Inference for conditioned Galton-Watson trees from their Harris path

Romain Azaïs, Alexandre Genadot, Benoît Henry

Inference for conditioned Galton-Watson trees from their Harris path

Romain Azaïs[a], Alexandre Genadot[b], and Benoît Henry[c]

[a] Inria Nancy – Grand Est, Team BIGS and Institut Élie Cartan de Lorraine, Nancy, France
[b] Institut de Mathématiques de Bordeaux and Inria Bordeaux – Sud Ouest, Team CQFD
[c] Institut Élie Cartan de Lorraine, Nancy, France and Inria Nancy – Grand Est, Team TOSCA

Abstract

Tree-structured data naturally appear in various fields, particularly in biology where plants and blood vessels may be described by trees, but also in computer science because `XML` documents form a tree structure. This paper is devoted to the estimation of the relative scale of ordered trees that share the same layout. The theoretical study is achieved for the stochastic model of conditioned Galton-Watson trees. New estimators are introduced and their consistency is stated. A comparison is made with an existing approach of the literature. A simulation study shows the good behavior of our procedure on finite-sample sizes. An application to the analysis of revisions of Wikipedia articles is also considered through real data.

1 Introduction

Many data are naturally modeled by an ordered tree structure: from blood vessels in biology to `XML` files in computer science through the secondary structure of RNA in biochemistry. The statistical analysis of a dataset of hierarchical records is thus of a great interest. In this paper, we consider hierarchical data sharing the same layout, like plants of a same species or pages of a same website do. Detecting differences in these tree structures can not be tackled by estimating their common form but their relative scale. Our aim is to propose a new method to estimate this scale parameter relying on a theoretical study for Galton-Watson trees conditioned on their number of nodes.

A Galton-Watson tree is the genealogy tree of a population starting from one initial ancestor (the root) in which each individual gives birth to a random number of children according to the same probability distribution, independently on each other. Any ordered tree may be encoded by its Harris path which returns height of nodes in depth-first order (see Subsection 2.2, Algorithm 1 and Figure 1). Aldous [2, Theorem 23] stated the following asymptotic property of the Harris path $\mathcal{H}[\tau_n]$ of a Galton-Watson tree τ_n conditioned on having n nodes,

$$\left(\frac{\mathcal{H}[\tau_n](2nt)}{\sqrt{n}},\ t\in[0,1]\right)\xrightarrow{(d)}\left(\frac{2}{\sigma}\mathbb{e}_t,\ t\in[0,1]\right), \tag{1}$$

when n goes to infinity, whenever the birth distribution is 1 on average with standard deviation σ and $\mathbb{e}$ denotes the normalized Brownian excursion. This means that conditioned Galton-Watson trees asymptotically share a common form (the so-called *continuum random tree*) given by the Brownian excursion, and can be differentiated only by the scale parameter of interest σ^{-1}. This class of random trees is thus well-adapted to the problem at hand. It should be already noticed that, even if a Galton-Watson tree is generated from a sequence of i.i.d. random variables, this is not the case for the conditioned structure. Estimation of σ^{-1}

is thus not trivial. We would like to emphasize that several main classes of random trees can be seen as conditioned Galton-Watson trees [7, 12], e.g., Motzkin trees from the uniform offspring distribution on the set $\{0, 1, 2\}$, Catalan trees from the offspring distribution $(0.25, 0.5, 0.25)$ on $\{0, 1, 2\}$ or Cayley trees from the Poisson offspring distribution. One also refers the reader to [3, 3.1 Galton-Watson trees] for an enumeration of some specific parameterizations. To sum up, conditioned Galton-Watson trees model a large variety of random hierarchical structures.

We consider two strategies for estimating the scale parameter σ^{-1} from a forest of conditioned Galton-Watson trees. Both rely on the idea motivated by the weak convergence (1) that, on average, the normalized Harris paths of the forest should look like the expected process $(2\sigma^{-1}E(t), t \in [0, 1])$ at least asymptotically, where $E(t) = \mathbb{E}[\mathbb{e}(t)]$. The parameter σ^{-1} can thus be expressed as the solution of a least square problem. Our first method consists in computing the least square estimator of σ^{-1} from the concatenation of the Harris paths of the forest. We establish two results of convergence in Subsection 4.2. For only one Galton-Watson tree τ_n conditioned on having n nodes, this estimator of σ^{-1} is given (see Subsection 3.1) by

$$\widehat{\lambda}[\tau_n] = \frac{\langle \mathcal{H}[\tau_n](2n\cdot), E\rangle}{2\sqrt{n}\|E\|_2^2}.$$

By virtue of the weak convergence (1), one may remark (see Corollary 6) that

$$\widehat{\lambda}[\tau_n] \xrightarrow{(d)} \sigma^{-1}\Lambda_\infty,$$

where $\Lambda_\infty = \frac{\langle \mathbb{e}, E\rangle}{\|E\|_2^2}$. Actually, the aforementioned least square estimator only exploits the average behavior of Λ_∞ (in other words, the average asymptotic behavior of Harris paths) and not its complete distribution. Our second strategy takes into account the shape of the distribution of Λ_∞: we estimate σ^{-1} by the parameter x that aligns the theoretical distribution of $x\Lambda_\infty$ and the empirical measure of the $\widehat{\lambda}[\tau_{n_i}^i]$'s in terms of Wasserstein distance. Convergence results are stated in Subsection 4.3. We point out that the theoretical properties of Λ_∞ are far from obvious. In particular, we establish by Malliavin calculus that Λ_∞ is absolutely continuous with respect to the Lebesgue measure in Proposition 8, which is required in some proofs.

The statistical framework investigated here (forest of conditioned Galton-Watson trees and estimation of functions of σ) has only been considered in a recent paper [3]. The authors of [3] exploit a result providing the asymptotic distribution of the height of a uniformly sampled node in the tree (see [3, Proposition 4]) to construct estimators of the variance σ^2 and develop asymptotic tests. For the sake of comparison, we rely on the estimation strategy chosen in this article to provide a new estimator of σ^{-1}. We compare these alternative approaches from both theoretical and numerical points of view. In particular, we show in Subsection 3.1 that the variances of our estimators are approximately 4 times lower than the one of the estimator based on this competitive approach of the literature. Our results are better in terms of dispersion because the estimators take into account all the behavior of the tree and not only the behavior of a randomly chosen node. We also point out that the theoretical setting of [3] is slightly different because investigations are directly based on infinite trees (i.e., *continuum random trees*) and not on large but finite trees. Furthermore, we would like to emphasize that, to the best of our knowledge, Harris paths (and more generally coding processes such as Łukasiewicz walks or contour processes) have never been considered to perform statistical analysis of ordered trees.

The application of our estimators on simulated and real data in Section 5 appears to be a non trivial task, in particular because it requires important preliminary computations. For this reason and to provide a turnkey solution, we have developed a `Matlab` toolbox that enables users to quickly and easily apply our methods to data. This toolbox as well as a detailed user documentation can be found at the webpage: `http://agh.gforge.inria.fr`. The numerical experiments presented in Subsection 5.2 show that both our estimators and the approach developed in [3] are intrinsically biased because of the approximation of the Harris paths of finite trees by the average Brownian excursion. Indeed, we empirically observe on simulation examples that Harris paths weakly converge to the Brownian excursion from below (see Figure 5). As a

consequence, we introduce a numerical correction of this negative bias, also implemented in the toolbox. The simulation study illustrates the good behavior of the corrected estimates on finite-sample sizes.

Visualizing the evolution of historical hierarchical data is a difficult issue in particular because such object has no representation in a Euclidean space. This problem occurs in the study of the sequence of revisions of a given Wikipedia article. Indeed, the famous free Internet encyclopedia allows its users (the *Wikipedians*) to edit almost any articles. Starting from the creation of a given article, the history of revisions is accessible and can be investigated to understand how the contributors agree on its structure, or to automatically detect vandalism[1] [1, 18]. *IBM's History Flow* is a visualization tool for documents in various stages of their development which has been applied to Wikipedia articles [23, 24]. We think that our method may be a complementary tool to this famous technique. Indeed the structure of `HTML` documents, such as Wikipedia articles, may be encoded by an ordered tree structure (see Figure 11). Furthermore, all the Wikipedia webpages share the same layout and thus can be differentiated by their relative scale. In Subsection 5.3, we apply our estimators to the analysis of two Wikipedia articles. We highlight that Wikipedia articles undergo "running in" period before reaching some kind of steady state in which the contributors had agreed on the structure of the article. In addition, we show that our technics may be used to detect improper editions of an article.

The organization of the paper is as follows. Section 2 is devoted to the formulation of the problem at hand: definition of conditioned Galton-Watson trees in Subsection 2.1, definition of Harris paths in Subsection 2.2, asymptotic behavior of Harris paths of conditioned Galton-Watson trees in Subsection 2.3. The two estimation procedures are presented in Section 3, while Section 4 focuses on the results of convergence. We point out that different regimes of convergence may be considered because size of the forest and sizes of trees can be chosen large. Precisions on this topic are provided in Subsection 4.1. Simulation techniques for conditioned Galton-Watson trees, numerical experiments and application to real data are presented in Section 5. In this preliminary version, all the figures of the numerical experiments section have been deferred to the end of the article.

2 Conditioned Galton-Watson trees

2.1 Definition

Trees are connected graphs with no cycles. A rooted tree τ is a tree in which one node has been distinguished as the root, denoted by $r(\tau)$ (always drawn at the bottom of the tree in this paper). In this case, the edges are assigned a natural orientation, away from the root towards the leaves. One obtains a directed rooted tree in which there exists a parent-child relationship: the parent of a node v is the first vertex met on the path to the root starting from v. The length of this path (in number of nodes) is called the height $h(v)$ of v. The set $c(v)$ of children of a vertex v is the set of nodes that have v as parent. An ordered or plane tree is a rooted tree in which an ordering has been specified for the set of children of each node, conventionally drawn from left to right. In this paper we consider ordered rooted trees simply referred to as trees. In addition, for any node v, $\tau[v]$ denotes the subtree of τ composed of v and all of its descendants in τ.

Intuitively, a Galton-Watson tree can be seen as a tree encoding the dynamic of a population generated from some offspring distribution μ on $\mathbb{N}$. A Galton-Watson tree τ with offspring distribution μ is a random ordered rooted tree constructed recursively as follows.

- $\diamond$ The number of children $\#c(r(\tau))$ emanating from the root is a random variable with law μ. The first generation consists thus in $\#c(r(\tau))$ vertices.
- $\diamond$ Assume that the n^{th} generation of children has been constructed and consists in a list of vertices $\mathcal{V}_n$. Then, the generation $n+1$ is constructed such that $\{\#c(v) : v \in \mathcal{V}_n\}$ is a collection of independent random variables with law μ.

[1] It frequently happens that malicious people willingly disrupt the content of an article, for instance, for political or ideological reasons.

The asymptotic behavior of Galton-Watson trees may exhibit different regimes depending on the average number of children per capita,

$$\overline{\mu} = \sum_{k \geq 0} k\mu(k).$$

- ◇ The subcritical case: $\overline{\mu} < 1$. In this case, the average number of nodes is finite. This means that the population almost-surely extincts.
- ◇ The critical case: $\overline{\mu} = 1$. The fact that the offspring distribution μ is critical also ensures the almost-sure finiteness of the tree, except when $\mu(1) = 1$. When $\mu(1) < 1$, in contrary to the sub-critical case, the expected number of nodes is infinite.
- ◇ The supercritical case: $\overline{\mu} > 1$. In this case, the number of vertices is infinite with positive probability.

We use the notation $\mathrm{GW}_n(\mu)$ for the distribution of Galton-Watson trees with offspring distribution μ conditioned on having n nodes.

Remark 1 *In this paper, we will always state our results in terms of critical Galton-Watson trees. However, this is not really a restriction since, as noted in [20, 6.3 Brownian asymptotics for conditioned Galton-Watson trees], for any offspring distribution μ, there exists a critical law μ' such that*

$$\mathrm{GW}_n(\mu) \overset{(d)}{=} \mathrm{GW}_n(\mu').$$

In particular, this means that the average number of children $\overline{\mu}$ is not identifiable from conditioned Galton-Watson trees without some additional assumptions on μ.

2.2 From ordered trees to Harris paths

The Harris walk $\mathcal{H}[\tau]$ of an ordered rooted tree τ is defined from the depth-first search algorithm and the notion of height of nodes already presented in Subsection 2.1. Depth-first search is an algorithm for traversing a tree which one explores as far as possible along each branch before backtracking. The version of the algorithm used to define the Harris walk of a tree is presented in Algorithm 1.

Function `DFS`(τ, $l = \emptyset$)**:**
  Data: an ordered tree τ
  Result: vertices of τ in depth-first order
  add $r(\tau)$ to l
  for v **in** $c(r(\tau))$ **do**
    if *r(t[v])* **is not in** l **then**
      call `DFS` ($t[v]$,l)
      add again $r(\tau)$ to l
  return l

Algorithm 1: Recursive depth-first search.

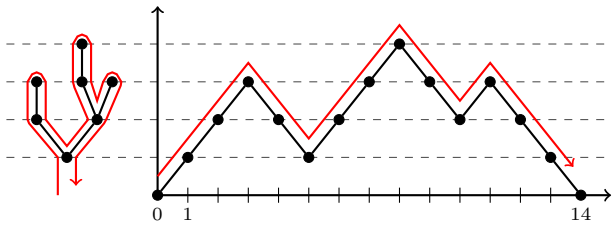

Figure 1: Construction of the Harris path (right) from 0 to $2n = 14$ as the contour of an ordered tree (left) with $n = 7$ nodes.

