APPLICATIONES MATHEMATICAE
47,1 (2020), pp. 29–44

Krzysztof Bartoszek (Linköping)

LIMIT DISTRIBUTION OF THE QUARTET BALANCE INDEX FOR ALDOUS'S ($\beta \geq 0$)-MODEL

Abstract. This paper builds on T. Martínez-Coronado, A. Mir, F. Rosselló and G. Valiente's 2018 work, introducing a new balance index for trees. We show that this balance index, in the case of Aldous's ($\beta \geq 0$)-model, converges weakly to a distribution that can be characterized as the fixed point of a contraction operator on a class of distributions.

1. Introduction. Phylogenetic trees (connected graphs without cycles that have a distinguished node, called "root", interpreted as the "start" of the tree) are key to evolutionary biology. However, they are not easy to summarize or compare as it may not be obvious how to tackle their topologies, understood as the internal branching structure. Therefore, many summary indices have been proposed in order to "project" a tree into $\mathbb{R}$. Such indices aim to quantify some property of the tree, and one of the most studied properties is the symmetry of the tree. Tree symmetry is commonly captured by a balance index. Multiple balance indices have been proposed, including Sackin's [S72], Colless' [C82] or the total cophenetic index [MRR]. A compact introduction to phylogenetics, containing in particular a list of tree asymmetry measures, can be found in [F04, pp. 562–564]. The present work concerns a newly proposed balance index, the quartet index QI [MCMRV].

One of the reasons for introducing summary indices for trees is to use them for significance testing—whether the tree comes from a given probabilistic model. Obtaining the distribution (for a given number n of contem-

2020 *Mathematics Subject Classification*: Primary 05C80, 60F05; Secondary 60J85, 62P10, 92B10, 92D15.

Key words and phrases: balance index, contraction method, phylogenetic tree, tree shape, weak convergence.

Received 10 January 2019; revised 10 May 2019.

Published online 29 July 2019.

DOI: 10.4064/am2385-6-2019

porary species, i.e. leaves of the tree, or in the limit as $n \to \infty$) of indices is usually difficult and is often done only for the "simplest" Yule (pure-birth [Y24]) tree case and sometimes for the uniform model (see e.g. [A91, SM01]).

Using the contraction method, central limit theorems were found for various balance indices, like the total cophenetic index (Yule model case [B18]) and jointly for Sackin's and Colless' indices (in the Yule and uniform model cases [BFJ06]). Furthermore, in [BF06] it was shown that Sackin's index has the same weak limit as the number of comparisons of the Quicksort algorithm [H62], both after normalization of course.

In [CF10] the number of occurrences of patterns in a tree are considered, where a pattern is understood as "any subset of the set of all phylogenetic trees of fixed size k". For a tree with n leaves such a pattern will satisfy the recursion

$$X_{n,k} \overset{\mathcal{D}}{=} X_{L_n,k} + X^*_{n-L_n,k}$$

where $X_{n,k}$, $X^*_{n,k}$ and L_n are independent, $X_{n,k} \overset{\mathcal{D}}{=} X^*_{n,k}$ and L_n is the size of the left subtree branching from the root. For the Yule and uniform models the authors of [CF10] derived central limit theorems (normal limit distribution) with Berry–Esseen bounds and Poisson approximations in the total variation distance. The above description is rather abstract but can be restated in a more direct way. The term n is the number of leaves of the tree (i.e. nodes of degree 1). The pattern of fixed size k is a generic term, but in [CF10, Table 1] concrete examples are given: k-pronged nodes, k-caterpillars, or nodes with minimal clade size k.

In the present manuscript we will consider the number of fully balanced subtrees with $k = 4$ leaf nodes. However, in our case the recursion will be of a non-homogeneous form, hence the results from [CF10] do not carry over. The random variable $X_{n,k}$ is the number of occurrences of the given pattern (of size k) in a tree of size n. In principle the index k could be dropped at this description level, but we kept it here for consistency with [CF10].

Even though the pure-birth model seems to be very widespread in the phylogenetics community, more complex models need to be studied, especially in the context of tree balance. From [RS13, Lemma 4] it can be deduced that Yule trees have to be rather balanced—as the maximum quartet weight (the maximum of the number of randomly placed marks along branches over induced subtrees on four leaves) is asymptotically proportional to the expectation of the tree's height.

In this work, using the contraction method, we show convergence in law of the (scaled and centred) quartet index and derive a representation (as a fixed point of a particular contraction operator) of the weak limit. Remarkably, this is possible not only for the Yule tree case but also for Aldous's more general β-model (in the $\beta \geq 0$ regime).

The paper is organized as follows. In Section 2 we introduce Aldous's β-model and the quartet index. In Section 3 we prove our main result, Theorem 3.1, via the contraction method. When studying the limit behaviour of recursive-type indices for pure-birth binary trees one finds that for each internal node the leaves inside its clade are uniformly split into subclades as the node splits. However, in Aldous's β-model this is not the case, the split is according to a BetaBinomial distribution, and a much finer analysis is required to show weak convergence, as $n \to \infty$, of the recursive-type index to the fixed point of the appropriate contraction. Theorem 3.1 is not specific to the quartet index but covers a more general class of models, where each internal node split divides its leaf descendants according to a BetaBinomial distribution (with $\beta \geq 0$). In Section 4 we apply Theorem 3.1 to the quartet index and characterize its weak limit. In Section 5 we illustrate the results with simulations. