correct conditional law by the branching properties of $T_{t'}^{\infty}$. □

$G_{\text{end}}^{\text{mod}}$ has the law of two independent copies of $T_{t'}^{\infty}$. Therefore $G_{\text{end}}^{\text{mod}} \in \Omega^+$ a.s. Let $G_{\text{end}}^{\text{modk}}$ denote the components of $\tilde{v}_1$ and $\tilde{v}_2$ in $\Phi(G_{\text{end}}^{\text{mod}})$, the forbidden degree version of G^{mod}. For any $i \leq K_n$, $j \in \{1, 2\}$, let $\mathscr{D}_{j,i}^-$ be the union of the sets $D_{j',i'}$ where (j', i') is strictly smaller than (j, i) for the lexicographic order. The set $\mathscr{D}_{j,i}^-$ is the union of the sets $D_{j',i'}$ constructed before $D_{j,i}$. Let $\mathscr{D}_{j,i}^+ = \mathscr{D}_{j,i}^- \cup D_{j,i}$. The following lemma gives a few useful observations about the sets $D_{j,i}$ and $E_{j,i}$ and the graph G^{mod} produced by the algorithm and allows to link them with the component of v_1 and v_2 in $G_{n,t'}^k$:

Lemma 3.46.

1. *Every pair of vertices of $\mathscr{D}_{2,K_n}^+$ that are neighbors in G_{end}^{mod} are also neighbors in $G_{n,t'}^k$ and in G_{end}^{modk}, and the edge between them has the same label in $G_{n,t'}^k$ and G_{end}^{modk}.*

2. *Every vertex of $D_{j,i}$ belongs to the component of v_j in $G_{n,t'}^k$ and in G_{end}^{modk}.*

3. *For $i \leq K_n$ and $j \in \{1, 2\}$, every neighbor of vertices of $E_{j,i}$ in $G_{n,t'}^k$ (resp. in G_{end}^{modk}) belong to either $D_{j,i+1}$ or $E_{j,i-1}$.*

The first and third observations come from the definition of $D_{j,i}$. The second observation is consequence of the first observation of Lemma 3.46 and the property 4 of Lemma 3.45.

The first two observations guarantee that the subgraph of $G_{\text{end}}^{\text{mod}}$ restricted to vertices of $\mathscr{D}^+_{2,K_n}$ can be seen as a subgraph of $G^k_{n,t'}$. The last property guarantees that the degree of vertices of $E_{j,i}$ is the same in this subgraph and in $G^k_{n,t'}$.

3.5.4 Finding a path in $G^k_{n,t}$

Definition 3.21. For every $i < K_n$ and $j \in \{1,2\}$, $\tilde{E}_{j,i}$ is the set of vertices w of $E_{j,i}$ such that no edge is added to w between t' and t in $G^\infty_{n,\cdot}$. Let $\tilde{E}_{j,K_n} = E_{j,K_n}$.

The sets $\hat{E}_{j,i}$ are defined by induction:

- For $j \in \{1,2\}$, $\hat{E}_{j,0} = \tilde{E}_{j,0}$.
- For $j \in \{1,2\}$, $i \in \{1,\dots,K_n\}$, $\hat{E}_{j,0} = \{u \in \tilde{E}_{j,i} : p(u) \in \hat{E}_{j,i-1}\}$, where $p(u)$ denotes the parent of u in $G_{\text{end}}^{\text{modk}}$.

The set $\hat{E}_{j,i}$ is the subset of vertices u of $\tilde{E}_{j,i}$ such that no edge is added between t' and t to any vertex in the path from $\tilde{v}_j$ to u (including u if $i < K_n$, excluding u if $i = K_n$). By Lemma 3.46, any neighbor in $G_{\text{end}}^{\text{modk}}$ of a vertex v of $E_{j,i}$ is either in $D_{j,i+1}$ or in $D_{j,i-1}$. Therefore the following lemma holds:

Lemma 3.47. *Starting from G_{end}^{modk}, remove every elements of $\cup D_{j,i} \setminus \tilde{E}_{j,i}$. Let $\hat{G_{end}^{modk}}$ be the union of the components of v_1 and v_2 in the resulting graph. Then the set of vertices in the balls of radius K_n centered at v_1 and v_2 in $\hat{G_{end}^{modk}}$ is $\cup \hat{E}_{j,i}$.*

Lemma 3.48.

1. *For $j \in \{1,2\}$, every vertex of the ball of radius $K_n - 1$ in $\hat{G_{end}^{modk}}$ is in the component of v_j in the graph $G^k_{n,u}$ for any $u \in [t',t]$.*
2. *For $j \in \{1,2\}$, if w of $\hat{E}_{j,K_n}$ is such that*
   - *the degree of w in G_{end}^{modk} is smaller than $k-2$, and*
   - *at most one edge is added to w in $G^\infty_{n,\cdot}$ between t' and t*

   then w is in the component of v_j in $G^k_{n,u}$ for any $u \in [t',t]$.

By Lemma 3.46, all the elements of $\cup_i \tilde{E}_{j,i}$ and the edges between them in $G_{\text{end}}^{\text{modk}}$ are present in $G^k_{n,t'}$. Therefore $\hat{G_{\text{end}}^{\text{modk}}}$ is a subgraph of $G^k_{n,t'}$. As by definition of $\tilde{E}_{j,i}$ no edge is added to any vertex in the balls of radius $K_n - 1$ between t' and t, no vertex of the balls of radius K_n reaches the forbidden degree between t' and t and therefore the ball of radius $K_n - 1$ is a subgraph of $G^k_{n,u}$ for any $u \in [t',t]$.

If w is a non-saturated vertex of $\hat{G^{\text{modk}}_{\text{end}}}$ such that at most one edge is added to w in $G^{\infty}_{n,\cdot}$ between t' and t, the edge between w and its parent in $G^{\text{modk}}_{\text{end}}$ is present in $G^{k}_{n,u}$ for any $u \in [t', t]$.

Theorem 3.49. *For any $\eta > 0$, for ϵ_1 and ϵ_2 small enough and n large enough, with probability at least $(a^k_t)^2 - \eta$, there exists a vertex w_1 in $\hat{E}_{1,K_n}$ and a vertex w_2 in $\hat{E}_{2,K_n}$ such that:*

- *The vertices w_1 and w_2 are non-saturated in $G^k_{n,t'}$.*
- *There is an edge added between time t' and time t between w_1 and w_2 in $G^{\infty}_{n,\cdot}$.*
- *No other edge is added between time t' and time t to either w_1 or w_2 in $G^{\infty}_{n,\cdot}$.*

Theorem 3.49 implies Lemma 3.37, as by Lemma 3.48, w_1 and w_2 are respectively in the component of v_1 and v_2 in $G^k_{n,t}$ and the edge between v_1 and v_2 is present in $G^k_{n,t}$.

To prove this result, the following facts will be proven:

1. For ϵ_2 small enough, with probability larger than $(a^k_t)^2 - \eta$, for all j, there are at least $n^{\frac{2}{3}}$ non saturated vertices in $|\hat{E}_{j,K_n}|$.
2. The probability that the previous condition happens without the existence of w_1 and w_2 satisfying the conditions of Theorem 3.49 converges to 0.

It will be done by finding a subgraph of $\hat{G^{\text{modk}}_{\text{end}}}$ that is a branching process with offspring law close to the law of $T^k_{t'}$. Therefore, the probability of survival of each component is close to $a^k_{t'}$, and conditionally on survival, the components will be large enough.

Lemma 3.50. *Let $z_n = n^{1-\frac{\epsilon_2}{2}}$. With high probability, at any point of the algorithm $|A^G| \leq z_n$.*

Every element of A^G belongs to a ball of radius b_n centered at an element of $\cup_{j,i} D_{j,i}$ in $G^{\infty}_{n,t'}$. Moreover, by Lemma 3.23, if $GE_1 = 1$, there exists $x_n = n^{o(1)}$ such that every ball of radius b_n in $G^{\infty}_{n,t'}$ contains less than x_n vertices. Therefore,

$$GE_1 |A^G_{\text{end}}| \leq x_n \sum_{1 \leq j \leq 2, 0 \leq i \leq K_n} |D_{j,i}|.$$

As every element of $D_{j,i}$ belongs to the ball of radius i centered at $\tilde{v}_j$ in $G^{\text{modk}}_{\text{end}}$, and $G^{\text{modk}}_{\text{end}}$ has the same law as $T^k_{t'}$:

$$\mathbb{E}\Big(\sum_{1 \leq j \leq 2, 0 \leq i \leq K_n} |D_{j,i}|\Big) \leq 2\mathbb{E}(|B_{K_n}(T^k_{t'}, \emptyset)|)$$

where $B_{K_n}(T_{t'}^k, \emptyset)$ is the ball of radius K_n centered at the root $\emptyset$. As $T_{t'}^k$ is a (two-stages) supercritical branching process, by Lemma 3.41:

$$\frac{|B_i(T_{t'}^k, \emptyset)|}{\rho_{t'}^i} \xrightarrow[i\to\infty]{} W.$$

with W a random variable with a finite expectation. As $K_n = \frac{(1-\epsilon_2)\ln n}{\ln(\rho_{t'})}$, this limit implies that

$$\begin{aligned}
\sum_{1\leq j\leq 2, 0\leq i\leq K_n} \mathbb{E}|D_{j,i}| &\leq n^{1-\epsilon_2+o(1)} \\
\mathbb{E}(GE_1|A_{\text{end}}^G|) &\leq x_n n^{1-\epsilon_2+o(1)} \\
&= n^{1-\epsilon_2+o(1)}
\end{aligned}$$

Therefore, with high probability, $|A_{\text{end}}^G| \leq z_n$. Let GE_2 be the random variable equal to 1 if this inequality holds and $GE_1 = 1$. Otherwise, let $GE_2 = 0$. As A^G is an increasing subset, $|A^G| \leq |A_{\text{end}}^G|$ at any point of the algorithm.

Lemma 3.51. *With high probability, all the vertices in the balls of radius $K_n + b_n$ centered at $\tilde{v}_1$ and $\tilde{v}_2$ in G_{end}^{mod} have less than $\log n$ children. Let GE_3 be the random variable equal to 1 if this property holds and $GE_2 = 1$. Otherwise, let $GE_3 = 0$.*

Proof. The expected number of vertices in the sphere of radius i centered at $\tilde{v}_j$ in G^{mod} is t'^i. Therefore the expected total volume of each ball of radius $K_n + b_n$ centered at $\tilde{v}_1$ and $\tilde{v}_2$ is equal to $\frac{t'^{K_n+b_n}-1}{t'-1} = o(n^{\alpha_1})$ for some $\alpha_1 > 0$. Therefore it is smaller than n^{α_1} with high probability by Markov's inequality. We assume that n is large enough such that $\log n > 2t'$. The law of the degree of a given vertex is a Poisson random variable of parameter t', therefore the probability that one of the first n^{α_1} vertices of each component on $G_{\text{end}}^{\text{mod}}$ (in the breadth-first order) has degree larger than $\log n$ is smaller than

$$\begin{aligned}
2n^{\alpha_1} \sum_{i=\lfloor \log n \rfloor}^{\infty} e^{t'} \frac{t'^i}{i!} &\leq 2n^{\alpha_1} e^{-t'} \frac{t'^{\log_n}}{\lfloor \log_n \rfloor !} \underbrace{\sum_{i\geq 0} \frac{t'^i}{(2t')^i}}_{=2} \\
&= 4n^{\alpha_1} e^{-t'} \frac{t'^{\log_n}}{\lfloor \log_n \rfloor !} \\
&\sim 4e^{-t'} n^{\alpha_1} n^{\log t'} \frac{1}{\sqrt{2\pi \lfloor \log n \rfloor}} \left(\frac{e}{\lfloor \log n \rfloor} \right)^{\lfloor \log n \rfloor} \\
&\leq \alpha_2 n^{\alpha_3 - \log\log n} \\
&\to 0
\end{aligned}$$

With α_2 and α_3 two constants. The Stirling's approximation is used for the factorial. □

Lemma 3.52. *With high probability, no initialisation fails.*

As A is initially empty, it is not possible that $v_1 \in A$ before the first initialisation. If $GE_2 = 1$, the size of the set A is smaller than z_n before the second initialisation. As v_2 is a uniform vertex independent of $\mathscr{F}_{1,\infty}$, the probability that v_2 belongs to A before its initialisation, conditionally on $\mathscr{F}_{1,\infty}$ before the second initialisation is equal to $\frac{|A^G|}{n}$. This quantity is smaller than $\frac{z_n}{n}$ if $GE_2 = 1$, therefore

$$\begin{aligned}
\mathbb{P}(v_2 \in A_{1,\infty}) &\leq \mathbb{P}(GE_2 = 0) + \mathbb{P}(GE_2 = 1 \text{ and } v_2 \in A_{1,\infty}) \\
&= \mathbb{P}(GE_2 = 0) + \mathbb{E}(GE_2 \frac{|A_{1,\infty}|}{n}) \\
&\leq \mathbb{P}(GE_2 = 0) + \frac{z_n}{n} \\
&= o(1)
\end{aligned}$$

A vertex is added to A^C for the first time during the construction of $S_{j,i+1}$ if one of the following event happens:

1. Two vertices of $S_{j,i}$ are linked by an edge in $G^{\infty}_{n,t'}$.
2. A vertex of $S_{j,i}$ is linked to a vertex of $A^G_{j,i} \setminus B^G_{j,i}$ by an edge in $G^{\infty}_{n,t'}$.
3. Two vertices of $S_{j,i}$ are linked by an edge to the same vertex $w \notin S_{j,i-1}$ in $G^{\infty}_{n,t'}$.
4. There is a multiple edge in $G^{\infty}_{n,t'}$ between a vertex w of $S_{j,i}$ and a vertex in $V^G \setminus A_{j,i}$.
5. When a vertex w is probed, the Poisson point process of intensity $\frac{|A^G|}{n}$ creating dummy vertices is non-empty.

By Lemma 3.45, conditionally on the σ-algebra $\mathscr{F}_{j,i}$, the labels of the edges in $G^{\text{mod}}_{n,t'}$ with no endpoint in $B_{j,i}$ are an i.i.d. family of Poisson point processes if intensity $\frac{1}{n}$ on $[0, t']$. Therefore, conditionally on $\mathscr{F}_{j,i}$:

1. the probability of the first event is smaller than $\binom{|S_{j,i}|}{2}\mathbb{P}(Poi(\frac{t'}{n}) \geq 1)$;
2. The probability of the second event is smaller than $|S_{j,i}||A^G_{i,j}|\mathbb{P}(Poi(\frac{t'}{n}) \geq 1)$;
3. the probability of the third event is smaller than $n\binom{|S_{j,i}|}{2}\mathbb{P}(Poi(\frac{t'}{n}) \geq 1)^2$ (there are less than n possible choices for w);

4. the probability of the fourth event is smaller than $|S_{j,i}|n\mathbb{P}(Poi(\frac{t'}{n}) \geq 2)$;

5. the probability of the last event is smaller than $\frac{t'|S_{j,i}||A^G_{j,i+1}|}{n}$.

Recall that $GE_3 = 1$ implies that $|S_{j,i}| \leq x_n$ and $|A^G_{j,i+1}| \leq z_n$. Let $y_n = n^{-\frac{\epsilon_2}{3}}$. Either $GE_3 = 0$ or the sum of these probabilities is smaller than $\frac{y_n}{2b_n}$, for n large enough. By union bound, the probability than one of these events happens for $i \leq b_n$ during one of the initialisations and $GE_2 = 3$ is smaller than $y_n = o(1)$.

Let GE_4 be the random variable equals to 1 if no initialisation fails and $GE_3 = 1$. Otherwise, let $GE_4 = 0$.

3.5.5 Construction of an included branching process

The goal of this part is to find a branching process simultaneously included in $G^{\text{mod}}_{\text{end}}$, $G^k_{n,t'}$ and $G^k_{n,t}$ with law close to T^k_t.

3.5.5.a Construction of the included branching process for time t'

The idea is to find a two-stages branching process whose set of vertices is a subset of $\cup\tilde{E}_{j,i}$ and whose law will be close to T^k_t. In order to do this, some bounds of the probability that a given vertex w is in $D_{j,i} \setminus \tilde{E}_{j,i}$ will be needed. For technical reasons, $w \in D_{j,i} \setminus E_{j,i}$ and $w \in E_{j,i} \setminus \tilde{E}_{j,i}$ will be treated separately. Let us add the following families of random variables, that will be used as extra randomness when needed, such that, conditionally on $\mathscr{F}_{\text{end}}$, these families are i.i.d. families of uniform random variables on $[0,1]$.

- $(U^a_w)_{w \in G^{\text{mod}}_{\text{end}}}$;
- $(U^b_{w,w'})_{w \leq w' \in G^{\text{mod}}_{\text{end}}}$;
- $(U^c_{w,w'})_{w \leq w' \in G^{\text{mod}}_{\text{end}}}$;
- $(U^d_w)_{w \in G^{\text{mod}}_{\text{end}}}$;

These random variables will be used to construct independent variables, by using the following straightforward construction:

Lemma 3.53. *If U and V are random variables, $\mathscr{F}$ a σ-algebra and $p \in [0,1)$ such that:*

- *$V \in \{0,1\}$ a.s.*
- *$\mathbb{P}(V = 1|\mathscr{F}) < p$ a.s.*

- *U is a uniform random variable on $[0,1]$, independent of $\mathscr{F}$.*

Then the random variable V^{ind} equals to 1 if either $V=1$ or $U < \frac{p-\mathbb{P}(V=1|\mathscr{F})}{\mathbb{P}(V=0|\mathscr{F})}$ is a Bernoulli random variable of parameter p, larger than V, independent of $\mathscr{F}$.

For any $w \in G_{\text{end}}^{\text{mod}}$, let us define the following sets:

- $\Gamma_w^a = \{w' \in G_{\text{end}}^{\text{mod}}; w' < w\}$.
- $\Gamma_{w+}^a = \{w' \in G_{\text{end}}^{\text{mod}}; w' \leq w\} = \Gamma_w^a \cup \{w\}$.
- $\Gamma_w^b = \{(w', w'') \in {G_{\text{end}}^{\text{mod}}}^2; w' \leq w'' < w\}$.
- $\Gamma_{w+}^b = \{(w', w'') \in {G_{\text{end}}^{\text{mod}}}^2; w' \leq w'' \leq w\}$.
- $\Gamma_w^c = \{(w', w'') \in {G_{\text{end}}^{\text{mod}}}^2; w' \leq w'', w' < w, \text{generation}(w'') \leq \text{generation}(w)\}$.
- $\Gamma_{w+}^c = \{(w', w'') \in {G_{\text{end}}^{\text{mod}}}^2; w' \leq w'', w' \leq w, \text{generation}(w'') \leq \text{generation}(w)\}$.
- $\Gamma_w^d = \{w' \in G_{\text{end}}^{\text{mod}} : \text{generation}(w') + (0, b_n) \leq \text{generation}(w)\}$.

These sets will denote the set of uniform variables that have been used via Lemma 3.53 when w is probed or fake-probed. For every $w \in B_{\text{end}}^G$, let $\mathscr{F}_w^{aug}$ be the σ-algebra generated by $\mathscr{F}_w$, $(U_{w'}^a)_{w' \in \Gamma_w^a}$, $(U_{(w',w'')}^b)_{(w',w'') \in \Gamma_w^b}$, $(U_{(w',w'')}^c)_{(w',w'') \in \Gamma_w^c}$ and $(U_{w'}^d)_{w' \in \Gamma_w^d}$.

For every (j,i), $i \geq b_n$, let $\mathscr{F}_{j,i}^{aug}$ be the σ-algebra generated by $\mathscr{F}_{j,i}$, $(U_w^a)_{w \in \mathscr{B}_{j,i-1}}$, $(U_{w,w'}^b)_{(w,w') \in \mathscr{B}_{j,i-1}}$, $(U_{w,w'}^c)_{(w,w') \in \mathscr{B}_{j,i-1}}$ and $(U_w^d)_{w \in \mathscr{B}_{j,i-b_n}}$.

For every vertex $w \in B_{\text{end}}^G$, let $\mathscr{F}_{w+}^{aug}$ be the σ-algebra generated by $\mathscr{F}_w$, $(U_{w'}^a)_{w' \in \Gamma_{w+}^a}$, $(U_{(w',w'')}^b)_{(w',w'') \in \Gamma_{w+}^b}$, $(U_{(w',w'')}^c)_{(w',w'') \in \Gamma_{w+}^c}$ and $(U_{w'}^d)_{w' \in \Gamma_{w+}^d}$.

For any w probed or fake-probed during the construction of $E_{j,i}$ (therefore such that $\text{generation}(w) = (j, i+b_n-1)$), $\mathscr{F}_{j,i+b_n-1}^{aug} \subset \mathscr{F}_w^{aug} \subset \mathscr{F}_{w+}^{aug} \subset \mathscr{F}_{j,i+b_n}^{aug}$.

Lemma 3.54. *For every $w \in B_{end}$, every $\mathscr{F}_{end}$-measurable random variable X, the distribution of X is the same conditionally on $\mathscr{F}_w$ (resp. $\mathscr{F}_{w+}$) and conditionally on $\mathscr{F}_w^{aug}$ (resp. $\mathscr{F}_{w+}^{aug}$).*

For every $i \in \{a,b,c,d\}$, every $w \in B_{\text{end}}^G$, Γ_w^i is $\mathscr{F}_w$-measurable and Γ_{w+}^i is $\mathscr{F}_{w+}$-measurable. Lemma 3.54 comes from the fact that the law of every U-variables is independent of $\mathscr{F}_{\text{end}}$. Lemma 3.54 simply means that any event that can be expressed without the extra randomness is independent of the extra randomness.

In order to construct the included branching process, one needs to study when a vertex belongs to $\cup D_{j,i} \setminus E_{j,i}$. Assuming $GE_1 = 1$, a vertex w_1 is in $D_{j,i} \setminus E_{j,i}$ for some (i,j) if a b_n-1-child or a b_n-child of w_1 is added to A^C during the construction of $D_{j,i+1}$. Let w_2 be a b_n-1-child of w_1. During the construction of $D_{j,i+1}$, only b_n-1-children of elements of $D_{j,i}$ are probed.

- The vertex w_2 is added to A^C while probed if one of the following happens:
  - There exists a multiple edge between w_2 and another vertex in $G^{\infty}_{n,t'}$.
  - When w_2 is probed, $\mu^{dummy}_{w_2}$ is not empty.
  - There is an edge between w_2 and a vertex w_3 of $A^G_{w_2} \setminus B^G_{w_2}$.
- The vertex w_2 is added to A^C while another vertex w_3 is probed if there is an edge between w_2 and w_3 in $G^{\infty}_{n,t'}$.
- A child w_3 of w_2 is added to A^C while w_2 is probed if there is a multiple edge between w_2 and w_3 in $G^{\infty}_{n,t'}$.
- A child w_3 of w_2 is added to A^C while another vertex w_4 is probed if there is an edge between w_4 and w_3 in $G^{\infty}_{n,t'}$.

For any $(i,j) \in \{0,\dots,K_n\}\times\{1,2\}$, let $\Delta_{j,i+b_n-1}$ be the set of b_n-1-children of elements of $D_{j,i}$. For any i, j, $\Delta_{j,i+b_n-1}$ is a subset of $S_{j,i+b_n-1}$. Then w_2 or one of its children is added to A^C during the construction of $D_{j,i+1}$ if at least one of the following events happens:

1. while w_2 is probed, $\mu^{dummy}_{w_2}$ is not empty;
2. there is a multiple edge in $G^{\infty}_{n,t'}$ between w_2 and vertex not in $A^G_{w_2}$;
3. there is an edge between w_2 and a vertex of A^G not in $\Delta_{j,i+b_n-1}$ and not a children of an element of $\Delta_{j,i+b_n-1}$;
4. there is an edge between w_2 and a vertex $w_3 \in \Delta_{j,i+b_n-1}$;
5. there are a vertex $w_3 \neq w_2$ in $\Delta_{j,i+b_n-1}$ and $w_4 \notin A^G_{j,i+b_n-1}$, such that there are edges between w_2 and w_4 and between w_3 and w_4 in $G^{\infty}_{n,t'}$.

For any (i,j) and any vertices w_2 and w_3 in $\Delta_{j,i+b_n-1}$, let $X^a_{w_2}$ be equal to 1 if any of the first three events occurs, $X^b_{w_2,w_3}$ be equal to 1 if the fourth event occurs and $X^c_{w_2,w_3}$ be equal to 1 if the fifth event occurs. This classification corresponds to the following criteria:

a. one vertex of $\Delta_{j,i+b_n-1}$ is added to A^C;

b. two vertices $w_2 < w_3$ of $\Delta_{j,i+b_n-1}$ are added to A^C, when w_2 is considered;

c. two vertices $w_2 < w_3$ of $\Delta_{j,i+b_n-1}$ are added to A^C, when w_3 is considered.

$\mathscr{F}_w^a$ be the σ-algebra generated by $\mathscr{F}_w$ and $G_{w+}^{\text{mod}-}$. $\mathscr{F}_w^a$ represents the knowledge obtained by the algorithm before w is probed, and the set of labels of edges outgoing from w (without knowing their other endpoints). Let $\mathscr{F}_w^{aug,a}$ be the σ-algebra generated by $\mathscr{F}_w^{aug}$ and $G_{w+}^{\text{mod}-}$, *i.e.* $\mathscr{F}_w^a$ with the knowledge of the appropriate extra-randomness.

For any $w \in \cup_{j,i}\Delta_{j,i+b_n-1}$, let GE_w^c be equal to 1 if for every $w' < w$ the degree of w' in $G_{\text{end}}^{\text{mod}}$ is smaller than $\log n$, and let GE_w^a be equal to 1 if the degree of w is smaller than $\log n$ and $|A_w^G| \leq z_n$. For any w, $GE_w^a \geq GE_4$ and $GE_w^c \geq GE_4$. GE_w^a is $\mathscr{F}_w^a$-measurable, and GE_w^c is $\mathscr{F}_{w-}$-measurable.

Let $\eta_n^a := \frac{(\ln n + t')z_n + \ln^2 n}{n}$, $\eta_n^b := \frac{t'}{n}$ and $\eta_n^c := \frac{t' \ln n}{n}$. For n large enough, all these variables are in $[0,1]$.

Lemma 3.55. *For any $w \in \cup\Delta_{j,i+b_n-1}$, conditionally on $\mathscr{F}_w^{aug,a}$, the families (X_w^a), $(X_{w,w'}^b)_{w' \in \Delta_{j,i+b_n-1}, w<w'}$, $(X_{w',w}^c)_{w' \in \Delta_{j,i+b_n-1}, w'<w}$ are independent Bernoulli variables. Moreover:*

- $\mathbb{P}(GE_w^a X_w^a = 1 | \mathscr{F}_w^{aug,a}) \leq \eta_n^a$ *a.s.;*
- *For all $w' \in \Delta_{j,i+b_n-1}$, $w < w'$, $\mathbb{P}(X_{w,w'}^b = 1 | \mathscr{F}_w^{aug,a}) \leq \eta_n^b$ a.s.;*
- *For all $w' \in \Delta_{j,i+b_n-1}$, $w' < w$, $\mathbb{P}(GE_w^c X_{w,w'}^c = 1 | \mathscr{F}_w^{aug,a}) \leq \eta_n^c$ a.s.;*

Proof. By Lemma 3.54, it is sufficient to prove Lemma 3.55 with $\mathscr{F}_w^a$ instead of $\mathscr{F}_w^{aug,a}$.

$\mathscr{F}_{j,i+b_n-1}$, $G_w^{\text{mod}-}$, GE_w^c are measurable with respect to $\mathscr{F}_w$. The only difference between $G_w^{\text{mod}-}$ and $G_{w+}^{\text{mod}-}$ is the set of edges outgoing from w. Let denote in this proof by e_w the set of labels of outgoing edges. Therefore $\mathscr{F}_w^a$ is the σ-algebra generated by e_w and $\mathscr{F}_w$.

$V^G \setminus B_w^G$ can be decomposed in these disjoint subsets:

- $V_w^{a1} = V^G \setminus A_w^G$;
- V_w^{a2} is the subset of vertices in A_w^G, but not in $\Delta_{j,i+b_n-1}$ or children of elements of $\Delta_{j,i+b_n-1}$;
- $V_w^b = \Delta_{j,i+b_n-1} \cap A_w^G$;
- For any $w' \in \Delta_{j,i+b_n-1}$, $w' < w$, $V_{w,w'}^c$ is the subset of vertices that are in A_w^G and are children of w'.

Conditionally on $\mathscr{F}_w$:

- μ_w^{dummy} is a Poisson point process of intensity $\frac{|A_w^G|}{n}$ on $[0,t']$,
- $(\Pi_{(w,w')})_{w' \in V_w^G \setminus B_w^G}$ is an i.i.d. family of Poisson point processes of intensity $\frac{1}{n}$ on $[0,t']$, independent of μ_w^{dummy}.

e_w is the union of μ_w^{dummy} and the $n - |A_w^G|$ processes associated to elements on V_w^{a1}. Therefore, conditionally on $\mathscr{F}_w$, e_w is a Poisson point process of intensity 1 on $[0, t']$.

- The variable $(X_w^a, G_{w+}^{\text{mod}})$ depends only on μ_w^{dummy} and on $(\Pi_{(w,w')})_{w' \in V_w^{a1} \cup V_w^{a2}}$;
- for $w' \in \Delta_{j,i+b_n-1}$, $w < w'$, $X_{w,w'}^b$ depends only on the $\Pi_{(w,w')}$;
- for $w' \in \Delta_{j,i+b_n-1}$, $w > w'$, $X_{w,w'}^c$ depends only on the $(\Pi_{(w,w'')})_{w'' \in V_{w,w'}^c}$,

therefore conditionally on $\mathscr{F}_w$, the following variables are independent:

- $(X_w^a, G_{w+}^{\text{mod}-})$,
- $(X_{w,w'}^b)$ for $w' \in \Delta_{j,i+b_n-1}$, $w < w'$
- $(X_{w,w'}^c)$ for $w' \in \Delta_{j,i+b_n-1}$.

This implies that conditionally on $\mathscr{F}_w$ and $G_{w+}^{\text{mod}-}$ (*i.e.* conditionally on $\mathscr{F}_w^a$), the following variables are independent:

- X_w^a,
- $(X_{w,w'}^b)$ for $w' \in \Delta_{j,i+b_n-1}$, $w < w'$
- $(X_{w,w'}^c)$ for $w' \in \Delta_{j,i+b_n-1}$.

The conclude the proof, we now prove that the parameters of these variables are respectively smaller than η_n^a, η_n^b and η_n^c.