Remark 2 *In Algorithm 1, each node v appears $\#c(v) + 1$ times. Starting from the root of a tree τ, the result is thus a sequence of length*

$$\sum_{v \in \tau} (\#c(v) + 1) = \#\tau + \sum_{v \in \tau} \#c(v) = 2\#\tau - 1,$$

because the root is the only vertex not to be counted.

The Harris walk $\mathcal{H}[\tau]$ of τ is defined as a sequence of integers indexed by the set $\{0, \dots, 2\#\tau\}$ as follows:

- $\mathcal{H}[\tau](0) = \mathcal{H}[\tau](2\#\tau) = 0$,
- for $1 \leq k \leq 2\#\tau$, $\mathcal{H}[\tau](k) = h(v) + 1$ where v is the k^{th} node in depth-first traversal of τ.

The Harris process is then defined as the linear interpolation of the Harris walk (see example in Figure 1). Note that, as displayed in Figure 2, the tree can be recovered from its Harris process such that the correspondence is one to one.

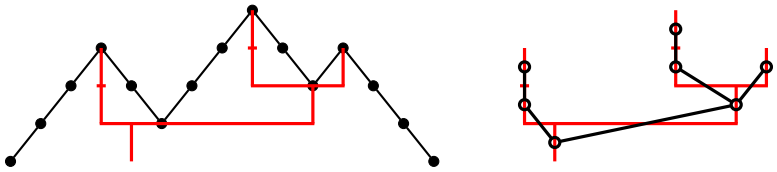

Figure 2: The ordered tree of Figure 1 in its Harris path (left): each vertical axis represents a node of the original structure (right). A common picture helping to see how to recover the tree from the contour is to imagine putting glue under the contour and then crushing the contour horizontally such that the inner parts of the contour which face each others are glued.

2.3 Asymptotic behavior of Harris paths

Let $\tau_n \sim \mathrm{GW}_n(\mu)$ with $\overline{\mu} = 1$. The variance of the offspring distribution μ is denoted by σ^2,

$$\sigma^2 = \sum_{k \geq 1} (k-1)^2 \mu(k).$$

We focus on the asymptotic behavior of the Harris process $\mathcal{H}[\tau_n](2n\cdot)$ when n tends to infinity. The convergence in distribution has been stated by Aldous [2, Theorem 23].

Theorem 3 *When n goes to infinity, we have*

$$\left(\frac{\mathcal{H}[\tau_n](2nt)}{\sqrt{n}},\ t \in [0,1] \right) \xrightarrow{(d)} \left(\frac{2}{\sigma} \mathbb{e}_t,\ t \in [0,1] \right),$$

where $\mathbb{e}$ is a standard Brownian excursion, the convergence holding in law in the space $\mathcal{C}([0,1], \mathbb{R})$.

Let us simply recall that a standard Brownian excursion is a Brownian motion conditioned on being positive and on taking the value 0 at time 1. The density of $\mathbb{e}_t$, for $0 \leq t \leq 1$, is given in [22, XI. 3. Bessel Bridges] and writes

$$\forall\, x \in \mathbb{R},\ f_{\mathbb{e}_t}(x) = \sqrt{\frac{2}{\pi}} \frac{x^2}{\sqrt{t(1-t)}^3} \exp\left(-\frac{x^2}{2t(1-t)} \right) \mathbb{1}_{\mathbb{R}_+}(x).$$

From this, we can compute some simple functionals of the excursion. For instance, we have,

$$\forall\, 0 \leq t \leq 1, \quad E_t = \mathbb{E}[\mathbb{e}_t] = 4\sqrt{\frac{t(1-t)}{2\pi}} \quad \text{and} \quad \mathbb{E}\left[\mathbb{e}_t^2\right] = 3t(1-t). \tag{2}$$

The easiest way to simulate a Brownian excursion is certainly from its identity in law with a three-dimensional Bessel bridge [22, Theorem XII.4.2], which is simply the Euclidean norm of a three-dimensional Brownian bridge,

$$(\mathbb{e}_t,\ t \in [0,1]) \overset{(\mathrm{d})}{=} \left(\sqrt{\sum_{i=1}^{3} \left(B_t^i - t B_1^i \right)^2},\ t \in [0,1] \right), \tag{3}$$

where the B^i's are three independent Brownian motions. The convergence presented in Theorem 3 also holds in expectation [8, Theorem 1].

Theorem 4 *When n goes to infinity, we have,*

$$\forall\, 0 \leq t \leq 1, \quad \mathbb{E}\left[\frac{\mathcal{H}[\tau_n](2nt)}{\sqrt{n}}\right] \longrightarrow \frac{2}{\sigma} E_t,$$

where the function $(E_t,\ 0 \leq t \leq 1)$ has been defined in (2).

Remark 5 *Theorem 3 establishes that, in the asymptotic regime, the shape of a conditioned Galton-Watson tree is given by the normalized Brownian excursion, regardless of the offspring distribution μ. However, there is one scale parameter given by the inverse of the standard deviation of μ. As a consequence, when μ is unknown, the only quantity of interest that one may access by asymptotic inference is σ^{-1}. In the next part, we shall focus on the estimation of σ^{-1}.*

3 Estimation procedure

3.1 Adequacy of the Harris path with the expected contour

Let $\tau_n \sim \mathrm{GW}_n(\mu)$ with $\overline{\mu} = 1$. We assume that the offspring distribution μ is unknown. By virtue of Theorem 4, the asymptotic average behavior of the normalized Harris process $(n^{-1/2}\mathcal{H}[\tau_n](2nt),\ 0 \leq t \leq 1)$ is given by $(2\sigma^{-1}E_t,\ 0 \leq t \leq 1)$, where σ^{-1} is obviously also unknown. We propose to estimate σ^{-1} by minimizing the $\mathbb{L}^2$-error defined by

$$\lambda \mapsto \left\| \frac{\mathcal{H}[\tau_n](2n\cdot)}{\sqrt{n}} - 2\lambda E \right\|_2^2,$$

where, and in all the sequel, $\mathbb{L}^2 = \mathbb{L}^2([0,1],\mathbb{R})$ and its usual norm is denoted $\|\cdot\|_2$ for the sake of readability. The solution of this least-square problem is well-known and is given by

$$\widehat{\lambda}[\tau_n] = \frac{\langle \mathcal{H}[\tau_n](2n\cdot), E\rangle}{2\sqrt{n}\|E\|_2^2}, \tag{4}$$

where $\langle\cdot,\cdot\rangle$ is the scalar product of $\mathbb{L}^2$.

Corollary 6 *When n goes to infinity, we have*

$$\widehat{\lambda}[\tau_n] \xrightarrow{(d)} \sigma^{-1}\Lambda_\infty,$$

where the random variable Λ_∞ is defined by

$$\Lambda_\infty = \frac{\langle e, E\rangle}{\|E\|_2^2}. \tag{5}$$

Proof. The result directly follows from Theorem 3 because the functional $x \mapsto \langle x, E\rangle$ is continuous on $\mathcal{C}([0,1])$. □

Remark 7 *The convergence in distribution stated in Corollary 6 seems quite unsatisfactory because this means that $\widehat{\lambda}[\tau_n]$ is not a consistent estimator of σ^{-1} and the least-square strategy thus looks like inadequate. Nevertheless, one can not expect a stronger convergence from the observation of only one stochastic process within a finite window of time. This is why one may only focus on the estimation of the parameter of interest σ^{-1} from a forest of conditioned Galton-Watson trees. This statistical framework is also the one considered in [3].*

Computing $\widehat{\lambda}[\tau_n]$ is a first step in the estimation of the inverse standard deviation from a large number of conditioned Galton-Watson trees. As a consequence, the distribution of the limit variable Λ_∞ is of first importance.

Proposition 8 *The random variable Λ_∞ admits a density f_{Λ_∞} with respect to the Lebesgue measure. Furthermore,*

$$\mathbb{E}[\Lambda_\infty] = 1 \quad \textit{and} \quad \mathbb{V}\mathrm{ar}(\Lambda_\infty) = \frac{1}{\|E\|_2^4} \int_0^1 \int_0^1 g(s,u)\, E_s\, E_u \,\mathrm{d}s\,\mathrm{d}u \;-\; 1, \tag{6}$$

where the mapping $g : [0,1]^2 \to \mathbb{R}_+$ is defined from

$$g(t,u) = \begin{cases} \frac{2}{\pi}\left[3\sqrt{t(u-t)(1-u)} + (2t(1-u)+u(1-t))\arcsin\left(\sqrt{\frac{t(1-u)}{u(1-t)}}\right)\right] & \textit{if } 0 \le t \le u \le 1, \\ g(u,t) & \textit{otherwise.} \end{cases}$$

Proof. The existence of a density was already known [16, 17] for the random variable $\int_0^1 \mathbb{e}_s \mathrm{d}s$ but to the best of our knowledge no paper investigates the existence of a density for Λ_∞. In these papers the study is performed thanks to the analysis of the double Laplace transform

$$\lambda \mapsto \int_0^\infty \exp(-\lambda t)\mathbb{E}\left[\exp\left(-t\int_0^1 \mathbb{e}_s \mathrm{d}s\right)\right]\mathrm{d}t.$$

Thanks to the Feynman-Kac formula, the authors express this quantity in terms of Airy functions. Then, they inverse the Laplace transform via analytical methods. Unfortunately, their method does not extend to our case. Indeed, in their case, an expression of the double Laplace transform given above is derived from the Feynman-Kac formula for standard Brownian motion which tells us that the function

$$u(t,x) = \mathbb{E}_x\left[f(B_t)\exp\left(\int_0^t B_s \mathrm{d}s\right)\right], \quad \forall\,(t,x) \in \mathbb{R}_+ \times \mathbb{R},$$

is solution of the partial differential equation

$$\begin{cases} \partial_t u(t,x) = \frac{1}{2}\Delta u(t,x) + x u(t,x) & \forall\, x \in \mathbb{R},\ t \in \mathbb{R}_+, \\ u(0,x) = f(x) & \forall\, x \in \mathbb{R}. \end{cases}$$

In this case, taking the Laplace transform in time of u leads to an ordinary differential equation whose solution can be expressed in terms of Airy functions [11]. In our problem, this partial differential equation becomes inhomogeneous in time which prevents us to use this Laplace transform. As a consequence, we think that one can not obtain information by this method. This is why we propose a proof using Malliavin calculus and the representation of the Brownian excursion as a three-dimensional Bessel bridge (3) to show that Λ_∞ admits a density. We consider the probability space $(\mathcal{C}([0,1],\mathbb{R}^3), \mathcal{F}, \mathbb{W})$, where $\mathcal{C}([0,1],\mathbb{R}^3)$ is endowed with the uniform topology, $\mathcal{F}$ is the corresponding Borel σ-field and $\mathbb{W}$ is the Wiener measure. Let T be the continuous linear operator defined by

$$\begin{array}{rccc} T: & \mathcal{C}([0,1],\mathbb{R}^3) & \to & \mathcal{C}([0,1],\mathbb{R}^3), \\ & \varphi & \mapsto & (T\varphi(s) = \varphi_s - s\varphi_1)\,. \end{array}$$

Let also Γ be the following function,

$$\Gamma : \varphi \mapsto \int_0^1 |\varphi(s)|_2 \frac{E_s}{\|E\|_2^2}\mathrm{d}s.$$

where $|x|_2$ denotes the Euclidian norm on $\mathbb{R}^3$. With these notations and (3), we have that the pushforward measure of $\mathbb{W}$ through the application

$$F : \varphi \mapsto \Gamma(T\varphi),$$

is the law of Λ_∞. In other words, the random variable F is equal in distribution to Λ_∞. Now for every φ in $\mathcal{C}([0,1],\mathbb{R}^3)$ such that $Leb\left(\overline{\{t\in\mathbb{R}_+ \,:\, \varphi(t)=0\}}\right)=0$, we have that Γ is Fréchet differentiable at point φ. Moreover, the derivative at such point φ is given by

$$\begin{array}{rccc} D_\varphi\Gamma: & (\mathcal{C}([0,1],\mathbb{R}^3) & \to & \mathbb{R}, \\ & h & \mapsto & \int_0^1 \frac{(\varphi(s),h(s))}{|\varphi(s)|_2}\frac{E_s}{\|E\|_2^2}\,\mathrm{d}s, \end{array}$$

where $(\cdot,\cdot)$ denotes the Euclidean scalar product on $\mathbb{R}^3$. Indeed, some straightforward manipulations give

$$\int_0^1\left[|\varphi(s)+h(s)|_2-|\varphi(s)|_2-\frac{(\varphi(s),h(s))}{|\varphi(s)|_2}\right]\frac{E_s}{\|E\|_2^2}\mathrm{d}s=\int_0^1\left[\frac{|h(s)|_2^2+(\varphi(s),h(s))\left(1-\frac{|\varphi(s)+h(s)|_2}{|\varphi(s)|_2}\right)}{|\varphi(s)+h(s)|_2+|\varphi(s)|_2}\right]\frac{E_s}{\|E\|_2^2}\mathrm{d}s.$$

Now, Cauchy-Schwarz inequality entails

$$\begin{aligned}\left|\int_0^1\left[|\varphi(s)+h(s)|_2-|\varphi(s)|_2-\frac{(\varphi(s),h(s))}{|\varphi(s)|_2}\right]\mathrm{d}s\right| &\le \frac{3\sqrt{\pi}}{2\sqrt{2}}\int_0^1\left[\frac{|h(s)|_2^2+|h(s)|_2\left||\varphi(s)|_2-|\varphi(s)+h(s)|_2\right|}{|\varphi(s)+h(s)|_2+|\varphi(s)|_2}\right]\mathrm{d}s \\ &\le \frac{3\sqrt{\pi}}{2\sqrt{2}}\|h\|_\infty\int_0^1\left[\frac{|h(s)|_2+\left||\varphi(s)|_2-|\varphi(s)+h(s)|_2\right|}{|\varphi(s)+h(s)|_2+|\varphi(s)|_2}\right]\mathrm{d}s.\end{aligned}$$

Now, since

$$\int_0^1\left[\frac{|h(s)|_2+\left||\varphi(s)|_2-|\varphi(s)+h(s)|_2\right|}{|\varphi(s)+h(s)|_2+|\varphi(s)|_2}\right]\mathrm{d}s$$

is well-defined (because the integrand is bounded by 2) and goes to zero as $\|h\|_\infty$ goes to zero, this proves that $D_\varphi\Gamma$ is the Fréchet derivative of Γ at point φ. The functional T being linear, F is also Fréchet differentiable with Fréchet derivative given by

$$\begin{array}{rccc} D_\varphi F: & (\mathcal{C}([0,1],\mathbb{R}^3) & \to & \mathbb{R}, \\ & h & \mapsto & \int_0^1 \frac{(T\varphi(s),Th(s))}{|T\varphi(s)|_2}\frac{E_s}{\|E\|_2^2}\,\mathrm{d}s. \end{array}$$

Moreover, let h be an element of $\mathbb{L}^2([0,1],\mathbb{R}^3)$, we have

$$\left|F\left(\omega+\int_0^{\cdot}h(s)\mathrm{d}s\right)-F(\omega)\right|\le\frac{3\pi}{4}\int_0^1\left\{\left|\int_0^t h(s)\mathrm{d}s\right|_2+t\left|\int_0^1 h(s)\mathrm{d}s\right|_2\right\}E_t\mathrm{d}t.$$

But in the right hand side of the last inequality, we have, using Jensen's inequality,

$$\begin{aligned}\int_0^1\left\{\sqrt{\sum_{i=1}^3\left(\int_0^t h_s^i\mathrm{d}s\right)^2}+t\sqrt{\sum_{i=1}^3\left(\int_0^1 h_s^i\mathrm{d}s\right)^2}\right\}E_t\,\mathrm{d}t &\le \int_0^1\sqrt{\sum_{i=1}^3\left(\int_0^1(h_s^i)^2\mathrm{d}s\right)}(1+t)E_t\,\mathrm{d}t \\ &= \int_0^1\|h\|_{\mathbb{L}^2([0,1],\mathbb{R}^3)}(1+t)E_t\,\mathrm{d}t.\end{aligned}$$

From this, using the results of [19, p. 35], we have that F belongs to the space $\mathbb{D}^{1,2}$, which is the domain of the Malliavin operator D in $\mathbb{L}^2([0,1],\mathbb{R}^3)$ (see [19, pp. 25–27] for more details). Finally, it follows that the Malliavin derivative of F is given by

$$DF(\omega) = \left(\int_0^1 \frac{(\omega_s - s\omega_1)E_s}{|\omega_s - s\omega_1|_2 \|E\|_2^2} (\mathbb{1}_{s>u} - s)\mathrm{d}s,\ u \in [0,1] \right) \in \mathbb{L}^2([0,1],\mathbb{R}^3).$$

Now, since DF is not zero in $\mathbb{L}^2([0,1],\mathbb{R}^3)$ for $\mathbb{W}$-almost every ω, we get, together with [19, Theorem 2.1.2], the existence of a density for F with respect to the Lebesgue measure. The calculus of the variance is derived from the joint density of the couple $(\mathbb{e}_t, \mathbb{e}_s)$ for $(s,t) \in [0,1]^2$, which is given in [22, XI. 3. Bessel Bridges]. □

It should be noted that $\widehat{\lambda}[\tau_n]$ is somehow asymptotically unbiased because its weak limit is σ^{-1} on average by (6) and Corollary 6. The expression (6) of the variance of Λ_∞ is an explicit but quite intractable formula. Nevertheless, it may be at least evaluated numerically to compute the variance of Λ_∞. Otherwise, we can also use Monte Carlo simulations to produce a sample with same law as Λ_∞ to achieve this task. Both methods lead to

$$\mathbb{V}\mathrm{ar}(\Lambda_\infty) \simeq 0.0690785.$$

At this point, it is quite interesting to compare our approach to the one developed in [3]. The authors of [3] construct estimators for the variance of the offspring distribution of a forest of conditioned critical Galton-Watson trees. Their strategy relies on the distance to the root of a uniformly sampled node v of the considered tree $\tau_n \sim \mathrm{GW}_n(\mu)$,

$$\delta[\tau_n] = \frac{h(v)}{\sqrt{n}}. \tag{7}$$

Using Theorem 3, it has been shown that $\delta[\tau_n]$ converges in law, when the number of nodes n goes to infinity, towards $\sigma^{-1}\Delta_\infty$ where the random variable Δ_∞ follows the Rayleigh distribution with scale 1 [3, Proposition 4] with density,

$$\forall\, x \in \mathbb{R}_+,\ f_{\Delta_\infty}(x) = x \exp\left(-\frac{1}{2}x^2\right).$$

We emphasize that $\delta[\tau_n]$ is somehow biased because $\mathbb{E}[\Delta_\infty] = \sqrt{\frac{\pi}{2}} \neq 1$. Nevertheless, one may avoid this issue by considering the quantity

$$\widehat{\delta}[\tau_n] = \sqrt{\frac{2}{\pi}}\delta[\tau_n] \tag{8}$$

that converges to $\sigma^{-1}\sqrt{\frac{2}{\pi}}\Delta_\infty$ which is σ^{-1} on average. As a consequence, $\widehat{\lambda}[\tau_n]$ and $\widehat{\delta}[\tau_n]$ are two quantities directly computable from the tree τ_n and that may be used to construct an estimator of the inverse standard deviation of interest. We propose to compare them from their respective asymptotic dispersion which should be as small as possible in order to get an accurate estimator. A first comparison may be done by computing the variances of Λ_∞ and $\sqrt{\frac{2}{\pi}}\Delta_\infty$. One has

$$\mathbb{V}\mathrm{ar}\left(\sqrt{\frac{2}{\pi}}\Delta_\infty\right) \simeq 0.2732395 \qquad \text{and} \qquad \mathbb{V}\mathrm{ar}(\Lambda_\infty) \simeq 0.0690785.$$

This difference in the dispersions is quite apparent in Figure 3 where the densities of $\sqrt{\frac{2}{\pi}}\Delta_\infty$ and Λ_∞ have been displayed. Consequently, one may expect better results in terms of dispersion from our approach.

3.2 Estimation strategies

In this section, we detail two ideas in order to estimate σ^{-1} from a forest of conditioned Galton-Watson trees. A forest is defined as a tuple of trees. Let N be a positive integer. In this section, we consider a forest $\mathcal{F}$ made of N independent trees $\tau^1, \ldots, \tau^N$ with respective sizes $n_1, \ldots, n_N$ and respective laws $\mathrm{GW}_{n_1}(\mu), \ldots, \mathrm{GW}_{n_N}(\mu)$.

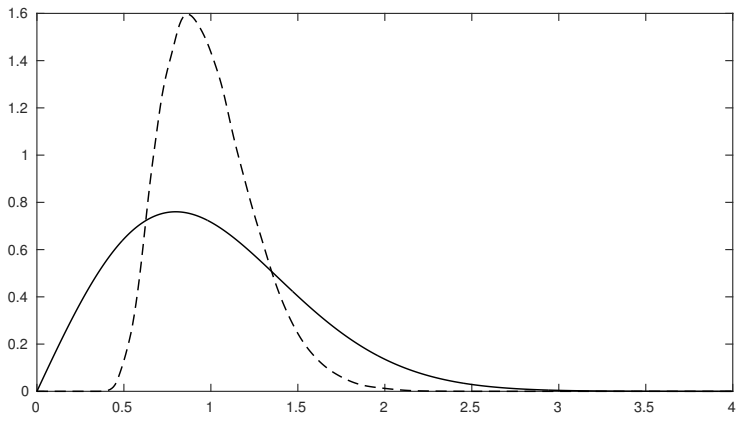

Figure 3: Densities of $\sqrt{\frac{2}{\pi}}\Delta_\infty$ (full line) where Δ_∞ follows the Rayleigh distribution given by $f(x) = \frac{\pi}{2}x\exp\left(-\frac{\pi x^2}{4}\right)$ for $x \in \mathbb{R}_+$ and of Λ_∞ (dashed line) estimated from 1 000 000 simulated Brownian excursions.

3.2.1 Least-square estimation

This first strategy lies on the goodness of fit between the Harris path of the forest with the expected limiting contour. This adequacy is measured thanks to an $\mathbb{L}^2([0,N],\mathbb{R})$-norm. More precisely, we denote $(\mathcal{H}[\mathcal{F}](t),\ t \in [0,N])$ the Harris path of the forest $\mathcal{F}$. This process is defined by

$$\forall\, 0 \le t \le N,\ \mathcal{H}[\mathcal{F}](t) = \sum_{i=1}^{N} \frac{1}{\sqrt{n_i}} \mathcal{H}[\tau^i](2n_i(t-i+1))\mathbb{1}_{[i-1,i)}(t).$$

The Harris path of a forest is simply the concatenation of the Harris paths of each tree, in the natural order. We propose to estimate σ^{-1} by $\widehat{\lambda}_{ls}[\mathcal{F}]$ that minimizes the $\mathbb{L}^2([0,N],\mathbb{R})$-error

$$\lambda \mapsto \|\mathcal{H}[\mathcal{F}](\cdot) - \lambda H(\cdot - \lfloor\cdot\rfloor)\|^2_{\mathbb{L}^2([0,N],\mathbb{R})}.$$

As aforementioned in (4), $\widehat{\lambda}_{ls}[\mathcal{F}]$ can be explicitly computed. Indeed, on can check that

$$\widehat{\lambda}_{ls}[\mathcal{F}] = \frac{\langle \mathcal{H}[\mathcal{F}](\cdot), H(\cdot - \lfloor\cdot\rfloor)\rangle}{\|H(\cdot - \lfloor\cdot\rfloor)\|_2^2}. \tag{9}$$

Interestingly, $\widehat{\lambda}_{ls}[\mathcal{F}]$ is only the average of the quantities $\widehat{\lambda}[\tau^i]$ (defined in (4)),

$$\widehat{\lambda}_{ls}[\mathcal{F}] = \frac{1}{N}\sum_{i=1}^{N} \widehat{\lambda}[\tau^i].$$

Thus, according to Theorems 4 and 6, one can expect that $\widehat{\lambda}_{ls}[\mathcal{F}]$ tends to σ^{-1} in some sense, when both N and n_i go to infinity, by virtue of the law of large numbers.

3.2.2 Estimation by minimal Wasserstein distance

In the preceding method, we did not use our knowledge of the limiting distribution of the random variable of type $\lambda[\tau^n]$. In order to take this into account, one may want to test the goodness of fit between the empirical measure $\widehat{\mathcal{P}}$ defined by

$$\widehat{\mathcal{P}} = \frac{1}{N}\sum_{i=1}^{N} \delta_{\widehat{\lambda}[\tau^i]} \tag{10}$$

and the the law of Λ_∞. Using Wasserstein metrics to align distributions is rather natural since it corresponds to the transportation cost between two probability laws. In particular, this feature appears to be useful in a statistical framework [5, 10]. In our case, $\widehat{\mathcal{P}}$ is expected to look in some sense like $\sigma^{-1}\Lambda_\infty$ in the asymptotic regime of an infinite forest of infinite trees. That is why, we propose to estimate σ^{-1} with the real number λ which minimizes the distance between $\widehat{\mathcal{P}}$ and $\lambda\Lambda_\infty$. More precisely, our estimator $\widehat{\lambda}_W[\mathcal{F}]$ is defined by

$$\widehat{\lambda}_W[\mathcal{F}] = \underset{\lambda>0}{\arg\min}\, d_W\left(\widehat{\mathcal{P}}\,,\, \mathbb{P}_{\lambda\Lambda_\infty}\right), \tag{11}$$

where d_W denotes the Wasserstein distance of order 2 and $\mathbb{P}_{\lambda\Lambda_\infty}$ denotes the law of $\lambda\Lambda_\infty$.

The Wasserstein distance of order 2, denoted $d_W(\nu_1,\nu_2)$, between two measures ν_1 and ν_2 can be defined from their cumulative distribution functions F_1 and F_2 as follows,

$$d_W(\nu_1,\nu_2) = \|F_1^{-1} - F_2^{-1}\|_2. \tag{12}$$

Let $\widehat{F}$ be the cumulative function of the empirical measure $\widehat{\mathcal{P}}$, while $F_{\lambda\Lambda_\infty}$ stands for the cumulative function of the random variable $\lambda\Lambda_\infty$. As a consequence of (12), one has

$$\begin{aligned} d_W\left(\frac{1}{N}\sum_{i=1}^N \delta_{\widehat{\lambda}[\tau^i]}\,,\, \mathbb{P}_{\lambda\Lambda_\infty}\right)^2 &= \int_0^1 \left(\widehat{F}^{-1}(s) - F^{-1}_{\lambda\Lambda_\infty}(s)\right)^2 \mathrm{d}s \\ &= \int_0^1 \left(\widehat{F}^{-1}(s) - \lambda F^{-1}_{\Lambda_\infty}(s)\right)^2 \mathrm{d}s, \end{aligned}$$

thanks to the fact that $F^{-1}_{\lambda\Lambda_\infty} = \lambda F^{-1}_{\Lambda_\infty}$. It follows that minimizing the Wasserstein distance boils down to solve a least-square minimization problem. Hence, it comes that

$$\begin{aligned} \widehat{\lambda}_W[\mathcal{F}] &= \frac{\langle \widehat{F}^{-1}, F^{-1}_{\Lambda_\infty}\rangle}{\|F^{-1}_{\Lambda_\infty}\|_2^2} \\ &= \frac{1}{\|F^{-1}_{\Lambda_\infty}\|_2^2}\sum_{i=1}^N \widehat{\lambda}[\tau^{(i)}] \int_{\frac{i-1}{N}}^{\frac{i}{N}} F^{-1}_{\Lambda_\infty}(s)\mathrm{d}s, \end{aligned}$$

where $(\widehat{\lambda}[\tau^{(i)}])_{1\le i\le N}$ denotes the order statistic associated with the family $(\widehat{\lambda}[\tau^i])_{1\le i\le N}$.