For any element y of e_w, let α_y be equal to 0 if $y \in \mu_w^{dummy}$ and equal to w' if $y \in \Pi_{(w,w')}$. Conditionally on $\mathscr{F}_w^a$, the family $(\alpha_y)_{y \in e_w}$ is an i.i.d. family, with $\mathbb{P}(\alpha_y = 0 | \mathscr{F}_w^a) = \frac{|A_w^G|}{n}$ and $\mathbb{P}(\alpha_y = v | \mathscr{F}_w^a) = \frac{1}{n}$ for any $v \in V^G \setminus A_w^G$. Therefore, conditionally on $\mathscr{F}_w$ and $G_{w+}^{\text{mod}-}$, the probability that there is a multiple edge between w and an element of $V^G \setminus A_w^G$ is smaller than $(n - |A_w^G|)\mathbb{P}(\text{Binom}(|e_w|, \frac{1}{n}) \geq 2) \leq \frac{|e_w|^2}{n}$ and the probability that there is an element y of e_w such that $\alpha_y = 0$ is smaller than $\frac{|e_w||A_w^G|}{n}$. If $GE_w^a = 1$, then $|e_w| \leq \log n$ and $|A_w^G| \leq z_n$. Conditionally on $\mathscr{F}_w^a$, the probability that there is at least an edge between w and an element of V_w^{a2}, conditionally on $\mathscr{F}_w$, is smaller than $\frac{t'|V_w^{a2}|}{n} \leq \frac{t'|A_w^G|}{n}$. If $GE_w^a = 1$, this quantity is smaller than $\frac{t' z_n}{n}$. By summing the previous results, we obtain that $\mathbb{P}(GE_w^a X_w^a = 1 | \mathscr{F}_w^a) \leq \eta_n^a$.

$X_{w,w'}^b$ is equal to 1 if $\Pi_{(w,w')} \neq \emptyset$. Therefore $\mathbb{P}(X_{w,w'}^b | \mathscr{F}_w^a) = e^{-\frac{t'}{n}} \frac{t'}{n} = \eta_n^b$.

$X_{w',w}^c$ is equal to 1 if $\cup_{w'' \in V_{w,w'}^c} \Pi_{(w,w'')} \neq \emptyset$. The probability of this event is smaller than $\frac{t'|V_{w,w'}^c|}{n}$. If $GE_w^c = 1$, then $|V_{w,w'}^c| \leq \log n$ and therefore $\mathbb{P}(GE_w^c X_{w',w}^c) \leq \frac{t' \log n}{n} = \eta_n^c$. □

The following constructions use the tool explained in Lemma 3.53.

For any $w \in \cup_{j,i}\Delta_{j,i+b_n-1}$, let $X_w^{a,indep} = 1$ if either $GE_w^a X_w^a = 1$ or $U_w^a \leq \frac{\eta_n^a - \mathbb{P}(GE_w^a X_w^a = 1 | \mathscr{F}_w^{aug,a})}{\mathbb{P}(GE_w^a X_w^a = 0 | \mathscr{F}_w^{aug,a})}$. Otherwise, let $X_w^{a,indep} = 0$.

For any $(w,w') \in \cup_{j,i}\Delta^2_{j,i+b_n-1}$ with $w < w'$, let $X_{w,w'}^{b,indep} = 1$ if either $X_{w,w'}^b = 1$ or $U_{w,w'}^b \leq \frac{\eta_n^b - \mathbb{P}(X_{w,w'}^b = 1 | \mathscr{F}_w^{a,aug})}{\mathbb{P}(X_{w,w'}^b = 0 | \mathscr{F}_w^{a,aug})}$. Otherwise, let $X_{w,w'}^{b,indep} = 0$.

For any $(w,w') \in \cup_{j,i}\Delta^2_{j,i+b_n-1}$ with $w' < w$, let $X_{w',w}^{c,indep} = 1$ if either $GE_w^c X_{w',w}^c = 1$ or $U_{w',w}^c \leq \frac{\eta_n^c - \mathbb{P}(GE_w^c X_{w',w}^c = 1 | \mathscr{F}_w^{a,aug})}{\mathbb{P}(GE_w^c X_{w',w}^c = 0 | \mathscr{F}_w^{a,aug})}$. Otherwise let $X_{w',w}^{c,indep} = 0$.

Corollary 3.56. *For all $w \in \Delta_{j,i+b_n-1}$ and $\bar{w} \in \{w, w+\}$, conditionally on $\mathscr{F}_{j,i-1}^{aug}$ and $G_{\bar{w}}^{mod-}$, the families of random variables $(X_{w'}^{a,indep})_{w' \in \Delta_{j,i+b_n-1} \cap \Gamma_{\bar{w}}^1}$, $(X_{w',w''}^{b,indep})_{(w',w'') \in \Gamma_{\bar{w}}^2 \cap (\Delta_{j,i+b_n-1})^2}$, $(X_{w',w''}^{c,indep})_{(w',w'') \in \Gamma_{\bar{w}}^c \cap (\Delta_{j,i+b_n-1})^2}$ are independent i.i.d. families of Bernoulli random variables of respective parameters η_n^a, η_n^b and η_n^c.*

The version with $w+$ is equivalent to the version with w by substituting w by the next vertex of $\Delta_{j,i+b_n-1}$. Corollary 3.56 is proven by induction. Corollary 3.56 holds for w if w is the first element of $\Delta_{j,i+b_n-1}$ (for the breadth-first order), as the families are empty. Assuming that Corollary 3.56 holds for a vertex w, let F be the σ-algebra generated by the random variables $(X_{w'}^{a,indep})_{w' \in \Delta_{j,i+b_n-1} \cap \Gamma_w^1}$, $(X_{w',w''}^{b,indep})_{(w',w'') \in \Gamma_w^2 \cap (\Delta_{j,i+b_n-1})^2}$, $(X_{w',w''}^{c,indep})_{(w',w'') \in \Gamma_w^c \cap (\Delta_{j,i+b_n-1})^2}$. All these variables are $\mathscr{F}_w^{aug}$-measurable, therefore conditionally on $\mathscr{F}_{j,i+b_n-1}^{aug}$, F and $G_{w+}^{\text{mod}-}$:

- (X_w^a), $(X_{w,w'}^b)_{w' \in \Delta_{j,i}, w<w'}$, $(X_{w',w}^c)_{w' \in \Delta_{j,i}, w'<w}$ are independent Bernoulli variables, of respective parameters smaller than η_n^a, η_n^b and η_n^c, by Lemma 3.55;
- $(X_w^{a,indep})$, $(X_{w,w'}^{b,indep})_{w' \in \Delta_{j,i}, w<w'}$, $(X_{w',w}^{c,indep})_{w' \in \Delta_{j,i}, w'<w}$ are independent Bernoulli variables, of respective parameters equal to η_n^a, η_n^b and η_n^c, by Lemma 3.53.

And therefore Corollary 3.56 holds for $w+$. By applying Corollary 3.56 to the last element of $\Delta_{j,i}$, one obtains:

Corollary 3.57. *Conditionally on the σ-algebra $\mathscr{F}_{j,i+b_n-1}$ and the unlabelled graph $G_{j,i+b_n}^{mod-}$, the families of random variables $(X^{a,indep})_{w \in \Delta_{j,i+b_n-1}}$, $(X_{w,w'}^{b,indep})_{w<w' \in \Delta_{j,i+b_n-1}}$, $(X_{w,w'}^{c,indep})_{w<w' \in \Delta_{j,i+b_n-1}}$ are independent i.i.d. families of Bernoulli random variables of respective parameters η_n^a, η_n^b and η_n^c.*

For any $w \in D_{j,i}$, let $X_w^{a,parent} = 1$ if there exists w', a $b_n - 1$-child of w such that $X_{w'}^{a,indep} = 1$. For any $w_1, w_2 \in D_{j,i}$, let $X_{w_1,w_2}^{bc,parent} = 1$ if

there exists w_1', a b_n-1-child of w and w_2' a b_n-1-child of w_2 such that $X^{b,indep}_{w_1',w_2'}=1$ or $X^{c,indep}_{w_1',w_2'}=1$. Let $GE_{j,i}$ be the random variable equal to 1 if all the balls of radius b_n-1 centered at vertices of $D_{j,i}$ in G^{mod} contains at most x_n vertices and $|A^G_{j,i}|\le z_n$. Otherwise, let $GE_{j,i}=0$. $GE_{j,i}$ is $\mathscr{F}_{j,i+b_n-1}$-measurable, and $GE_{j,i}\ge GE_4$.

Corollary 3.58. *Conditionally on $\mathscr{F}^{aug}_{j,i+b_n-1}$ and $G^{mod-}_{j,i+b_n}$, the families of random variables $(X^{a,parent}_w)_{w\in D_{j,i}}$ and $(X^{bc,parent}_{w,w'})_{w,w'\in D_{j,i}}$ are independent. Moreover,*

$$\begin{aligned}\mathbb{E}(GE_{j,i}X^{a,parent}_w|\mathscr{F}^{aug}_{j,i+b_n-1},G^{mod-}_{j,i+b_n}) &\le x_n\eta^a_n\\ \mathbb{E}(GE_{j,i}X^{bc,parent}_{w,w'}|\mathscr{F}^{aug}_{j,i+b_n-1},G^{mod-}_{j,i+b_n}) &\le x_n^2(\eta^b_n+\eta^c_n)=:\eta^{bc,parent}_n.\end{aligned}$$

Lemma 3.59. *Let I be an ordered finite set, and let $(X_{i,i'})_{i,i'\in I,i\le i'}$ be a family of independent Bernoulli random variables of parameters smaller than η. For $i<i'$, let $X_{i',i}=X_{i,i'}$. Let $\mathscr{F}_2$ be a σ-algebra. Then there exists a family of random variables $(X^{bound}_i)_{i\in I}$ such that:*

- *$(X^{bound}_i)_{i\in I}$ is an i.i.d. family of Bernoulli random variables of parameter $\min(1,2\sqrt{\eta|I|})$.*
- *For any i, $i'\in I$, $X_{i,i'}\le X^{bound}_i$.*

For any Y $\mathscr{F}_2$-measurable, $\mathbb{E}(Y|(X_{i,i'})_{i,i'\in I})=\mathbb{E}(Y|(X_{i,i'})_{i,i'\in I},(X^{bound}_i)_{i\in I})$.

Lemma 3.59 allows to take an i.i.d. family of Bernoulli variables indexed by two elements of I and bound it by an i.i.d. indexed family of Bernoulli variables indexed by one element of I. The bound $2\sqrt{\eta|I|}$ does not seem to be optimal, but is sufficient for our use. As we need a specific version of Lemma 3.59, the proof will be done in our particular case. A proof of Lemma 3.59 can be found in Section 3.5.7.

Let $\eta^{parent}_n=x_n\eta^a_n+2\sqrt{z_n\eta^{bc,parent}_n}=o(1)$. We assume that n is large enough such that $\eta^{parent}_n\le 1$. As a consequence, $z_n\eta^{bc,parent}_n\le 1$.

For any $w\in D_{j,i}$, let $X^{parent}_w=1$ if either $GE^{parent}_{j,i}X^{a,parent}_w=1$ or there exists $w'\in D_{j,i}$ such that $GE^{parent}_{j,i}X^{bc,parent}_{w,w'}=1$. Otherwise, let $X^{parent}_w=0$. Let us define $(X^{parent,indep}_w)_{w\in D_{j,i}}$ as $X^{parent,indep}_w=1$ if either

- $X^{parent}_w=1$;
- $U^d_w\le\frac{\eta^{parent}_n-\mathbb{P}(X^{parent}_w=1|\mathscr{F}^{aug}_{j,i-1},G^{\mathrm{mod}-}_{j,i},(X^{parent,indep}_{w'})_{w'<w,w'\in D_{j,i}})}{\mathbb{P}(X^{parent}_w=0|\mathscr{F}^{aug}_{j,i+b_n-1},G^{\mathrm{mod}-}_{j,i+b_n},(X^{parent,indep}_{w'})_{w'<w,w'\in D_{j,i}})}$.

If neither condition happens, let $X^{parent,indep}_w=0$.

This family of variables is defined by induction on w. Let $\mathscr{F}^{parent,indep}_w$ be the σ-algebra generated by $\mathscr{F}^{aug}_{j,i-1}$, $G^{\mathrm{mod}-}_{j,i}$ and $(X^{parent,indep}_{w'})_{w'<w,w'\in D_{j,i}}$.

Lemma 3.60. *For all $w \in D_{j,i}$,*

$$\mathbb{P}(X_w^{parent} = 1 | \mathscr{F}_w^{parent,indep}) \leq x_n \eta_n^a + \sqrt{z_n \eta_n^{bc,parent}}.$$

Proof. Let $p_w = \frac{\eta_n^{parent} - \mathbb{P}(X_w^{parent}=1|\mathscr{F}_w^{parent,indep})}{\mathbb{P}(X_w^{parent}=0|\mathscr{F}_w^{parent,indep})}$.

Lemma 3.60 is proven by induction. We assume that Lemma 3.60 holds for every $w' \in D_{j,i} < w$. Therefore for every $w' < w$:

$$\begin{aligned} p_{w'} &\geq \frac{\sqrt{z_n \eta_n^{bc,parent}}}{\mathbb{P}(X_w^{parent} = 0 | \mathscr{F}_w^{parent,indep})} \\ &\geq \sqrt{z_n \eta_n^{bc,parent}} \end{aligned} \tag{3.8}$$

$$\begin{aligned} X_w^{parent} &\leq X_w^{a,parent} + \sum_{w' \in D_{j,i}, w' < w} X_{w',w}^{bc,parent} + \\ &\quad + \sum_{w' \in D_{j,i}, w' > w} X_{w,w'}^{bc,parent} \\ \mathbb{E}(X_w^{parent} | \mathscr{F}_w^{parent,indep}) &\leq \mathbb{E}(X_w^{a,parent} | \mathscr{F}_w^{parent,indep}) + \\ &\quad + \sum_{w' \in D_{j,i}, w' < w} \mathbb{E}(X_{w',w}^{bc,parent} | \mathscr{F}_w^{parent,indep}) + \\ &\quad + \sum_{w' \in D_{j,i}, w' > w} \mathbb{E}(X_{w,w'}^{bc,parent} | \mathscr{F}_w^{parent,indep}) \end{aligned} \tag{3.9}$$

As $D_{j,i}$ is $\mathscr{F}_w^{parent,indep}$-measurable. Each term of the right-hand-side of the inequality (3.9) will be bound separately.

$$\begin{aligned} x_n \eta_n^a &\geq \mathbb{E}\left(X_w^{a,parent} \middle| \mathscr{F}_{j,i}^{aug}, G_{j,i}^{\text{mod}-} \right) \\ &= \mathbb{E}\left(X_w^{a,parent} \middle| \begin{array}{l} \mathscr{F}_{j,i}^{aug}, G_{j,i}^{\text{mod}-}, (X_{w'}^{a,parent})_{w'<w, w' \in D_{j,i}}, \\ (X_{w',w''}^{bc,parent})_{w'<w'' \in D_{j,i}} \end{array} \right) \\ &= \mathbb{E}\left(X_w^{a,parent} \middle| \begin{array}{l} \mathscr{F}_{j,i}^{aug}, G_{j,i}^{\text{mod}-}, (X_{w'}^{a,parent})_{w'<w \in D_{j,i}}, \\ (X_{w',w''}^{bc,parent})_{w'<w'' \in D_{j,i}}, (U_{w'}^d)_{w' \in D_{j,i}} \end{array} \right). \end{aligned} \tag{3.10}$$

The first inequality comes from Corollary 3.58, the second equality comes from the conditional independence of the variables $X_w^{a,parent}$ and $X_{w,w'}^{bc,parent}$ in Corollary 3.58 and the last equality comes from the conditional independence of $(U_w^d)_{w \in D_{j,i+1}}$. As $\mathscr{F}_w^{parent,indep}$ is included in the σ-algebra used in (3.10), we have:

$$\mathbb{E}(X_w^{a,parent} | \mathscr{F}_w^{parent,indep}) \leq x_n \eta_n^a. \tag{3.11}$$

We are now going to prove that for all w, w':

$$\mathbb{E}(X_{w,w'}^{bc,parent}|\mathscr{F}_{w}^{parent,indep}) \leq \sqrt{\frac{\eta_n^{bc,parent}}{z_n}} \text{ a.s.} \tag{3.12}$$

Let $w' < w \in D_{j,i}$. Let $\mathscr{F}_{w',w}^{parent}$ be the σ-algebra generated by $\mathscr{F}_{j,i-1}^{aug}$, $G_{j,i}^{\text{mod}-}$, $X_{w^1\in D_{j,i}}^{a,parent}$, $(X_{w^1,w^2}^{bc,parent})_{(w^1,w^2)\neq(w',w),\, w^1<w^2\in D_{j,i}}$, $(U_{w^1}^d)_{w^1\in D_{j,i}\setminus\{w'\}}$. As $\mathscr{F}_{w}^{parent,indep}$ is included in the σ-algebra generated by $\mathscr{F}_{w,w'}^{parent}$ and $X_{w'}^{parent,indep}$, we will prove:

$$\mathbb{E}(X_{w',w}^{bc,parent}|\mathscr{F}_{w',w}^{parent}, X_{w'}^{parent,indep}) \leq \sqrt{\frac{\eta_n^{bc,parent}}{z_n}} \text{ a.s.} \tag{3.13}$$

as (3.12) is a consequence of (3.13). There are two possibilities. If $X_{w'}^{parent,indep} = 0$, then $X_{w',w}^{bc,parent} = 0$ and (3.13) holds. Otherwise, if $X_{w'}^{parent,indep} = 1$:

$$\begin{aligned}
\mathbb{P}(X_{w',w}^{bc,parent} = 1|\mathscr{F}_{w',w}^{parent}, X_{w'}^{parent,indep} = 1) &= \frac{\mathbb{P}(X_{w',w}^{bc,parent} = 1, X_{w'}^{parent,indep} = 1|\mathscr{F}_{w',w}^{parent})}{\mathbb{P}(X_{w'}^{parent,indep} = 1|\mathscr{F}_{w',w}^{parent})} \\
&= \frac{\mathbb{P}(X_{w',w}^{bc,parent} = 1|\mathscr{F}_{w',w}^{parent})}{\mathbb{P}(X_{w'}^{parent,indep} = 1|\mathscr{F}_{w',w}^{parent})} \\
&\leq \frac{\mathbb{P}(X_{w',w}^{bc,parent} = 1|\mathscr{F}_{w',w}^{parent})}{\mathbb{P}(U_{w'}^d \leq p_{w'}|\mathscr{F}_{w',w}^{parent})} \\
&= \frac{\mathbb{P}(X_{w',w}^{bc,parent} = 1|\mathscr{F}_{w',w}^{parent})}{\mathbb{E}(p_{w'}|\mathscr{F}_{w',w}^{parent})}
\end{aligned} \tag{3.14}$$

By using Corollary 3.58 and the conditional independence of $(U_w^d)_{w\in D_{j,i}}$ we have:

$$\begin{aligned}
\mathbb{P}(X_{w',w}^{bc,parent} = 1|\mathscr{F}_{w',w}^{parent}) &= \mathbb{P}(X_{w',w}^{bc,parent} = 1|\mathscr{F}_{j,i+b_n-1}^{aug}, G_{j,i+b_n}^{\text{mod}-}) \\
&\leq \eta_n^{bc,parent}
\end{aligned} \tag{3.15}$$

By using (3.8) and (3.15) in (3.14), the equation (3.13) holds, and therefore equation (3.12) is proven for all $w' < w$.

Let $w' > w$. By Corollary 3.58 and the conditional independence of the $(U_v^d)_{v\in D_{j,i}}$:

$$\begin{aligned}
\mathbb{E}\left(X_{w,w'}^{bc,indep} \,\middle|\, \begin{array}{c} \mathscr{F}_{j,i+b_n-1}^{aug}, G_{j,i+b_n}^{\text{mod}-}, X_{w^1\in D_{j,i}}^{a,parent}, \\ (X_{w^1,w^2}^{bc,parent})_{w^1\leq w, w_1\leq w_2\in D_{j,i}}, \\ (U_{w^1}^d)_{w^1\neq w\in D_{j,i}} \end{array}\right) &= \mathbb{E}(X_{w,w'}^{bc,indep}|\mathscr{F}_{j,i+b_n-1}^{aug}, G_{j,i+b_n}^{\text{mod}-}) \\
&\leq \eta_n^{bc,parent} \\
&\leq \sqrt{\frac{\eta_n^{bc,parent}}{z_n}}
\end{aligned}$$

as by hypothesis $\eta^{bc,parent} z_n \leq 1$. As $\mathscr{F}_w^{parent,indep}$ is measurable with respect to this σ-algebra, the bound also holds conditionally on $\mathscr{F}_w^{parent}$, proving inequality (3.12) for $w' > w$.

By summing the bounds obtained in (3.11) and (3.12), we obtain:

$$\begin{aligned}\mathbb{E}(GE_{j,i}X_w^{parent}|\mathscr{F}_w^{parent,indep}) &\leq x_n\eta_n^a + z_n\sqrt{\frac{\eta_n^{bc,parent}}{z_n}} \\ &= x_n\eta_n^a + \sqrt{z_n\eta_n^{bc,parent}}.\end{aligned}$$

□

Corollary 3.61. *Conditionally on $\mathscr{F}_{j,i+b_n-1}^{aug}$ and $G_{j,i+b_n}^{mod-}$, $(X_w^{parent,indep})_{w\in D_{j,i+b_n}}$ is an i.i.d. family of Bernoulli variables of parameter η_n^{parent}.*

Corollary 3.61 uses the construction of Lemma 3.53 to obtain i.i.d. variables.

Definition 3.22. For every $i \in \mathbb{N}$, $j \in \{1,2\}$, $w \in S_{j,i}\backslash D_{j,i}$, let $X_w^{parent,indep} = 1$ if $U_w^d \leq \eta_n^{parent}$. Otherwise, let $X_w^{parent,indep} = 0$.

Definition 3.22 allows us to extend the family $(X_w^{parent,indep})_{w\in D_{j,i}}$ to vertices not in a $D_{j,i}$, in such a way that the family is still i.i.d:

Lemma 3.62. *Conditionally on $F_{j,i+b_n-1}^{aug}$ and $G_{j,i+b_n}^{mod-}$, $(X_w^{parent,indep})_{w\in S_{j,i}}$ is an i.i.d. family of Bernoulli variables of parameter η_n^{parent}.*

By construction, $(X_w^{parent,indep})_{w\in S_{j,i}\backslash D_{j,i}}$ is an i.i.d. family of Bernoulli variables of parameter η_n^{parent}, independent of $(X_w^{parent,indep})_{w\in D_{j,i}}$ and $G_{j,i+b_n}^{\text{mod}-}$, conditionally on $F_{j,i+b_n-1}^{aug}$.

Lemma 3.63. *Conditionally on G_{end}^{mod-}, $(X_w^{parent,indep})_{w\in G_{end}^{mod-}}$ is an i.i.d. family of Bernoulli variables of parameter η_n^{parent}.*

The proof is done by proving by induction the following two claims:

- $C_{j,i}^1$: Conditionally on $G_{j,i+b_n}^{\text{mod}-}$, $(X_w^{parent,indep})_{w\in\mathscr{B}_{j,i}}$ is an i.i.d. family of Bernoulli variables of parameter η_n^{parent}.
- $C_{j,i}^2$: Conditionally on $G_{j,i+b_n+1}^{\text{mod}-}$, $(X_w^{parent,indep})_{w\in\mathscr{B}_{j,i}}$ is an i.i.d. family of Bernoulli variables of parameter η_n^{parent}.

Lemma 3.64. *Conditionally on $\mathscr{F}_{j,i+b_n}^{aug}$, the law of $G_{j,i+b_n+1}^{mod-}$ can be described as follows:*

- *Let* $(P_w)_{w\in S_{j,i+b_n}}$ *be an i.i.d. family of Poisson point processes of intensity* 1 *on* $[0,t']$.

- *Starting with* G^{mod-}*, for each vertex* $w \in S_{j,i+b_n}$*, for each element* y *of* P_w *graft an edge labelled by* y *between* w *and a new vertex. The resulting graph is* $G^{mod-}_{j,i+b_n+1}$.

Proof. By Lemma 3.43, this construction gives the law of $G^{\mathrm{mod}}_{j,i+b_n+1}$ conditionally on $\mathscr{F}_{j,i+b_n}$. By Lemma 3.54, this construction therefore gives the law of $G^{\mathrm{mod}}_{j,i+b_n+1}$ conditionally on $\mathscr{F}_{j,i+b_n}$. □

1. $C^1_{1,0}$ comes from Lemma 3.62 for $j=1$ and $i=0$.

2. Let $(j,i) \in \{1,2\}\times\mathbb{N}$. The graph $G^{\mathrm{mod}-}_{j,i+b_n}$ and the variables $(X_w^{parent,indep})_{w\in\mathscr{B}_{j,i}}$ are $\mathscr{F}^{aug}_{j,i+b_n}$-measurable. Therefore, Lemma 3.64 entails that, $C^1_{j,i}$ implies $C^2_{j,i}$.

3. For any j, i, $C^2_{j,i}$ implies $C^1_{j,i+1}$ by Lemma 3.62.

4. The graph $G^{\mathrm{mod}-}_{j,\infty}$ is obtained from $G^{\mathrm{mod}-}_{j,K_n+b_n}$ by grafting i.i.d. copies of $T^\infty_{t'}$ to every vertex of S_{j,K_n+b_n}. The family $(X_w^{parent,indep})_{w\in G^{\mathrm{mod}-}_{j,\infty}\setminus G^{\mathrm{mod}-}_{j,K_n+b_n}}$ is constructed as an i.i.d. family of Bernoulli variables of parameter η_n^{parent} via Definition 3.22. Therefore for all j, C^2_{j,K_n} implies $C^2_{j,\infty}$.

5. $C^1_{1,\infty}$ implies $C^1_{2,0}$ from Lemma 3.62 for $j=2$ and $i=0$.

To summarize the result so far:

- GE^4 is Bernoulli variable, such that $\mathbb{P}(GE^4=0) \xrightarrow[n\to\infty]{} 0$.

- Conditionally on $G^{\mathrm{mod}-}_{\mathrm{end}}$, $(X_w^{parent,indep})_{w\in G^{\mathrm{mod}-}_{\mathrm{end}}}$ is an i.i.d. family of Bernoulli variables of parameter η_n^{parent}.

- $\eta_n^{parent} \xrightarrow[n\to\infty]{} 0$.

- If $GE^4=1$, then $w \in D_{j,i}\setminus E_{j,i}$ implies $X_w^{parent,indep}=1$.

- If $GE^4=1$, no initialisation fails.

This summary ends the first part of the proof of the construction of the included branching process, dealing with the vertices in $D_{j,i}\setminus E_{j,i}$. We now need to deal with the vertices in $E_{j,i}\setminus\tilde{E}_{j,i}$, *i.e.* the vertices that might reach the forbidden degree between time t' and t.

3.5.5.b Extension of the included branching process to the time t

As $G^{\infty}_{n,t'}$ and $G^{\infty}_{n,[t',t]}$ are independent, conditionally on $\mathscr{F}^{aug}_{\text{end}}$ the sets of labelled edges in $G^{\infty}_{n,[t',t]}$ is an i.i.d. family of Poisson point processes of intensity $\frac{1}{n}$ on $[t',t]$. We now remove remaining vertices adjacent to an edge added between t' and t, except if this edge is added between two vertices of the K_nth generation.

- Let $E^{t'-}$ be the set of vertices w of $\cup_{i<K_n:j\in\{1,2\}} D_{j,i}$ such that $X^{parent,indep}_w = 0$;
- let E^{t',K_n} be the set of vertices w of $D_{1,K_n} \cup D_{2,K_n}$ such that $X^{parent}_w = 0$;
- let $E^{t'+} = E^{t'-} \cup E^{t',K_n}$.
- For each w in $E^{t'-}$, let $X^{[t',t]}_w = 1$ if at least on edge is added to w in $G^{\infty}_{n,t}$ between t' and t.
- For each w in E^{t',K_n}, let $X^{[t',t]}_w = 1$ if there is at least one edge added between w and an element of $V^G \setminus E^{t',K_n}$.

The variable $X^{[t',t]}_w$ is used to know if w might be removed because of an edge added to w between time t' and t.

We split the set of edges according to how many endpoints belong to $E^{t'+}$.

- For any $(w,w') \in (E^{t'+})^2 \setminus (E^{t',K_n})^2$, let $X^{[t',t],a}_{w,w'} = 1$ if at least one edge is added between w and w' between t' and t. Otherwise, let $X^{[t',t],a}_{w,w'} = 0$.
- For any $(w,w') \in (E^{t',K_n})^2$ let $X^{[t',t],a}_{w,w'} = 0$.
- For any $w \in E^{t'+}$, let $X^{[t',t],b}_w = 1$ if at least one edge is added between w and an element of $V^G \setminus E^{t'+}$ between time t' and t. Otherwise, let $X^{[t',t],b}_w = 0$.

For any $w \in E^{t'+}$,

$$X^{[t',t]} = \max_{w'\in E^{t'+}} (X^{[t',t],a}_{w,w'}, X^{[t',t],b}_w).$$

Conditionally on $\mathscr{F}^{aug}_{\text{end}}$, $(X^{[t',t],a}_{w,w'})_{(w,w')\in E^{t'+}}$ is a family of independent Bernoulli variables of parameter $1 - exp(-\frac{t-t'}{n}) \leq \frac{\epsilon_1}{n}$ and is independent of $(X^{[t',t],b}_w)_{w\in E^{t'+}}$. The set $E^{t'+}$ is included in A^G_{end}, therefore if $GE_4 = 1$, then $|E^{t'+}| \leq z_n$. As GE_4 is $\mathscr{F}^{aug}_{\text{end}}$-measurable, Lemma 3.59 allows to construct $(X^{[t',t],a}_w)_{w\in E^{t'+}}$ such that, if $GE^4 = 1$:

- For all w, w', $X_{w,w'}^{[t',t],a} \leq X_w^{[t',t],a}$;
- Conditionally on $\mathscr{F}_{\text{end}}^{aug}$, $(X_w^{a,[t',t]})_{w\in E^{t'+}}$ is an i.i.d. family of Bernoulli variables of parameter $2\sqrt{\frac{\epsilon_1}{n}z_n}$, independent of $((X_w^{[t',t],b})_{w\in E^{t'+}}, \mathscr{F}_{\text{end}}^{aug}, G_{n,t'}^{\infty})$.

Conditionally on $\mathscr{F}_{\text{end}}^{aug}$, $(X_w^{[t't],b})_{w\in E^{t'-}}$ is an independent i.i.d. family of Bernoulli variables of parameter smaller than $\epsilon_1\frac{n-|E^{t'+}|}{n} \leq \epsilon_1$. Therefore conditionally on $\mathscr{F}_{\text{end}}^{aug}$, $\max(X_w^{[t',t],a}, X_w^{[t',t],b}) \geq GE_4X_w^{[t',t]}$ is an independent family of Bernoulli variables of parameter smaller than $2\sqrt{\frac{\epsilon_1}{n}z_n}+\epsilon_1 =: \eta_n^{[t',t]}$. It should be noticed that $\eta_n^{[t',t]} = \epsilon_1 + o(1)$.

Let $(U_w^e)_{w\in G_{\text{end}}^{\text{mod}}}$ be a family of random variables such that conditionally on $(\mathscr{F}_{\text{end}}^{aug}, G_{n,t}^{\infty}, (X_w^{[t',t],a})_{w\in E^{t'+}})$, the family $(U_w^e)_{w\in G^{\text{mod}}}$ is i.i.d. uniform variables on $[0;1]$. For all $w \in G_{\text{end}}^{\text{mod}} \setminus E^{t'+}$, let $X_w^{[t',t],a} = X^{[t',t],b} = 0$.