Remark 9 *We point out the fact that there is no problem of definition in the above quantities because both $\widehat{F}^{-1}$ and $F^{-1}_{\Lambda_\infty}$ belong to $\mathbb{L}^2$. In the first case, this follows from the fact that $\widehat{F}^{-1}$ is bounded (because $\widehat{\mathcal{P}}$ has compact support). For $F^{-1}_{\Lambda_\infty}$, this comes from the uniform sampling principle which entails that*

$$\int_0^1 F^{-1}_{\Lambda_\infty}(u)^2\ \mathrm{d}u = \mathbb{E}[\Lambda_\infty^2].$$

Remark 10 *The proposed methodology consists in identifying the best parameter λ that allows to align distributions $\widehat{\mathcal{P}}$ and $\mathbb{P}_{\lambda\Lambda_\infty}$. The Wasserstein distance is well-adapted to this problem because it is computed from the inverse cumulative distribution functions together with the fact that $F^{-1}_{\lambda\Lambda_\infty} = \lambda F^{-1}_{\Lambda_\infty}$. As a consequence, one may get the optimal parameter $\widehat{\lambda}_W[\mathcal{F}]$ from only a numerical estimate of $F^{-1}_{\Lambda_\infty}$. The same trick does not hold for maximum likelihood method: one can not express the likelihood of $\lambda\Lambda_\infty$ as a function of the two variables λ and f_{Λ_∞}. Thus this alternative method is not adequate without an explicit formula for f_{Λ_∞}, which is always out of our reach.*

4 Main results

4.1 Statistical framework

4.1.1 Increasing sequences of random forest

Before going further, the statistical framework needs to be precisely formulated. In the sequel, the set of integer sequences is denoted $\mathbb{S}$. For any positive real number A, we denote by $\mathbb{S}_A$ the subset of $\mathbb{S}$ defined by

$$\mathbb{S}_A = \left\{ u \in \mathbb{S} \,:\, \min_{i \geq 1} u_i \geq A \right\}.$$

In addition, for any sequence u in $\mathbb{S}$ and any positive integer N, $\vec{u}_N$ is the multi-integer made of the N first components of u, that is

$$\vec{u}_N = (u_1, \ldots, u_N).$$

Moreover, for any multi-integer $\mathbf{n}$ in $\bigcup_{n \geq 1} \mathbb{N}^n$, we denote by $\ell(\mathbf{n})$ its number of components and by $\mathfrak{m}(\mathbf{n})$ its minimal value, that is

$$\mathfrak{m}(\mathbf{n}) = \min_{1 \leq i \leq \ell(\mathbf{n})} \mathbf{n}_i.$$

Somehow, in the forests we are about to consider, $\mathfrak{m}(\mathbf{n})$ shall refer to the size of the smallest tree whereas $\ell(\mathbf{n})$ shall refer to the size of the forest.

Now, let us introduce our probabilistic framework. Let $(\tau_n^k)_{n,k \geq 1}$ be a family of conditioned Galton-Watson trees such that, for a given n, the family $(\tau_n^k)_{k \geq 1}$ is i.i.d. $\mathrm{GW}_n(\mu)$. From this family, we define, for any multi-integer $\mathbf{n} = (n_1, \ldots, n_N)$, the random forest $\mathcal{F}_{\mathbf{n}}$ made of the trees $(\tau_{n_1}^1, \ldots, \tau_{n_N}^N)$.

The idea of this construction is to consider increasing (in the sense of inclusion) sequences of random forests. Indeed, assume we are given a sequence $(u_n)_{n \geq 1}$ of integers (corresponding with the sizes of our trees), then the N first trees of the forest $\mathcal{F}_{\vec{u}_{N+1}}$ are the same as the trees of the forest $\mathcal{F}_{\vec{u}_N}$.

4.1.2 Considered asymptotic regimes

In this context, several asymptotic regimes may be considered. In the present paper, we consider the following asymptotics: when $\ell(\mathbf{n})$ goes to infinity referred to as the *large forest* regime and when $\mathfrak{m}(\mathbf{n})$ goes to infinity namely the *large trees* regime.

To be crystal clear, let us precise what we mean by saying that something converges as $\mathfrak{m}(\mathbf{n})$ (or $\ell(\mathbf{n})$) goes to infinity. Let f be an application from $\bigcup_{n \geq 1} \mathbb{N}^n$ into some metric space $(\mathcal{E}, d)$ (of course, what we are about to say trivially extends to any topological space). We say that f converges to some element e of $\mathcal{E}$ as $\mathfrak{m}(\mathbf{n})$ ($\ell(\mathbf{n})$, respectively) goes to infinity if

$$\forall\, \varepsilon > 0,\ \exists\, A \in \mathbb{R}_+,\ \forall\, \mathbf{n} \in \cup_{n \geq 1} \mathbb{N}^n, \quad \mathfrak{m}(\mathbf{n}) > A \ (\ell(\mathbf{n}) > A, \text{respectively}) \ \Rightarrow\ d(f(\mathbf{n}), \mathbf{e}) < \varepsilon.$$

Corollary 13 and Proposition 15 deal with the *large trees* regime, whereas Proposition 13, Lemma 16 and Proposition 17 focus on the *large forest* regime.

4.2 Least square estimation

This first result focuses on the *large trees* regime.

Corollary 11 *When $\mathfrak{m}(\mathbf{n})$ goes to infinity, we have*

$$\widehat{\lambda}_{ls}[\mathcal{F}_{\mathbf{n}}] \xrightarrow{(d)} \sigma^{-1} \frac{1}{\ell(\mathbf{n})} \sum_{i=1}^{\ell(\mathbf{n})} \Lambda_{\infty,i}, \tag{13}$$

where the $\Lambda_{\infty,i}$'s are N independent copies of Λ_∞. Furthermore, when $\mathfrak{m}(\mathbf{n})$ goes to infinity, we have

$$\mathbb{E}\left[\widehat{\lambda}_{ls}[\mathcal{F}_{\mathbf{n}}]\right] \longrightarrow \sigma^{-1}.$$

Proof. The first convergence is a direct consequence of the independence of the family $(\tau^i_{\mathbf{n}_i})_{1\leq i\leq \ell(\mathbf{n})}$ and the fact that each one converges to a random variable with law Λ_∞ by Corollary 6. Now, its remains to prove the second statement. Since the family $(\tau^i_{\mathbf{n}_i})_{1\leq i\leq \ell(\mathbf{n})}$ is made of independent random variables, its follows from Theorem 4 and the definition (4) of $\widehat{\lambda}[\tau_{n_i}]$ that the proof of this last statement boils down to prove that, when n goes to infinity,

$$\int_0^1 \mathbb{E}\left[\frac{\mathcal{H}[\tau_n](2ns)}{\sqrt{n}}\right] E_s \,\mathrm{d}s \longrightarrow \frac{2}{\sigma}\int_0^1 E_s^2 \,\mathrm{d}s,$$

where τ_n is some tree with law $\mathrm{GW}_n(\mu)$. It is known from [8, Lemma 4] that, for any positive integer n and real number $0<t<1$,

$$\forall x\in\mathbb{R}_+,\ \mathbb{P}\left(\frac{\mathcal{H}[\tau_n](2nt)}{\sqrt{n}} > x\right) \leq \frac{C}{t}\exp\left(\frac{-Dx}{\sqrt{t}}\right). \tag{14}$$

From this last estimate, one can easily shows that $\mathbb{E}\left[\frac{\mathcal{H}[\tau_n](2n\cdot)}{\sqrt{n}}\right]$ is uniformly bounded (w.r.t. n) by an integrable function. Finally, the result follows from Theorem 4 and the dominated convergence theorem. □

Remark 12 *It is worth noting that the limit appearing in the right hand side of (13) is unbiased, that is*

$$\mathbb{E}\left[\sigma^{-1}\frac{1}{\ell(\mathbf{n})}\sum_{i=1}^{\ell(\mathbf{n})}\Lambda_{\infty,i}\right] = \sigma^{-1}.$$

Moreover, (6) and the law of large numbers entail that this same limit almost surely converges to σ^{-1} as $\ell(\mathbf{n})$ goes to infinity.

The following result states a stronger convergence when $\ell(\mathbf{n})$ goes to infinity before $\mathfrak{m}(\mathbf{n})$. The spirit of this result is that, given an increasing sequence of random forests, the least square estimator can not be too far from σ^{-1} as soon as the sizes of the trees are large enough. In particular, due to the weakness of the convergence of conditioned Galton-Watson trees given in Theorem 4, one can not expect a stronger result of convergence.

Proposition 13 *We have,*

$$\forall\,\epsilon>0,\ \exists\, A\in\mathbb{N},\ \forall\, u\in\mathbb{S}_A,\ \mathbb{P}\left(\limsup_{N\to\infty}\left|\widehat{\lambda}_{ls}[\mathcal{F}_{\vec{u}_N}]-\sigma^{-1}\right|<\epsilon\right)=1.$$

Proof. We begin the proof by showing that the family $\left(\widehat{\lambda}[\tau_n^k]\right)_{n,k\geq 1}$ has uniformly bounded fourth moments. By Jensen's inequality, there exists a positive constant c such that

$$\begin{aligned}\mathbb{E}\left[\left(\widehat{\lambda}[\tau_n^i]\right)^4\right] &\leq c\int_0^1 \mathbb{E}\left[\left(\frac{\mathcal{H}[\tau_n^i](2ns)}{\sqrt{n}}\right)^4\right]\mathrm{d}s\\ &= 4c\int_0^1\int_{\mathbb{R}_+} x^3\,\mathbb{P}\left(\frac{\mathcal{H}[\tau_n^i](2ns)}{\sqrt{n}}>x\right)\mathrm{d}x\,\mathrm{d}s.\end{aligned}$$

Finally, using again equation (14) gives the desired bound,

$$\mathbb{E}\left[\widehat{\lambda}[\tau_n^i]^4\right] \leq \frac{12\,c\,C}{D^4}. \tag{15}$$

From this point we consider a sequence u of integers. This sequence corresponds to the sizes of the trees in our increasing sequence of random forests $(\mathcal{F}_{\vec{u}_N})_{N\geq 1}$. We recall according to the definitions given in the beginning of this section that the random forest $\mathcal{F}_{\vec{u}_N}$ is composed of the trees $(\tau^1_{u_1},\dots,\tau^N_{u_N})$.

Let $m^i_{u_i}$ be the expectation of $\widehat{\lambda}[\tau^i_{u_i}]$. It is worth noting that this expectation depends only on the integer u_i. Now, using the uniform bound on the fourth moment, it is easy to show using standard methods that

$$\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\lambda}[\tau^i_{u_i}]-m^i_{u_i}\right)\xrightarrow{a.s.}0, \tag{16}$$

when N goes to infinity. Moreover, using Theorem 4, we have that $m^i_{u_i}$ converges to σ^{-1} as u_i goes to infinity, from which it follows that for any $\epsilon>0$, there exists an integer A such that

$$\left|m^i_{u_i}-\sigma^{-1}\right|<\epsilon, \tag{17}$$

whenever $u_i>A$. Finally, letting all the u_i's be greater than A, we have that there exists a measurable set Ω_u , with mass 1, such that, using (16) and (17), for all ω in this set,

$$\limsup_{N\to\infty}\left|\frac{1}{N}\sum_{i=1}^{N}\widehat{\lambda}[\tau^i_{u_i}](\omega)-\sigma^{-1}\right|\leq\limsup_{N\to\infty}\frac{1}{N}\left|\sum_{i=1}^{N}\widehat{\lambda}[\tau^i_{u_i}](\omega)-m^i_{u_i}\right|+\limsup_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\left|m^i_{u_i}-\sigma^{-1}\right|\leq\epsilon.$$

□

Remark 14 *According to the proof of the preceding theorem, it would be very interesting to control the rate of convergence in Theorem 4. Indeed, this would enable us to get controls of the error in the convergence stated in Proposition 13 given in terms of the smallest tree in the increasing sequence of random forests.*

4.3 Estimation by minimal Wasserstein distance

As in the preceding section, we begin by looking at the convergence in $\mathfrak{m}(\mathbf{n})$ (*large trees* regime).

Proposition 15 *When $\mathfrak{m}(\mathbf{n})$ goes to infinity, we have*

$$\widehat{\lambda}_W[\mathcal{F}_{\mathbf{n}}]\xrightarrow{(d)}\frac{1}{\sigma\|F^{-1}_{\Lambda_\infty}\|_2^2}\sum_{i=1}^{\ell(\mathbf{n})}\Lambda_{\infty,(i)}\int_{\frac{i-1}{\ell(\mathbf{n})}}^{\frac{i}{\ell(\mathbf{n})}}F^{-1}_{\Lambda_\infty}(s)\mathrm{d}s,$$

where the $\Lambda_{\infty,(i)}$'s are N independent copies of Λ_∞ sorted in increasing order. In addition, the limit is asymptotically unbiased. Indeed, when $\ell(\mathbf{n})$ goes to infinity,

$$\frac{1}{\sigma\|F^{-1}_{\Lambda_\infty}\|_2^2}\mathbb{E}\left[\sum_{i=1}^{\ell(\mathbf{n})}\Lambda_{\infty,(i)}\int_{\frac{i-1}{\ell(\mathbf{n})}}^{\frac{i}{\ell(\mathbf{n})}}F^{-1}_{\Lambda_\infty}(s)\mathrm{d}s\right]\longrightarrow\frac{1}{\sigma}.$$

Proof. The convergence in distribution is straightforward from Corollary 6 and standard methods on order statistics. We now prove that the estimator is asymptotically unbiased. In order to lighten the notation, let us set

$$N=\ell(\mathbf{n}).$$

It is well known, since Λ_∞ has a density, that, for any $1\leq i\leq N$, one has (see for instance [6])

$$\mathbb{E}\left[\Lambda_{\infty,(i)}\right]=N\binom{N-1}{i-1}\int_0^\infty xF_{\Lambda_\infty}(x)^{i-1}(1-F_{\Lambda_\infty}(x))^{N-i}f_{\Lambda_\infty}(x)\mathrm{d}x.$$

Hence,

$$\mathbb{E}\left[\sum_{i=1}^{N}\Lambda_{\infty,(i)}\int_{\frac{i-1}{N}}^{\frac{i}{N}}F_{\Lambda_\infty}^{-1}(s)\mathrm{d}s\right]$$
$$=N\int_0^\infty x f_{\Lambda_\infty}(x)\sum_{i=1}^{N}\binom{N-1}{i-1}F_{\Lambda_\infty}(x)^{i-1}(1-F_{\Lambda_\infty}(x))^{N-i}\int_0^{\frac{1}{N}}F_{\Lambda_\infty}^{-1}\left(s+\frac{i-1}{N}\right)\mathrm{d}s\ \mathrm{d}x.$$

This rewrites thanks to the right inverse sampling principle as

$$\mathbb{E}\left[\sum_{i=1}^{N}\Lambda_{\infty,(i)}\int_{\frac{i-1}{N}}^{\frac{i}{N}}F_{\Lambda_\infty}^{-1}(s)\mathrm{d}s\right]=\int_0^1 F_{\Lambda_\infty}^{-1}(y)K_n\left(F_{\Lambda_\infty}^{-1}\right)(y)\ \mathrm{d}y,$$

where K_n is defined for all function φ in $\mathbb{L}^2$ by

$$K_n(\varphi)(y)=\sum_{i=1}^{N}\binom{N-1}{i-1}y^{i-1}(1-y)^{N-i}\int_0^{\frac{1}{N}}\varphi\left(s+\frac{i-1}{N}\right)\mathrm{d}s,\quad\forall\, y\in[0,1].$$

The operators K_n are known as Berstein-Kantorovich operators which were introduce in the 30's by Kantorovich in order to extend the properties of Berstein polynomials to non-continuous functions [13]. In particular, it is known that, for all φ in $\mathbb{L}^2$, $K_n(\varphi)$ converges strongly to φ in $\mathbb{L}^2$ (see [15] for an old but practical reference).

Now, according to Cauchy-Schwarz inequality we have that

$$\left|\int_0^1 F_{\Lambda_\infty}^{-1}(y)K_n\left(F_{\Lambda_\infty}^{-1}\right)(y)\ \mathrm{d}y-\int_0^1 F_{\Lambda_\infty}^{-1}(y)^2\ \mathrm{d}y\right|\le\|F_{\Lambda_\infty}^{-1}\|_2^2\int_0^1\left(K_n(F_{\Lambda_\infty}^{-1})(y)-F_{\Lambda_\infty}^{-1}(y)\right)^2\mathrm{d}y.$$

But since $K_n(F_{\Lambda_\infty}^{-1})$ converges to $F_{\Lambda_\infty}^{-1}$ in $\mathbb{L}^2$, we finally obtain

$$\mathbb{E}\left[\sum_{i=1}^{N}\Lambda_{\infty,(i)}\int_{\frac{i-1}{N}}^{\frac{i}{N}}F_{\Lambda_\infty}^{-1}(s)\mathrm{d}s\right]\longrightarrow\|F_{\Lambda_\infty}^{-1}\|_2^2,$$

when N goes to infinity. This gives the result. □

In addition, we have the same kind of strong convergence result for this estimator. It lies on the fact that the empirical measure $\widehat{\mathcal{P}}$ defined in (10) must be close (in Wasserstein distance) to the law of $\sigma^{-1}\Lambda_\infty$ as soon as the trees are large enough. More precisely, we have the following lemma.

Lemma 16 *Let $\mathcal{P}$ be the law of $\sigma^{-1}\Lambda_\infty$. Let also $\mathcal{P}_{\mathbf{n}}$ be the empirical distribution defined for any multi-integer $\mathbf{n}$ by*

$$\mathcal{P}_{\mathbf{n}}=\sum_{i=1}^{\ell(\mathbf{n})}\delta_{\widehat{\lambda}[\tau_{\mathbf{n}_i}^i]}.$$

Then, the following statement holds,

$$\forall\,\epsilon>0,\ \exists\, A\in\mathbb{N},\ \forall\, u\in\mathbb{S}_A,\quad\mathbb{P}\left(\limsup_{N\to\infty}d_W\left(\mathcal{P}_{\vec{u}_N},\mathcal{P}\right)<\epsilon\right)=1.$$

Proof. Let $\mathbf{n}$ be a multi-integer. Let Π_δ being the canonical projection of $\mathbb{R}$ on $[-\delta,\delta]$, for a positive real number δ. We have

$$d_W\left(\mathcal{P}_{\mathbf{n}}(\omega),\mathcal{P}\right)\le d_W\left(\mathcal{P}_{\mathbf{n}}(\omega),\Pi_\delta\mathcal{P}_{\mathbf{n}}(\omega)\right)+d_W\left(\Pi_\delta\mathcal{P}_{\mathbf{n}}(\omega),\Pi_\delta\mathcal{P}\right)+d_W\left(\mathcal{P},\Pi_\delta\mathcal{P}\right),\tag{18}$$

where $\Pi_\delta\mu$ denotes the image measure of μ by Π_δ. To obtain the desired result, we need to control each of the three terms in the right hand side of (18).

◇ **Third term.** First, it is clear, for any probability measure μ, that Π_δ is a transport of μ on $\Pi_\delta\mu$ which needs not to be optimal [4, 2. Generalities on Kantorovich transport distances]. Hence,

$$d_W\left(\mu, \Pi_\delta\mu\right) \leq \sqrt{\int_{\mathbb{R}} |x - \Pi_\delta(x)|^2 \, \mu(dx)}.$$

It follows, since $x \mapsto x^2$ is integrable with respect to $\mathcal{P}$, that δ can be chosen in order to have

$$d_W\left(\mathcal{P}, \Pi_\delta\mathcal{P}\right) \leq \sqrt{\mathbb{E}\left[\left(\Lambda_\infty\right)^2 \mathbb{1}_{|\Lambda_\infty|>\delta}\right]} < \frac{\varepsilon}{3}. \tag{19}$$

◇ **First term.** On the other hand, following the same lines as in the proof of Proposition 13, one can shows that

$$\forall\, \epsilon > 0,\ \exists\, A \in \mathbb{N},\ \forall\, u \in \mathbb{S}_A, \quad \mathbb{P}\left(\limsup_{N\to\infty} \left| \frac{1}{N}\sum_{i=1}^{N} \widehat{\lambda}[\tau^i_{u_i}]^2 \mathbb{1}_{|\widehat{\lambda}[\tau^i_{u_i}]|>\delta} - \mathbb{E}\left[\Lambda^2_\infty \mathbb{1}_{|\Lambda_\infty|>\delta}\right] \right| < \epsilon \right) = 1. \tag{20}$$

This bound allows us to control the first term in the right hand side of (18) since

$$\begin{aligned} d_W\left(\mathcal{P}_{\mathbf{n}}\left(\omega\right), \Pi_\delta\mathcal{P}_{\mathbf{n}}\left(\omega\right)\right) &\leq \sqrt{\int_{\mathbb{R}} |x - \Pi_\delta(x)|^2 \, \mathcal{P}_{\mathbf{n}}(\omega)(dx)} \\ &\leq \sqrt{\frac{1}{N}\sum_{i=1}^{N} \widehat{\lambda}[\tau^i_{n_i}](\omega)^2 \mathbb{1}_{|\widehat{\lambda}[\tau^i_{n_i}](\omega)|>\delta}}. \end{aligned}$$

Hence, it remains to control the second term.

◇ **Second term.** Since $\Pi_\delta\mathcal{P}_{\mathbf{n}}\left(\omega\right)$ and $\Pi_\delta\mathcal{P}$ are compactly supported measure, for any multi-integer $\mathbf{n}$, we have the following duality formula for the first order Wasserstein distance (which we denote W_1),

$$W_1\left(\Pi_\delta\mathcal{P}_{\vec{n}}\left(\omega\right), \Pi_\delta\mathcal{P}\right) = \sup_{\phi \in Lip_1([-\delta,\delta])} \left\{ \left| \int_{\mathbb{R}} \phi(x) (\Pi_\delta\mathcal{P}_{\mathbf{n}}\left(\omega\right)(\mathrm{d}x) - \Pi_\delta\mathcal{P}(\mathrm{d}x)) \right| \right\},$$

where $Lip_1\left([-\delta,\delta]\right)$ denotes the set of 1-Lipschitz continuous function on $[-\delta,\delta]$. Since, $[-\delta,\delta]$ is compact, $Lip_1\left([-\delta,\delta]\right)$ is separable endowed with the uniform topology. This implies the existence of a countable family $(f_k)_{k\geq 1}$ which is dense. Using again the method of the proof of Proposition 13, one can show

$$\forall\, \epsilon > 0,\ \exists\, A \in \mathbb{N},\ \forall\, u \in \mathbb{S}_A, \quad \mathbb{P}\left(\limsup_{N\to\infty} |\mathcal{P}_{\vec{u}_N} f_k - \mathcal{P} f_k| < \epsilon\right) = 1, \tag{21}$$

where $\mathcal{P}f$ denotes $\int_{\mathbb{R}} f(x)\mathcal{P}(dx)$. Now, the density of $(f_k)_{k\geq 1}$ entails that for any function f in $Lip_1\left([-\delta,\delta]\right)$, one can finds a function f_k such that $\|f_k - f\|_\infty < \varepsilon$, for any positive ε. Hence, (21) holds for any function in $\mathcal{C}_K$ on the same event. Moreover, since $\Pi_\delta\mathcal{P}_{\mathbf{n}}\left(\omega\right)$ and $\Pi_\delta\mathcal{P}$ are compactly supported measures,

$$d_W\left(\Pi_\delta\mathcal{P}_{\mathbf{n}}\left(\omega\right), \Pi_\delta\mathcal{P}\right) \leq C\sqrt{W_1\left(\Pi_\delta\mathcal{P}_{\mathbf{n}}\left(\omega\right), \Pi_\delta\mathcal{P}\right)},$$

which implies

$$\forall\, \epsilon > 0,\ \exists\, A \in \mathbb{N},\ \forall\, u \in \mathbb{S}_A, \quad \mathbb{P}\left(\limsup_{N\to\infty} d_W\left(\Pi_\delta\mathcal{P}_{\vec{u}_N}\left(\omega\right), \Pi_\delta\mathcal{P}\right) < \epsilon\right) = 1. \tag{22}$$

To end, using (19), (20) and (22) in (18) leads to the result. □

Finally, we get the following result in the *large forest* regime.

Proposition 17 *We have,*

$$\forall\, \epsilon > 0,\ \exists\, A \in \mathbb{N},\ \forall\, u \in \mathbb{S}_A, \quad \mathbb{P}\left(\limsup_{N\to\infty} \left| \widehat{\lambda}_W[\mathcal{F}_{\vec{u}_N}] - \frac{1}{\sigma} \right| < \epsilon \right) = 1.$$

Proof. By the Cauchy-Schwarz inequality, the convergence of this estimator is conditioned to the convergence of the Wasserstein distance in the following manner,

$$\begin{aligned} \left| \widehat{\lambda}_W[\mathcal{F}_{\mathbf{n}}] - \frac{1}{\sigma} \right| &= \frac{\left| \langle \widehat{F}^{-1} - \sigma^{-1} F_{\Lambda_\infty}^{-1}, F_{\Lambda_\infty}^{-1} \rangle \right|}{\| F_{\Lambda_\infty}^{-1} \|_2^2} \\ &\leq \frac{\left\| \widehat{F}^{-1} - \sigma^{-1} F_{\Lambda_\infty}^{-1} \right\|_2 \left\| F_{\Lambda_\infty}^{-1} \right\|_2}{\| F_{\Lambda_\infty}^{-1} \|_2^2} \\ &= \frac{d_W\left(\mathcal{P}_{\mathbf{n}}\,,\,\mathcal{P}\right)}{\| F_{\Lambda_\infty}^{-1} \|_2}. \end{aligned}$$

The result finally arises from the preceding Lemma concerning the convergence of the empirical measure in the sense of the Wasserstein distance. □

5 Numerical illustration

In this preliminary version, all the figures of this section have been deferred to the end of the article.

5.1 Simulation of conditioned Galton-Watson trees

In order to illustrate our estimation techniques on Galton-Watson forests, we need to make some numerical experiments. However, simulation of conditioned Galton-Watson trees is a difficult problem of independent importance. In this section, we briefly present an algorithm due to Devroye [7] allowing to achieve this aim. Given an integer n and a distribution μ on the set $\{0, \ldots, K\}$, this algorithm provides, in two steps, the simulation of the Łukasciewicz walk $\mathcal{L}[\tau_n]$ of a tree τ_n with distribution $\mathrm{GW}_n(\mu)$. Three more steps are required to obtain the corresponding Harris path $\mathcal{H}[\tau_n]$ through other coding processes (see for example [9]).