Let $X_w^{[t',t],indep} = 1$ if either:

- $X_w^{[t',t],a} = 1$;
- $X_w^{[t',t],b} = 1$;
- $U_w^e \leq \frac{\eta_n^{[t',t]}-p'_w}{1-p'_w}$ with $p'_w = \mathbb{P}(X_w^{[t',t],a} = 1 \text{ or } X_w^{[t',t],b} = 1|\mathscr{F}_{\text{end}}^{aug})$.

Then conditionally on $\mathscr{F}_{\text{end}}^{aug}$, $(X_w^{[t',t],indep})_{w\in G^{\text{mod}}}$ is an i.i.d. family of Bernoulli variables of parameter $\eta_n^{[t',t]}$. Therefore conditionally on $G^{\text{mod}-}$, $(X_w^{parent,indep}), (X^{[t',t],indep})$ are two independent families of Bernoulli variables of parameters η_n^{parent} and $\eta_n^{[t',t]}$, and therefore their maximum is an i.i.d. family of Bernoulli variables of parameter $\eta_n^{[0,t]} = 1-(1-\eta_n^{[t',t]})(1-\eta_n^{parent}) = \epsilon_1 + o(1)$.

3.5.5.c Use of the included branching process

The subgraph of $G_{\text{end}}^{\text{modk}}$ with only vertices w such that $X_w^{parent,indep} = X_w^{[t',t],final} = 0$ has same law as an independent percolation of parameter $1-\eta_n^{[0;t]}$ on two independent copies of $T_{t'}^k$.

For n large enough, $\eta_n^{[0,t]} \leq 2\epsilon_1$. For this reason, in the following part, we will study the percolation of parameter $1-2\epsilon_1$. The resulting graph is also a two-stages multitype branching process.

Lemma 3.65. *Let $G_{end}^{modk,\epsilon_1}$ be the two-stages multitype branching process obtained by doing a percolation of parameter $1-2\epsilon_1$ on G_{end}^{modk}. Let a_{t',ϵ_1}^k be the probability of survival, and ρ_{t',ϵ_1} the spectral radius associated to $G_{end}^{modk,\epsilon_1}$. Then:*

$$a_{t',\epsilon_1}^k \xrightarrow[\epsilon_1\to 0, t'\to t]{} a_t^k; \tag{3.16}$$

$$\liminf_{\epsilon_1 \to 0, t' \to t} \rho_{t',\epsilon_1} \geq \rho_t. \tag{3.17}$$

The proof of Lemma 3.65 will be done in Section 3.5.6.

Recall that $K_n = \frac{(1-\epsilon_2)\ln n}{\ln(\rho_{t'})}$. Let $\epsilon > 0$. When ϵ_1 and ϵ_2 tends to 0, $(\rho_{t',\epsilon_1})^{\frac{1}{1-\epsilon_2}}$ converges to $\rho_t > (\rho_t)^{\frac{2}{3}}$. We assume that ϵ_1 and ϵ_2 are small enough so that $(\rho_{t',\epsilon_1})^{1-\epsilon_2} > (\rho_{t'})^{\frac{2}{3}}$, and that $a^k_{t',\epsilon_1} \geq a^k_t - \epsilon$.

The graph $G^{\text{modk},\epsilon_1}_{\text{end}}$ is a two-stages branching process. Therefore, conditionally on survival, the size of the ith generation grows as $(\rho_{t',\epsilon_1})^i$, by Lemma 3.38:

$$\frac{|S^i(G^{\text{modk},\epsilon_1}_{\text{end}})|}{(\rho_{t',\epsilon_1})^i} \xrightarrow[i\to\infty]{} W \text{ a.s.}$$

where S^i denotes the sphere of radius i and W is a random variable, a.s. positive if $G^{\text{modk},\epsilon_1}_{\text{end}}$ is infinite. Therefore, for any $1 < \alpha < \rho_{t',\epsilon_1}$:

$$\mathbb{P}(|S^i(G^{\text{modk},\epsilon_1}_{\text{end}})| \geq \alpha^i) \xrightarrow[i\to\infty]{} a^k_{t',\epsilon_1}.$$

By choosing α close enough to ρ_{t',ϵ_1}, $n^{\frac{2}{3}} = o(\alpha^{K_n})$, and therefore the probability that the K_nth generation of $G^{\text{modk},\epsilon_1}_{\text{end}}$ contains at least $n^{\frac{2}{3}}$ is larger than $a^k_{t',\epsilon_1} - \epsilon$.

Conditionally on the first K_n generations of $G^{\text{modk},\epsilon_1}_{\text{end}}$, the offspring of each vertex w of the K_nth generation of $G^{\text{modk},\epsilon_1}_{\text{end}}$, seen as a subgraph of $G^{\text{modk}}_{\text{end}}$ is described by the functions f_i. By Lemma 3.31, for $t' \leq t$, these density functions are bounded from below, let say by $\delta > 0$. Therefore, by the law of large numbers, with high probability when n tends to infinity, if the K_nth generation of $G^{\text{modk},\epsilon_1}_{\text{end}}$ contains at least $n^{\frac{2}{3}}$ vertices, then at least $\frac{\delta}{2}n^{\frac{2}{3}}$ of them have no children in $G^{\text{modk}}_{\text{end}}$. As a consequence, for n large enough and ϵ_1 and ϵ_2 small enough, the following events happen simultaneously with probability larger than $(a^k_t - \epsilon)^2 - \epsilon$:

- The first and second component of $G^{\text{modk},\epsilon_1}_{\text{end}}$ are infinite;
- The K_nth generation of these components contains at least $n^{\frac{2}{3}}$ vertices.
- Among these vertices, at least $2n^{\frac{3}{5}}$ have no children in $G^{\text{modk}}_{\text{end}}$, and are therefore not saturated in $G^k_{n,t'}$.

If these conditions are satisfied, let H_1 (resp. H_2) be the set of the first $2n^{\frac{3}{5}}$ (in the breadth-first order) vertices of the generation K_n of the first (resp. second) component of $G^{\text{modk},\epsilon_1}_{\text{end}}$ with no children in $G^{\text{modk}}_{\text{end}}$. By construction, any element w of H_j satisfies the following properties:

- There is a path from v_j to w in $G^k_{n,t'}$, such that no edge is added to any vertex of this path between t' and t, except possibly to w.

- No edge is added between w and an element of $V^G \setminus E^{t'-}$ between t' and t.

Let $\mathscr{F}^{aug2}$ be the σ-algebra generated by $\mathscr{F}^{aug}$, $(U_w^e)_{w\in G^{\mathrm{mod}_{\mathrm{end}}}}$ and $(\Pi_{(w,w')} \cap [t',t])_{(w,w')\in (V^G)^2\setminus (E^{t',K_n})^2}$.

H_1 and H_2 are $\mathscr{F}_{\mathrm{end}}^{aug2}$-measurable. Conditionally on $\mathscr{F}_{\mathrm{end}}^{aug2}$, the sets of labels of edges between elements of E^{t',K_n} with labels in $[t',t]$ are independent Poisson point processes of intensity $\frac{1}{n}$ on $[t',t]$. Remove any vertex w in $H_1 \cup H_2$, such that at least an edge is added between w and element of $E^{t',K_n} \setminus (H_1 \cup H_2)$. These events are independent over the vertices w, and of probability smaller than ϵ_1. Let $\tilde{H}_1$ (resp. $\tilde{H}_2$) be the set of remaining vertices w of H_1 (resp. H_2). Assuming $\epsilon_1 < \frac{1}{2}$, by the law of large numbers, with high probability $|\tilde{H}_1| \geq n^{\frac{3}{5}}$ and $|\tilde{H}_2| \geq n^{\frac{3}{5}}$.

For any $(w_1, w_2) \in \tilde{H}_1 \times \tilde{H}_2$, exactly one edge is added between t' and t between w_1 and w_2 with probability $e^{-\frac{\epsilon_1}{n}}\frac{\epsilon_1}{n} =: a_n$. These events are independent, therefore conditionally on $|\tilde{H}_1| \geq n^{\frac{3}{5}}$, $|\tilde{H}_2| \geq n^{\frac{3}{5}}$, the probability that no simple edge is added between an element of $\tilde{H}_1$ and an element of $\tilde{H}_2$ is smaller than:

$$(1-a_n)^{n^{\frac{6}{5}}} \xrightarrow[n\to\infty]{} 0.$$

Moreover, the probability that there exists a vertex $w_1 \in H_1 \cup H_2$ such that two edges toward two different vertices of $H_1 \cup H_2$ are added between t' and t is smaller than $(4n^{\frac{3}{5}})^3(\frac{\epsilon_1}{n})^2 = o(1)$ by union bound.

All these results implies that if $|H_1| \geq 2n^{\frac{3}{5}}$ and $|H_2| \geq 2n^{\frac{3}{5}}$, then with high probability there exist $w_1 \in \tilde{H}_1$ and $w_2 \in \tilde{H}_2$ such that:

- An edge is added between w_1 and w_2 between t' and t.
- No other edge is added to either w_1 or w_2 between t' and t.

In that case there exists a path from v_1 to v_2 in $G_{n,t}^k$, through w_1 and w_2, and this concludes the proof of Lemma 3.37.

3.5.6 Proof of Lemmata 3.42 and 3.65

For all $s \leq t$ and $\epsilon \geq 0$, let $T_{s,\epsilon}^k$ be the component of the root obtained by taking a percolation of parameter $1-2\epsilon$ on T_s^k. $T_{s,\epsilon}^k$ is a two-stages multitype branching process, with one offspring law for the root, and another offspring law for all the other vertices. By definition, the spectral radius of a two-stages branching process is the spectral radius of the associated one-stage branching process. Let $(T_{s,\epsilon}^{k+,y})_{y\in[0,t']}$ denote the one-stage branching process associated to $T_{s,\epsilon}^k$ starting at a non root vertex, where y denotes the type of the first vertex. Let $T_{s,\epsilon}^{k+,\epsilon,y}$ be the branching process obtained by doing

a percolation of parameter $1-2\epsilon$ on $T_s^{k+,y}$. With the notation of section 3.5.2.a, let $M_{s,\epsilon}$ be the operator M for the branching process $T_{s,\epsilon}^{k+,\epsilon,\cdot}$. If $\epsilon = 0$, M_s will be used to denote $M_{s,0}$. As M corresponds to the expected number of vertices of each type, $M_{s,\epsilon} = (1-2\epsilon)M_s$ and

$$\rho_{s,\epsilon} = (1-2\epsilon)\rho_s.$$

Therefore Lemma 3.42 implies the limit (3.17) in Lemma 3.65.

For $s \leq t$, let $m_s(y,z)$ denote the expected density of the number of children of type z that $T_s^{k+,y}$ has. With the functions g_i introduced in Lemma 3.32, for all $y, z \in [0,s]$, :

$$m_s(y,z) = \sum_{i=1}^{k-2} \frac{1}{i-1!} \underbrace{\int_0^s \int_0^s \dots \int_0^s}_{i-1 \text{ times}} g_i(t,y,z,x_2,x_3,\dots,x_i) dx_2 dx_3 \dots dx_i.$$

By the boundedness and continuity property of the functions g_i shown in Lemma 3.31, this means that $s \to M_s$ is a continuous application, and therefore that its spectral radius is a upper semi-continuous function of s, by [New51], proving Lemma 3.42.

For any $s \leq t$ and $\epsilon \geq 0$, let $q_{s,\epsilon}(x)$ be the extinction probability of the one-stage multitype branching process $T_{s,\epsilon}^{k+,\epsilon,x}$ starting with a vertex of type x. As previously, to simplify notations, let $q_s = q_{s,0}$. By Lemma 3.39, $q_{s,\epsilon}$ is the smaller positive solution of:

$$\varphi_{s,\epsilon}(f) = f$$

where $\varphi_{s,\epsilon}$ is the operator defined by

$$\varphi_{s,\epsilon} f(y) = \mathbb{E}_y(\prod_{i=1}^{Z^\epsilon} f(X_i^\epsilon))$$

where $(X_1^\epsilon, \dots, X_{Z^\epsilon}^\epsilon)$ has the law of the types of the vertices of the first generation of $T_{s,\epsilon}^{k+,x}$. By the percolation properties, $\varphi_{s,\epsilon}$ can be computed as:

$$\varphi_{s,\epsilon} f(x) = \mathbb{E}_x(\prod_{i=1}^{Z} f(X_i)^{B_i})$$

where $(X_1, \dots, X_Z)$ has the law of the types of the vertices in the first generation of $T_s^{k+,x}$, and $(B^i)_{i\in\mathbb{N}}$ is an i.i.d. family of Bernoulli variables equal to 0 with probability $1-2\epsilon$. Therefore, for $s=t$ and $\epsilon = 0$:

$$\begin{aligned} \varphi_{t,0} f(y) &= g^0(t,y) + \int_{x_1=0}^t g^1(t,y,x_1) f(x_1) dx_1 + \dots + \\ &+ \frac{1}{k-2!} \int_{x_1,x_2\dots x_{k-2}=0}^t g^{k-2}(t,y,x_1,\dots,x_{k-2}) f(x_1)\dots f(x_{k-2}) dx_1, \dots dx_{k-2}. \end{aligned} \tag{3.18}$$

For any , $\delta \in [0,1]$, let $q^\delta = \delta \mathbf{1} + (1-\delta)q_t$. For any $\epsilon \geq 0$, $s \in [0,t]$, $\delta \in [0,1]$, $x \in [0,t]$, let

$$l(\epsilon, s, \delta, x) = (\varphi_{s,\epsilon} q^\delta)(x) - q^\delta(x).$$

As T_t^k is supercritical by hypothesis, q_t^k is not uniformly equal to 1, as we are working in the supercritical case. As all the functions g are positive and continuous by Lemma 3.31, q_t^k is continuous and not equal to 1 at any point. For any x, $\delta \to q^\delta(x)$ is therefore a strictly increasing positive linear function. Therefore for any $i \geq 2$ and any $x_1 \dots x_i$, $\delta \to q^\delta(x_1) q^\delta(x_2) \dots q^\delta(x_i)$ is a strictly convex function of $\delta \in [0,1]$, and therefore so is:

$$\delta \to h^i(t, y, \delta) := \int_{x_1, x_2 \dots x_i = 0}^{t} g^i(t, y, x_1, \dots, x_i) q^\delta(x_1) \dots q^\delta(x_i) dx_1, \dots dx_i.$$

By equation (3.18), for a fixed y, $l(0,t,\delta,y)$ is a polynomial in δ, of degree at most $k-2$, and its coefficients can be expressed as integrals of the functions q_t^k and g_i. As these functions are continuous, the coefficient of the polynomial $\delta \to l(0,t,\delta,y)$ are continuous in y and its derivative $(\delta, y) \to \frac{\partial l(0,t,\delta,y)}{\partial \delta}$ is a continuous function of (δ, y) and is negative for $\delta = 0$ for all y. As $[0,1] \times [0,t]$ is a compact set, there exists $\eta > 0$ and $\delta_0 > 0$ such that, for all $y \in [0,t]$ and all $0 \leq \delta \leq \delta_0$,

$$\begin{aligned} \frac{\partial l(0,t,\delta,y)}{\partial \delta} &< -\eta \\ l(0,t,\delta,y) &< -\eta\delta \qquad \text{as } l(0,t,0,y) = 0 \end{aligned}$$

Let $\delta \in (0, \delta_0)$. By Lemma 3.31, $(\epsilon, s, y) \to l(\epsilon, s, \delta, y)$ is a continuous function on the compact set $\{(\epsilon, s, \delta, y) : \epsilon \in [0, \frac{1}{2}], s \in [0,t], y \in [0,s]\}$, and therefore uniformly continuous. There exists $\epsilon_{\max} > 0$ and $s_{\min} < t$ such that, for all $\epsilon \in [0, \epsilon_{\max}]$, $s \in [s_{\min}, t]$, $y \in [0,s]$:

$$\begin{aligned} |l(\epsilon, s, \delta, y) - l(0, t, \delta, y)| &\leq \eta\delta \\ l(\epsilon, s, \delta, y) &\leq 0 \\ (\varphi_{s,\epsilon} q^\delta)(y) &\leq q^\delta(y) \end{aligned}$$

As this inequality holds for all y, by Corollary 3.40, $q^\delta \geq q_{s,\epsilon}$. As the law of the set of labels of the vertices of the first generation of $T_{s,\epsilon}^k$ converges to the law of the set of labels of the vertices of the first generation of T_t^k, and as δ can be chosen arbitrarily small, this inequality implies that

$$\liminf_{\epsilon \to 0, s \to t} \tilde{q}_{t',\epsilon}^k \leq \tilde{q}_{t,0}^k$$

where $\tilde{q}_{s,\epsilon}^k$ is the probability of extinction of $T_t^{k,\epsilon}$. The other side of the limits is a consequence of the following facts:

- by definition, $\tilde{q}^k_{s,\epsilon}$ is the limit the non-decreasing sequence $(\mathbb{P}(|T^{k,\epsilon}_s| \leq n))_{n\geq 0}$,
- the application $(s,\epsilon) \to T^{k,\epsilon}_s$ is continuous for the local limit,
- The event $|T| \leq n$ is continuous for the local limit.

Therefore the equation (3.16) of Lemma 3.65 holds (the probability of survival $a^k_{s,\epsilon}$ is equal to $1 - \tilde{q}^k_{s,\epsilon}$).

3.5.7 Proof of Lemma 3.59

Proof. If $2\sqrt{\eta|I|} \geq 1$, Lemma 3.59 is straightforward, so w.l.o.g. we can assume that $2\sqrt{\eta|I|} \leq 1$. In particular, $\eta|I| \leq 1$.

For any i, let $X_i = \max_j X_{i,j}$.

Let $(U_i)_{i\in I}$ be an independent i.i.d. family of uniform variables. For any i, let $X^{indep}_i \in \{0,1\}$ be defined by $X^{indep}_i = 1$ if and only if either:

- $X_i = 1$,
- or $U_i \leq \frac{2\sqrt{\eta|I|} - \mathbb{P}(X_i=1|X^{indep}_1,\ldots,X^{indep}_{i-1})}{\mathbb{P}(X_i=0|X^{indep}_1,\ldots,X^{indep}_{i-1})}$.

Lemma 3.66. *For any $i \in I$, $P(X_i = 1|(X^{indep}_{i'})_{i'<i}) \leq \sqrt{\eta|I|}$.*

If Lemma 3.66 holds, by Lemma 3.53, $(X^{indep}_i)_{i\in I}$ is an i.i.d. family of Bernoulli variables of parameter $2\sqrt{\eta|I|}$.

Proof. Lemma 3.66 is proven by induction. By union bound,

$$\begin{aligned}\mathbb{E}(X_i|(X^{indep}_{i'})_{i'<i}) &\leq \sum_{i'\in I} \mathbb{E}(X_{i,i'}|(X^{indep}_{i'})_{i'<i}) \\ &= \sum_{j\in I, j<i} \mathbb{E}(X_{i,j}|(X^{bound}_{i'})_{i'<i}) + \sum_{j\in I, j\geq i} \mathbb{E}(X_{i,j}|(X^{bound}_{i'})_{i'<i})\end{aligned}$$

For any $j \in I$, let $\mathscr{F}_{i,j}$ be the σ-algebra generated by $(X_{i_1,i_2})_{i_1\, i_2\in I,\,(i_1,i_2)\neq(i,j)}$ and $(U_{i_1})_{i_1\in I, i_1\neq j}$. For all $i_1 \notin \{i,j\}$, $X^{indep}_{i_1}$ is $\mathscr{F}_{i,j}$-measurable. Conditionally on $\mathscr{F}_{i,j}$, U_j and $X_{i,j}$ are independent, U_j is a uniform variable and $X_{i,j}$ is a Bernoulli random variable of parameter smaller than η. If $j < i$:

$$\mathbb{E}(X_{i,j}|(X^{indep}_{i'})_{i'<i}) = \mathbb{E}(\mathbb{E}(X_{i,j}|\mathscr{F}_{i,j}, X^{indep}_j)|(X^{indep}_{i'})_{i'<i}).$$

There are two possibilities. If $X_j^{indep} = 0$, then $X_{i,j} = 0$. Otherwise, if $X_j^{indep} = 1$:

$$\begin{aligned}
\mathbb{E}(X_{i,j}|\mathscr{F}_{i,j}, X_j^{indep} = 1) &= \frac{\mathbb{P}(X_{i,j} = 1, X_j^{indep} = 1|\mathscr{F}_{i,j})}{\mathbb{P}(X_j^{indep} = 1|\mathscr{F}_{i,j})} \\
&\leq \frac{\mathbb{P}(X_{i,j} = 1|\mathscr{F}_{i,j})}{\mathbb{P}(U_j \leq p_j|\mathscr{F}_{i,j})} \\
&= \frac{\mathbb{P}(X_{i,j} = 1|\mathscr{F}_{i,j})}{\mathbb{E}(p_j|\mathscr{F}_{i,j})}
\end{aligned}$$

where $p_j = \frac{2\sqrt{\eta|I|} - \mathbb{P}(X_j = 1|X_1^{indep}, \ldots, X_{j-1}^{indep})}{\mathbb{P}(X_j = 0|X_1^{indep}, \ldots, X_{j-1}^{indep})}$. By induction,

$$\begin{aligned}
\sqrt{\eta|I|} &\leq 2\sqrt{\eta|I|} - \mathbb{P}(X_j = 1|X_1^{indep}, \ldots, X_{j-1}^{indep}) \\
&\leq p_j \\
\mathbb{E}(X_{i,j}|\mathscr{F}_{i,j}, X_j^{indep} = 1) &\leq \frac{\eta}{\sqrt{\eta|I|}} \\
&= \sqrt{\frac{\eta}{|I|}}
\end{aligned}$$

Therefore, almost surely $\mathbb{E}(X_{i,j}|\mathscr{F}_{i,j}, X_j^{indep}) \leq \sqrt{\frac{\eta}{|I|}}$, and therefore $\mathbb{E}(X_{i,j}|(X_{i'}^{indep})_{i'<i}) \leq \sqrt{\frac{\eta}{|I|}}$.

If $j \geq i$, $\mathbb{E}(X_{i,j}|\mathscr{F}_{i,j}) \leq \eta \leq \frac{\eta}{\sqrt{\eta|I|}} = \sqrt{\frac{\eta}{|I|}}$, and therefore $\mathbb{E}(X_{i,j}|(X_{i'}^{indep})_{i'<i}) \leq \sqrt{\frac{\eta}{|I|}}$, as all the $X_{i'}^{indep}$, for $i' < i$ are $\mathscr{F}_{i,j}$-measurable.

By summing over I, one obtains Lemma 3.59.

□

□

Chapter 4

The height of the Lyndon tree

Contents

We consider the set $\mathcal{L}_n$ of n-letters long Lyndon words on the alphabet $\mathcal{A} = \{0, 1\}$. For a random uniform element L_n of the set $\mathcal{L}_n$, the binary tree $\mathfrak{L}(L_n)$ obtained by successive standard factorizations of L_n and of the factors produced by these factorizations is the *Lyndon tree* of L_n. We prove that the height H_n of $\mathfrak{L}(L_n)$ satisfies

$$\lim_n \frac{H_n}{\ln n} = \Delta^\star,$$

in which the constant $\Delta^\star$ is solution of an equation involving large deviation rate functions related to the asymptotics of Eulerian numbers ($\Delta^\star \simeq 5.092\ldots$). The convergence is the convergence in probability of random variables.

4.1 Introduction

4.1.1 Lyndon words and Lyndon trees

We recall some notations of [Lot97] for readability. For an alphabet $\mathcal{A}$, $\mathcal{A}^n$ is the set of n-letters words, and the language, i.e. the set of finite words,

$$\{\emptyset\} \cup \mathcal{A} \cup \mathcal{A}^2 \cup \mathcal{A}^3 \cup \ldots,$$

is denoted by $\mathcal{A}^\star$. The length of a word $w \in \mathcal{A}^\star$ is denoted by $|w|$. A total order, $\prec$, on the alphabet $\mathcal{A}$, induces a corresponding lexicographic order, again denoted by $\prec$, on the language $\mathcal{A}^\star$: the word w_1 is smaller than the word w_2 (for the lexicographic order, $w_1 \prec w_2$) at one of the following conditions: either w_1 is a prefix of w_2, or there exist words p, v_1, v_2 in $\mathcal{A}^\star$ and letters $a_1 \prec a_2$ in $\mathcal{A}$, such that $w_1 = pa_1v_1$ and $w_2 = pa_2v_2$. For any factorization $w = uv$ of w, vu is called a rotation of w, and the set $\langle w \rangle$ of rotations of w is called the *necklace* of w. A word w is primitive if $|w| = \#\langle w \rangle$.

The notion of *Lyndon word* has many equivalent definitions, to be found, for instance, in [Lot97].

Definition 4.1 (Lyndon word)**.** A word w is Lyndon if w is primitive and is the smallest element of $\langle w \rangle$.

Figure 4.1: a. $\mathcal{A} = \{\mathtt{a}, \mathtt{b}\}$ and $\mathfrak{L}(\mathtt{a}^3\mathtt{b}^4)$, b. $\mathcal{A} = \{1, 2, \ldots, 9\}$ and $\mathfrak{L}(174352698)$.

Example. The word $w = \mathtt{aabaab}$ is the smallest in its necklace

$$\langle w \rangle = \{\mathtt{aabaab}, \mathtt{abaaba}, \mathtt{baabaa}\}$$

but is not Lyndon; `baac` is not Lyndon, nor `acba` or `cbaa`, but `aacb` is Lyndon.

Here is a recursive characterization of Lyndon words:

Proposition 4.1. *One-letter words are Lyndon. A word w with length $n \geq 2$ is a Lyndon word if and only if there exists two Lyndon words u and v such that $w = uv$ and $u \prec v$.*

Among such decompositions of w, the decomposition with the longest second factor (or suffix) v is called the *standard* decomposition.

Example. $\mathtt{0011} = (\mathtt{001})(\mathtt{1}) = (\mathtt{0})(\mathtt{011})$ is a Lyndon word with two such decompositions. The latter is the standard decomposition.

The set of Lyndon words is denoted by $\mathcal{L}$, and we set $\mathcal{L}_n = \mathcal{L} \cap \mathcal{A}^n$. The Lyndon tree (cf. [HR03], also called standard bracketing tree by some authors, e.g. [Bar90]) of the Lyndon word w is a binary tree obtained by iteration of the standard decomposition:

Definition 4.2 (Lyndon tree)**.** For $w \in \mathcal{L}$, the Lyndon tree $\mathfrak{L}(w)$ of w is a labeled finite binary tree defined as follows:

- if $|w| = 1$, $\mathfrak{L}(w)$ has a unique node labeled w, and no edges;
- if (u, v) is the standard decomposition of w, then $\mathfrak{L}(w)$ is the binary tree with label w at its root, $\mathfrak{L}(u)$ as its left subtree and $\mathfrak{L}(v)$ as its right subtree.

Remark. The labels of the leaves of a Lyndon tree are letters. Also, the label of an internal node is the concatenation of the labels of its two children, and, if $|w| = n$, $\mathfrak{L}(w)$ is a rooted binary tree with n leaves, and $n - 1$ internal nodes. In general, the height of a rooted tree $\mathfrak{T}$, denoted $h(\mathfrak{T})$, is the maximal distance between the root of $\mathfrak{T}$ and one of its leaves.

4.1.2 Result

The asymptotic behavior of the size of the right and left subtrees of $\mathfrak{L}(L_n)$, for n large, have been studied in [BCN05, CA10], for L_n a random element of $\mathcal{L}_n$. The height $h(\mathfrak{L}(L_n))$ of $\mathfrak{L}(L_n)$ is of interest for analysis of algorithms and cryptanalysis , cf. [SSM92, SR03, BCN05], but it seems to have resisted analysis up to now.

For a 2-letter alphabet, say $\mathcal{A} = \{0, 1\}$, and for $n \geq 1$, let L_n denote a uniform random word in $\mathcal{L}_n$. Let $(A(n, k))_{n,k}$ denote the Eulerian numbers, i.e. $A(n, k)$ is the number of permutations σ of n symbols having exactly k descents (k places where $\sigma(i) \geq \sigma(i + 1)$) . Set

$$\begin{aligned}
\Xi(\theta) &= \lim_n \frac{1}{n} \ln(A(n, \lfloor \theta n \rfloor)/n!), \\
\Psi(\lambda, \mu, \nu) &= \ln\left(\frac{(1+\mu)^{1+\mu}}{\mu^\mu} \; \frac{(e\lambda \ln 2)^\nu \ln 2}{\nu^\nu 2^\lambda} \right) + \; \Xi(\lambda - \mu), \\
\Delta^\star &= \sup_{\lambda,\mu,\nu>0} \frac{(1+\nu+\mu)\ln 2 + \Psi(\lambda,\mu,\nu)}{\lambda(\ln 2)^2} \\
&= 5.092\ldots
\end{aligned}$$

See Lemma 4.13 or [GK94, p. 299] for an expression of Ξ. We shall prove that:

Theorem 4.2.

$$\frac{h(\mathfrak{L}(L_n))}{\ln n} \xrightarrow{\mathbb{P}} \Delta^\star.$$

Conditionally given their lengths, the two factors of the standard decomposition of a uniform Lyndon word are not independent, and they are not uniform Lyndon words either, which seems to preclude a recursive approach to the proof of this Theorem. We shall rather use a coupling method: in Section 4.2 we sketch the main steps of the construction, on the same probability space, of a random Lyndon tree, and of two well studied trees, the binary search tree of a random uniform permutation, and a Yule tree, in such a way that the height of the Lyndon tree is closely related to some statistics of the two other trees. Then Theorem 4.2 follows from a large deviation result presented in Section 4.3.

4.2 Coupling results

4.2.1 Reduction to a Bernoulli source

If the word u is primitive but is not Lyndon, the Lyndon tree $\mathfrak{L}(u)$ of u is the Lyndon tree of the unique Lyndon word in the necklace $\langle u \rangle$ of u, in short $\mathfrak{L}(u)$ is the Lyndon tree of the Lyndon word of u. If u is periodic, we define the Lyndon word of u as the word $\mathtt{0}^{|u|-1}\mathtt{1}$, and $\mathfrak{L}(u)$ is defined accordingly. Then the following algorithm:

- let W_∞ be an infinite word of uniformly random characters, obtained through the binary expansion of a number U uniformly distributed on $[0, 1]$;
- let W_n be the word W_∞ truncated after n letters, and let L_n be the Lyndon word of W_n.

produces a n-letters long random Lyndon word L_n. Conditionally, given that W_n is primitive, this random Lyndon word L_n is uniform on $\mathcal{L}_n$, but the unconditional distribution of L_n fails to be uniform due to the small probability that W_n is periodic. However, the total variation distance between the probability distribution of L_n and the uniform distribution on $\mathcal{L}_n$ is $\mathcal{O}\left(2^{-n/2}\right)$ (cf. e.g. [CA10, Lemma 2.1]), thus any property that holds true asymptotically almost surely with respect to either distribution, holds true a.a.s. for both. From now on, we shall consider that L_n is produced by the previous algorithm.