- ◇ **Simulation of numbers of children.** The multinomial distribution of parameters $(\mu_k)_{0\leq k\leq K}$ and n may be defined by its probability mass function,

$$\mathbb{P}(N_0 = n_0, \ldots, N_K = n_K) = \begin{cases} \dfrac{n!}{n_0! \ldots n_K!} \mu_0^{n_0} \ldots \mu_K^{n_K} & \text{if} \quad \displaystyle\sum_{k=0}^{K} n_k = n, \\ 0 & \text{else.} \end{cases}$$

  Simulation of the multinomial distribution presents no difficulty. By rejection sampling, we simulate multinomial random variables until obtaining a sequence $(N_k)_{0\leq k\leq K}$ satisfying

$$\sum_{k=0}^{K} k N_k = n - 1.$$

  We define the sequence $(\zeta_i)_{1\leq i\leq n}$ from

$$(\zeta_i)_{1\leq i\leq n} = (\underbrace{0, \ldots, 0}_{N_0}, \underbrace{1, \ldots, 1}_{N_1}, \ \ldots\ , \underbrace{K, \ldots, K}_{N_K}).$$

  Let $(\xi_i)_{1\leq i\leq n}$ be a sequence obtained as a random permutation of $(\zeta_i)_{1\leq i\leq n}$. A suitable technique for random shuffling is presented in [14, Algorithm P (p.139)]. The sequence $(\xi_i)_{1\leq i\leq n}$ represents the vertices's numbers of children in the depth-first search order.

◇ **Computation of Łukasciewicz walk.** Let L be the process defined by $L(0) = 0$ and,

$$\forall\, 0 \leq k \leq n-2, \quad L(k+1) = L(k) + \xi_{k+1} - 1.$$

Set $l = 1 + \arg\min\{L(k) \ : \ 0 \leq k \leq n-1\}$. Then there exists a tree τ_n with n nodes whose Łukasciecwicz walk is defined by

$$\mathcal{L}[\tau_n](k) = \begin{cases} L(l+k) + \min L - 1 & \text{if} \quad 0 \leq k \leq n-1-l, \\ L(k-n+l) + \min L - 1 & \text{if} \quad n-l \leq k \leq n-1. \end{cases}$$

◇ **From Łukasiewicz walk to height process.** Now, we compute the corresponding height process [9, eq.(2)],

$$\forall\, 0 \leq k \leq n-1, \quad \mathfrak{H}[\tau_n](k) = \#\left\{0 \leq j \leq k-1 \ : \ \mathcal{L}[\tau_n](j) = \min_{j \leq l \leq n} \mathcal{L}[\tau_n](l)\right\}.$$

◇ **From height process to contour process.** Let $(b_k)_{0 \leq k \leq n-1}$ be the sequence defined from $b_k = 2k - \mathfrak{H}[\tau_n](k)$ if $0 \leq k \leq n-1$ and $b_n = 2(n-1)$. Then the b_i's are sorted in increasing order. The contour process $\mathcal{C}[\tau_n](k)$ is defined for any $0 \leq k \leq 2n-2$ by [9, eq.(1)]

$$\mathcal{C}[\tau_n](k) = \begin{cases} \mathfrak{H}[\tau_n](i) - (k - b_i) & \text{if} \quad \exists\, 0 \leq i \leq n-2, \quad b_i \leq k < b_{i+1} - 1, \\ k - b_{i+1} + \mathfrak{H}[\tau_n](i+1) & \text{if} \quad \exists\, 0 \leq i \leq n-2, \quad b_{i+1} - 1 \leq k < b_{i+1}, \\ \mathfrak{H}[\tau_n](b_{n-1}) - (k - b_{n-1}) & \text{if} \quad b_{n-1} \leq k \leq b_n. \end{cases}$$

◇ **From contour process to Harris path.** The Harris path is only a small modification of the contour process, defined by $\mathcal{H}[\tau_n](0) = \mathcal{H}[\tau_n](2n) = 0$ and

$$\forall\, 1 \leq k \leq 2n-1, \ \mathcal{H}[\tau_n](k) = \mathcal{C}[\tau_n](k-1) + 1.$$

5.2 Inference for a forest of binary conditioned Galton-Watson trees

The aim of this section is to analyze the finite-sample behavior of both estimators introduced in this paper by means of numerical experiments. The theoretical study achieved in Section 4 shows that we can expect to obtain good numerical results, at least for large trees and/or a large forest. To this goal, we consider a forest of independent conditioned Galton-Watson trees with common critical birth distribution μ such that $\mu(k) = 0$ for $k \geq 3$. Such a distribution satisfies the following linear system of equations,

$$\begin{cases} \mu(0) + \mu(1) + \mu(2) &= 1 \\ \mu(1) + 2\mu(2) &= 1 \\ \mu(1) + 4\mu(2) - 1 &= \sigma^2 \end{cases}$$

which is equivalent to

$$\mu(0) = \mu(2) = \frac{\sigma^2}{2} \qquad \text{and} \qquad \mu(1) = 1 - \sigma^2.$$

In other words, μ is entirely characterized by its variance σ^2. Simulations of Galton-Watson trees $\mathrm{GW}_n(\mu)$ are performed with the method provided in Subsection 5.1.

Let $\mathcal{F} = (\tau^i)_{1 \leq i \leq N}$ be a forest of N independent trees such that, for any $1 \leq i \leq N$, $\tau^i \sim \mathrm{GW}_{n_i}(\mu)$ for some integer n_i. From the Harris process of each tree τ^i, one first computes the quantity

$$\widehat{\lambda}\left[\tau^i\right] \quad = \quad \frac{\langle \mathcal{H}[\tau^i](2n_i \cdot), E\rangle}{2\sqrt{n_i}\|E\|_2^2},$$

where E is known and defined in (2). Then, we propose to estimate σ^{-1} in the two following ways, where $(\widehat{\lambda}[\tau^{(i)}])_{1 \leq i \leq N}$ denotes the order statistic associated to the family $(\widehat{\lambda}[\tau^i])_{1 \leq i \leq N}$.

Least Squares	Wasserstein
$\widehat{\lambda}_{ls}[\mathcal{F}] = \frac{1}{N}\sum_{i=1}^{N} \widehat{\lambda}\left[\tau^i\right]$	$\widehat{\lambda}_{W}[\mathcal{F}] = \frac{1}{\|F^{-1}_{\Lambda_\infty}\|_2^2}\sum_{i=1}^{N} \widehat{\lambda}\left[\tau^{(i)}\right] \int_{\frac{i-1}{N}}^{\frac{i}{N}} F^{-1}_{\Lambda_\infty}(s)\mathrm{d}s$

Remark 18 *In order to compute $\widehat{\lambda}_W[\mathcal{F}]$, we need to be able to perform computations using the function $F^{-1}_{\Lambda_\infty}$. Unfortunately, in view of the theoretical study of Λ_∞ made in Proposition 3.1, one can not expect to have an explicit expression for this function. In the following of this section, we use a numerical estimation of $F^{-1}_{\Lambda_\infty}$ by Monte Carlo simulations. To achieve this goal, we perform simulations of Λ_∞ thanks to formula (5) by simulating Brownian excursions thanks to (3). In order to ensure that the error made on $F^{-1}_{\Lambda_\infty}$ does not propagate too much in our results, $F^{-1}_{\Lambda_\infty}$ is estimated from one million simulations of Λ_∞.*

The theoretical investigations of Section 4 establish that our estimators are unbiased in the *large trees* regime $\mathfrak{m}(\mathbf{n}) \to \infty$. Nevertheless, the problem is not as simple when working with finite trees. A clear illustration of this comes from the numerical evaluations of the average Harris processes of finite trees. Indeed, the numerical study of Figure 4 shows that the average Harris processes of small trees seem to be lower than the limiting Harris process. Hence, the quantities $\widehat{\lambda}[\tau^i]$ are expected to underestimate the target σ^{-1}. But any estimator based on the asymptotic behavior of conditioned Galton-Watson trees is expected to present such a bias. In particular, we state in our numerical experiments that the estimator proposed in [3] presents the same bias.

The natural question arising from the preceding comments is: how is the bias of a conditioned Galton-Watson tree related to its size and/or the unknown parameter σ ? The numerical study presented in Figure 5 shows that the quantity $\eta(n) = \sigma^{-1}\mathbb{E}[\widehat{\lambda}[\tau_n]]^{-1}$, where $\tau_n \sim \mathrm{GW}_n(\mu)$, seems close to be uncorrelated to σ at least when σ is large enough. This allows us to construct a bias corrector independent on the unknown standard deviation σ. In addition, the dependency on n may be modeled by the relation $\eta(n) = 1 - (a\sqrt{n} + b)^{-1}$. The coefficients appearing in η may be estimated from simulated data,

$$\widehat{\eta}(n) = 1 - (0.504273\sqrt{n} + 0.9754839)^{-1}$$

(see Figure 5 again). The correction is obviously expected to be better for large values of σ. Finally, we construct the following corrected versions of the estimators $\widehat{\lambda}_{ls}[\mathcal{F}]$ and $\widehat{\lambda}_W[\mathcal{F}]$.

Corrected Least Squares	Corrected Wasserstein
$\widehat{\lambda}^c_{ls}[\mathcal{F}] = \frac{1}{N}\sum_{i=1}^{N} \widehat{\eta}(\#\tau^i)\widehat{\lambda}\left[\tau^i\right]$	$\widehat{\lambda}^c_{W}[\mathcal{F}] = \frac{1}{\|F^{-1}_{\Lambda_\infty}\|_2^2}\sum_{i=1}^{N} \widehat{\eta}\left(\#\tau^{(i)}\right) \widehat{\lambda}\left[\tau^{(i)}\right] \int_{\frac{i-1}{N}}^{\frac{i}{N}} F^{-1}_{\Lambda_\infty}(s)\mathrm{d}s$

In light of the previous comments, computing the estimators proposed in this paper is not an easy task. According to Remark 18, this needs to perform an important number of simulations of Λ_∞ in order to get an accurate approximation of $F^{-1}_{\Lambda_\infty}$. Moreover, to be able to correct the aforementioned bias, one needs to perform many simulations of finite trees. Together with this work, we propose a `Matlab` toolbox which already includes these preliminary computations and allows to directly and quickly compute our estimators for forests. This toolbox as well as its documentation and the scripts used in this paper are available at the webpage: `http://agh.gforge.inria.fr`.

For improved comparability, we also compute the estimator $\widehat{\lambda}_{un}[\mathcal{F}]$ of σ^{-1} based on the work [3] (see Subsection 3.1) given by

$$\widehat{\lambda}_{un}[\mathcal{F}] = \frac{1}{N}\sum_{i=1}^{N} \widehat{\delta}[\tau^i],$$

where $\widehat{\delta}[\tau^i]$ is defined (see equations (7) and (8)) from a node v randomly chosen in τ^i by

$$\widehat{\delta}[\tau^i] = \frac{\sqrt{2}\,h(v)}{\sqrt{\pi \# \tau^i}}.$$

The estimator $\widehat{\lambda}_{un}[\mathcal{F}]$ is expected to present the bias due to the approximation of Harris paths by their expected limit. We correct it by the aforementioned method,

$$\widehat{\lambda}^{c}_{un}[\mathcal{F}] = \frac{1}{N}\sum_{i=1}^{N} \widehat{\eta}(\#\tau^i)\widehat{\delta}[\tau^i].$$

In Figures 6, 7 and 8, estimators $\widehat{\lambda}_{ls}[\mathcal{F}]$, $\widehat{\lambda}_{W}[\mathcal{F}]$ and $\widehat{\lambda}_{un}[\mathcal{F}]$ are denoted by "LSE", "Wasserstein" and "Uniform node" (or "UN" in short), respectively.

The study of Figure 6 shows that for values of σ greater than 0.5, the bias correction works properly. Moreover, it also shows that the approach developed in [3] presents the same kind of bias as ours. In the case of small parameter σ, the bias correction is not as accurate. This was expected because the bias corrector does not fit as well to the bias curve for small values of sigma as it does for greater values of σ.

Since we have an estimation procedure which seems to work, the natural further study is to see how the quality of our estimators varies as the characteristics of the forest change. We begin by looking at the variations when the sizes of the trees increase. A priori, the sizes of the trees in the considered forest should not have influence on the dispersion of the estimators. Indeed, our estimation strategy is based on the approximation of the Harris path of a finite tree by its limit. As a consequence, the size parameter only governs the quality of this approximation. Whatever the sizes of the trees, the dispersion will be given by the variance of the limit distribution Λ_∞. As expected Figure 7 shows that the dispersion of the estimators does not change as the sizes of the trees change when σ takes great values. Similarly, as shown in Figure 8, for small values of σ, the sizes of the trees do not influence the dispersion of the estimator. However, Figure 8 also shows that the sizes of the trees have a positive influence on the bias of the estimators.
Finally, Figure 9 shows the variation of the dispersion of the least-square estimator as the size of the forest changes. It appears to be consistent with the theoretical tolerence intervals given by the central limit theorem. Similar results have been obtained from the Wasserstein method (see Figure 10).

5.3 Real data analysis: history of Wikipedia webpages

The aim of this section is to show that the methodology developed in this paper can be used to analyze the history of some real hierchical data. More precisely, we focus on the evolution over time of a given webpage on the World Wide Web. `HTML` is the standard markup language for creating webpages. Documents encoded in a markup language naturally presents a tree structure: the area delimited by opening and closing tags represents a node of the tree; the children of this node are given by the tags directly found in this area in the order they appear (see Figure 11 for an example of `HTML` document and the corresponding ordered tree structure). It should be noted that the ordered tree representing an `HTML` document does not take into account the text between tags but only the hierchical structure.

Wikipedia is probably the most famous free Internet encyclopedia. It allows its users to create and edit almost any article. All past changes are listed in reverse-chronological order and are accessible from the current version of the Wikipedia webpage. Consequently, each article forms a time series composed of hundreds of revisions. The analysis of this chronological dataset is difficult because of the complex structure of the data which has no representation in an Euclidean state space. We propose to apply the strategy presented in this paper to investigate this question and obtain informations on the history of articles. Wikipedia webpages are obviously not conditioned Galton-Watson trees, but they share the same structure with a typical layout. Thus articles at hand might not be differentiated by considering their shape but some scale parameter, as it is the case for conditioned Galton-Watson trees. We claim that the quantity $\widehat{\lambda}[\tau]$, where τ is the underlying tree of a given webpage, is a good estimate of its relative scale and may be used to represent the revision history.

We begin with the English version of the Wikipedia article *Gravitational wave*[2]. This article has been edited 2001 times by 810 Wikipedians since its creation on September 3rd 2001 (information acquired on August

[2]Wikipedia article *Gravitational wave*: https://en.wikipedia.org/wiki/Gravitational_wave

11 2016). For each month since January 2005, we compute $\widehat{\lambda}_{ls}$ from the forest of the versions revised during this month. If no revision has been found during this period, $\widehat{\lambda}_{ls}$ is equal to the estimate of the previous month, and recursively. Figure 12 displays the evolution of $\widehat{\lambda}_{ls}$ over time. First, we remark two spikes (a), negative in May 2007, and (b), positive in May 2016. Both these spikes correspond to massive vandalism of the article on May 7 2007 (addition of 720 pointless sections with random text) and May 23 2016 (complete deletion of the article) by malicious people. Indeed, if we do not consider these two vandalized webpages in our estimation, we obtain the graph of Figure 13 (left) that has no spikes. In addition, we observe in Figure 13 (left) that the time series of $\widehat{\lambda}_{ls}$ has roughly two regimes (c) and (d). The first period (c) corresponds to the "running in" required to find the adequate structure of the article. In this period, the webpage is subject to major changes that are most often additions of new sections or paragraphs but may be deletions of inappropriate content. When a good structure arises, the webpage is then slowly broadened during the second regime (d). It should be remarked in Figure 13 (right) that two important modifications occur in the period (d): (e) between July 2013 and April 2014 and (f) in February 2016. The (e) period is related to major changes in the webpage (mainly addition of references and reorganization of some sections) especially following advances in this field. The second event (f) corresponds to extensive adding following the announce of the first observation of gravitational waves using the Advance LIGO detectors.

We perform the same methodology on the history of the Wikipedia article *Chocolate*[3] (see Figure 14). This article has been edited 6332 times by 3105 Wikipedians since its creation on November 13 2001 (information acquired on August 11 2016). All the spikes observed on the graph of Figure 14 correspond to acts of vandalism (deletion of substantial content). For the sake of example, we highlight two major events (a) (in May 2008) and (b) (in June 2010) occuring during the "running in" period (c): (a) corresponds to important additions in the article (sections *Etymology*, *Holydays* and *Manufacturers* have been added), while (b) is related to the creation of the parallel article *Health effects of chocolate* leading to deletion of the corresponding sections in the main article.

For both examples, we empirically observe that, when $\widehat{\lambda}_{ls}$ decreases, some content has been added to the webpage, and conversely, when $\widehat{\lambda}_{ls}$ increases, some parts of the article have been removed. Our analysis shows that, starting from their creation, these Wikipedia articles are broadened over time after a long "runing in" period used to unconsciously find the adequate structure. One may also detect vandalism on Wikipedia articles by identifying spikes a posteriori. Vandalism is usually removed by dedicated individuals who patrol Wikipedia webpages, but this is an onerous task with a rate of 10 edits per second[4] and around 7% of edits have been estimated to be vandalism [21]. Vandalism detection is often based on a combination of various indicating features [1, 18]. Our algorithm might be used as a new feature for identifying acts of vandalism on the structure of the article.

References

[1] Adler, B. T., de Alfaro, L., Mola-Velasco, S. M., Rosso, P., and West, A. G. *Wikipedia Vandalism Detection: Combining Natural Language, Metadata, and Reputation Features.* Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 277–288.

[2] Aldous, D. The Continuum Random Tree III. *Ann. Probab. 21*, 1 (01 1993), 248–289.

[3] Bharath, K., Kambadur, P., Dey, D., Arvin, R., and Baladandayuthapani, V. Inference for large tree-structured data. *Preprint arXiv:1404.2910* (2014).

[4] Bobkov, S., and Ledoux, M. One-dimensional empirical measures, order statistics and Kantorovich transport distances. *Preprint* (2014).

[3] Wikipedia article *Chocolate*: https://en.wikipedia.org/wiki/Chocolate

[4] Wikipedia statistics (last consulted on August 11 2016): https://en.wikipedia.org/wiki/Wikipedia:Statistics

[5] Czado, C., and Munk, A. Nonparametric validation of similar distributions and assessment of goodness of fit. *Journal of the Royal Statistical Society: Series B 60*, 1 (1998), 223–241.

[6] David, H. A., and Nagaraja, H. N. *Order statistics*, third ed. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2003.

[7] Devroye, L. Simulating size-constrained Galton-Watson trees. *SIAM Journal on Computing 41*, 1 (2012), 1–11.

[8] Drmota, M., and Marckert, J.-F. Reinforced weak convergence of stochastic processes. *Statist. Probab. Lett. 71*, 3 (2005), 283–294.

[9] Duquesne, T. A limit theorem for the contour process of conditioned Galton-Watson trees. *Ann. Probab. 31*, 2 (04 2003), 996–1027.

[10] Gallón, S., Loubes, J.-M., and Maza, E. Statistical properties of the quantile normalization method for density curve alignment. *Mathematical Biosciences vol. 242*, 2 (Apr. 2013), pp. 129–142.

[11] Janson, S. Brownian excursion area, Wright's constants in graph enumeration, and other Brownian areas. *Probability Surveys 4* (2007), 80–145.

[12] Janson, S. Simply generated trees, conditioned Galton-Watson trees, random allocations and condensation. *Probab. Surveys 9* (2012), 103–252.

[13] Kantorovitch, L. V. Sur certains développements suivant les polynômes de la forme de S. Bernstein. *C. R. Acad. Sci. URSS* (1930).

[14] Knuth, D. Seminumerical algorithms, 2nd edn, vol. 2 of the art of computer programming, 1981.

[15] Lorentz, G. G. G. *Bernstein polynomials.* Mathematical Expositions, no. 8. University of Toronto Press, Toronto, 1953.

[16] Louchard, G. Kac's formula, Levy's local time and Brownian excursion. *J. Appl. Probab. 21*, 3 (1984), 479–499.

[17] Louchard, G., and Janson, S. Tail estimates for the Brownian excursion area and other Brownian areas. *Electron. J. Probab. 12* (2007), no. 58, 1600–1632.

[18] Mola-Velasco, S. Wikipedia vandalism detection through machine learning: Feature review and new proposals. In *CLEF 2010 LABs and Workshops, Notebook papers* (September 2010).

[19] Nualart, D. *The Malliavin calculus and related topics*, second ed. Probability and its Applications (New York). Springer-Verlag, Berlin, 2006.

[20] Pitman, J. *Combinatorial stochastic processes*, vol. 32. Springer Science & Business Media, 2006.

[21] Potthast, M. Crowdsourcing a Wikipedia vandalism corpus. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2010), SIGIR '10, ACM, pp. 789–790.

[22] Revuz, D., and Yor, M. *Continuous Martingales and Brownian Motion.* Grundlehren der mathematischen Wissenchaften A series of comprehensive studies in mathematics. Springer, 1999.

[23] Viégas, F. B., Wattenberg, M., and Dave, K. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2004), CHI '04, ACM, pp. 575–582.

[24] Viegas, F. B., Wattenberg, M., Kriss, J., and van Ham, F. Talk before you type: Coordination in Wikipedia. In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences* (Washington, DC, USA, 2007), HICSS '07, IEEE Computer Society, pp. 78–.

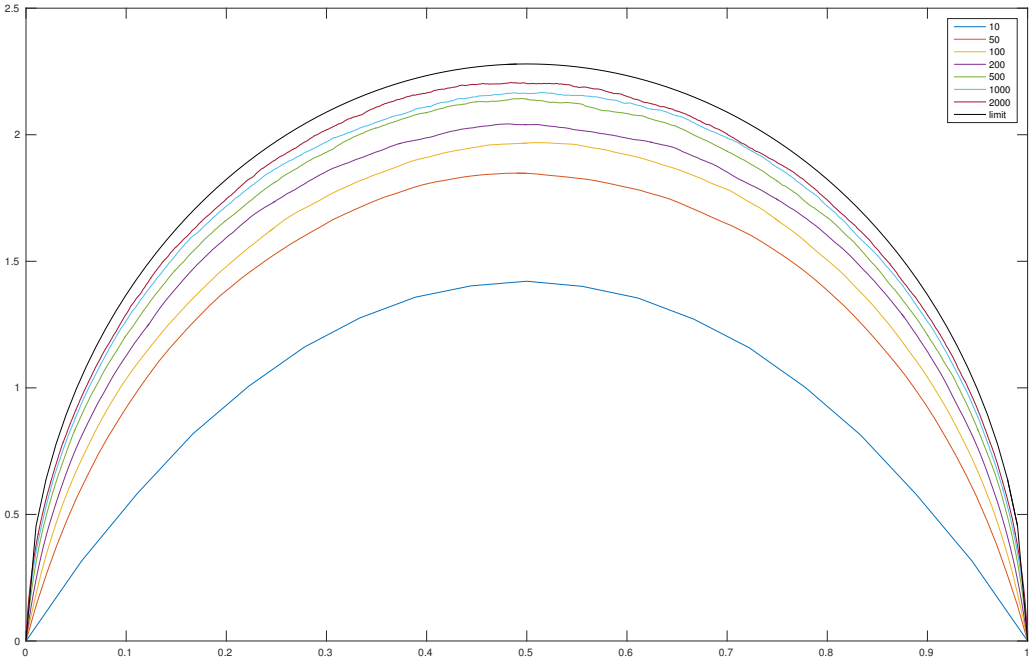

Figure 4: Estimated mean Harris processes of binary conditioned Galton-Watson trees with size n and $\sigma = 0.7$ calculated from 2000 trees for each value of n.

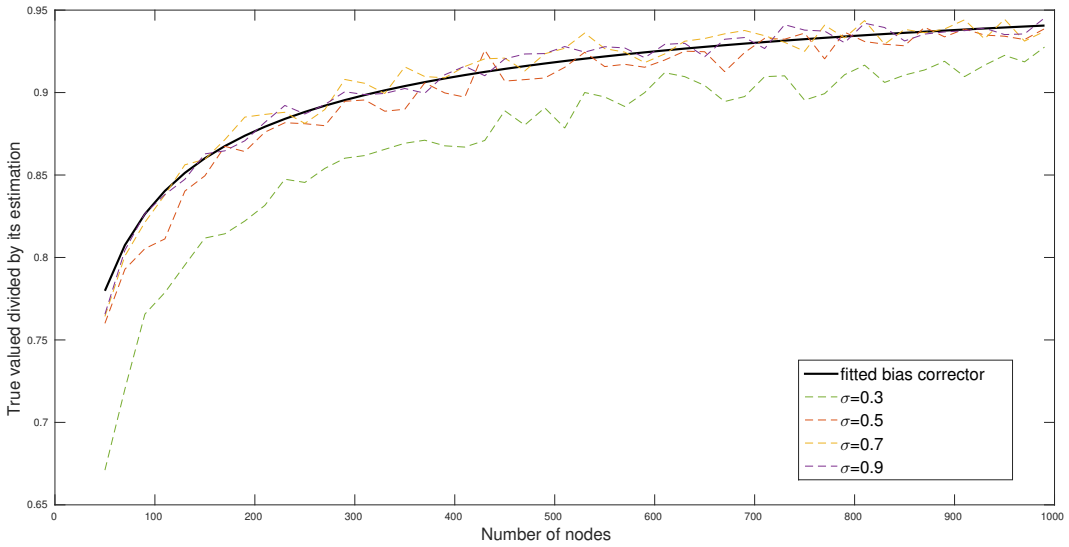

Figure 5: Estimation of the quantity $\eta(n) = \sigma^{-1}\mathbb{E}[\widehat{\lambda}[\tau_n]]^{-1}$, where $\tau_n \sim \mathrm{GW}_n(\mu)$, for different values of σ and different numbers of nodes n, together with the fitted bias corrector function $\widehat{\eta}$. Estimations have been made by Monte Carlo method with samples of 2000 trees for each couple (n, σ).

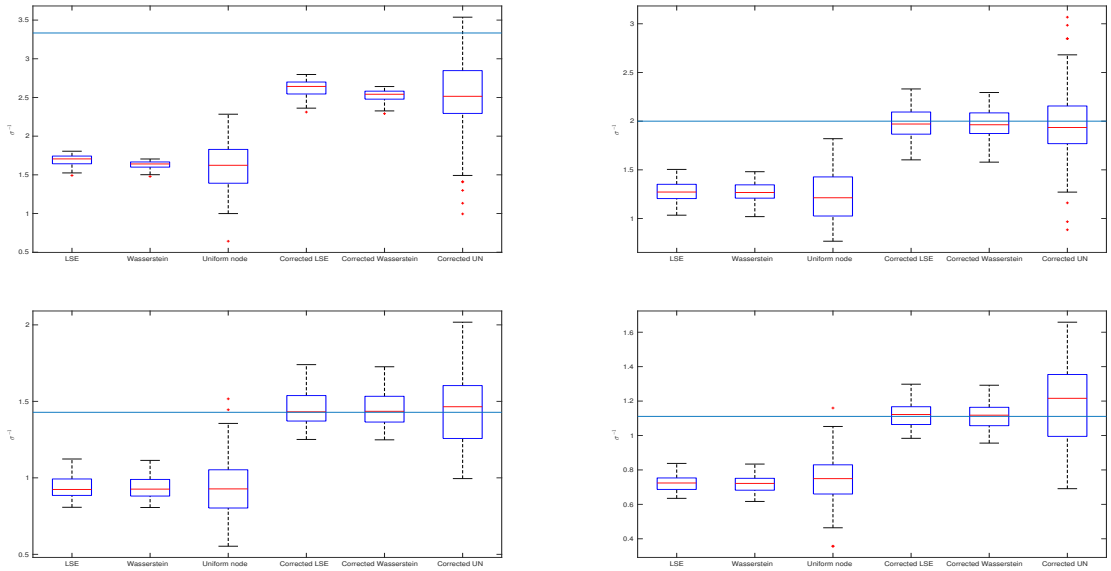

Figure 6: Estimation and bias correction for forests of 10 trees with 20 nodes for σ equals to 0.3 (top, left) 0.5 (top, right), 0.7 (bottom, left) and 0.9 (bottom right). Boxplots have been drawn from 100 replicates each.