4.2.2 Poissonization

In the first steps of the recursive construction of $\mathfrak{L}(L_n)$, the sizes of the factors of the successive standard decompositions are predicted by the positions of the longest runs of `0`'s, and the structure of the top levels of $\mathfrak{L}(L_n)$ is given by the lexicographic comparisons between the suffixes of L_n beginning at these longest runs. But when n is large, the number of runs of `0`'s is typically $n/4$, and several among these runs are tied for the title of the longest run. Actually the lengths of the runs behave pretty much as a sample of $n/4$ i.i.d. geometric random variables with parameter $1/2$, and, according to [BSW94], for any strictly increasing sequence n_k such that $\lim_k \log_2 n_k - \lfloor \log_2 n_k \rfloor = \alpha \in [0, 1)$, the probability p_{m,n_k} that $m \geq 1$ among the n_k elements of such a sample are tied for the maximum is given, approximately, by

$$p_{m,n_k} \simeq \sum_{j\in\mathbb{Z}} e^{-2^{\alpha+j}} \frac{(2^{\alpha+j-1})^m}{m!}. \tag{4.1}$$

Thus the number of ties does not converge in distribution, but has a set of limit distributions indexed by $\alpha \in \mathbb{R}/\mathbb{Z}$.

Such a complex behavior does not bode well, so we shall rather analyze a transform of this problem, in the form of the Lyndon tree of a word with random length. Consider the finite word W^ℓ formed by a letter `1` followed by the truncation of W_∞ at the position τ_ℓ of the ℓth `0` in the first run of ℓ consecutive `0`'s of W_∞. Then W^ℓ is primitive, and L^ℓ denotes the Lyndon word of W^ℓ, *i.e.*:

$$W^\ell = 1\underbrace{010110\ldots1}_{\text{prefix of } W_\infty}\overbrace{000000}^{\ell 0s} \quad \text{and} \quad L^\ell = \overbrace{000000}^{\ell 0s}1\underbrace{010110\ldots1}_{\text{prefix of } W_\infty}.$$

If $\hat{\tau}_\ell$ is the position of the last 1 before τ_ℓ, L^ℓ is the concatenation of $0^\ell 1$ and of the truncation of the word W_∞ at position $\hat{\tau}_\ell$.

Now, there exists a unique longest run of 0's in W^ℓ as well as in L^ℓ, and this run is ℓ letters long, to be compared with the behavior revealed by (4.1). Moreover, if Z_k denotes the number of runs longer than $\ell-k-1$, then $Z_0 = 1$ and $(Z_j)_{0\le j\le \ell-1}$ is a Galton-Watson process with offspring distribution $2^{-k}\mathbb{1}_{k\ge 1}$, so that Z_j has a geometric distribution with parameter 2^{-j}, see for instance [Dev92]. The family tree of this Galton-Watson process gives a lot of information on $\mathfrak{L}(L^\ell)$, leading ultimately to the proof of Theorem 4.3 below. The replacement of $\mathfrak{L}(W_n)$ by $\mathfrak{L}(L^\ell)$, motivated by (4.1), has deeper consequences, initially unexpected to us: embedded in $\mathfrak{L}(L^\ell)$, appears a Yule family tree (see Section 4.2.5.b), and with this Yule tree come, besides several useful Galton-Watson trees, some Poisson point processes that leads to the Poisson-like formula (4.4). The asymptotic analysis of (4.4) finally leads to the computation of $\Delta^\star$. Thus the replacement of $\mathfrak{L}(W_n)$ by $\mathfrak{L}(L^\ell)$ can be seen as some kind of *poissonization*.

Note that τ_ℓ, the length of L^ℓ up to one unit, has a geometric distribution of order ℓ , cf. [BK02, p. 10], and, typically, grows exponentially fast with ℓ:

$$\mathbb{E}[|L^\ell|] = 2^{\ell+1} - 1.$$

Thus, expectedly, the typical height of the Lyndon tree of L^ℓ grows linearly with ℓ:

Theorem 4.3. $\frac{h(\mathfrak{L}(L^\ell))}{\ell} \xrightarrow[\ell\to\infty]{\mathbb{P}} \Delta^\star \ln 2.$

In Section 4.4 we deduce Theorem 4.2 from Theorem 4.3: choosing $\ell(n) = \log_2 n - \varepsilon_n$ in such a way that $\mathbb{P}\left(\tau_{\ell(n)} \ge n\right)$ is small, and that, with a large probability, $W^{\ell(n)}$ is a factor of W_n, we compare carefully $\mathfrak{L}(W^{\ell(n)})$ and $\mathfrak{L}(W_n)$. We set

$$\Delta^\bullet = \Delta^\star \times \ln 2.$$

Sections 4.2 and 4.3 are devoted to the proof of Theorem 4.3.

4.2.3 Reduction to a skeleton

In the top levels of the tree $\mathfrak{L}(L^\ell)$, the successive standard decompositions of the Lyndon word L^ℓ, at the smallest suffixes of L^ℓ, split the word L^ℓ at the longest runs of 0's. For ℓ large enough, the longest runs are sparse enough to preserve some degree of independence between the factors. This is not true anymore at the lowest levels of the tree $\mathfrak{L}(L^\ell)$. For this reason, it is easier to split the study of the Lyndon tree in two parts: the first one focuses on the top of the tree, where the runs of 0's are still above a threshold a_ℓ, and the second part studies a forest of *shrubs* at the bottom of the tree, each of them labeled with a factor of L^ℓ that contains only runs of 0's shorter than the threshold. The top part is a tree itself (a subtree of $\mathfrak{L}(L^\ell)$), and each

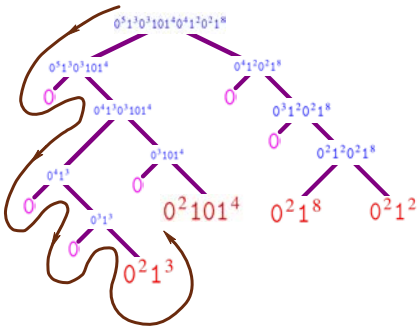

Figure 4.2: For $L^5 = 0^5 1^3 0^3 101^4 0^4 1^8 0^2 1^2$, the tree $\mathfrak{T}_5 = \mathfrak{L}^2(L^5)$ has 6 needles and 4 blades. In brown, its contour traversal.

shrub of the forest at the bottom of $\mathfrak{L}(L^\ell)$ is rooted at (or grafted on) a leaf of the top tree. We follow here the same path as [BD07], our shrubs playing the same rôle as their *spaghetti-like subtrees*. Let us define by induction the tree above the threshold k, with $k \geq 1$:

Definition 4.3 (Top tree)**.** If w denotes a Lyndon word, $\mathfrak{L}^k(w)$ is a finite labeled binary tree built recursively according to the following set of instructions:

- if w has one factor 0^k or less (thus 0^{k+1} is not a factor of w), $\mathfrak{L}^k(w)$ is a single node, labeled w;
- otherwise, let (u, v) be the standard decomposition of w. Then the root of $\mathfrak{L}^k(w)$ has label w, the left subtree of $\mathfrak{L}^k(w)$ is $\mathfrak{L}^k(u)$ and the right subtree is $\mathfrak{L}^k(v)$.

$\mathfrak{L}^k(w)$ is called the top tree associated to w, with threshold k.

Set

$$\mathfrak{T}_\ell = \mathfrak{L}^{a_\ell}(L^\ell).$$

The threshold a_ℓ depends on ℓ. It has to be large enough for the top tree to retain the independence properties between the factors, but small enough that we can handle the shrubs, though they lack these nice independence properties. We shall assume that

$$a_\ell \uparrow \infty \quad \text{and} \quad a_\ell = o(\ell).$$

4.2.4 A binary search tree

Let $\mathfrak{s}_k(w)$ denote the suffix of the finite word w with length $|w| - k + 1$, and let $\sigma_w(k)$ be the rank of $\mathfrak{s}_k(w)$ once the sequence $(\mathfrak{s}_j(w))_{1 \leq j \leq |w|}$ is sorted in increasing lexicographic order. Then a word w is Lyndon if and only if $\sigma_w(1) = 1$, and in this case, according to [HR03], $\mathfrak{L}(w)$ is the binary

search tree of the permutation σ_w. Note that according to [SF96, p. 362], it should rather be called the heap-ordered tree of σ_w, see Section 4.2.5.b and Figure 4.6. The asymptotic behavior of the height of the binary search tree of a random uniform permutation (not different from that of the corresponding heap-ordered tree, cf. [SF96, p. 364, Th. 7.1]) is well studied [Dev86, Rob10], but the distribution of σ_{L_n} or of σ_{L^ℓ} is all but uniform. In this Section, we produce a coupling between σ_{L^ℓ} and a random uniform permutation. In the next Sections we inspect the relations between the heap-ordered trees of these two random permutations.

Let us take a closer look at $\mathfrak{T}_\ell$: observe that if $w \in \mathcal{L}$, then $\mathtt{0}w \in \mathcal{L}$, and the two factors of the standard factorization of $\mathtt{0}w$ are $\mathtt{0}$ and w. Thus either a leaf v of $\mathfrak{T}_\ell$ has label $\mathtt{0}$, and v is called a *needle*, or the label of v is a factor of w beginning with $\mathtt{0}^{a_\ell}\mathtt{1}$, and v is called a *blade*. The number N_ℓ of blades of $\mathfrak{T}_\ell$ has a geometric distribution with parameter $2^{-\ell+a_\ell}$, and the set of blades has a natural order related to the contour traversal (see Figure 4.2), that allows to identify it to $[\![1, N_\ell]\!]$. Note that the number of needles has a simple expression in terms of a Galton-Watson process with geometric offspring distribution. In the analysis of the shape of $\mathfrak{T}_\ell$, the configuration of the needles is a special concern, and the following bound will help at some point. Let $g(v)$ denote the number of needles on the path between a blade v and the root $\emptyset$ of $\mathfrak{L}^k(w)$, and let $M(w)$ be the length of the longest run of $\mathtt{0}$'s in $\langle w \rangle$. Then

Lemma 4.4. *For any blade v of $\mathfrak{L}^k(w)$, $g(v) \leq M(w) - k$.*

Proof. For any interior node ν of $\mathfrak{L}^k(w)$, let $g(\nu)$ denote the natural extension of g to the interior nodes of $\mathfrak{L}^k(w)$ and let $m(\nu)$ be the length of the longest run of $\mathtt{0}$'s in the label $f(\nu)$ of ν. Then, if ν is not a needle of $\mathfrak{L}^k(w)$, $g(\nu) + m(\nu)$ does not decrease on the edge towards the root, for

- either the father μ of ν has the label $\mathtt{0}f(\nu)$, in which case $g(\nu) = 1 + g(\mu)$ and $m(\nu) = -1 + m(\mu)$;
- or $f(\mu) \neq \mathtt{0}f(\nu)$, in which case $g(\nu) = g(\mu)$ and $m(\nu) \leq m(\mu)$.

Thus $m(\emptyset) + g(\emptyset) = M(w) \geq m(v) + g(v) = g(v) + k$. □

We shall need some notations: in the contour traversal of $\mathfrak{T}_\ell$, there exists a sequence of $n_v - a_\ell \geq 0$ needles between a blade v and the previous blade (or between v and the root, if there exists no previous blade). The concatenation, starting at this sequence of needles, included, of the labels of the leaves in the order of the contour traversal of $\mathfrak{T}_\ell$, is a suffix $\mathfrak{s}(v)$ of L^ℓ that can be written $\mathtt{0}^{n_v}\mathtt{1}\,t_v$, the run $\mathtt{0}^{n_v}$ being maximal in the sense that $\mathtt{0}^{n_v+1}\mathtt{1}\,t_v$ is not a suffix of L^ℓ.

The words of the sequence $(t_v)_{1 \leq v \leq N_\ell}$ have different lengths, being proper suffixes of each other, so they are all different, and we can give a reformulation of the algorithm that produces $\mathfrak{T}_\ell$, or more generally $\mathfrak{L}^k(w)$, in terms of

the family $T_\ell = ((n_v, t_v))_{1\le v\le N_\ell}$ of the blades (with labels $\mathtt{0}^k\mathtt{1}\,t_v$), in which only the n_v's and the relative order of the t_v's matter. With this reformulation of the algorithm, a slight perturbation of the t_v's produces a new tree, $\mathfrak{S}_\ell$, that is easier to handle than $\mathfrak{T}_\ell$ due to its property of independence of labels, but that has essentially the same *profile* as $\mathfrak{T}_\ell$ (i.e. it has the same repartition of blades with respect to the height). Let $\epsilon^{(j)}$ denote the sequence of integers defined, for $j \in I$, by

$$\epsilon_i^{(j)} = \delta_{i,j}.$$

For $\mathcal{R}$ a totally ordered set, $\mathbb{N}_0 \times \mathcal{R}$ inherits a lexicographic order, $\prec$, from $\mathcal{R}$: $(n,t) \prec (m,u)$ if $n > m$ or if $n = m$ and $t < u$. Let $B = (l_i, r_i)_{1\le i\le N}$ (resp. $L = (l_i)_{1\le i\le N}$, $R = (r_i)_{1\le i\le N}$) be a finite sequence of elements of $\mathbb{N}_0 \times \mathcal{R}$ (resp. $\mathbb{N}_0$, $\mathcal{R}$), with no repetitions in the sequence R. Assume that $l_j \ge k$ for each j, and that (l_1, r_1) is the smallest element of B, for $\prec$.

Definition 4.4. The Lyndon tree $\mathfrak{L}^k(B)$ is defined by induction by:

1. If $N = 1$ and $l_1 = k$, $\mathfrak{L}^k(B)$ has no edge and its unique vertex, with label (k, r_1), is a blade.
2. Otherwise, consider the new sequence B' formed from $L - \epsilon^{(1)}$ and R and let i_0 denote the index of the smallest element in B', for $\prec$.
   (a) If $i_0 = 1$, then $\mathfrak{L}^k(B)$ is the binary tree with a needle (labeled $\mathtt{0}$) as its left child and $\mathfrak{L}^k(B')$ as its right child.
   (b) If $i_0 \ge 2$, then the binary tree $\mathfrak{L}^k(B)$ has $\mathfrak{L}^k((l_i, r_i)_{1\le i\le i_0-1})$ for left subtree, and $\mathfrak{L}^k((l_{i+i_0}, r_{i+i_0})_{0\le i\le N-i_0})$ for right subtree.

Remark. Since $\sum(l_i + 1)$ is strictly decreasing at each recursive call to instruction (2), and since the l_i's are not allowed to drop under level k, $\mathfrak{L}^k(B)$ is well-defined as long as the r_i's are distinct. The N blades of $\mathfrak{L}^k(B)$ are labeled $(k, r_i)_{1\le i\le N}$, and, during the contour traversal, they appear in this order.

Remark. For $\mathcal{R} = \{t_v \,|\, 1 \le v \le N_\ell\}$ and $T_\ell = ((n_v, t_v))_{1\le v\le N_\ell}$ defined in this section,

$$\mathfrak{L}^{a_\ell}(T_\ell) = \mathfrak{T}_\ell,$$

or more precisely, the shapes are the same, but the labels are different. When the label of some node of $\mathfrak{L}^{a_\ell}(T_\ell)$ is (k, t_v), the corresponding label of $\mathfrak{T}_\ell$ is the prefix of $\mathtt{0}^k\mathtt{1}t_v$ that stops with the last $\mathtt{1}$ before the next occurrence of $\mathtt{0}^k$.

For the analysis of $\mathfrak{T}_\ell$, the fact that the t_v's are suffixes of t_1, precluding any form of independence, is bothering. In order to fix the problem, in T_ℓ, we replace the sequence $(t_v)_{1\le v\le N_\ell}$ with a new sequence $(s_v)_{1\le v\le N_\ell}$ of infinite binary words, close to the t_v's but independent, defined as follows:

let $(\zeta_i)_{i\in\mathbb{N}}$ be an i.i.d. sequence of uniform infinite words, independent of T_ℓ and let s_v be the concatenation of p_v, the prefix formed by the first a_ℓ letters of t_v, with ζ_v. When t_{N_ℓ} is shorter than a_ℓ letters, p_{N_ℓ} is completed with the appropriate number of $\mathtt{0}$'s, before the concatenation with ζ_{N_ℓ}. This way, we obtain a new sequence $S_\ell = ((n_v, s_v))_{1\le v\le N_\ell}$, and we set

$$\mathfrak{S}_\ell \;=\; \mathfrak{L}^{a_\ell}(S_\ell)\,.$$

Differences between $\mathfrak{T}_\ell$ and $\mathfrak{S}_\ell$ occurs scarcely, only when at least a_ℓ letters are used to distinguish two suffixes, so that $\mathfrak{T}_\ell$ and $\mathfrak{S}_\ell$ are close, in a sense stated precisely in Proposition 4.7. We have:

Proposition 4.5. *The probability distribution of S_ℓ is given by:*

1. *$n_1 = \ell$, and for $v \ge 2$, $n_v - a_\ell$ is a geometric random variable with parameter $\frac{1}{2}$, conditioned to be smaller than $\ell - a_\ell$;*
2. *s_v is a copy of W_∞;*
3. *For all v, n_v and s_v are independent;*
4. *N_ℓ is geometric with parameter $2^{a_\ell-\ell}$;*
5. *N_ℓ and the sequence $(n_v, s_v)_{v\in\mathbb{N}}$ are independent;*
6. *$(n_v, s_v)_{v\in\mathbb{N}}$ is a sequence of i.i.d. random variables.*

In terms of words, this can be rephrased as follows:

Proposition 4.6. *The sequence of words $\tilde{S}_\ell = (\mathtt{0}^{n_v-a_\ell}\mathtt{1}s_v)_{2\le v\le N_\ell}$, followed by the word $\mathtt{0}^{n_1-a_\ell}\mathtt{1}s_1$, is distributed as a sequence of copies of W_∞, observed until the first occurrence of the prefix $\mathtt{0}^{\ell-a_\ell}$, this first occurrence $\mathtt{0}^{n_1-a_\ell}\mathtt{1}s_1$ being eventually truncated of any $\mathtt{0}$ in excess of $\mathtt{0}^{\ell-a_\ell}\mathtt{1}\ldots$, so that $n_1 = \ell$.*

Proof of Proposition 4.5. Consider the word $W'_\infty = \mathtt{0}^\ell\mathtt{1}W_\infty$ and let x_j, $j \ge 1$, be the jth letter in W'_∞, let $\tilde{n}_k$ denote the length of the kth maximal run of $\mathtt{0}$'s longer than $a_\ell - 1$ in W'_∞, let τ_k be the position of the letter $\mathtt{1}$ ending this kth run of $\mathtt{0}$'s, so that $x_{\tau_k-1}x_{\tau_k}$ ends the kth occurrence of the pattern $\mathtt{0}^{a_\ell}\mathtt{1}$ in W'_∞. Let $\tilde{N}_\ell$ be the number of runs of $\mathtt{0}$'s longer than $a_\ell - 1$ before the second run longer than $\ell - 1$ occurs, and let $\tilde{p}_k = x_{\tau_k+1}x_{\tau_k+2}\ldots x_{\tau_k+a_\ell}$. Then $(x_{\ell+1+j})_{j\ge1}$ is a Bernoulli process, the τ_j's are stopping times for the related filtration, and

$$\tilde{N}_\ell = N_\ell, \quad (\tilde{n}_v, \tilde{p}_v)_{1\le v\le N_\ell} = (n_v, p_v)_{1\le v\le N_\ell},$$

by definition. But since $\tau_k + a_\ell + 1 \le \tau_{k+1}$, $(\tilde{n}_v, \tilde{p}_v)_{v\ge1}$ is an i.i.d. sequence, with $\tilde{p}_v$ uniform on $\{\mathtt{0},\mathtt{1}\}^{a_\ell}$ and independent of $\tilde{n}_v$, and $\tilde{n}_v - a_\ell$ geometric. This entails the six points of Proposition 4.5. □

Due to Proposition 4.6, conditionally given N_ℓ, the ranks of the terms of the sequence $\tilde{S}_\ell = (\mathtt{0}^{n_v - a_\ell}\mathtt{1}s_v)_{2\le v\le N_\ell}$ with respect to the lexicographic order form a uniform permutation on $N_\ell - 1$ symbols. This permutation is quite close to the random non-uniform permutation induced by the family $(\mathfrak{s}(v))_{2\le v\le N_\ell}$ of suffixes of L^ℓ. As a consequence, the subtree $\mathfrak{U}_\ell$ of $\mathfrak{S}_\ell$ induced by the root and the blades, once the needles and the first blade erased, forms the *binary search tree* of a uniform permutation, a well studied random tree: for instance, a coupling between the Yule process and the binary search tree leads to a precise analysis of the depths of the leaves of the binary search tree, see [CKMR05]. The depths of blades in $\mathfrak{S}_\ell$, though they depend on their depths in $\mathfrak{U}_\ell$, are also affected by the positions of the needles, and we need to tweak the arguments of [CKMR05] in order to include the needles in their analysis.

In Section 4.2.5 we prove that the coupling between $\mathfrak{T}_\ell$ and $\mathfrak{S}_\ell$ is tight enough that the depths of leaves of $\mathfrak{T}_\ell$ and $\mathfrak{S}_\ell$ share the same asymptotic behavior, at some level of detail, see Proposition 4.7 below. The proof of Proposition 4.7 relies on a coupling between $\mathfrak{S}_\ell$ and a Yule process, described in Section 4.2.5.c. This coupling is also the key to the analysis of the depths of blades in $\mathfrak{S}_\ell$, which are expressed as functionals of the Yule process, see Section 4.2.6.

Set $\mathfrak{p}_{N_\ell} = t_{N_\ell}$, and if $v < N_\ell$, set $t_v = \mathfrak{p}_v\mathtt{0}^{n_{v+1}}\mathtt{1}t_{v+1}$, so that $(\mathtt{0}^{n_v}\mathtt{1}\mathfrak{p}_v)_{1\le v\le N_\ell}$ is a factorization of L^ℓ and $(\mathtt{0}^{a_\ell}\mathtt{1}\mathfrak{p}_v)_{1\le v\le N_\ell}$ is a sequence of Lyndon words. One obtains the Lyndon tree $\mathfrak{L}(L^\ell)$ when one grafts each shrub $\mathfrak{t}(v) = \mathfrak{L}(\mathtt{0}^{a_\ell}\mathtt{1}\mathfrak{p}_v)$ on $\mathfrak{T}_\ell$, $\mathfrak{t}(v)$ replacing the corresponding blade v of $\mathfrak{T}_\ell$.

Remark. Due to Proposition 4.7, the tree $\mathfrak{A}_\ell$ obtained by grafting the shrubs $\mathfrak{t}(v)$'s on the corresponding blades of $\mathfrak{S}_\ell$ (rather than $\mathfrak{T}_\ell$) is very close in height to $\mathfrak{L}(L^\ell)$. But $\mathfrak{t}(v)$ depends on $\mathfrak{S}_\ell$ only through the prefix p_v of $\mathfrak{p}_v$ and p_v is short compared to $\mathfrak{p}_v$ when a_ℓ grows (for $|p_v| = a_\ell$ while $\mathbb{E}\left[|\mathfrak{p}_v|\right] \simeq 2^{a_\ell}$). This has a crucial consequence for the study of the heights H_v's of the shrubs $\mathfrak{t}(v)$'s at the bottom of the tree: in Section 4.3.2 we shall see that $(H_v)_{1\le v\le N_\ell}$ behave like a sample of independent geometric random variables with parameter $1/2$, essentially *independent* from $\mathfrak{S}_\ell$.

4.2.5 The distance between $\mathfrak{T}_\ell$ and $\mathfrak{S}_\ell$

For a blade $v \in [\![1, N_\ell]\!]$, let h_v (resp. $\tilde{h}_v$) be its height in $\mathfrak{S}_\ell$ (resp. in $\mathfrak{T}_\ell$). Set

$$d_v = \left|h_v - \tilde{h}_v\right|, \qquad D_\ell = \max_{1\le v\le N_\ell} d_v.$$

Then

Proposition 4.7.

$$\lim_\ell \mathbb{P}\left(D_\ell \ge \frac{\ell}{\sqrt{a_\ell}}\right) = 0.$$

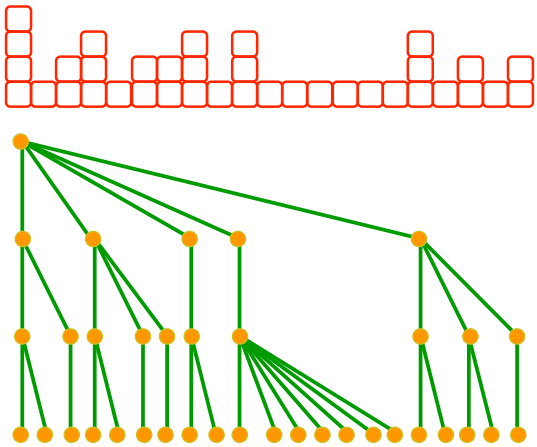

Figure 4.3: Here $\ell = 4$ and $a_\ell = 1$. Top figure: a sample $s = (n_v - a_\ell + 1)_{2 \le v \le N_\ell} = 12312231311111131212$ of the geometric distribution, stopped before the first value larger than 3, value that is reduced to $n_1 = 4$, and is used as a prefix of s. Below, the corresponding Galton-Watson tree $\mathfrak{GW}_\ell$.

The proof of Proposition 4.7 uses branching random walks arguments, as in [Big77], and a coupling of $\mathfrak{S}_\ell$ with a Yule process. For $1 \le i < j \le N_\ell$, the case when $t_i \prec t_j$ while $s_i \succ s_j$ is called an *inversion* between i and j. In order to bound D_ℓ, we need to track inversions. Let $\varpi_\ell = 2^{-a_\ell}$ and let U_i be the real number with dyadic expansion s_i (according to point (2) of Proposition 4.5, U_i is uniformly distributed on $[0, 1]$). We have:

Lemma 4.8. *If there is an inversion between i and j then s_i and s_j coincide on the first a_ℓ letters, and U_i and U_j are in the same dyadic interval with width ϖ_ℓ.*

Proof. Since t_i and t_j are at least a_ℓ-letters long by definition and since t_i (resp. t_j) coincides with s_i (resp. s_j) on the first a_ℓ letters, then the outcome of the comparisons $t_i \prec\succ t_j$ and $s_i \prec\succ s_j$ can be different only if it is decided after the first a_ℓ letters, which entails, by definition of the lexicographic order, that the 4 words have the same a_ℓ-letters long prefix. □

4.2.5.a A Galton-Watson process

Lemma 4.8 allows to compute an upper bound on D_ℓ that depends only on $\mathfrak{S}_\ell$, and functionals of $\mathfrak{S}_\ell$ are more tractable than functionals of $\mathfrak{T}_\ell$ for several reasons. One of them is the description of $\mathfrak{S}_\ell$ in terms of a Galton-Watson process with geometric offspring distribution: due to points (1), (4) and (5) of Proposition 4.5, $(n_{v+1} - a_\ell + 1)_{1 \le v \le N_\ell - 1}$ is distributed as a sample of i.i.d. geometric random variables, observed until the last time when all the terms of the sequence are smaller than $\ell - a_\ell$. Then the first terms in

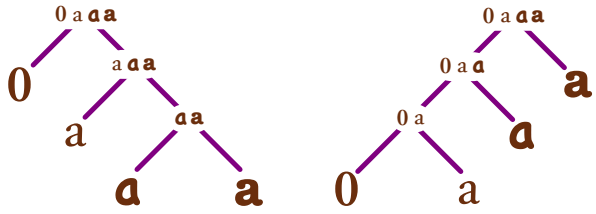

Figure 4.4: $\mathcal{S}_{k,w}$ and $\mathcal{T}_{k,w}$ in the worst case scenario, when the 3 scions in $\mathcal{O}_{k,w}$ are in the same dyadic interval, and $(t_j)_j$ is increasing while $(s_j)_j$ is decreasing, resulting in $(d(\omega)|\omega \in \mathcal{O}_{k,w}) = (1, 1, 2)$.

T_ℓ or in S_ℓ, $(n_1, t_1) = (\ell, t_1)$ and (n_1, s_1), are seen as the ancestors, and the indices v such that $n_v \geq \ell - k$ form generation k. More precisely, if $n_v \geq \ell - k$, and if the next index that belongs to generation k is w, the offspring of $(\ell - k, t_v)$ at generation $k+1$ is formed by $(\ell - k - 1, t_v)$ and by the nodes (n_j, t_j) or (n_j, s_j) such that $v < j < w$ and $n_j = \ell - k - 1$. Let us call the set

$$\begin{aligned}\mathcal{F}_{k,v} &= \{(\ell - k, t_v), (\ell - k - 1, t_v)\} \cup \{(n_j, t_j) \,|\, v < j < w \text{ and } n_j = \ell - k - 1\} \\ &= \{(\ell - k, s_v), (\ell - k - 1, s_v)\} \cup \{(n_j, s_j) \,|\, v < j < w \text{ and } n_j = \ell - k - 1\}\end{aligned}$$

the *family* of $(\ell - k, t_v)$ or of $(\ell - k, s_v)$. Due to the memoryless property of the geometric distribution, the probability p_n that $(\ell - k, s_v)$ has $O_{k,v} = n$ children satisfies

$$\mathbb{P}\left(O_{k,v} = n\right) = 2^{-n}\, \mathbb{1}_{n \geq 1},$$

see for instance [Dev92, page 601] for an explanation. This process stops at generation $\ell - a_\ell$.

We call $\mathfrak{GW}_\ell$ the family tree of the Galton-Watson process described in this paragraph. The tree $\mathfrak{GW}_\ell$ depends only on the sequence $(n_j)_j$, and can be seen as the tree induced by some nodes of $\mathfrak{S}_\ell$, or of $\mathfrak{T}_\ell$, indifferently, including the blades. The differences between $\mathfrak{S}_\ell$ and $\mathfrak{T}_\ell$ can be analyzed at the level of the binary subtrees $\mathcal{S}_{k,w}$ (resp. $\mathcal{T}_{k,w}$) induced by the family $\mathcal{F}_{k,w}$ in $\mathfrak{S}_\ell$ (resp. in $\mathfrak{T}_\ell$), for the only comparisons that involve the t_j's or the s_j's are inside the families, the other comparisons being settled by inspections of the first terms n_j's of the labels, that are the same in S_ℓ and in T_ℓ.