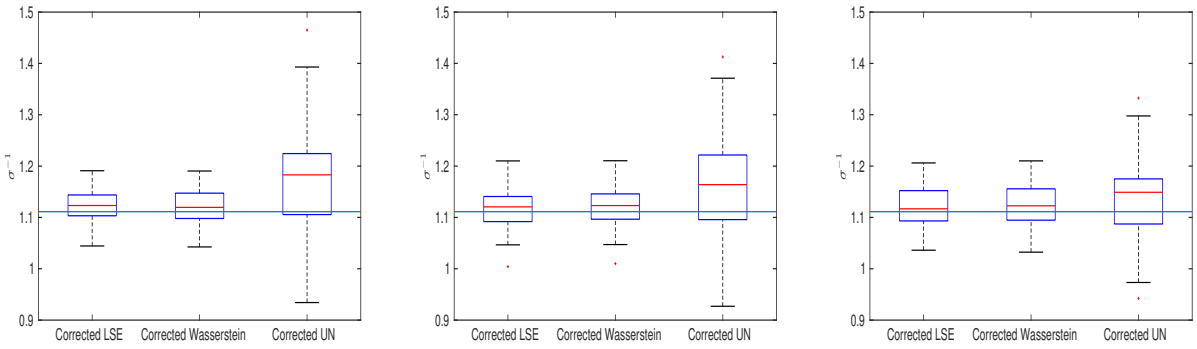

Figure 7: Influence of the size of the trees for σ equals to 0.9: tree sizes varying from 20 nodes (left), 50 nodes (center), to 100 nodes (right). Forests of 50 trees. Boxplots have been drawn from 100 replicates each.

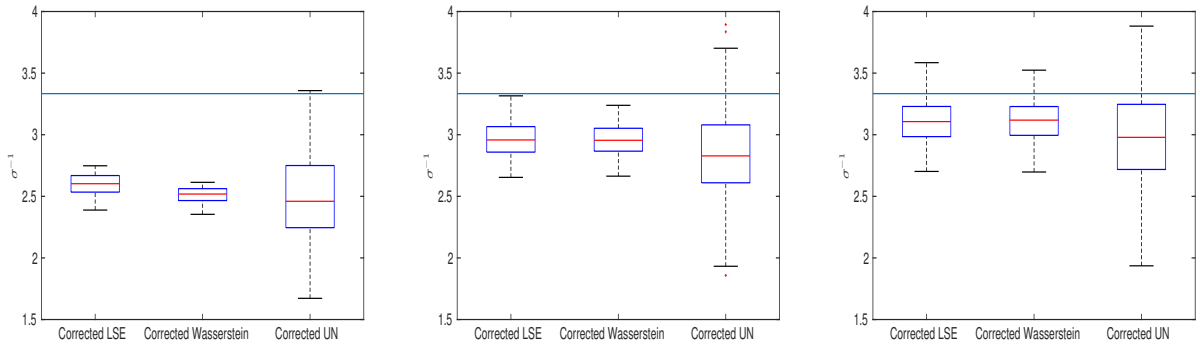

Figure 8: Influence of the size of the trees for σ equals to 0.3: tree sizes varying from 20 nodes (left), 50 nodes (center), to 100 nodes (right). Forests of 50 trees. Boxplots have been drawn from 100 replicates each.

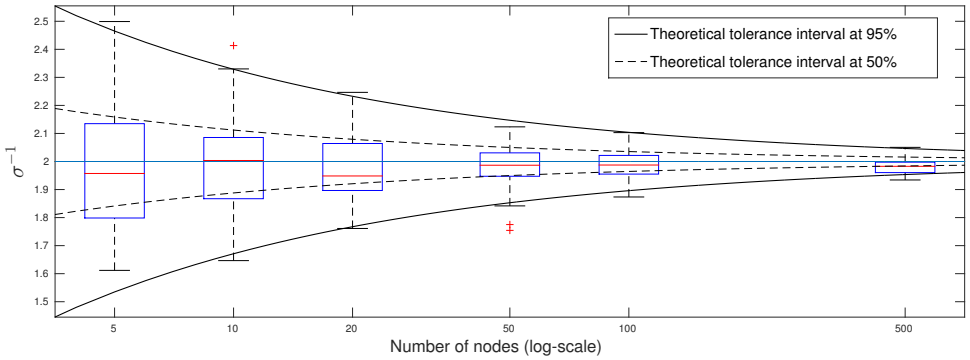

Figure 9: Least-square estimation of σ^{-1} for different sizes of forests ($\sigma = 0.5$ with trees of size 20). Boxplots have been drawn from 100 replicates each.

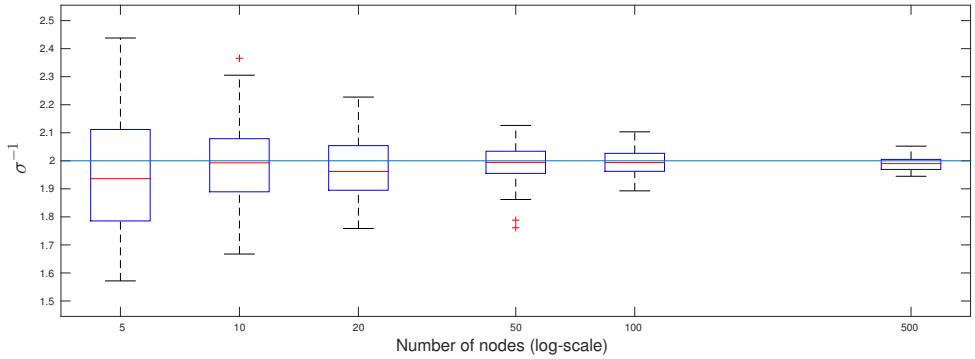

Figure 10: Wasserstein estimation of σ^{-1} for different sizes of forests ($\sigma = 0.5$ with trees of size 20). Boxplots have been drawn from 100 replicates each.

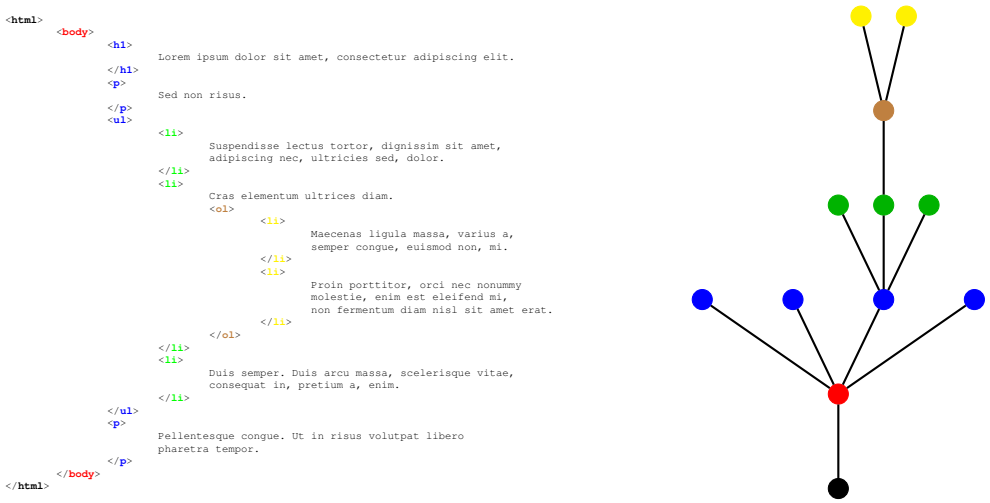

Figure 11: Underlying ordered tree structure (right) present in an `HTML` document (left). Each level in the tree is colored in the same way as the corresponding tags in the document. Natural order from top to bottom in the `HTML` document corresponds to left-to-right order in the tree.

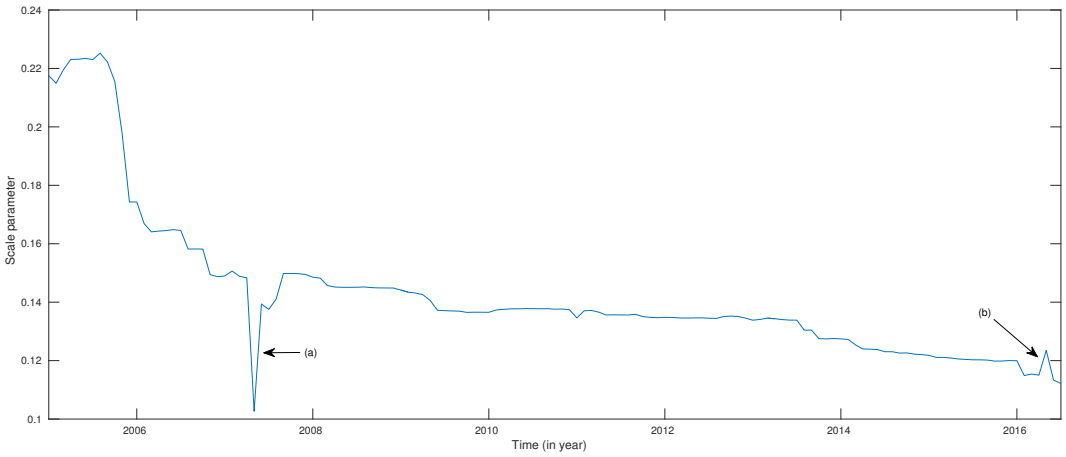

Figure 12: History of $\widehat{\lambda}_{ls}$ between January 2005 and June 2016 for the Wikipedia article *Gravitational wave*. Events (a) and (b) are related to vandalism.

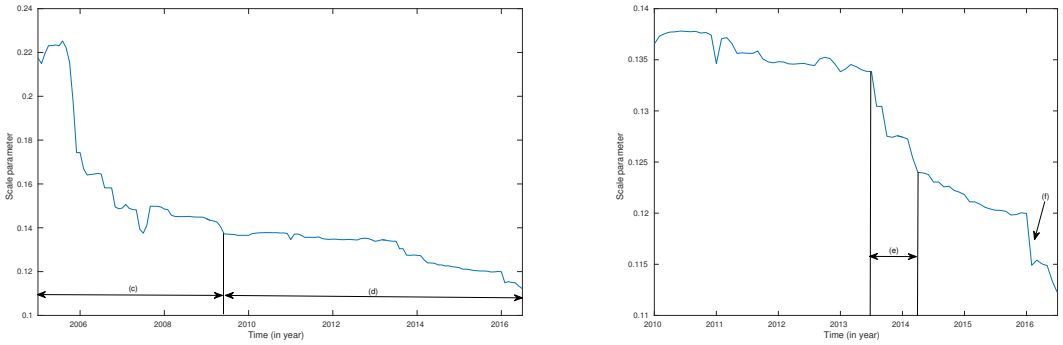

Figure 13: History of $\widehat{\lambda}_{ls}$ for the Wikipedia article *Gravitational wave* without taking into account the two vandalism pages related to (a) and (b) between 2005 and 2016 (left) and 2010 and 2016 (right).

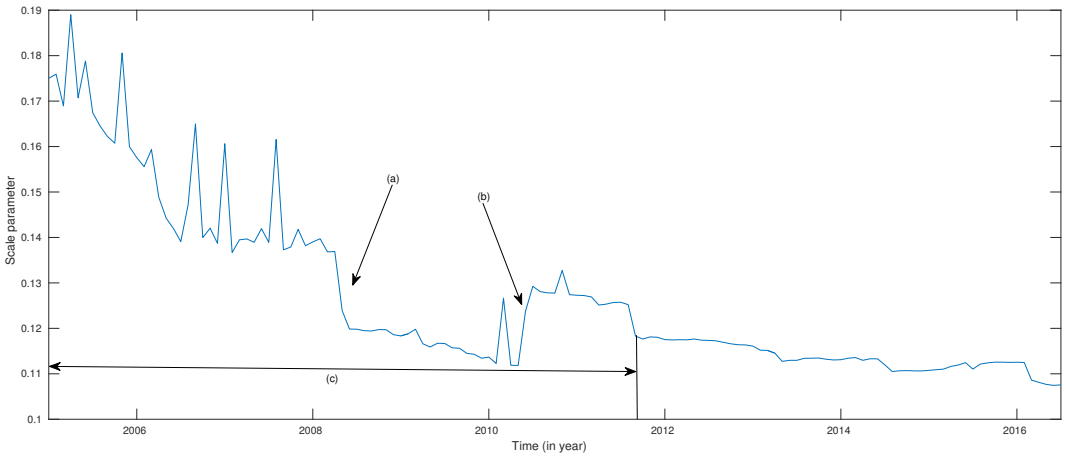

Figure 14: History of $\widehat{\lambda}_{ls}$ between January 2005 and June 2016 for the Wikipedia article *Chocolate*. All spikes are related to vandalism.