The subtrees $\mathcal{S}_{k,w}$ and $\mathcal{T}_{k,w}$ have one needle, the other leaves beginning with one of the elements of the offspring $\mathcal{O}_{k,w} = \mathcal{F}_{k,w} \setminus \{(\ell - k, t_w)\}$. For such a leaf $\omega = (\ell - k - 1, t_j) \in \mathcal{O}_{k,w}$, let $d(\omega)$ denote the modulus of the difference between the height of ω in $\mathcal{T}_{k,w}$ and the height of $(\ell - k - 1, s_j)$ in $\mathcal{S}_{k,w}$. Then a bound for d_v is given by the sum of the $d(\omega)$ on the path between the blade v and the root. Thus the Galton-Watson tree $\mathfrak{GW}_\ell$ underlies a branching random walk with positive steps $d(\omega)$'s, and the rightmost position of this branching random walk gives an upper bound for

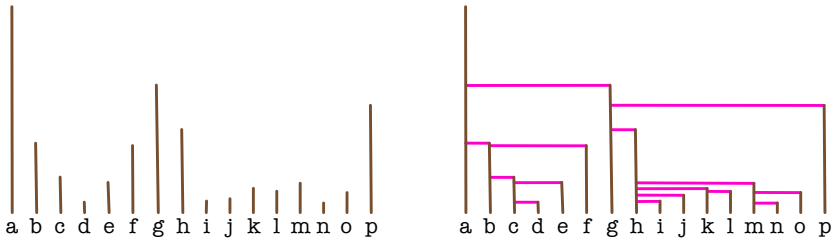

Figure 4.5: A sample of exponential random variables, until hitting t, and the related tree $\mathfrak{E}_t$ (in which the pink horizontal lines are to be seen as vertices rather than as edges).

D_ℓ. Note that the independence of the steps $d(\omega)$ is questionable at this stage of the proof, see Section 4.2.5.c.

4.2.5.b A Yule process

According to Lemma 4.8, there are differences between $\mathcal{T}_{k,w}$ and $\mathcal{S}_{k,w}$ only if some of the U_j's involved in the comparisons at step (2) of Algorithm 4.4 are in the same dyadic interval with width $\varpi_\ell = 2^{-a_\ell}$. An additional condition is that the results of these comparisons change the leader, i.e. the smallest element in B', at step (2) of Algorithm 4.4. But the leader does not change if it belongs to a different dyadic interval. It turns out that the family $\mathcal{F}_{k,v}$ can be seen as a Galton-Watson process itself, elements of the different dyadic intervals being the generations. Such nested Galton processes with geometric progeny typically appear in Yule processes. In this section, we define a Yule process in terms of the sequence T_ℓ, and in Section 4.2.5.c we shall see how a bound for D_ℓ can be derived from this Yule process. In Section 4.2.6, we show how to represent the height of a blade of $\mathfrak{S}$ in term of this Yule process, a representation that is essential for the computation of the height of the Lyndon tree, through the large deviation results of Section 4.3.

A *Yule process* $\mathcal{Y}$ (cf. [AN, page 109] or [Ald01]) models a population in which each individual lives forever, and gives birth to a new individual according to a Poisson process with rate ρ. We assume that the population starts at time 0 with a single individual, called the *ancestor*. One can keep track of the history of the population through the *Yule tree* $\mathfrak{Y}$ [CKMR05], a family tree of the population, in which a vertical life line is drawn downward, for each individual, starting at an ordinate given by *minus* the date of birth, on the left of the life line of its father, and is connected to the life line of its father by a dotted horizontal line. Let $\mathfrak{Y}_t$ denote the family tree $\mathfrak{Y}$ truncated at time t, i.e. at ordinate $-t$.

This representation comes handy for the description of a correspondence between the Yule tree and a sample $Y = (Y_n)_{n\geq 1}$ of i.i.d. exponential

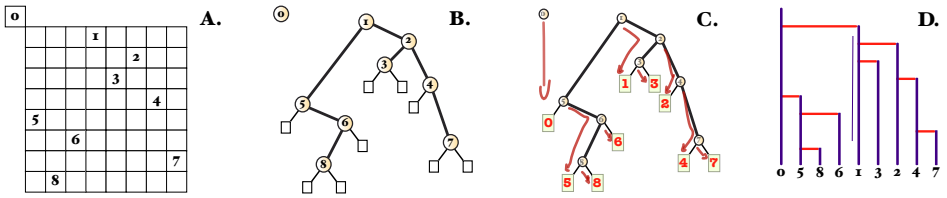

Figure 4.6: A. The permutation $\sigma = (5, 8, 6, 1, 3, 2, 4, 7)$. B. The heap-ordered tree $\mathfrak{H}_\sigma$. C. The Lyndon tree $\mathfrak{L}(\mathtt{0}\sigma)$ built from $\mathfrak{H}_\sigma$. D. The Lyndon tree $\mathfrak{L}(\mathtt{0}\sigma)$ built from $f(\mathtt{0}\sigma)$, here $f(x) = 9 - x$.

random variables with rate ρ. Consider the sequence $Z^{(t)}$ defined by:

$$T_t = \inf\{n \geq 1 \,|\, Y_n \geq t\}, \quad Z^{(t)} = (Z_k^{(t)})_{0 \leq k \leq T_t - 1} = (t, Y_1, Y_2, \ldots, Y_{T_t - 1}).$$

Then picture each term $Z_k^{(t)}$ of the sequence $Z^{(t)}$ by a vertical line of the corresponding length $Z_k^{(t)}$, drawn at abscissa $T_t - k$, and connect, through an horizontal line, the top of each line k, but the leftmost, to the next line on its left whose height tops the height $Z_k^{(t)}$, to obtain a family tree $\mathfrak{E}_t$.

Proposition 4.9. *$\mathfrak{E}_t$ and $\mathfrak{Y}_t$ have the same probability distribution.*

We do not know a reference for Proposition 4.9, that is part of the folklore on the topic. For the binary search tree, however, there exists a well studied analog: the shape of the *heap-ordered tree*[1] $\mathfrak{H}_\sigma$ of a random permutation σ has the same distribution as the shape of the related binary search tree. If we see σ as a word on the alphabet $[\![\mathtt{n}]\!]$, $\mathfrak{H}_\sigma$ is the Lyndon tree of the word $\mathtt{0}\sigma$ and can also be obtained through the construction by vertical lines that leads to $\mathfrak{E}_t$, if one starts with the sample $(f(0), f(\sigma(1)), f(\sigma(2)), \ldots, f(\sigma(n)))$ (in which f is any positive strictly decreasing function) rather than with $Z^{(t)}$, cf. Fig. 4.6.

4.2.5.c Representation of $\mathfrak{S}_\ell$ in terms of a Yule process

Let us denote by U_v the real number whose dyadic development is $\mathtt{0}^{n_v - a_\ell}\mathtt{1}s_v$. Due to Proposition 4.6,

$$U = (U_2, U_3, \ldots, U_{N_\ell}, U_1)$$

is distributed as a sequence of uniform random variables observed until a term belongs to $[0, 2^{-\ell + a_\ell}]$, this last term being eventually multiplied by a power of 2, so as to belong to $]2^{-\ell + a_\ell - 1}, 2^{-\ell + a_\ell}]$. As a consequence, the sequence X image of U by the mapping

$$X_i = -\log_2 U_i$$

[1] See for instance [SF96, p. 362].

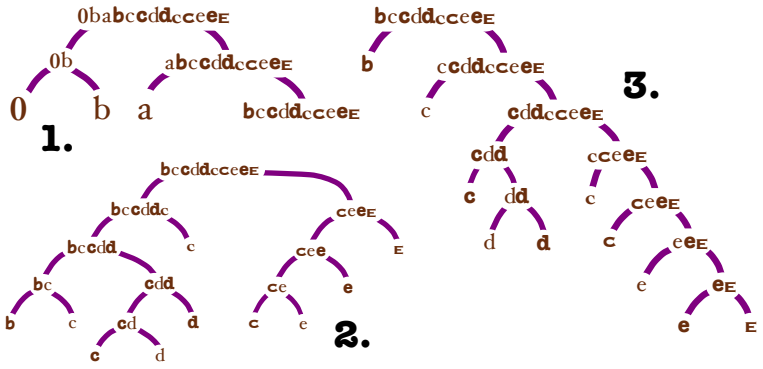

Figure 4.7: 1. $\mathcal{S}^{(2)}$ and $\mathcal{T}^{(2)}$ are equal in this example in which the blades of the family belong to 5 dyadic intervals `a`, `b`, `c`, `d`, `e` sorted in lexicographic order: the only `a` being interleaved between the 2 `b`'s, comparisons between the 2 `b`'s do not matter. Then we show two possible shapes (2. and 3.) for the subtree below `bccddcceee`, depending on the outcomes of comparisons $t_i \gtrless t_j$ and $s_i \gtrless s_j$. The fluctuation $d(\mathtt{E})$ is $7-2=(k_c-1)_+ + (k_e-1)_+$, without $(k_d-1)_+$, for no `d` is involved in this part of the tree.

is distributed as a sequence of exponential random variables observed until a term belongs to $[\ell - a_\ell, +\infty[$, this last term being eventually shifted by an integer, so as to belong to $[\ell - a_\ell, \ell - a_\ell + 1[$. Then the construction of Section 4.2.5.b, based on the sequence $(\ell - a_\ell, X_2, X_3, \ldots, X_{N_\ell})$, gives a Yule family tree $\mathfrak{Y}_{\ell - a_\ell}$ with lifetime $\ell - a_\ell$, and with intensity $\rho = \ln 2$. Then the tree induced by the points of $\mathfrak{Y}_{\ell-a_\ell}$ whose distance to the root is an integer between 0 and $\ell - a_\ell$ is $\mathfrak{GW}_\ell$. Also, removing the nodes of $\mathfrak{GW}_\ell$ splits $\mathfrak{Y}_{\ell-a_\ell}$ into connected components $\mathfrak{Y}^{(x)}$, one for each interior node $x = (\ell - k, s_v)$ of $\mathfrak{GW}_\ell$, connected components that are independent and distributed as $\mathfrak{Y}_1$ (still with intensity $\ln 2$). The set of leaves of $\mathfrak{Y}^{(x)}$ is $\mathcal{O}_{k,v}$.

4.2.5.d Proof of Proposition 4.7

For some element (n_j, s_j), $j > v$, of $\mathcal{O}_{k,v}$, $n_j = \ell - k - 1$. Also, s_j is the dyadic expansion of

$$\tilde{U}_j = 2^{\ell - a_\ell - k} U_j - 1 = 2^{\ell - a_\ell - k - X_j} - 1,$$

and $\tilde{U}_j$ belongs to the dyadic interval $[m2^{-a_\ell}, (m+1)2^{-a_\ell})$ if and only if

$$\tilde{X}_j = \{X_j\} = X_j - \ell + k + a_\ell + 1 \in I_m,$$

in which the intervals

$$I_m \;=\; \left[-\log_2\left(\tfrac{1}{2} + (m+1)2^{-a_\ell - 1}\right) \;;\; -\log_2\left(\tfrac{1}{2} + m2^{-a_\ell - 1}\right)\right)$$

satisfy

$$\biguplus_{0\leq m<2^{a_\ell}} I_m \;=\; [0,1).$$

In turn, the points of $\mathfrak{Y}^{(k,w)}$ with ordinates

$$y_m \;=\; -\log_2\left(\tfrac{1}{2}+m2^{-a_\ell-1}\right), \quad 0\leq m<2^{a_\ell},$$

induce Galton-Watson subtrees $\mathfrak{GW}_{(k,w)}$ whose offspring distribution changes at each generation, and, at generation m, is a geometric distribution with parameter

$$p_m \;=\; \frac{\frac{1}{2}+m2^{-a_\ell-1}}{\frac{1}{2}+(m+1)2^{-a_\ell-1}} \;=\; 1-\frac{1}{2^{a_\ell}+m+1}.$$

Incidentally, $\mathfrak{Y}^{(k,w)}$ is distributed as $\mathfrak{Y}_1$ and it is built from the sample of exponential random variables $\tilde{X}_j$, for $(\ell-k-1,s_j)\in\mathcal{O}_{k,w}$. Keeping in mind Lemma 4.8, note that differences between $\mathfrak{T}_\ell$ and $\mathfrak{S}_\ell$ occur only at the level of the families of $\mathfrak{GW}_{(k,w)}$, because, if (n_i,s_i) and (n_j,s_j) are not in the same family but $\tilde{U}_i$ and $\tilde{U}_j$ are in the same dyadic interval, then

- either $n_j\neq n_i$, and the comparisons between (n_j,s_j) and (n_i,s_i), or between (n_j,t_j) and (n_i,t_i), have the same output, for only the n_j' s are involved,

- either $n_j=n_i$, but in the sequence S_ℓ, between (n_j,s_j) and (n_i,s_i), there exists some element (n_k,s_k) such that $n_k>n_i$, and the standard factorization will cut the label just before (n_k,s_k) in both trees, or before a similar element, independently of an eventual inversion between i and j,

- or (n_i,s_i) and (n_j,s_j) are in the same family of $\mathfrak{GW}_\ell$, i.e. $n_j=n_i=\ell-k-1$, and (n_i,s_i) and (n_j,s_j) are in the same subtree $\mathfrak{GW}^{(k,w)}$ but in the sequence S_ℓ, between (n_j,s_j) and (n_i,s_i), there exists some element (n_k,s_k) such that $\tilde{X}_k$ is in some interval I_{m_1} while $\tilde{X}_i$ and $\tilde{X}_j$ are in some I_{m_2} with $m_2>m_1$. Then (n_i,s_i) and (n_j,s_j) land in different subtrees before an eventual inversion between them can produce a difference between $\mathfrak{T}_\ell$ and $\mathfrak{S}_\ell$.

Inside such a family of $\mathfrak{GW}^{(k,w)}$, inversion may occur between any couple of leaves, and may produce differences between the corresponding subtrees $\mathfrak{T}^{(k,w)}$ and $\mathfrak{S}^{(k,w)}$ of $\mathfrak{T}_\ell$ and $\mathfrak{S}_\ell$, so that the operation $\mathfrak{L}$ can produce any binary tree, as far as we know, only the number f of leaves of these 2 binary subtrees of $\mathfrak{T}^{(k,w)}$ and $\mathfrak{S}^{(k,w)}$ being given: thus the maximal depth of a leave is $f-1$ and the minimal depth is 1, if $f>1$. In any case the difference between the depth of a leave in $\mathfrak{GW}^{(k,w)}$ and in the corresponding subtree of

$\mathfrak{T}_\ell$ is bounded by $(f-2)_+$. Note that in a family at level m, the probability of difference between $\mathfrak{T}_\ell$ and $\mathfrak{S}_\ell$ is thus bounded by

$$\mathcal{O}\left((1-p_m)^2\right) = \mathcal{O}\left(2^{-2a_\ell}\right),$$

which is pretty small. There is an exception, in which the bound is $(f-1)_+$ rather than $(f-2)_+$, when the only needle of $\mathfrak{S}^{(k,w)}$ happens, at some level m_0 that depends on the label of the ancestor of $\mathfrak{Y}^{(k,w)}$. This precludes the Markov property of branching random walks but stochastic monotonicity alleviates the problem.

Let us give a formal argument: consider

$$m_\ell(\theta) \;=\; \mathbb{E}\left[\sum_{1\le v\le N_\ell} e^{\theta d_v}\right],$$

and assign to each blade v the position $-d_v$. Let $Z_t^{(n)}$ denote the number of blades of generation n with a position to the left of t. Then, as in [Big77, page 634],

$$\mathbb{P}\left(D_\ell \ge t\right) = \mathbb{P}\left(Z_{-t}^{(\ell)} \ge 1\right) \;\le \mathbb{E}[Z_{-t}^{(\ell)}] \;\le e^{-\theta t} m_\ell(\theta). \tag{4.2}$$

Set

$$\begin{aligned} p = 1-2^{-a_\ell} \;&=\; 1-\varepsilon \\ &\le \min\left\{p_m \,|\, 0\le m\le 2^{a_\ell}-1\right\} \\ F(p,\theta) = \sum_{k\ge 1} k e^{\theta(k-2)_+} p(1-p)^{k-1} \;&= p(1-e^{-\theta}) + \frac{pe^{-\theta}}{\left(1-e^\theta\varepsilon\right)^2} \\ G(p,\theta) = \sum_{k\ge 1} k e^{\theta(k-1)_+} p(1-p)^{k-1} \;&= \frac{p}{\left(1-e^\theta\varepsilon\right)^2} \end{aligned}$$

The key to the proof of Proposition 4.7, in the spirit of [Big77, Corollary (3.4)], is the next Lemma:

Lemma 4.10. *For $\theta \ge 0$,*

$$m_\ell(\theta) \;\le\; \left(\tfrac{G(1-\varepsilon,\theta)}{F(1-\varepsilon,\theta)}\; F(1-\varepsilon,\theta)^{1/\varepsilon}\right)^{\ell-a_\ell}.$$

Given Lemma 4.10, setting $t = \frac{\ell}{\sqrt{a_\ell}}$ in (4.2), and

$$e^{2\theta_\ell} = 2^{a_\ell} = \tfrac{1}{\varepsilon},$$

and using

$$\begin{aligned} F(p,\theta_\ell) &= 1+\varepsilon+3\varepsilon\sqrt{\varepsilon}+2\varepsilon^2+2\varepsilon^2\sqrt{\varepsilon}+\dots \\ G(p,\theta_\ell) &= 1+2\sqrt{\varepsilon}+2\varepsilon+2\varepsilon\sqrt{\varepsilon}+\dots \end{aligned}$$

one obtains

$$\begin{aligned}\mathbb{P}\left(D_\ell \geq \tfrac{\ell}{\sqrt{a_\ell}}\right) &\leq e^{-\theta_\ell t} m_\ell(\theta_\ell) \\ &\leq \exp\left(-\ell\sqrt{a_\ell}\ln\sqrt{2} + \mathcal{O}(\ell)\right),\end{aligned}$$

which ends the proof of Proposition 4.7.

Proof of Lemma 4.10. Let us consider the words of $\{0,1\}^{a_\ell}$ as an increasing sequence $\left\{w_0 \prec w_1 \prec \cdots \prec w_{-1+\frac{1}{\varepsilon}}\right\}$. Given $k \in [\![0, \ell - a_\ell]\!]$ and $m \in [\![0, \frac{1}{\varepsilon}[\![$, the factorizations of the sequences S_ℓ and T_ℓ in subsequences that start with some term lexicographically smaller than $(k+a_\ell, w_m)$ and end just before the next one are the same, due to Proposition 4.8. These factorizations transfer to the sequence $(X_j)_{1\leq j\leq N_\ell}$: here the factors are subsequences that begin with some X_j larger than $k+y_m$ and end just before the next one. The next level of factorization (fragmentation) is described by the subtrees of $\mathfrak{Y}_\ell$ with leaves at level $k+y_{m+1}$ (i.e. at time $\ell - k - y_{m+1}$) and ancestors at level $k+y_m$. Each of these subtrees $\mathfrak{h}$ is in bijection with binary subtrees of $\mathfrak{S}_\ell$ and of $\mathfrak{T}_\ell$ with the same number $\mathcal{O}_\mathfrak{h}$ of leaves (needles excluded), but eventually different shapes, leading to different depths of their leaves, relatively to the 2 subtrees. We argued that $\mathcal{O}_\mathfrak{h}$ has a geometric distribution with parameter $p_m \geq p$, independently of the other subtrees, and that for each leaf, the difference in depth is bounded by $(\mathcal{O}_\mathfrak{h} - 2)_+$, in the absence of a needle, and by $(\mathcal{O}_\mathfrak{h} - 1)_+$, in presence of a needle. The uniform bound $(\mathcal{O}_\mathfrak{h} - 1)_+$ would be easier to handle, fitting perfectly in the context of branching random walks, but it is too crude for our purposes.

In each subtree $\mathfrak{Y}^{(k,w)}$, there are 2^{a_ℓ} levels of subtrees of type $\mathfrak{h}$, and there exists exactly one needle, in some subtree located on the right at a level $m(w)$ such that $\{X_w\} \in I_{m(w)}$: this level $m(w)$ depends on the levels of the tree $\mathfrak{Y}_\ell$ that are closer to the root of $\mathfrak{Y}_\ell$. Let us bound the differences in depth of each leaf of a subtree $\mathfrak{h}$ at level $m(w)$ of $\mathfrak{Y}^{(k,w)}$ by $d_\mathfrak{h} = (\mathcal{O}_\mathfrak{h} - 1)_+$, and let us bound by $d_\mathfrak{h} = (\mathcal{O}_\mathfrak{h} - 2)_+$ the differences in depth of each leaf of the subtrees $\mathfrak{h}$ of $\mathfrak{Y}^{(k,w)}$ at other levels than $m(w)$. Let us call $\mathfrak{Y}_k$ the part of $\mathfrak{Y}_\ell$ that is at a distance of root smaller or equal than k. Then, given $\mathfrak{Y}_k$, the conditional distribution of $\mathfrak{Y}^{(k,w)}$ endowed with elementary displacements $d_\mathfrak{h}$ from the roots to the leaves of each subtree $\mathfrak{h}$, is that of a non homogeneous branching random walk. Thus, if $\delta_j^{(k,w)}$ denotes the difference in depth for the leaf j between $\mathfrak{S}^{(k,w)}$ and $\mathfrak{T}^{(k,w)}$, then

$$\mathbb{E}\left[\sum_j e^{\theta\delta_j^{(k,w)}} \,\middle|\, \mathfrak{Y}_k\right] \leq \tfrac{G(p_{m(w)},\theta)}{F(p_{m(w)},\theta)} \prod_{m=0}^{2^{a_\ell}-1} F(p_m,\theta).$$

Note that F and G are decreasing in p, for the geometric distribution is stochastically decreasing in its parameter. Thus, by stochastic monotonicity,

$$\mathbb{E}\left[\sum_j e^{\theta\delta_j^{(k,w)}} \,\middle|\, \mathfrak{Y}_k\right] \leq \tfrac{G(1-\varepsilon,\theta)}{F(1-\varepsilon,\theta)}\ F(1-\varepsilon,\theta)^{1/\varepsilon}.$$

Now, let us denote by $d_i^{(k)}$ the difference in depth of a leaf $i=(k,w)$ between $\mathfrak{L}^{\ell-k}(T_\ell)$ and $\mathfrak{L}^{\ell-k}(S_\ell)$, and assume that j is a generic leaf of $\mathfrak{S}^{(k,w)}$ or $\mathfrak{T}^{(k,w)}$, indifferently. Then

$$d_j^{(k+1)} \leq d_i^{(k)} + \delta_j^i,$$

and

$$\begin{aligned}\mathbb{E}\left[\sum_j e^{\theta d_j^{(k+1)}}\right] &\leq \mathbb{E}\left[e^{\theta d_i^{(k)}} \sum_j e^{\theta\delta_j^i}\right]\\ &= \mathbb{E}\left[e^{\theta d_i^{(k)}}\ \mathbb{E}\left[\sum_j e^{\theta\delta_j^i} \,\middle|\, \mathfrak{Y}_k\right]\right]\\ &\leq \mathbb{E}\left[e^{\theta d_i^{(k)}}\right]\ \tfrac{G(1-\varepsilon,\theta)}{F(1-\varepsilon,\theta)}\ F(1-\varepsilon,\theta)^{1/\varepsilon}.\end{aligned}$$

Lemma 4.10 follows by induction. □

4.2.6 Depths of leaves of the Lyndon tree in terms of the Yule process

As observed for instance in [CKMR05, Lemma 2.1], $\mathfrak{Y}_{\ell-a_\ell}$ seen as a planar tree, with no edge length, or with edge length 1, is a random binary search tree. That is to say, $\mathfrak{U}_\ell$ is a random binary search tree, for the planar tree structures of $\mathfrak{Y}_{\ell-a_\ell}$ and $\mathfrak{U}_\ell$ depend only on the relative order of the words $\mathtt{0}^{n_v-a_\ell}\mathtt{1}s_v$ for $\mathfrak{U}_\ell$, and of the real numbers X_v for $\mathfrak{Y}_{\ell-a_\ell}$, through the same algorithm, and the mapping $\mathtt{0}^{n_v-a_\ell}\mathtt{1}s_v \longrightarrow X_v$ is monotone. Thus the depth of the blade v in the binary search tree $\mathfrak{U}_\ell$ induced by $\mathfrak{S}_\ell$ is the depth of the leaf v corresponding to X_v in $\mathfrak{Y}_{\ell-a_\ell}$.

Now, the difference between the depth of the blade v in $\mathfrak{U}_\ell$ and its depth in $\mathfrak{S}_\ell$ is the number of needles on the path to the root, whose expression is given by Proposition 4.11 below. For a given blade v of $\mathfrak{S}_\ell$, consider the marked point process Π_v formed by the vertices of $\mathfrak{Y}_{\ell-a_\ell}$ on the path from v to the root, the leaf v and the root excluded. This path is naturally identified with $[0,\ell-a_\ell]$, the leaf being at 0 and the root at $\ell-a_\ell$, for instance, and the mark being 0 or 1, respectively, according to the side of the branching, say left or right, leading to a decomposition

$$\Pi_v \ = \ \Pi_v^{(0)} \cup \Pi_v^{(1)}.$$

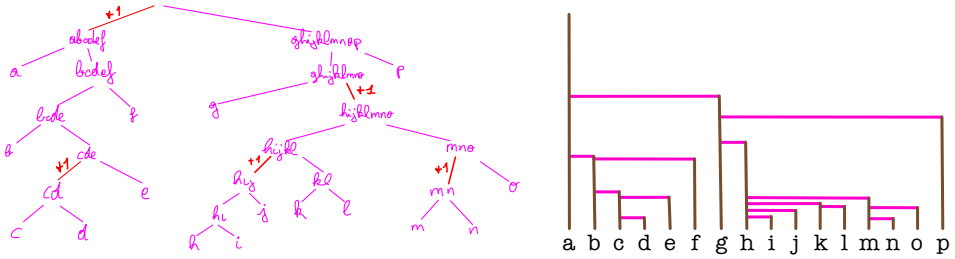

Figure 4.8: Trees $\mathfrak{S}_\ell$ and $\mathfrak{Y}_{\ell-a_\ell}$. In $\mathfrak{S}_\ell$, the sign $+1$ marks the presence of a needle, while in $\mathfrak{Y}_{\ell-a_\ell}$ the lengths of the sticks are rather $2^{Y_i} = 1/U_i$: a needle occurs above the branchings `ab`, `cd`, `hm`, `mn` because, in $\mathfrak{Y}_{\ell-a_\ell}$, the corresponding first stick is at least twice longer than the second, and above `hj` because the factor is larger than 4.

By convention, the mark for both points 0 and $\ell - a_\ell$ is one, and unless mentioned otherwise, they are not included in the point processes. For a point process $\pi = \{\xi_1 < \xi_2 < \cdots < \xi_{k-1} < \xi_k\}$ in some interval $[0, m]$, $G(\pi)$ denotes

$$G(\pi) \;=\; \sum_{r=1}^{k+1} \lfloor \xi_r - \xi_{r-1} \rfloor, \tag{4.3}$$

in which $\xi_{k+1} = m$ and $\xi_0 = 0$. We have

Proposition 4.11. *For* $1 \le v \le N_\ell$,

$$0 \le h_v - (\#\Pi_v + 1 + G(\Pi_v^{(1)})) \le 1.$$

Proof. Here $\#\Pi_v + 1$ accounts for the depth of v in $\mathfrak{U}_\ell$, $\lfloor \xi_1 \rfloor$ is $n_v - a_\ell$, and one can check that $\lfloor \xi_r - \xi_{r-1} \rfloor = a$ if the corresponding labels w_r and w_{r-1} satisfy $(n_{w_r} - a, s_{w_r}) \prec (n_{w_{r-1}}, s_{w_{r-1}})$ but $(n_{w_r} - a - 1, s_{w_r}) \succ (n_{w_{r-1}}, s_{w_{r-1}})$, in which case the corresponding edge in $\mathfrak{U}_\ell$ is obtained by erasing a needles of $\mathfrak{S}_\ell$. In order for the relation in Proposition 4.11 to be exact, the point ξ_{k+1} at the top should be X_1, while it is defined by $\xi_{k+1} = m = \ell - a_\ell$ here. The inequality $0 \le X_1 - \ell + a_\ell < 1$ entails that $0 \le h_v - (\#\Pi_v + 1 + G(\Pi_v^{(1)})) \le 1$. □

The special rôle of $\Pi_v^{(1)}$ in Proposition 4.11 reflects the asymmetry of the Lyndon tree. Since $G(\Pi_v^{(1)})$ tends to be large when $\#\Pi_v^{(1)}$ is small, and small when $\#\Pi_v^{(1)}$ is large, the difference between the height and the saturation level should be smaller for the Lyndon tree than for the binary search tree.

From now on, we shall assume that the depth h_v of a leaf v of $\mathfrak{S}_\ell$ is given by

$$h_v = \#\Pi_v + 1 + G(\Pi_v^{(1)}),$$

since this does not affect the limit in Theorem 4.3.

4.3 Proof of Theorem 4.3

In Section 4.3.1 we prove the key formula (4.4) about the Yule tree $\mathfrak{Y}_\ell$. In fine, through the couplings between $\mathfrak{T}_\ell$ and $\mathfrak{S}_\ell$ on one hand, between $\mathfrak{Y}_\ell$ and $\mathfrak{S}_\ell$ on the other hand (cf. Propositions 4.7 and 4.11), Proposition 4.12 allows to describe the profile of $\mathfrak{T}_\ell$, i.e. the repartition of the leaves of $\mathfrak{T}_\ell$ according to their depth. Section 4.3.2 presents a sketch of the proof of Theorem 4.3. The rest of Section 4.3 is devoted to a detailed proof of Theorem 4.3.

4.3.1 A many-to-one formula

A leaf v of $\mathfrak{Y}_\ell$ is said to be *of type* (m, n, A) if its left (resp. right) depth in $\mathfrak{Y}_\ell$ is m (resp. n) and if its point process $\Pi_v^{(1)}$ belongs to A. Let $\pi_{\ell,m,n,A}$ denote the average number of leaves of type (m, n, A) in $\mathfrak{Y}_\ell$ and let $\mathbb{U}_{n,\ell}$ denote the uniform probability distribution on the simplex $\{0 < \xi_1 < \xi_2 < \cdots < \xi_n < \ell\}$. Then

Proposition 4.12.

$$\pi_{\ell,m,n,A} \;=\; \frac{(\ell \ln 2)^{m+n} 2^{-\ell}}{m!n!}\; \mathbb{U}_{n,\ell}(A). \tag{4.4}$$

Up to a factor 2^{m+n}, the right hand of (4.4) is the probability that two independent Poisson processes on $[0, \ell]$, with intensity $\rho/2 = \ln\sqrt{2}$, $\Pi^{(0)}$ (resp. $\Pi^{(1)}$), have m (resp. n) points, and that $\Pi^{(1)}$ belongs to A. This could be seen as an elementary instance of the many-to-one formula for branching random walks, cf. [HH09].

Proof. Consider the probability p that a random direction of a random Yule tree $\mathfrak{Y}_\infty$ produces a leaf of type (m, n, A). On one hand, conditioning first on $\mathfrak{Y}_\infty$, the probability is the number of such leaves times the probability 2^{-m-n} of the path to one of these leaves, so that, going to expectations,

$$p = 2^{-m-n}\pi_{\ell,m,n,A}.$$

On the other hand, consider the random walk, on $\mathfrak{Y}_\infty$, of a particle with life time ℓ: the particle chooses *left* or *right* at birth times that form Poisson process with intensity $\rho = \ln 2$ on $[0, \ell]$, and due to the colouring theorem (cf. Kingman, Ch. 5) the times when the particle goes left (resp. right) form two independent Poisson processes $\Pi^{(0)}$ and $\Pi^{(1)}$ with intensity $\rho/2$. Then p is the probability that $\#\Pi^{(0)} = m$, $\#\Pi^{(1)} = n$, and $\Pi^{(1)} \in A$. Thus:

$$p = \frac{(\ell \ln 2)^m 2^{-m-\ell/2}}{m!}\; \frac{(\ell \ln 2)^n 2^{-n-\ell/2}}{n!}\; \mathbb{U}_{n,\ell}\,(A)\,.$$

□

When A_g is the set of point processes Π on $[0,\ell]$ such that $G(\Pi) = g$, we set

$$\pi_{\ell,m,n,A_g} \;=\; \pi_{\ell,m,n,g}.$$

4.3.2 Sketch of proof

To sketch the final argument, that follows [BD07], it will be convenient to think of $\Psi(\lambda,\mu,\nu)$ as the following limit:

$$\Psi(\lambda,\mu,\nu) \;=\; \lim_n n^{-1}\;\ln\left(\pi_{\ell,m,n,g}\right), \tag{4.5}$$

when the sequence $(\lambda_n,\nu_n,\mu_n) = (\ell/n, m/n, g/n)$ converges to (λ,ν,μ). Thus the number of blades of type (m,n,g) in $\mathfrak{S}_\ell$ would be, approximately:

$$e^{n\Psi(\lambda,\mu,\nu)} = e^{\ell\Psi(\lambda,\mu,\nu)/\lambda}.$$

Results in the direction of formula (4.5), sufficient for our purposes, are proven in Section 4.3.3. On each blade v of type (m,n,g) of $\mathfrak{S}_\ell$, with depth

$$m+n+g = \frac{1+\nu+\mu}{\lambda}\;\ell,$$

we graft the shrub $\mathfrak{t}(v)$. In Section 4.3.4, we prove that the maximum height of a set of k such shrubs behaves like the maximum of a sample of k independent geometric random variables with parameter $1/2$, i.e. the maximum is essentially $\log_2 k$. As a consequence, the total height of the highest leaf in any shrub that is grafted on a blade of type (m,n,g) in $\mathfrak{S}_\ell$ is approximately

$$m+n+g+\;\log_2\left(e^{n\Psi(\lambda,\mu,\nu)}\right) \;=\; \tfrac{(1+\nu+\mu)\ln 2\;+\;\Psi(\lambda,\mu,\nu)}{\lambda\ln 2}\;\ell.$$

Set

$$\Delta(\lambda,\mu,\nu) = ((1+\nu+\mu)\ln 2 \;+\; \Psi(\lambda,\mu,\nu))/\lambda\ln 2.$$

Then the highest leaf in the tree $\mathfrak{A}_\ell$ is approximately $\Delta^\bullet\ell$ high, in which

$$\Delta^\bullet \;=\; \sup_{\lambda,\mu,\nu>0}\Delta(\lambda,\mu,\nu). \tag{4.6}$$

The supremum is obtained for $\nu = 2\lambda\ln 2$, leading to

$$\Delta^\bullet = \sup_{\lambda,\mu>0}\Delta(\lambda,\mu,2\lambda\ln 2).$$

Finally, due to Proposition 4.7, the same is true when the shrubs are grafted on the blades of $\mathfrak{T}_\ell$, producing $\mathfrak{L}(L^\ell)$ instead of $\mathfrak{A}_\ell$.

Remark. After scaling to recover the Lyndon tree based on a n-letters long Lyndon word, one finds that the maximal height occurs in a shrub grafted on a leaf of $\mathfrak{S}_{\log_2 n}$ whose left (resp. right) depth in $\mathfrak{U}_{\log_2 n}$ is approximately $2\ln n$ (resp. $1.62..\ln n$). This can be compared with the height $4.31..\ln n$ of $\mathfrak{U}_{\log_2 n}$, induced at equal parts by left (resp. right) depths $2.15..\ln n$. The contribution of needles is approximately $0.86..\ln n$, leading to a height $4.48..\ln n$ for the leaf of $\mathfrak{S}_{\log_2 n}$ on which the shrub with the highest top is grafted. Finally, this shrub is approximately $0.61..\ln n$ high.

4.3.3 Asymptotic behavior of $n^{-1}\ln \pi_{\ell,m,n,g}$

4.3.3.a Heuristic considerations

For a finite or infinite sequence of non negative numbers $b = (b_j)_{j\in I\subset\mathbb{N}_0}$, let us denote

$$|b| \;=\; \sum_{j\in I} b_j, \quad \langle b\rangle \;=\; \sum_{j\in I} j b_j, \quad \text{and} \quad \mathcal{H}(b) \;=\; -\sum_{j\in I} b_j \ln(b_j),$$

and if the entries are integers, set:

$$b! \;=\; \prod b_i! \quad \text{and} \quad \binom{|b|}{b} \;=\; \frac{|b|\,!}{\prod b_i!},$$

whenever they are defined. If $|b| = 1$, $\mathcal{H}(b)$ is the Shannon entropy of b.

Under $\mathbb{U}_{n,\ell}$, rather than the ξ_j's, we shall consider the vector γ of gaps between the order statistics, defined by

$$\gamma_j = \xi_j - \xi_{j-1},\ 1 \le j \le n+1, \text{ and } \gamma = (\gamma_j)_j,$$

with the convention that $\xi_0 = 0$ and $\xi_{n+1} = \ell$. The random vector γ is uniformly distributed on the simplex

$$\mathcal{D}_{n,\ell} \;=\; \left\{ \sum_{j=1}^{n+1} x_j = \ell \quad \text{and} \quad \forall j,\ x_j \ge 0 \right\},$$

and its distribution is denoted $\mathbb{U}_{n,\ell}$ again. Also, set $\rho = (\rho_j)_{1\le j\le n+1}$, in which

$$\rho_j \;=\; \lfloor \gamma_j \rfloor.$$

Then $G = |\rho|$. The probability $\mathbb{U}_{n,\ell}(\rho = r)$ depends only on $|r|$ and is given by:

$$\mathbb{U}_{n,\ell}\,(\rho = r) = n!\,\ell^{-n}\mathbb{P}\left(\ell - |r| - 1 < \sum_{j=1}^{n} U_j < \ell - |r| \right),$$

in which the U_i's are a sequence of i.i.d. random variables uniform on $[0,1]$, $\mathbb{P}(\dots)$ is the Lebesgue measure of the domain $\{\rho = r\}$, and $\ell^n/n!$ is the Lebesgue measure of $\mathcal{D}_{n,\ell}$. In order to take advantage of the symmetric rôle of the ρ_j's, let $\beta = (\beta_j)_{j\geq 0}$ be defined by

$$\beta_j = \#\left\{1 \leq i \leq n+1 \,|\, \rho_i = j\right\},$$

so that

$$|\beta| = n+1, \text{ and } \quad \langle\beta\rangle = |\rho| = G(\xi).$$

Then

$$\mathbb{U}_{n,\ell}\left(\beta = b\right) \;=\; \frac{n!}{\ell^n}\binom{|b|}{b}\,\mathbb{P}\left(\ell - \langle b\rangle - 1 < \sum_{j=1}^{n} U_j < \ell - \langle b\rangle\right),$$

and (4.4) becomes

$$\pi_{\ell,m,n,g} = \frac{\ell^m(\ln 2)^{m+n}}{m!\,2^\ell}\mathbb{P}\left(\ell - g - 1 < \sum_{j=1}^{n} U_j < \ell - g\right) \sum_{\langle b\rangle = g}\binom{|b|}{b}. \quad (4.7)$$

Set

$$p_{n,m} = \mathbb{P}\left(m - 1 < \sum_{j=1}^{n} U_j < m\right) \qquad \text{and} \qquad t_{n,g} = \sum_{\langle b\rangle = g, |b| = n+1}\binom{|b|}{b},$$

so that (4.7) can be written

$$\pi_{\ell,m,n,g} = \frac{\ell^m(\ln 2)^{m+n}}{m!\,2^\ell}\; p_{n,\ell-g}\; t_{n,g}.$$

From stronger results on the asymptotics of the Eulerian numbers (denoted here $(A(k,n))_{0\leq k\leq n}$), that can be found in [GK94][formulas (6.12) to (6.16), page 299][2], we know that

Lemma 4.13. *For $0 < \theta < 1$, the limit:*

$$\Xi(\theta) \;=\; \lim_n \tfrac{1}{n}\ln p_{n,\theta n}$$

exists and is given by

$$\Xi(\theta) \;=\; \ln\sinh\alpha - \alpha\coth\alpha + 1 - \ln\alpha,$$

in which α is given implicitly by $-\alpha^{-1} + 1 + \coth\alpha = 2\theta$.

[2] According to [Tan73], $A(k,n) = n!\; p_{n,k+1}$.

On the other hand, for distributions b such that

$$b_j \simeq c_j(n+1), \quad |c| = 1, \quad \text{and} \quad \langle c \rangle = \mu,$$

we expect that

$$\tfrac{1}{n} \ln \left(\binom{|b|}{b} \right) = \mathcal{H}(c) + o(1).$$

Following the lines of Lemma 8.3.1 in [Ash65, p. 238], one obtains that

$$\mathcal{H}(c) \leq \mathcal{H}\left(d^{(\mu)}\right) = (1+\mu)\ln(1+\mu) - \mu \ln \mu,$$

in which $d^{(\mu)}$ is the geometric distribution with expectation μ. As usual in large deviation theory, only the leading term of $t_{n,\mu n}$, provided by $c = d^{(\mu)}$, contributes to the limit of $n^{-1} \ln t_{n,\mu n}$, so that we expect the following behavior:

$$n^{-1} \ln t_{n,\mu n} \simeq \mathcal{H}\left(d^{(\mu)}\right).$$

Together with (4.4), this would lead to the following expression for Ψ:

$$\Psi(\lambda,\mu,\nu) = \mathcal{H}\left(d^{(\mu)}\right) + \Phi(\lambda,\nu) + \Xi(\lambda-\mu),$$

in which

$$\Phi(\lambda,\nu) = \lim_n \tfrac{1}{n} \ln \frac{\ell^m (\ln 2)^{m+n}}{m!\, 2^\ell} = \ln\left(\tfrac{(e\lambda \ln 2)^\nu \ln 2}{\nu^\nu 2^\lambda}\right).$$

The analysis of the contribution of the shrubs rests on more or less precise upper and lower bounds for $n^{-1} \ln(\pi_{\ell,m,n,g})$, to which the rest of the Section is devoted.

4.3.3.b Lower bound for $n^{-1} \ln \pi_{\ell,m,n,g}$

The lower bounds for $p_{n,g}$ and $t_{n,g}$ do not need to be very precise for our purposes: for $p_{n,g}$ the "lower bound" given by the existence of the limit in Lemma 4.13 is precise enough. With $\mu = g/n$, according to our heuristic considerations, the lower bound for $t_{n,g}$ provided by any term $\binom{|b|}{b}$ such that b is close to $d^{(\mu)}n$ should be good enough. The sequence $c^{(k)}$ defined by

- $c_j^{(k)} = d_j^{(\mu)}$ for $j \leq k$.
- $c_{k+1}^{(k)} = \sum_{j>k} d_j^{(\mu)} = 1 - \sum_{i \leq k} c_i^{(k)}$
- $c_i^{(k)} = 0$ for $i \geq k+2$.

satisfies

$$\begin{aligned}\mathcal{H}(c^{(k)}) - \mathcal{H}(d^{(\mu)}) &= \left(\tfrac{\mu}{1+\mu}\right)^{k+1}\left(k\ln\mu + (1+\mu)\ln\left(\tfrac{\mu}{1+\mu}\right)\right),\\ \langle c^{(k)}\rangle &= \mu\left(1-(\tfrac{\mu}{1+\mu})^{k+1}\right).\end{aligned}$$

Consider then the sequence of integers $b^{(n)}$ defined by

- $\forall i \leq k, b_i^{(n)} = \lfloor c_i^{(k)}(1+n)\rfloor$,
- $b_{k+1}^{(n)} = n+1-\sum_{i\leq k} b_i^{(n)}$,
- $b_i^{(n)} = 0$ pour $i \geq k+2$,

so that

$$\left|b_i^{(n)} - (n+1)nc_i^{(k)}\right| \leq \begin{cases} 1 & \text{if} \quad i \leq k,\\ k+1 & \text{if} \quad i = k+1,\\ 0 & \text{if} \quad i \geq k+2.\end{cases}$$

Thus,

$$\begin{aligned}\langle b^{(n)}\rangle &\leq (n+1)\langle c^{(k)}\rangle + (k+1)(b_{k+1}^{(n)} - (n+1)c_{k+1}^{(k)})\\ &\leq (n+1)\langle c^{(k)}\rangle + (k+1)^2\\ &\leq (n+1)\mu + (k+1)^2.\end{aligned}$$

On the other hand, since $b^{(n)}$ is obtained, from $(n+1)\,c^{(k)}$, by transport of mass from $i \leq k$ towards $k+1$, we have $\langle b^{(n)}\rangle \geq (n+1)\langle c^{(k)}\rangle$, and as a consequence

$$\langle b^{(n)}\rangle \geq (n+1)\mu\left(1-(\tfrac{\mu}{1+\mu})^{k+1}\right).$$

Due to the variations of $\varphi(x) = x\ln x - x$, and due to the inequality:

$$|\ln(n!) - \varphi(n)| \leq \ln(4n), \tag{4.8}$$

that holds for $n \geq 1$, we obtain that:

$$\begin{aligned}|\ln\binom{|b^{(n)}|}{b^{(n)}} - (1+n)\mathcal{H}(d^{(\mu)})| &\leq (n+1)|\mathcal{H}(c^{(k)}) - \mathcal{H}(d^{(\mu)})| + 4k\ln(5n)\\ &= \mathcal{O}(\ln^2 n),\end{aligned}$$

if one chooses $k, n \geq 5$, and also $k = c\ln n$ for c large enough, namely $c\ln(1+\frac{1}{\mu}) > 1$, so that $|\mathcal{H}(c^{(k)}) - \mathcal{H}(d^{(\mu)})|$ is small enough. Now b^n satisfies $\langle b^n\rangle = \mu n + \delta$, instead of μn. This is corrected by transferring a mass δ from 1 to 0, say, i.e. considering $b^n + (\delta, -\delta, 0, 0, \dots)$, causing thus a variation

$\delta\left(\ln(1+\frac{1}{\mu})+\Theta_\mu(\frac{1}{n})\right)$ of $\ln\binom{|b|}{b}$. Here δ is $\mathcal{O}(\ln^2 n)$, thus at least one term of the sum $t_{n,g}$ is equal to $\exp(n\mathcal{H}(d^{(\mu)})+\mathcal{O}(\ln^2 n))$, so that

$$t_{n,g} \geq \exp(n\mathcal{H}(d^{(\mu)})+\mathcal{O}(\ln^2 n)). \tag{4.9}$$

Owing to (4.8), we also have

$$\left| n\Phi(\lambda,\nu) - \ln \frac{\ell^m (\ln 2)^{m+n}}{m!\, 2^\ell} \right| \quad \leq \quad \ln(4m) = \ln(4\nu n). \tag{4.10}$$

Thus, owing to Lemma 4.13 and relations (4.9), (4.10), we obtain:

Proposition 4.14. *For $0<\varepsilon<0.05$, when T is large enough, the set*

$$L_{T,\varepsilon} = \{(m,n,g)\,|\,\pi_{T,m,n,g} > e^{\varepsilon n} \text{ and } m+n+g+\log_2 \pi_{T,m,n,g} > (\Delta^\bullet-\varepsilon)T\}$$

is not empty.

Proof. First we prove that, for any $\varepsilon \in (0, 0.05)$, the set

$$D_\varepsilon = \left\{(\lambda,\mu,\nu)\in(0,+\infty)^3 \mid \Psi(\lambda,\mu,\nu)>2\varepsilon \text{ and } \Delta(\lambda,\mu,\nu)>\Delta^\bullet-\varepsilon/2\right\}$$

contains a box. Consider the two functions

$$\tilde{\Delta}(\lambda,\mu)=\Delta(\lambda,\mu,2\lambda\ln 2) \text{ and } \tilde{\Psi}(\lambda,\mu)=\Psi(\lambda,\mu,2\lambda\ln 2).$$

When $\tilde{\Delta}$ reaches its maximum $\Delta^\bullet$, the corresponding value of $\tilde{\Psi}$, $\Psi^\bullet$, is positive. Actually, $\{\tilde{\Delta}\geq 5\ln 2\}$ entails $\{\tilde{\Psi}\geq 0.1\}$. The domain $\{(\lambda,\mu)|\tilde{\Delta}>5\ln 2\}$ is bounded and bounded away from $\lambda=0$. Both Ψ and Δ are Lipschitz continuous as functions of ν on any domain

$$\left\{(\lambda,\mu,\nu)|\tilde{\Delta}>5\ln 2 \text{ and } |\nu-2\lambda\ln 2|\leq\eta\right\}$$

with η not too large. This entails that D_ε contains a box. Thus it contains points of the form $(T,g,m)/n$ as long as T is large enough. If $(T,g,m)/n\in D_\varepsilon$, Lemma 4.13 and relations (4.9), (4.10) entail that if T is large enough then $(m,n,g)\in L_{T,\varepsilon}$. □

4.3.3.c Uniform upper bound on each term of $t_{n,g}$

For a sequence such that $|b|=n+1$ and $\langle b\rangle=\mu n$, the Markov inequality entails that

$$\sum_{i\leq\tau_n} b_i \quad \geq \quad \tfrac{\tau_n-\mu}{\tau_n}\, n.$$

Set

$$\mathcal{F}_{\mu,n} \quad = \quad \left\{ c \;\middle|\; c_i=0 \text{ for } i\geq\tau_n+1 \text{ and } n\geq|c|\geq\tfrac{\tau_n-\mu}{\tau_n}\, n, \langle c\rangle\leq\mu n \right\}.$$

Then

$$\inf_{\langle b\rangle=\mu n} b! \;\geq\; \inf_{c\in\mathcal{F}_{\mu,n}} c!$$

for the truncation, after the τ_nth term, of a sequence b such that $\langle b\rangle = \mu n$ produces an element c of $\mathcal{F}_{\mu,n}$ such that $b! \geq c!$.

Set $\tilde{c}_i = \frac{c_i}{n}$, so that $1 \geq |\tilde{c}| \geq \frac{\tau_n-\mu}{\tau_n}$. For $c \in \mathcal{F}_{\mu,n}$, thanks to (4.8), we have

$$\ln(c!) \geq n(\ln(n)-1)(1-\tfrac{\mu}{\tau_n}) \;-\; n\mathcal{H}\left(\tilde{c}\right) - (1+\tau_n)\ln(4n).$$

For $\tau_n = \sqrt{\mu n}$, this leads to:

$$\ln(c!) \geq n(\ln(n)-1) \;-\; n\mathcal{H}\left(\tilde{c}\right) - (1+2\sqrt{\mu n})\ln 4n.$$

Now, following the lines of Lemma 8.3.1 in [Ash65, p. 238], one obtains that

Proposition 4.15. *If* $|a| \leq 1$,

$$\begin{aligned}\mathcal{H}\left(a\right) &\leq (|a|+\langle a\rangle)\ln(|a|+\langle a\rangle) - 2|a|\ln|a| - \langle a\rangle\ln\langle a\rangle,\\ &\leq \mathcal{H}\left(d^{(\langle a\rangle)}\right) - 2|a|\ln|a|.\end{aligned}$$

Proof. Set $d_i = |a|\,\frac{\beta^i}{(1+\beta)^{i+1}}\,\mathbb{1}_{i\geq 0}$, so that $|d| = |a|$ and $\langle d\rangle = |a|\beta$. Then

$$\begin{aligned}\mathcal{H}\left(a\right) &= \sum_{i\geq 0} a_i \ln\tfrac{d_i}{a_i} - \sum_{i\geq 0} a_i(\ln\tfrac{|a|}{1+\beta} \,+\, i\ln\tfrac{\beta}{1+\beta})\\ &\leq -\sum_{i\geq 0} a_i(\ln\tfrac{|a|}{1+\beta} \,+\, i\ln\tfrac{\beta}{1+\beta})\\ &= -|a|\ln|a| + (|a|+\langle a\rangle)\ln(1+\beta) - \langle a\rangle\ln\beta.\end{aligned}$$

Choosing $\beta = \langle a\rangle/|a|$, i.e. $\langle d\rangle = \langle a\rangle$, leads to the first inequality. Finally, for $0 \leq y \leq 1$ and $x \geq 0$, $(y+x)\ln(y+x) \leq (1+x)\ln(1+x)$. □

Using that $\mu \longrightarrow \mathcal{H}\left(d^{(\mu)}\right)$ is increasing, we obtain, for $c \in \mathcal{F}_{\mu,n}$,

$$\begin{aligned}\mathcal{H}\left(\tilde{c}\right) &\leq \mathcal{H}\left(d^{(\langle\tilde{c}\rangle)}\right) - 2|\tilde{c}|\ln|\tilde{c}|\\ &\leq \mathcal{H}\left(d^{(\mu)}\right) + 2(1-|\tilde{c}|) \leq \mathcal{H}\left(d^{(\mu)}\right) + 2\sqrt{\tfrac{\mu}{n}},\end{aligned}$$

so that

$$\ln(c!) \geq n\ln n - n \;-\; n\mathcal{H}\left(d^{(\mu)}\right) - (1+2\sqrt{\mu n})\ln 12n.$$

This leads to

Proposition 4.16. *For a sequence such that* $|b| = n+1$ *and* $\langle b\rangle = \mu n = g$,

$$\ln\binom{|b|}{b} \;\leq\; n\,\mathcal{H}\left(d^{(\mu)}\right) \;+\; (3+2\sqrt{g})\ln(12n).$$

4.3.3.d Upper bound for $\#\{\langle b\rangle = \mu n\}$

As in Section 4.3.3.a, let us see each sequence b such that $\langle b\rangle = \mu n$ as the distribution $\delta(r)$ of a sequence of integers $r = (r_j)_{1\le j\le n+1}$ such that $|r| = \mu n$, in the sense that

$$b_j = \#\{1 \le i \le n+1 \,|\, r_i = j\}.$$

Now each sequence r can be partially sorted to form a new sequence $\sigma(r)$ as follows:

- the first terms of $\sigma(r)$ are the terms of r smaller than τ_n, sorted in increasing order;
- the other terms follow, and they retain their relative order in the sequence r.

Then $\delta\circ\sigma = \delta$ and the preimage of an element b such that $\langle b\rangle = \mu n$ has at least one element in $\sigma(\delta^{-1}(\{\langle b\rangle = \mu n\}))$. Thus

$$\#\{\langle b\rangle = \mu n\} \le \#\sigma(\delta^{-1}(\{\langle b\rangle = \mu n\})) \le (n+2)^{\tau_n+1} \times (\mu n)^{\mu n/\tau_n},$$

the first term of the product on the right because an increasing sequence is well described by its distribution, here $(b_j)_{0\le j\le\tau_n}$, the second term for the length of the second part of the sequence is shorter than $\mu n/\tau_n$ due to the Markov inequality, and each of its terms is smaller than μn and larger than τ_n. As long as $n \ge 2$, the choice $\tau_n = \sqrt{\mu n}$ leads to:

$$\#\{\langle b\rangle = g\} \;\le\; \mu^{\sqrt{g}}(n+2)^{1+2\sqrt{g}} \le (4ng)^{\sqrt{g}} \times 2n. \tag{4.11}$$

4.3.3.e Upper bound for $p_{n,\ell-g}$

Trite computations lead to the next Proposition:

Proposition 4.17. *For* $0.5 < \theta < 1$,

$$\mathbb{P}\left(\theta n < \sum_{j=1}^{n} U_j\right) \;=\; \mathbb{P}\left((1-\theta)n > \sum_{j=1}^{n} U_j\right) \;\le e^{n\Xi(\theta)}.$$

Proof. Set $S_n = \sum_{j=1}^n U_j$, $\varepsilon = 2\theta - 1 > 0$ and $t > 0$. From the Markov inequality, we obtain that:

$$\begin{aligned}\mathbb{P}\left(\theta n \le S_n\right) &\le \mathbb{E}\left[e^{t(2U-1)}\right]^n \; e^{-t\varepsilon n}\\ &\le e^{nh(t)},\end{aligned}$$

in which

$$h(t) = \ln\left(\frac{\sinh(t)\, e^{-t\varepsilon}}{t}\right).$$

Then

$$\frac{\partial h}{\partial t} = \coth(t) - \varepsilon - \frac{1}{t},$$

and since $t \to \coth(t) - \frac{1}{t}$ is continuous strictly increasing from -1 to 1, h is minimal when $t = \alpha$ in which α is defined by

$$\coth(\alpha) - \varepsilon - \alpha^{-1} = 0.$$

Finally

$$\begin{aligned} h(\alpha) &= \ln\sinh(\alpha) - \alpha\coth(\alpha) + 1 - \ln(\alpha) \\ &= \Xi(\theta). \end{aligned}$$

For $\theta < 0.5$ use that $1 - U_j$'s are uniform, and that $\Xi(\theta) = \Xi(1-\theta)$. □

Corollary 4.18. *For $n \geq 1$,*

$$\ln p_{n,\ell-g} \leq n(\Xi(\lambda - \mu) + \rho_n),$$

in which

$$\rho_n = \Xi(\lambda - \mu - \delta_n) - \Xi(\lambda - \mu),$$

$$0 \leq \delta_n = \left|(-\infty,\, \lambda - \mu - \tfrac{1}{2}] \cap [0, \tfrac{1}{n}]\right| \leq \tfrac{1}{n}.$$

4.3.3.f Conclusion

Combining Proposition 4.16 and (4.11), we obtain that, as long as $n \geq 2$:

$$\ln t_{n,g} \;\leq\; n\,\mathcal{H}\left(d^{(\mu)}\right) \;+\; (3 + 2\sqrt{g})\ln(24ng),$$

With (4.10), this gives

Proposition 4.19. *For $(\lambda, \mu, \nu) = (\ell, g, m)/n$, as long as $n \geq 2$, one has:*

$$\begin{aligned} r_{\ell,m,n,g}\ln 2 &= \ln \pi_{\ell,m,n,g} \;-\; \ell\,\tfrac{\Psi(\lambda,\mu,\nu)}{\lambda} \\ &\leq n\rho_n + \;(3 + 2\sqrt{g})\ln(24ng) + \ln(4m). \end{aligned}$$

4.3.4 Contribution of the shrubs

The proof of Theorem 4.3 has two final steps, contained in the next sections.

4.3.4.a Lower bound for $h(\mathfrak{L}(L^\ell))$

Proposition 4.20. *For any $\varepsilon > 0$,*

$$\lim_\ell \mathbb{P}\left(h(\mathfrak{L}(L^\ell)) \leq (\Delta^\bullet - \varepsilon)\ell\right) = 0.$$

Proof. Keeping in mind Section 4.2.5, we observe that

$$|h(\mathfrak{A}_\ell) - h(\mathfrak{L}(L^\ell))| \leq D_\ell.$$

Thus, according to Proposition 4.7, we only need to prove that

$$\lim_\ell \mathbb{P}\left(h(\mathfrak{A}_\ell) \leq (\Delta^\bullet - \varepsilon)\ell\right) = 0.$$

Working with $\mathfrak{A}_\ell$, we can use the representation of the depth of blades $\mathfrak{S}_\ell$ through a Yule process, as in Section 4.2.6. Let $\mathfrak{B}$ denote the infinite complete binary tree with the language $\mathcal{A}^\star = \{0,1\}^\star$ as set of vertices, and $\{(w, wa) \,|\, (w,a) \in \mathcal{A}^\star \times \mathcal{A}\}$ as edge set, $\emptyset$ being the root. We shall see the infinite family tree $\mathfrak{Y}$ of the Yule process as $\mathfrak{B}$ endowed with edge lengths that form an i.i.d. family of exponential random variables with parameter $\ln 2$. In this setting, we shall interpret the distance between $w \in \mathcal{A}^\star$ and the root $\emptyset$ as the date of the death of w, and the length of the last edge of the path from the root to w as the lifetime of w. An interior point of this last edge inherits the label w of the vertex sitting at the end of the edge.

For any $T > 0$, we shall consider a new tree $\mathfrak{D}_T$ defined as follows:

- the vertices of $\mathfrak{D}_T$ are the points of $\mathfrak{Y}$ at distance nT from the root (for any $n \geq 0$); they are almost surely interior points of some edges;
- a generic edge (x, y) of $\mathfrak{D}_T$ satisfies the following properties:
  - the distance between x and y in $\mathfrak{Y}$ is T;
  - the labels of x and y are of the form (w, ws) with $(w, s) \in \mathcal{A}^\star \times \mathcal{A}^\star$.

If points at distance nT from the root form the n-th generation of a population, then $\mathfrak{D}_T$ is the family tree of a Galton-Watson process with offspring distribution $2^{-T}(1-2^{-T})^{k-1}\,\mathbb{1}_{k\geq 1}$ and average offspring size 2^T. Each individual $w \in \mathcal{A}^\star$ living at time nT possesses, attached to him, the Yule family tree $\mathfrak{Y}^{(w)}$ with life time T that describes his progeny between times nT and $(n+1)T$. Consider then, for each descendent $v = ws$ of w living at time $(n+1)T$, its marked point process $\Pi_{w,ws}$ on $[0,T]$ induced by $\mathfrak{Y}^{(w)}$: this allows to decide whether ws is a descendent of w of type (m,n,A) or not. When ws is a descendent of w of type (m,n,A) in $\mathfrak{Y}^{(w)}$, we say that the edge (w, ws) of $\mathfrak{D}_T$ is open. Erasing the closed edges of $\mathfrak{D}_T$, one obtains a forest $\mathfrak{F}$ whose connected components are independent Galton-Watson trees

with average offspring size $\pi_{T,m,n,A}$. In the sequel, A is the set A_g of point processes Π on $[0,T]$ such that $G(\Pi) = g$, and we set

$$\pi_{T,m,n,A_g} \;=\; \pi_{T,m,n,g} \;=\; \pi_T.$$

With the help of Proposition 4.14, for T large enough, one can choose $(m_T, n_T, g_T) = (m, n, g) \in L_{T,\varepsilon}$, such that

$$\pi_T > e^{T\varepsilon/\lambda} > 1 \text{ and } \rho = m + n + g + \log_2 \pi_T > (\Delta^\bullet - \varepsilon)T. \tag{4.12}$$

As a consequence, the connected components of $\mathfrak{F}$ are supercritical. Since their offspring is bounded by a geometric distribution, according to the Kesten-Stigum Theorem, almost surely, one of the connected components, rooted at a random but finite distance $R \times T$ of the root of $\mathfrak{Y}$, is infinite and satisfies

$$\lim_k Z_k \; \pi_T^{-k} \;=\; C,$$

in which Z_k is the size of generation k of the said component, and C is random but almost surely positive. Finally, the number X_k of elements of type $(k-R)(m,n,g)$ at time kT in $\mathfrak{Y}$ is not smaller than Z_{k-R}, thus it satisfies a.s.

$$\liminf_k \tfrac{1}{k} \;\ln X_k \geq \ln \pi_T.$$

This entails that, for $\eta > 0$ and $c = -\eta + \ln \pi_T$,

$$\lim_k \mathbb{P}\left(X_k \leq e^{ck}\right) = 0.$$

Set $k = \lfloor (\ell - a_\ell)/T \rfloor$ and let S be the set of shrubs grafted on the blades of $\mathfrak{S}_\ell$ that belong to the progeny, at time $\ell - a_\ell$, of the X_k leaves of $\mathfrak{Y}_{kT}$. The set S contains $\tilde{Z}_\ell \geq X_k$ shrubs, and all the corresponding blades have a depth larger than

$$k \; (m + n + g),$$

according to Proposition 4.11. We shall prove that the highest among the $\tilde{Z}_\ell$ shrubs in S has a height M_ℓ given, approximately, by

$$M_\ell \simeq \log_2 \tilde{Z}_\ell \simeq \tfrac{ck}{\ln 2} \simeq \lfloor \ell/T \rfloor \log_2 \pi_T \simeq \tfrac{\Psi^\bullet \ell}{\ln 2},$$

then we shall use the fact that

$$h(\mathfrak{A}_\ell) \geq k \; (m + n + g) + M_\ell. \tag{4.13}$$

In general, a shrub contains at least one run of 1's, but with a probability 2^{-a_ℓ}, thus the height H_v of a shrub $\mathfrak{t}(v)$ is larger than the length of, say, its last run of 1's, i.e. stochastically larger than a geometric distribution with parameter $1/2$. Now, roughly speaking, the probability $p_{r,\alpha,a}$ that ar i.i.d.

shrubs are all lower than $(1-\alpha)\log_2 r$ is smaller than the probability that ar i.i.d. geometric random variables are all smaller than $(1-\alpha)\log_2 r$, i.e.

$$\begin{aligned} p_{r,\alpha,a} &\le (1-2^{-1-(1-\alpha)\log_2 r})^{ar} \\ &\le \exp(-r^{-(1-\alpha)}/2)^{ar} = \exp(-ar^{\alpha}/2), \end{aligned} \tag{4.14}$$

for $a > 0$, $0 < \alpha < 1$. However we are not interested here in a sample of $\tilde{Z}_\ell$ i.i.d. shrubs with the typical distribution, i.e. with the distribution associated with a Bernoulli word observed until the first occurence of $\mathtt{0}^{a_\ell}$: each shrub $\mathfrak{t}(v) \in S$ is selected according to the height $h_v \simeq m_v + n_v + g_v$ of the corresponding blade v, and h_v is determined by the operation of the algorithm $\mathfrak{L}^{a_\ell}$ on the sequence $S_\ell = ((n_w, s_w))_{1\le w\le N_\ell}$, see Definition 4.4, while $\mathfrak{t}(v) = \mathfrak{L}(\mathtt{0}^{a_\ell}\mathtt{1}\mathfrak{p}_v)$.

Now s_v and $\mathfrak{p}_v$ have the same a_ℓ-letters long prefix p_v, their respective suffixes, ζ_v and $\hat{\zeta}_v$, being independent. Furthermore $\hat{\zeta}_v$ is independent of S_ℓ, and as a consequence $\hat{\zeta}_v$ and h_v are independent. The probability that $\hat{\zeta}_v$ begins with the prefix $\mathtt{101}$ is thus $1/8$, and this insures that the last run of $\mathtt{1}$'s in $\mathfrak{p}_v$ is a factor of $\hat{\zeta}_v$ and does not depend on h_v. Let Z be the number of shrubs $\mathfrak{t}(v) \in S$ such that $\hat{\zeta}_v$ begins with $\mathtt{101}$: conditionally given $\tilde{Z}_\ell$, Z has a binomial distribution with parameters $\tilde{Z}_\ell$ and $1/8$. The maximal height $\mathfrak{H}_\ell$ in the sample of $\tilde{Z}_\ell$ shrubs is thus stochastically larger than the maximum of Z i.i.d. geometric random variables with parameter $1/2$, and satisfies, for any $\alpha > 0$,

$$\begin{aligned} \Pi_\ell &= \mathbb{P}\left(\mathfrak{H}_\ell \le (1-\alpha)ck/\ln 2\right) \\ &\le \mathbb{P}\left(Z \le \tfrac{e^{ck}}{9}\right) + \mathbb{P}\left(Z \ge \tfrac{e^{ck}}{9} \text{ and } \mathfrak{H}_\ell \le \tfrac{(1-\alpha)ck}{\ln 2}\right) \\ &\le \mathbb{P}\left(Z \le \tfrac{e^{ck}}{9}\right) + \exp(-e^{\alpha ck}/18), \end{aligned}$$

due to (4.14), and

$$\mathbb{P}\left(Z \le \tfrac{e^{ck}}{9}\right) \le \mathbb{P}\left(X_k \le e^{ck}\right) + \mathbb{P}\left(\mathrm{Bin}(e^{ck}, \tfrac{1}{8}) \le \tfrac{e^{ck}}{9}\right).$$

The probabilities on the right hand side vanish when ℓ grows, and so does Π_ℓ. Owing to (4.13),

$$\mathbb{P}\left(h(\mathfrak{A}_\ell) \le \gamma_\ell\right) \le \Pi_\ell,$$

in which

$$\gamma_\ell = k\ \left(m+n+g\ +\tfrac{(1-\alpha)c}{\ln 2}\right).$$

But due to $k = \lfloor(\ell - a_\ell)/T\rfloor$, and owing to our choice of (m,n,g), cf. (4.12),

$$\begin{aligned} \gamma_\ell &= \left(\rho - \tfrac{(1-\alpha)\eta + \alpha\ln\pi_T}{\ln 2}\right)k \\ &\ge (\Delta^\bullet - \varepsilon)(\ell - a_\ell) - \tfrac{(1-\alpha)\eta+\alpha\ln\pi_T}{\ln 2}\lfloor(\ell-a_\ell)/T\rfloor - (\Delta^\bullet-\varepsilon)T. \end{aligned}$$

Since α and η are arbitrary positive numbers, for a suitable choice,

$$\gamma_\ell \geq (\Delta^\bullet - 2\varepsilon)\ell,$$

for ℓ large enough, which concludes the proof. □

4.3.4.b Upper bound for $h(\mathfrak{L}(L^\ell))$

Proposition 4.21. *For any $\varepsilon > 0$,*

$$\lim_\ell \mathbb{P}\left(h(\mathfrak{L}(L^\ell)) \geq (\Delta^\bullet + \varepsilon)\ell\right) = 0.$$

Proof. Again, according to Proposition 4.7, we only need to prove that

$$\lim_\ell \mathbb{P}\left(h(\mathfrak{A}_\ell) \geq (\Delta^\bullet + \varepsilon)\ell\right) = 0.$$

A shrub $\mathfrak{t}(v)$ is the Lyndon tree of a Bernoulli word observed until the first occurence of $\mathtt{0}^{a_\ell}$, and as such it comes with a sequence of i.i.d. geometric random variables (the lengths of the runs of $\mathtt{0}$'s) observed until the hitting time of $[a_\ell, +\infty)$: this sequence yields a Galton-Watson tree $\mathfrak{GW}_{a_\ell}(v)$ with geometric offspring distribution, that has a_ℓ generations, as described by Figure 4.3. Based on $\mathfrak{GW}_{a_\ell}(v)$, one can define a branching random walk $\mathfrak{BGW}_{a_\ell}(v)$ as follows: when the offspring of an individual has size N, each of the N scions jumps N steps in the same direction, say to the left. Then the height H_v of $\mathfrak{t}(v)$ is stochastically smaller than the leftmost position M_v of $\mathfrak{BGW}_{a_\ell}(v)$ at generation a_ℓ: the height of a binary tree with n leaves is smaller than $n-1$ and then one has to account for at most one needle in each family of $\mathfrak{GW}_{a_\ell}(v)$, so N leaves plus one needle entails that the depth of each of the N scions is at most the depth of its father plus N. Assume that the positions of the members of generation a_ℓ for $\mathfrak{BGW}_{a_\ell}(v)$ form a point process $(X_i)_{i\in I}$. Then, as usual, for $0 \leq \theta < \ln 2$, and $x \geq 3$,

$$\mathbb{P}\left(M_v \geq x\right) \leq\ e^{-\theta x}\mathbb{E}\left[e^{\theta M_v}\right],$$

and

$$\begin{aligned}\mathbb{E}\left[e^{\theta M_v}\right] &\leq \mathbb{E}\left[\sum_{i\in I} e^{\theta X_i}\right] \\ &= \mathbb{E}\left[Ne^{\theta N}\right]^{a_\ell} \ =\ \left(\tfrac{2e^\theta}{(2-e^\theta)^2}\right)^{a_\ell}.\end{aligned}$$

It entails that

$$\begin{aligned}\mathbb{P}\left(M_v \geq xa_\ell\right) &\leq\ \left(\inf_{0\leq\theta<\ln 2} \tfrac{2e^{\theta(1-x)}}{(2-e^\theta)^2}\right)^{a_\ell} \\ &= \left(2^{-x-2}(x-1)^{1-x}(x+1)^{1+x}\right)^{a_\ell} \\ &\leq \left(2^{-x}(ex/2)^2\right)^{a_\ell},\end{aligned}$$

in which the equality of the second line holds for $x \geq 3$, while the last inequality holds for $x \geq 1$. Thus both the conditional probabilities given the first a_ℓ characters of the word, and, as a consequence, the conditional probability given the height of the root, satisfy:

$$\begin{aligned}\mathbb{P}\left(H_v \geq x a_\ell \mid .\ \right) &\leq 2^{a_\ell}\left(2^{-x}(ex/2)^2\right)^{a_\ell},\\ &\leq \left(2^{-x}e^2x^2\right)^{a_\ell}.\end{aligned}$$

As a consequence, the maximal height $\mathfrak{H}_S$ of any set S of shrubs selected according to the position of their roots in $\mathfrak{S}_\ell$ satisfies

$$\begin{aligned}\mathbb{P}\left(\tfrac{\mathfrak{H}_S}{\ell} \geq \xi/\ln 2\right) &\leq |S|\left(e^{-\xi\ell/a_\ell}(e\xi\ell/a_\ell \ln 2)^2\right)^{a_\ell}\\ &\leq e^{\ln|S|-\xi\ell+2a_\ell\ln(\xi\ell)},\end{aligned} \tag{4.15}$$

as long as $\xi\ell \geq 3a_\ell \ln 2$.

Let $Z_{\ell,m,n,g}$ denote the number of blades of type (ℓ,m,n,g) in $\mathfrak{S}_\ell$, and let $\mathfrak{H}_{\ell,m,n,g}$ (resp. $H_{\ell,m,n,g}$) denote the maximal height among the corresponding set of shrubs (resp. the height of the highest leaf of $\mathfrak{A}_\ell$ that has an ancestor among the $Z_{\ell,m,n,g}$ leaves of type (ℓ,m,n,g)). According to Propositions 4.7 and 4.11,

$$\begin{aligned}h(\mathfrak{L}(L^\ell)) \simeq h(\mathfrak{A}_\ell) &= \max\left\{H_{\ell,m,n,g} \,|\, m,n,g \in \mathbb{N}\right\}\\ &= \max\left\{m+n+g+\mathfrak{H}_{\ell,m,n,g} \,|\, m,n,g \in \mathbb{N}\right\}.\end{aligned}$$

As a consequence,

$$\mathbb{P}\left(h(\mathfrak{A}_\ell) \geq (\Delta^\bullet+2\varepsilon)\ell\right) \leq \sum_{m,n,g\in\mathbb{N}} \mathbb{P}\left(H_{\ell,m,n,g} \geq (\Delta^\bullet+2\varepsilon)\ell\right) \tag{4.16}$$

Set

$$\delta(\lambda,\mu,\nu) = \Delta^\bullet - \Delta(\lambda,\mu,\nu) \geq 0.$$

For an arbitrary choice of α, the Markov inequality yields that

$$\mathbb{P}\left(Z_{\ell,m,n,g} \geq\ 2^{\alpha\ell}\right) \leq e^{-\alpha\ell\ln 2+\ln\pi_{\ell,m,n,g}},$$

Thus, if S stands for the set of blades of type (ℓ,m,n,g) in $\mathfrak{S}_\ell$, (4.15) entails that, for $\varepsilon > 0$,

$$\mathbb{P}\left(H_{\ell,m,n,g} \geq (\Delta^\bullet+2\varepsilon)\ell\right) \leq e^{\alpha\ell\ln 2-\xi\ell+2a_\ell\ln(\xi\ell)} + \mathbb{P}\left(Z_{\ell,m,n,g} \geq\ 2^{\alpha\ell}\right)$$

in which

$$\xi = (\Delta^\bullet + 2\varepsilon - \tfrac{1+\mu+\nu}{\lambda})\ln 2\ = (\delta(.) + 2\varepsilon + \tfrac{\Psi(.)}{\lambda\ln 2})\ln 2.$$

The choice

$$\alpha = \Delta^\bullet + \varepsilon - \tfrac{1+\mu+\nu}{\lambda}$$

leads to

$$\mathbb{P}\left(H_{\ell,m,n,g} \geq (\Delta^\bullet + 2\varepsilon)\ell\right) \leq 2^{-\varepsilon\ell + 2a_\ell \log_2(\xi\ell)} + 2^{-(\varepsilon+\delta(.))\ell + r_{\ell,m,n,g}}, \quad (4.17)$$

provided that $\xi\ell \geq 3a_\ell \ln 2$, since we need (4.15) to hold. In view of (4.16) and (4.17), we need

- to reduce the domain of summation in (4.16),
- and to check that $r_{\ell,m,n,g}$ is uniformly $o(\ell)$ on the reduced domain.

First point: for $\eta > 0$, set

$$T_{\ell,\eta} = \left\{(m,n,g) \,|\, n \geq \eta\ell, \tfrac{m+n}{\ell \ln 2} \in \left[\tfrac{1}{3}, 5\right], \text{ and } \eta\ell + \ln \pi_{\ell,m,n,g} \geq 0\right\},$$

and note that $\#T_{\ell,\eta} \leq 25\ell^3$.

Lemma 4.22. *For η small enough, the probability that a leaf of $\mathfrak{A}_\ell$ higher than $(\Delta^\bullet + 2\varepsilon)\ell$ grows from a blade outside $T_{\ell,\eta}$ vanishes when ℓ grows, and*

$$\mathbb{P}\left(h(\mathfrak{A}_\ell) \geq (\Delta^\bullet + 2\varepsilon)\ell\right) \leq o(1) + \sum_{m,n,g \in T_{\ell,\eta}} \mathbb{P}\left(H_{\ell,m,n,g} \geq (\Delta^\bullet + 2\varepsilon)\ell\right).$$

Proof. First we prove that the average number $\pi_{\ell+}$ (resp. $\pi_{\ell-}$) of blades of type (m,n,g) with $m+n \geq 5\ell\ln 2$ (resp. with $m+n \leq (\ell \ln 2)/3$) vanishes when ℓ grows, and so does the probability that there exist such blades: due to (4.4),

$$\pi_{\ell+} = \sum_{k \geq 5\ell\ln 2} \frac{(\ell\ln 2)^k 2^{k-\ell}}{k!}, \qquad \pi_{\ell-} = \sum_{0 \leq k \leq (\ell\ln 2)/3} \frac{(\ell\ln 2)^k 2^{k-\ell}}{k!}.$$

If X follows the Poisson distribution with parameter $\ell \ln 2$, we have:

$$\pi_{\ell+} = \mathbb{E}\left[2^X 1_{X \geq 5\ell\ln 2}\right], \qquad \pi_{\ell-} = \mathbb{E}\left[2^X 1_{X \leq (\ell\ln 2)/3}\right],$$

and, for any $t > 0$,

$$\pi_{\ell+} \leq \mathbb{E}\left[2^X e^{t(X - 5\ell\ln 2)}\right] = \exp[(e^{t+\ln 2} - 1 - 5t)\ell \ln 2].$$

A suitable choice of t leads to $\lim_\ell \pi_{\ell+} = 0$, and similar inequalities yields that $\lim_\ell \pi_{\ell-} = 0$. In what follows, we consider only blades of type (m,n,g) with $a\ell \leq m+n \leq b\ell$, in which $a = (\ln 2)/3$ and $b = 5\ln 2$.

Set

$$\pi_{\ell,m,n} = \sum_g \pi_{\ell,m,n,g} = \frac{(\ell \ln 2)^{m+n} 2^{-\ell}}{m!n!}, \tag{4.18}$$

and let π_η be the expectation of the number of blades such that $\eta\ell + \ln \pi_{\ell,m,n,g} < 0$: it satisfies, for any $t \in (0,1)$,

$$\begin{aligned}
\pi_\eta \leq & \sum_{\substack{a\ell \leq m+n \leq b\ell, \\ (\ell-n)_+ \leq g \leq \ell}} \pi_{\ell,m,n,g}\, e^{-t(\eta\ell + \ln \pi_{\ell,m,n,g})} \\
= & \; e^{-t\eta\ell} \sum_{\substack{a\ell \leq m+n \leq b\ell, \\ (\ell-n)_+ \leq g \leq \ell}} \pi_{\ell,m,n,g}^{1-t} \\
\leq & \; \ell e^{-t\eta\ell} \sum_{a\ell \leq m+n \leq b\ell} \pi_{\ell,m,n}^{1-t} \\
\leq & \; b^{2t} \ell^{1+2t} e^{-t\eta\ell} \left(\sum_{a\ell \leq m+n \leq b\ell} \pi_{\ell,m,n} \right)^{1-t} \\
= & \; b^{2t} \ell^{1+2t} e^{-t\eta\ell} \mathbb{E}\left[2^X 1_{a\ell \leq X \leq b\ell} \right]^{1-t} \\
= & \; b^{2t} \ell^{1+2t} \exp(-\ell(t(\eta + b\ln 2) - b\ln 2)).
\end{aligned}$$

Choosing $1 > t > (1 + \frac{\eta}{b \ln 2})^{-1}$, we see that π_η vanishes as ℓ grows, and so does the probability of existence of blades such that $\eta\ell + \ln \pi_{\ell,m,n} < 0$ and $a\ell \leq m + n \leq b\ell$.

For the bounds we have in mind, Ξ has to be Lipshitz continuous around $\arg\max \Delta$, i.e. its argument $\lambda - \mu$ has to be bounded away from 0 and 1. This should be easy because Ξ goes to $-\infty$ when its argument is close to 0 or 1, leaving then little chance to Δ to be at its maximum. There is chance, however, for Δ to be large even if Ξ is close to $-\infty$: if n is very small, and thus λ very large, Ξ/λ could be small, and Δ could eventually be close to its maximum even though Ξ is close to $-\infty$. The next Lemma fixes this problem: consider the set $\tilde{S}_\eta$ of blades such that $\eta\ell + \ln \pi_{\ell,m,n} \geq 0$ and $n \leq \eta\ell$, for $0 < \eta < \ln 2$, and let $H_{\eta,\ell}$ denote the maximal height among the leaves of $\mathfrak{A}_\ell$ that have an ancestor in $\tilde{S}_\eta$. Then we have:

Lemma 4.23. *For any $\varepsilon > 0$, there exist $\eta > 0$ such that*

$$\lim_\ell \mathbb{P}\left(H_{\eta,\ell} \geq (1 + \ln 2 + \varepsilon)\ell \right) = 0.$$

Proof. Owing to (4.8) and (4.18),

$$0 \leq \; (\eta - \ln 2)\ell - \varphi(n) - \varphi(m) + (m+n)\ln(\ell \ln 2) + \ln 16mn.$$

Using that $c\varphi(\frac{x}{c}) = \varphi(x) - x\ln c$, for $c = \ell\ln 2$ we obtain:

$$\begin{aligned} 0 \leq\ & \eta\ell - c - c\varphi(\tfrac{m}{c}) - c\varphi(\tfrac{n}{c}) + 2\ln(2b\ell) \\ \leq\ & c\left(-1 - \varphi(\tfrac{m}{\ell\ln 2}) - \varphi(\tfrac{\eta}{\ln 2}) + \tfrac{\eta}{\ln 2} + \tfrac{2\ln(2b\ell)}{\ell\ln 2}\right) \end{aligned} \tag{4.19}$$

Since $x \longrightarrow -1 - \varphi(x)$ reaches its maximum, 0, when $x = 1$, then η small and ℓ large entails that $|m - \ell\ln 2|$ has to be small, for the left hand size to be nonnegative.

More precisely, for any $\varepsilon > 0$, one can choose $\eta > 0$, independent of m, n, ℓ, in such a way that, for ℓ large enough, (4.19) holds only if $m \in \ell(\ln 2 \pm \varepsilon)$, by strict convexity and smoothness of φ: solve $1 + \varphi(u) = 2(\frac{\eta}{\ln 2} - \varphi(\frac{\eta}{\ln 2}))$. For η small enough, the 2 solutions r_1 and r_2 are close enough to 1, so that $r_1 \leq \frac{m}{\ell\ln 2} \leq r_2$ entails $m \in \ell(\ln 2 \pm 2\varepsilon)$. Then one can choose ℓ large enough, so that $\frac{2\ln(2b\ell)}{\ell\ln 2} \leq \frac{\eta}{\ln 2} - \varphi(\frac{\eta}{\ln 2})$. Then the depth $m + n + g$ of a blade $v \in \tilde{S}$ is at most $\ell(\ln 2 + \varepsilon + \eta + 1)$, since $g \leq \ell$.

On the other hand, the expected number $\tilde{\pi}_\eta$ of blades in $\tilde{S}$ is smaller than $\exp(\ell(\eta(\ln\ln 2) - \varphi(\eta)))$: again let X, Y be i.i.d. and Poisson distributed with parameter $\ell\ln 2/2$, so that we have:

$$\tilde{\pi}_\eta \ =\ \mathbb{E}\left[2^{X+Y}1_{X\leq\eta\ell}\right] \ =\ 2^{\ell/2}\mathbb{E}\left[2^X 1_{X\leq\eta\ell}\right].$$

Again, for any $t > 0$, for instance for $t = \frac{\eta}{\ln 2}$,

$$\begin{aligned} \tilde{\pi}_\eta \leq\ & 2^{\ell/2}\mathbb{E}\left[2^X e^{t(\eta\ell - X)}\right] \ =\ \exp[\ell(t\eta + e^{-t}\ln 2)] \\ =\ & \exp[\ell(\eta(\ln\ln 2) - \varphi(\eta))]. \end{aligned} \tag{4.20}$$

Set

$$\tilde{\xi} = \eta(\ln\ln 2) - \varphi(\eta).$$

Then, due to (4.15), the probability that the taller shrub planted on a blade of $\tilde{S}$ is at least $\frac{3\tilde{\xi}\ell}{\ln 2}$ tall vanishes with ℓ: we have

$$\begin{aligned} \mathbb{P}\left(\mathfrak{H}_{\tilde{S}} \geq \tfrac{3\tilde{\xi}\ell}{\ln 2}\right) &\leq \mathbb{P}\left(|\tilde{S}| \geq e^{2\tilde{\xi}\ell}\right) + \mathbb{P}\left(\mathfrak{H}_{\tilde{S}} \geq \tfrac{3\tilde{\xi}\ell}{\ln 2} \text{ and } |\tilde{S}| \leq e^{2\tilde{\xi}\ell}\right) \\ &\leq e^{-\ell\tilde{\xi}} + e^{-\tilde{\xi}\ell + 2a_\ell\ln(3\tilde{\xi}\ell)}, \end{aligned}$$

as long as $\tilde{\xi}\ell \geq a_\ell\ln 2$. This leads to

$$\lim_\ell \mathbb{P}\left(H_{\eta,\ell} \geq (1 + \ln 2 + \varepsilon + \eta + \tfrac{3\tilde{\xi}}{\ln 2})\ell\right) = 0.$$

in which $\varepsilon + \eta + \frac{3\tilde{\xi}}{\ln 2}$ is arbitrarily close to 0 for η small. □

We proved that, with a large probability, there exists no blade with $\frac{m+n}{\ell\ln 2}$ outside $[a, b]$, and no blade with $\frac{m+n}{\ell\ln 2}$ inside $[a, b]$ but $\eta + \ln\pi_{\ell,m,n,g} < 0$. Then we proved that, with a large probability, a leaf of $\mathfrak{A}_\ell$ with depth larger than $(1 + \ln 2)\ell$ does not grow from a blade such that $\eta + \ln\pi_{\ell,m,n,g} \geq 0$ but $n \leq \eta\ell$. This concludes the proof since $\Delta^\bullet = 3.5... > 1 + \ln 2$. □

Inside $T_{\ell,\eta}$, we have $(5\ln 2)^{-1} \le \lambda \le \eta^{-1}$ and $0 \le \nu \le 5\lambda \ln 2$. We always have $0 \le \lambda - \mu \le 1$, but $\lambda - \mu$ has to be bounded away from 0 and 1, in order for $n\rho_n$ and $r_{\ell,m,n,g}$ to be small. This holds true due to

$$\begin{aligned} 0 &\le \eta\ell + \ln \pi_{\ell,m,n,g} \\ &= \eta\ell + n\Psi(\lambda,\mu,\nu) + r_{\ell,m,n,g}\ln 2 \\ &\le \eta\ell + n(\Psi(\lambda,\mu,\nu) + \rho_n)) + (3 + 2\sqrt{g})\ln(24ng) + \ln(4m). \end{aligned}$$

Dividing by ℓ, for ℓ large enough, this insures that on $T_{\ell,\eta}$,

$$\mathcal{H}\left(d^{(\mu)}\right) + \Phi(\lambda,\mu,\nu) + \Xi(\lambda - \mu - \delta_n)) \ge \tfrac{-2\eta}{5\ln 2},$$

in which all the terms on the left hand side are bounded on $T_{\ell,\eta}$, but eventually Ξ. Thus, for ℓ large enough, on $T_{\ell,\eta}$, Ξ is bounded away from $-\infty$, and $\lambda - \mu - \frac{1}{n}$ is bounded away from 0 and 1, which entails a Lipshitz condition on Ξ on the domain $T_{\ell,\eta}$. Thus, according to Proposition 4.19, on $T_{\ell,\eta}$, since $n \ge \eta\ell$, and $\lambda \le \eta^{-1}$,

$$\begin{aligned} \tfrac{1}{\ell} r_{\ell,m,n,g} &\le \tfrac{1}{\ell}(2n\rho_n + (6 + 4\sqrt{g})\ln(24ng) + 2\ln(4m)) \\ &\le \tfrac{10c\ln 2}{\eta\ell} + \tfrac{10(\ln(24)+2\ln\ell)}{\sqrt{\ell}} + \tfrac{2\ln(20\ell\ln 2)}{\ell} \end{aligned}$$

in which c is the Lipshitz coefficient for Ξ. Also

$$\begin{aligned} \xi &= (\Delta^\bullet + 2\varepsilon - \tfrac{1+\mu+\nu}{\lambda})\ln 2 \\ &\le (\Delta^\bullet + 2\varepsilon - \tfrac{\ln 2}{3})\ln 2, \end{aligned}$$

so that $2a_\ell \log_2(\xi\ell)$ and $r_{\ell,m,n,g}$ are uniformly $o(\ell)$ on the domain $T_{\ell,\eta}$. Finally

$$\begin{aligned} \xi &\ge -\eta - \tfrac{1}{\ell}\ln \pi_{\ell,m,n,g} + (\delta(.) + 2\varepsilon)\ln 2 + \tfrac{\Psi(\lambda,\mu,\nu)}{\lambda} \\ &= (\delta(.) + 2\varepsilon)\ln 2 - \eta - \tfrac{\ln 2}{\ell} r_{\ell,m,n,g} \end{aligned}$$

which ensures $\xi\ell \ge 3a_\ell \ln 2$, and as a consequence (4.15), on the domain $T_{\ell,\eta}$, for ℓ large enough. These inequalities yield that

$$\lim_\ell \sum_{(m,n,g)\in T_\ell} \mathbb{P}\left(H_{\ell,m,n,g} \ge (\Delta^\bullet + 2\varepsilon)\ell\right) = 0.$$

□

4.4 Depoissonization of the length

In this Section, we derive Theorem 4.2 from Theorem 4.3, i.e. we replace the random length $\tau_\ell + 1$ of W^ℓ by the deterministic length of W_n. A simple argument comes to mind: w.h.p. $\tau_{\log_2 n - \varepsilon_n} \le n - 1 \le \tau_{\log_2 n + \varepsilon_n}$ as long as

$$\varepsilon_n \to +\infty \text{ and } \varepsilon_n = o(\ln n)\,; \tag{4.21}$$

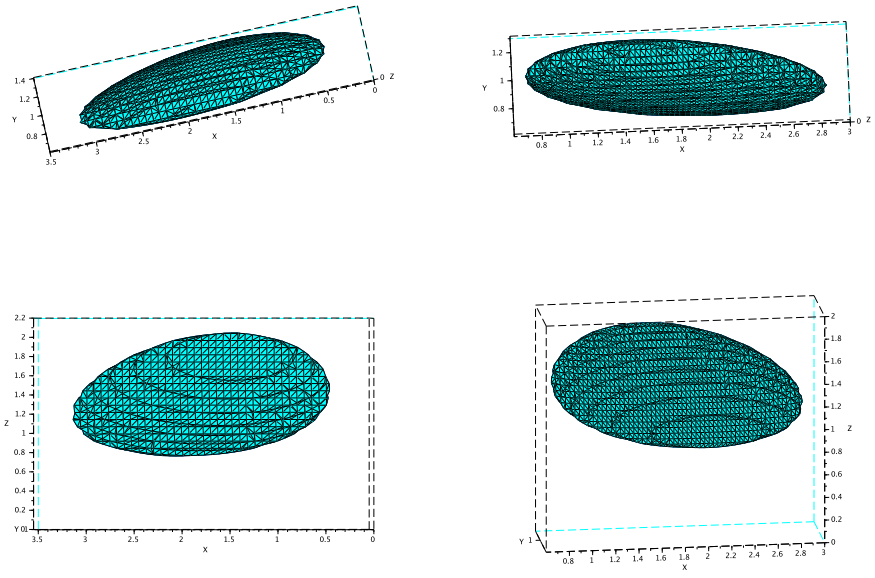

Figure 4.9: Domains $\Delta(\lambda, \mu, \nu) > 4.95$ and $\Delta(\lambda, \mu, \nu) > 5$, with $x = \coth(\lambda - \mu) - \frac{1}{\lambda - \mu}$, $y = \lambda$, and $z = \nu$.

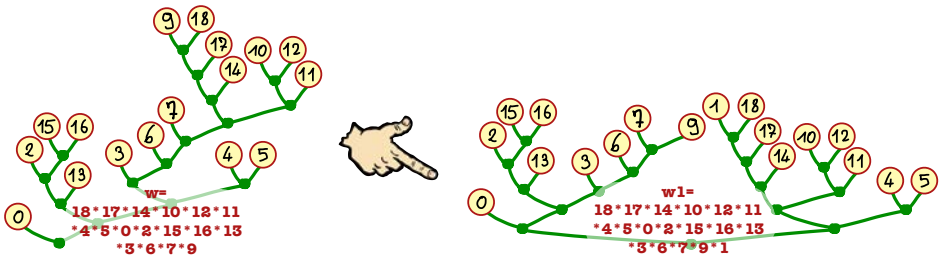

Figure 4.10: Here the alphabet is $\mathbb{N}$ and $6 = h(\mathfrak{L}(w1)) < h(\mathfrak{L}(w)) = 10$.

thus w.h.p. $h(\mathfrak{L}(W^{\log_2 n - \varepsilon_n})) \leq h(\mathfrak{L}(W_n)) \leq h(\mathfrak{L}(W^{\log_2 n + \varepsilon_n}))$. However, while $\ell \to h(\mathfrak{L}(W^\ell))$ is increasing, $n \to h(\mathfrak{L}(W_n))$ is not, see figure 4.10, and this line of proof fails. We do not even know whether $n \to h(\mathfrak{L}(W_n))$ is stochastically increasing.

In order to address these problems, let the integers α_n, β_n and γ_n be defined by

$$\alpha_n = \log_2 n - 3\varepsilon_n, \quad \beta_n = \lceil \log_2 n - 2\varepsilon_n \rceil, \quad \gamma_n = \lfloor \log_2 n \rfloor$$

in which ε_n meets conditions (4.21). Let us call *long run* of a word or of a necklace any maximal run that is at least α_n-letters long. Consider the infinite random word W_∞ and its n-letters long prefix W_n. A sequence $Z = (Z_i)_{i \geq 1}$ of i.i.d. factors of W_∞ is defined as follows: for $i \geq 2$, Z_i begins just after the end of Z_{i-1} and, counting from the position of the first `1` of Z_i, Z_i stops with the α_n-th `0` of the first long run after that first `1`; Z_1 stops according to the same rule, but starts with the first letter of W_∞. Each factor Z_i can be written $0^{X_i} Y_i$, in which

- Y_i has the same distribution as W^{α_n}, cf. section 4.2.2,
- $\mathbb{P}(X_i = k) = 2^{-k-1} \mathbb{1}_{k \geq 0}$,
- Y_i and X_i are independent.

The Lyndon word of Y_i, denoted L_i in what follows, is obtained by a rotation in which the right factor is 0^{α_n}; L_i has the same distribution as L^{α_n}, cf. section 4.2.2. The X_i's account for the overshoots of each run not shorter than α_n.

Definition 4.5. Let $(T_i)_{1 \leq i \leq k}$ be a finite sequence of planar trees, $\mathfrak{L}^+$ a planar tree with k marked leaves, naturally labeled by their rank (rank going clockwise, starting after the root), and σ a permutation of $\{1, 2 \dots k\}$. Then the tree obtained when one grafts to $\mathfrak{L}^+$ the tree T_i, its root being substituted to the leaf of $\mathfrak{L}^+$ with label $\sigma(i)$, for all i, is called the *permuted concatenation* of $(T_1, \dots, T_k)$ with underlying tree $\mathfrak{L}^+$.

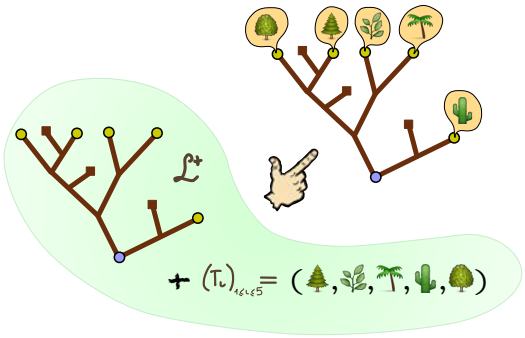

Figure 4.11: Permuted concatenation of $(T_i)_{1\le i\le 5}$. The underlying tree $\mathfrak{L}^+$ has 3 unmarked leaves and the permutation is circular.

For instance, for any integer-valued random variable N, $\mathfrak{L}(Z_1Z_2\dots Z_N)$ is a permuted concatenation of the trees $\mathfrak{L}(L_i)$. In this case the permutation is a circular one, and the underlying tree is a binary tree with N marked leaves, and $\sum_{i=1}^N X_i$ unmarked leaves (here these unmarked leaves are needles); thus the height of the underlying tree is at most $N-1+\sum_{i=1}^N X_i$.

Consider the prefix $W_{[n]}$ of W_∞ obtained by concatenation of the factors Z_i that are contained or overlap W_n, say

$$W_{[n]} = Z_1Z_2\dots Z_N.$$

4.4.1 Lower bound

According to [BK02, p. 111, (4.19)], the position of the α_n-th 0 of the rth long run has expectation $r2^{1+\alpha_n}-2$ and a variance smaller than $r(2+2^{2+2\alpha_n})$. So by Chebyshev's inequality, w.h.p. there exist at least $2^{2\varepsilon_n}$ long runs in W_n, which entails successively that $N\ge 3$, that $\mathfrak{L}(L_2)$ is a subtree of $\mathfrak{L}(W_n)$ and that $h(\mathfrak{L}(W_n))\ge h(\mathfrak{L}(L_2))$. For $\eta>0$, set

$$p_{n-} = \mathbb{P}\left(\tfrac{h(\mathfrak{L}(W_n))}{\ln n}\le \Delta^\star-\eta\right).$$

Thus

$$p_{n-}\le \mathbb{P}\left(N\le 2\right)+\mathbb{P}\left(\tfrac{h(\mathfrak{L}(L_2))}{\ln n}\le \Delta^\star-\eta\right).$$

On the other hand, the height H_i of $\mathfrak{L}(L_i)$ has the same distribution as $h(\mathfrak{L}(L^{\alpha_n}))$, thus, according to Theorem 4.3, for any $\varepsilon>0$, for instance if $0<\varepsilon<\eta$,

$$\lim_n \mathbb{P}\left(\tfrac{h(\mathfrak{L}(L_2))}{\alpha_n}\le (\Delta^\star-\varepsilon)\ln 2\right)=0$$

which, together with the previous inequality, gives $\lim_n p_{n-}=0$.

4.4.2 Upper bound

The upper bound for Theorem 4.2 follows at once from the next proposition:

Proposition 4.24. *For any $\eta > 0$, there exists a sequence (ε_n) that meets conditions* (4.21)*, and such that*

$$\lim_n \mathbb{P}\left(\frac{h(\mathfrak{L}(W_{[n]}))}{\ln n} \geq \Delta^\star + \eta\right) = 0, \tag{4.22}$$

and

$$\lim_n \mathbb{P}\left(h(\mathfrak{L}(W_n)) \geq h(\mathfrak{L}(W_{[n]}))\right) = 0. \tag{4.23}$$

Proof of (4.22)*.* Let us prove that the sequence

$$p_{n+} = \mathbb{P}\left(\frac{h(\mathfrak{L}(W_{[n]}))}{\ln n} \geq \Delta^\star + 2\eta\right)$$

vanishes. Since $\mathfrak{L}(W_{[n]})$ is a permuted concatenation of $(\mathfrak{L}(L_i))_{1\leq i\leq N}$, in which the underlying tree, $\mathfrak{L}^{\alpha_n}(W_{[n]})$, has $\sum_{1\leq i\leq N} X_i$ needles unmarked, and N marked leaves (or blades), Lemma 4.4 entails that:

$$h(\mathfrak{L}_{\alpha_n}(W_{[n]})) \leq N + \max_{1\leq i\leq N} X_i,$$

which yields the following bound:

$$h(\mathfrak{L}(W_{[n]})) \leq N + \max_{1\leq i\leq N} X_i + \max_{1\leq i\leq N} H_i.$$

Thus, if $t_n = o(\ln n)$ and if n is large enough,

$$p_{n+} \leq \mathbb{P}\left(N \geq t_n\right) + t_n \mathbb{P}\left(\tfrac{H_1}{\ln n} \geq \Delta^\star + \eta\right) + \mathbb{P}\left(\max_{1\leq i\leq t_n} X_i > 2\log_2 t_n\right).$$

Since $N-1$ is smaller than the number X of occurrences of the pattern $\mathtt{0}^{\alpha_n}$ in W_n, that satisfies

$$\mathbb{E}\left[X+1\right] = (n-\alpha_n+1)2^{-\alpha_n} + 1 \leq 2^{3\varepsilon_n},$$

we have

$$\mathbb{P}\left(N \geq 3^{3\varepsilon_n}\right) \leq (\tfrac{2}{3})^{3\varepsilon_n}.$$

According to Theorem 4.3, for $\eta > 0$,

$$\lim_n \mathbb{P}\left(\frac{h(\mathfrak{L}(L^{\gamma_n}))}{\gamma_n} \geq (\Delta^\star + \eta)\ln 2\right) = 0.$$

We can thus choose a sequence ε_n that fulfills conditions (4.21) in such a way that $3^{3\varepsilon_n} = o(\ln n)$ and

$$\lim_n 3^{3\varepsilon_n} \mathbb{P}\left(\frac{h(\mathfrak{L}(L^{\gamma_n}))}{\gamma_n} \geq (\Delta^\star + \eta)\ln 2\right) = 0.$$

But since $h(\mathfrak{L}(L^k))$ is stochastically increasing in k, we have

$$\lim_n 3^{3\varepsilon_n}\mathbb{P}\left(\frac{h(\mathfrak{L}(L^{\alpha_n}))}{\log_2 n} \geq (\Delta^\star+\eta)\ln 2\right) = 0.$$

Finally

$$\mathbb{P}\left(\max_{1\leq i\leq t_n} X_i > 2\log_2 t_n\right) \leq 1-(1-\tfrac{2}{t_n^2})^{t_n} \leq -t_n\ln(1-\tfrac{2}{t_n^2}),$$

and the choice $t_n = 3^{3\varepsilon_n}$ gives the desired result. □

Proof of (4.23). Let us call *huge run* of a word any maximal run of 0's of its necklace that is at least β_n-letters long. Let now N (resp. $\tilde{N}$) stand for the number of huge runs contained in the necklace of W_n (resp. of $W_{[n]}$); as in Section 4.4.1, w.h.p. N and $\tilde{N}$ grow at least like 2^{ε_n}. We shall prove that, w.h.p., $N=\tilde{N}$ and that, w.h.p., the Lyndon factors $\hat{L}_1,\hat{L}_2,\dots\hat{L}_N$ (resp. $\tilde{L}_1,\tilde{L}_2,\dots\tilde{L}_N$) induced by these huge runs have the same lexicographic order. The convention here is that $\hat{L}_1$ (or $\tilde{L}_1$ for that matter) begins with the first letter of the first huge run in W_n, and ends with the last letter before the second huge run in W_n.

This way, typically, $\hat{L}_N$ (resp. $\tilde{L}_N$) straddles the end of W_n (resp. $W_{[n]}$), for, in $\langle W_n\rangle$, or in $\langle W_{[n]}\rangle$ as well, w.h.p., the run of 0's that straddles the end of the word is not huge:

- in W_n the length of this run is stochastically smaller than the sum of two X_i's, thus the probability that it is huge is smaller than $2^{-\beta_n/2}$;
- since $W_{[n]}$ ends with a run of exactly α_n 0's, a run straddling the end of $W_{[n]}$ is huge only if the run of 0's at the beginning of W_∞ is longer than ε_n, i.e. with probability $2^{-\varepsilon_n}$.

Thus $N=\tilde{N}$ w.h.p., and $\hat{L}_i=\tilde{L}_i$ for $1\leq i\leq N-1$.

Also, w.h.p., there exists no huge run in the last $\beta_n+\sqrt{n}$ characters of W_n, the probability being bounded by

$$(1+\sqrt{n})2^{-\beta_n} = o(1),$$

and a $\sqrt{n}$-letters long pattern does not appear twice in W_n, the probability being bounded by

$$n^2 2^{-\sqrt{n}} = o(1).$$

It follows that, though different, $\hat{L}_N$ and $\tilde{L}_N$ begin with the same huge run, and have the same first $\sqrt{n}$ characters after their initial huge run. As a consequence, the outcome of the lexicographic comparison between $\hat{L}_N$ (resp. $\tilde{L}_N$) and $\hat{L}_i$, being known before reading the $\sqrt{n}$-th character of $\hat{L}_N$ (resp. $\tilde{L}_N$), is the same in both cases. So, $\mathfrak{L}(W_n)$ and $\mathfrak{L}(W_{[n]})$ are

permuted concatenations of $(\mathfrak{L}(\hat{L}_i))_{1\le i\le N}$ (resp. $(\mathfrak{L}(\tilde{L}_i))_{1\le i\le N}$) with the same underlying tree and the same permutation, and the height of their leaves are the same, but perhaps if the label of the leaf is a letter from $\hat{L}_N$ or $\tilde{L}_N$.

Finally, we prove that w.h.p. at least one of the highest leaves, say $v^\star$, of $\mathfrak{L}(W_n)$ is not labeled with a letter from the factor $\hat{L}_N$. This is due to the following facts:

1. if $1 \le x_n < y_n \le n$, then with a probability at least $\frac{y_n - x_n}{n}$, at least one of the highest leaves of $\mathfrak{L}(W_n)$ is between x_n and y_n;

2. we can choose x_n, y_n so that, simultaneously, $1 - \frac{y_n - x_n}{n} = o(1)$, and, w.h.p., $\hat{L}_N$ does not overlap $[\![x_n, y_n]\!]$.

Since the previous highly likely events insure that all the leaves from $\hat{L}_i$, $i < N$ have the same height in $\mathfrak{L}(W_n)$ and in $\mathfrak{L}(W_{[n]})$, the height of $v^\star$ in $\mathfrak{L}(W_{[n]})$ is $h(\mathfrak{L}(W_n))$, which entails (4.23).

Let us prove points (1) and (2). Let w be a primitive word of $\{\mathtt{0},\mathtt{1}\}^n$ and let $\langle w\rangle$ denote its necklace. Conditionally, given that $W_n \in \langle w\rangle$, W_n is uniform on $\langle w\rangle$. Let $\Pi = \{\Pi_i, 1 \le i \le k\} \subset [\![1, n]\!]$ be the set of positions of the labels of the highest leaves of $\mathfrak{L}(w)$. The ℓ-th element of $\langle w\rangle$ (obtained after the permutation of a ℓ-letters long suffix of w with the corresponding $n-\ell$-letters long prefix) fills the condition of point (1) iff $\ell \in [\![x_n, y_n]\!] - \Pi$. Now $[\![x_n, y_n]\!] - \Pi$ is an union of intervals of $\mathbb{Z}_n$ with the same width $y_n - x_n$ thus this reunion has exactly $y_n - x_n + 1$ elements if there exists only one highest leaf (this never happens for a binary tree), but else it has more elements. It follows that for any primitive word w,

$$\mathbb{P}\left(W_n \text{ meets condition (1)} | W_n \in \langle w\rangle\right) \ge \tfrac{y_n - x_n + 2}{n},$$

which takes care of point (1) for n large enough (recall e.g. [CA10, Lemma 2.1]).

For point (2), consider the position L (resp. R) of the $\mathtt{0}$ with rank β_n in the first huge run of W_∞ (resp. the position of the $\mathtt{0}$ ranked β_n in the first huge run of W_∞, *going backward* from the last letter of W_n). For R to be fully defined you need to see W_∞ as a *doubly* infinite sequence of random characters. If $\hat{L}_N$ overlaps $[\![x_n, y_n]\!]$ then $L - \beta_n \ge x_n$ or $R \le y_n$. Also L and $n + 1 - R$ follow the geometric distribution of order β_n (cf. [BK02, p. 17, (2.18)]) whose expectation (resp. variance) are

$$2^{1+\beta_n} - 2, \quad 2^{3+2\beta_n}(1 - (2\beta_n + 5)2^{-2-\beta_n} + 2^{-2-2\beta_n}),$$

with the consequence that, by Chebyshev's inequality, for n large enough,

$$\mathbb{P}\left(L \ge (1 + \rho_n\sqrt{2})2^{1+\beta_n}\right) \le \rho_n^{-2}.$$

One can choose $\rho_n \simeq 2^{\varepsilon_n}$ so that $x_n = (1+\rho_n\sqrt{2})2^{1+\beta_n} = n2^{-\varepsilon_n}$, and one can also take $n+1-y_n = x_n$, so that

$$\mathbb{P}\left(L \geq x_n\right) + \mathbb{P}\left(R \leq y_n\right) \leq 2\rho_n^{-2},$$

and $1 - \frac{y_n - x_n}{n} = 2^{1-\varepsilon_n} - \frac{1}{n}$, which proves point (2). □

Acknowledgements

The second author thanks Julien Clément for introducing him to the probabilistic analysis of Lyndon words and trees, 10 years ago[3], and Anne Briquet, Nicolas Broutin, Luc Devroye, Cyril Nicaud for renewing recently his interest after some years of unsuccessful attempts. Some discussions with Brigitte Chauvin were specially helpful.

[3] According to Julien Clément, the question of the height of the Lyndon tree was raised by Jean Berstel.

Bibliography

[Ald97] David Aldous. Brownian excursions, critical random graphs and the multiplicative coalescent. *Ann. Probab.*, 25(2):812–854, 1997.

[Ald01] David J. Aldous. Stochastic Models and Descriptive Statistics for Phylogenetic Trees, from Yule to Today. *Statistical Science*, 16(1):23–34, 02 2001.

[AN] Krishna B. Athreya and Peter E. Ney. *Branching processes*. Springer-Verlag, New York-Heidelberg. Die Grundlehren der mathematischen Wissenschaften, Band 196.

[AS04] David Aldous and J. Michael Steele. The Objective Method: Probabilistic Combinatorial Optimization and Local Weak Convergence. In *Probability on discrete structures*, volume 110 of *Encyclopaedia Math. Sci.*, pages 1–72. Springer, Berlin, 2004.

[Ash65] Robert Ash. *Information theory*. Interscience Tracts in Pure and Applied Mathematics, No. 19. Interscience Publishers John Wiley & Sons, New York-London-Sydney, 1965.

[BA99] Albert-László Barabási and Réka Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512, 1999.

[Bar90] Hélène Barcelo. On the action of the symmetric group on the Free Lie algebra and the partition lattice. *J. Combin. Theory Ser. A*, 55(1):93–129, 1990.

[BB09] François Baccelli and Bartlomiej Blaszczyszyn. *Stochastic Geometry and Wireless Networks, Volume I - Theory*, volume 1 of *Foundations and Trends in Networking Vol. 3: No 3-4, pp 249-449*. NoW Publishers, 2009. Stochastic Geometry and Wireless Networks, Volume II - Applications; see http://hal.inria.fr/inria-00403040.

[BC78] Edward A. Bender and E. Rodney Canfield. The asymptotic number of labeled graphs with given degree sequences. *J. Combinatorial Theory Ser. A*, 24(3):296–307, 1978.

[BCN05] Frédérique Bassino, Julien Clément, and Cyril Nicaud. The Standard Factorization of Lyndon Words: an Average Point of View. *Discrete Math.*, 290(1):1–25, 2005.

[BD07] Nicolas Broutin and Luc Devroye. An Analysis of the Height of Tries with Random Weights on the Edges. *Combinatorics, Probability and Computing*, 17(02):161–202, 2007.

[Big76] J. D. Biggins. The first- and last-birth problems for a multitype age-dependent branching process. *Advances in Appl. Probability*, 8(3):446–459, 1976.

[Big77] J. D. Biggins. Chernoff's Theorem in the branching random walk. *J. Appl. Probability*, 14(3):630–636, 1977.

[BJR07] Béla Bollobás, Svante Janson, and Oliver Riordan. The phase transition in inhomogeneous random graphs. *Random Structures Algorithms*, 31(1):3–122, 2007.

[BK02] N. Balakrishnan and Markos V. Koutras. *Runs and scans with applications*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], New York, 2002.

[Boh09] Tom Bohman. The Triangle-Free Process. *Adv. Math.*, 221(5):1653–1677, 2009.

[BR00] Béla Bollobás and Oliver Riordan. Constrained graph processes. *Electron. J. Combin.*, 7:Research Paper 18, 20, 2000.

[BS01] Itai Benjamini and Oded Schramm. Recurrence of distributional limits of finite planar graphs. *Electron. J. Probab.*, 6:no. 23, 13 pp. (electronic), 2001.

[BSW94] J. J. A. M. Brands, F. W. Steutel, and R. J. G. Wilms. On the number of maxima in a discrete sample. *Statist. Probab. Lett.*, 20(3):209–217, 1994.

[BvdHvL10] Shankar Bhamidi, Remco van der Hofstad, and Johan S. H. van Leeuwaarden. Scaling limits for critical inhomogeneous random graphs with finite third moments. *Electron. J. Probab.*, 15:no. 54, 1682–1703, 2010.

[BZ79] R. R. Bahadur and S. L. Zabell. Large deviations of the sample mean in general vector spaces. *Ann. Probab.*, 7(4):587–621, 1979.

[CA10] Philippe Chassaing and Elahe Zohoorian Azad. Asymptotic behavior of some factorizations of random words. *arXiv preprint arXiv:1004.4062*, 2010.

[CF03] Philippe Chassaing and Philippe Flajolet. Hachage, arbres, chemins & graphes. *Gaz. Math.*, (95):29–49, 2003.

[CKMR05] Brigitte Chauvin, Thierry Klein, Jean-Francois Marckert, and Alain Rouault. Martingales and Profile of Binary Search Trees. *Electron. J. probab*, 10(12):420–435, 2005.

[CM01] Philippe Chassaing and Jean-François Marckert. Parking functions, empirical processes, and the width of rooted labeled trees. *Electron. J. Combin.*, 8(1):Research Paper 14, 19 pp. (electronic), 2001.

[CR04] B. Chauvin and A. Rouault. Connecting Yule process, Bisection and Binary Search Tree via martingales. *J. Iranian Statistical Society*, 3:88–116, 2004.

[Dev86] L. Devroye. A Note on the Height of Binary Search Trees. *Journal of the ACM (JACM)*, 33(3):489–498, 1986.

[Dev92] Luc Devroye. A limit theory for random skip lists. *The Annals of Applied Probability*, 2(3):597–609, 08 1992.

[DM10] Amir Dembo and Andrea Montanari. Gibbs Measures and Phase Transitions on Sparse Random Graphs. *Brazilian Journal of Probability and Statistics*, 24(2):137–211, 2010.

[Drm01] M. Drmota. An analytic approach to the height of binary search trees. *Algorithmica*, 29(1-2):89–119, 2001. Average-case analysis of algorithms (Princeton, NJ, 1998).

[DS92] B. Drossel and F. Schwabl. Self-organized critical forest-fire model. *Phys. Rev. Lett.*, 69:1629–1632, Sep 1992.

[EK86] Stewart N. Ethier and Thomas G. Kurtz. *Markov Processes: Characterization and Convergence*. John Wiley and sons, 1986.

[ER59] P. Erdős and A. Rényi. On random graphs. I. *Publ. Math. Debrecen*, 6:290–297, 1959.

[ER60] P Erdős and A Rényi. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, pages 17–31, 1960.

[ESW95] Paul Erdős, Stephen Suen, and Peter Winkler. On the size of a random maximal graph. *Random Structures Algorithms*, 6(2-3):309–318, 1995.

[FS09] Philippe Flajolet and Robert Sedgewick. *Analytic combinatorics*. Cambridge University Press, Cambridge, 2009.

[Gil59] E. N. Gilbert. Random graphs. *Ann. Math. Statist.*, 30:1141–1144, 1959.

[GK94] Eldar Giladi and Joseph B. Keller. Eulerian number asymptotics. *Proc. Roy. Soc. London Ser. A*, 445(1924):291–303, 1994.

[Har63] Theodore E. Harris. *The Theory of Branching Processes*. Die Grundlehren der Mathematischen Wissenschaften, Bd. 119. Springer-Verlag, Berlin; Prentice-Hall, Inc., Englewood Cliffs, N.J., 1963.

[HH09] Robert Hardy and Simon C. Harris. A Spine Approach to Branching Diffusions with Applications to fflL^p-convergence of Martingales. In *Séminaire de Probabilités XLII*, volume 1979 of *Lecture Notes in Math.*, pages 281–330. Springer, Berlin, 2009.

[HR03] Christophe Hohlweg and Christophe Reutenauer. Lyndon words, permutations and trees. *Theoretical Computer Science*, 307(1):173–178, 2003.

[JŁR00] Svante Janson, Tomasz Łuczak, and Andrzej Rucinski. *Random Graphs*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience, New York, 2000.

[Jos14] Adrien Joseph. The component sizes of a critical random graph with given degree sequence. *The Annals of Applied Probability*, 24(6):2560–2594, 12 2014.

[Kin75] J. F. C. Kingman. The first birth problem for an age-dependent branching process. *Ann. Probability*, 3(5):790–801, 1975.

[LG10] Jean-François Le Gall. Itô's excursion theory and random trees. *Stochastic Process. Appl.*, 120(5):721–749, 2010.

[Lot97] M. Lothaire. *Combinatorics on words*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 1997.

[McK85] Brendan D. McKay. Asymptotics for symmetric 0-1 matrices with prescribed row sums. *Ars Combin.*, 19(A):15–25, 1985.

[MR95] Michael Molloy and Bruce Reed. A critical point for random graphs with a given degree sequence. In *Proceedings of the Sixth International Seminar on Random Graphs and Probabilistic Methods in Combinatorics and Computer Science, "Random Graphs '93" (Poznań, 1993)*, volume 6, pages 161–179, 1995.

[New51] J. D. Newburgh. The variation of spectra. *Duke Math. J.*, 18:165–176, 1951.

[NR06] Ilkka Norros and Hannu Reittu. On a conditionally Poissonian graph process. *Adv. in Appl. Probab.*, 38(1):59–75, 2006.

[OT01] Deryk Osthus and Anusch Taraz. Random maximal H-free graphs. *Random Structures Algorithms*, 18(1):61–82, 2001.

[Pit84] Boris Pittel. On growing random binary trees. *J. Math. Anal. Appl.*, 103(2):461–480, 1984.

[Pit94] Boris Pittel. Note on the heights of random recursive trees and random m-ary search trees. *Random Structures Algorithms*, 5(2):337–347, 1994.

[Ree03] Bruce Reed. The Height of a Random Binary Search Tree. *J. ACM*, 50(3):306–332 (electronic), 2003.

[Rob10] Matthew I. Roberts. Almost sure asymptotics for the random binary search tree. In *21st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms (AofA'10)*, Discrete Math. Theor. Comput. Sci. Proc., AM, pages 565–576. Assoc. Discrete Math. Theor. Comput. Sci., Nancy, 2010.

[RW92] A. Ruciński and N. C. Wormald. Random graph processes with degree restrictions. *Combin. Probab. Comput.*, 1(2):169–180, 1992.

[SÖ2] Bo Söderberg. A general Formalism for Inhomogeneous Random Graphs. *Phys. Rev. E*, 66(6):066121, Dec 2002.

[SF96] Robert Sedgewick and Philippe Flajolet. *An Introduction to the Analysis of Algorithms*. Addison-Wesley, 1996.

[Spe97] Joel Spencer. Enumerating graphs and Brownian motion. *Comm. Pure Appl. Math.*, 50(3):291–294, 1997.

[SR03] J. Sawada and F. Ruskey. Generating Lyndon brackets. An addendum to: "Fast algorithms to generate necklaces, unlabeled necklaces and irreducible polynomials over GF(2)". *J. Algorithms*, 46(1):21–26, 2003.

[SSM92] KG Subramanian, Rani Siromoney, and Lisa Mathew. Lyndon trees. *Theoretical Computer Science*, 106(2):373–383, 1992.

[Tan73] S. Tanny. A probabilistic interpretation of Eulerian numbers. *Duke Math. J.*, 40:717–722, 1973.

[Tur06] Tatyana S. Turova. Phase Transitions in Dynamical Random Graphs. *J. Stat. Phys.*, 123(5):1007–1032, 2006.

[Tur13] Tatyana S. Turova. Diffusion approximation for the components in critical inhomogeneous random graphs of rank 1. *Random Structures Algorithms*, 43(4):486–539, 2013.

Résumé

Cette thèse est consacrée à l'étude du comportement asymptotique de grands graphes et arbres aléatoires. Le premier modèle étudié est un modèle de graphe aléatoire inhomogène introduit par Bo Söderberg en 2002. Un chapitre de ce manuscrit est consacré à l'étude du comportement asymptotique de la taille des composantes connexes à proximité de la fenêtre critique, en le reliant à la longueur des excursions d'un mouvement brownien avec dérive parabolique, étendant les résultats obtenus par Aldous. Le chapitre suivant est consacré à l'étude d'un processus de graphes aléatoires proposé par Itai Benjamini, défini ainsi : les arêtes sont ajoutées indépendamment, à taux fixe. À chaque fois qu'un sommet atteint le degré k, toutes les arêtes adjacentes à ce sommet sont immédiatement supprimées. Ce processus n'est donc pas croissant, ce qui empêche d'utiliser directement un certain nombre d'approches usuelles. L'utilisation de limites locales, similaires à celles utilisées dans le PWIT, permet de montrer la présence (resp. l'absence) d'une composante géante à certaines étapes dans le cas $k \geq 5$ (resp. $k \leq 3$). Dans le cas $k = 4$, ces résultats permettent de caractériser la présence (resp. l'absence) d'une composante géante en fonction du caractère surcritique (resp. critique ou sous-critique) d'un processus de branchement associé. Dans le dernier chapitre est étudiée la hauteur d'un arbre de Lyndon associé à un mot de Lyndon choisi uniformément parmi les mots de Lyndon de longueur n, prouvant que cette hauteur est approximativement $c \ln n$, avec $c = 5,092\ldots$ la solution d'un problème d'optimisation. Afin d'obtenir ce résultat, nous couplons d'abord l'arbre de Lyndon à un arbre de Yule, que nous étudions ensuite à l'aide de techniques provenant des théories des marches branchantes et des grandes déviations.

Mots-clés: Probabilité, Graphes aléatoires, Limite locale, Transition de phase, Processus de branchement, Couplage.

Abstract

This thesis is dedicated to the study of the asymptotic behavior of some large random graphs and trees. First is studied a random graph model introduced by Bo Söderberg in 2002. One chapter of this manuscript is devoted to the study of the asymptotic behavior of the size of the connected components near the critical window, linking it to the lengths of excursion of a Brownian motion with parabolic drift. The next chapter talks about a random graph process suggested by Itai Benjamini, defined as follows: edges are independently added at a fixe rate. Whenever a vertex reaches degree k, all adjacent edges are removed. This process is non-increasing, preventing the use of some commonly used methods. By using local limits, in the spirit of the PWIT, we were able to prove the presence (resp. absence) of a giant component at some stages of the process when $k \geq 5$ (resp. $k \leq 3$). In the case $k = 4$, these results allows to link the presence (resp. absence) of a giant component to the supercriticality (resp. criticality or subcriticality) of an associated branching process. In the last chapter, the height of random Lyndon tree is studied, and is proven to be approximately $c \ln n$, in which $c = 5.092\ldots$ the solution of an optimization problem. To obtain this result, we couple the Lyndon tree with a Yule tree, then studied with the help of branching walks and large deviations.

Keywords: Probability, Random Graphs, Local Limits, Phase Transition, Branching Processes, Coupling.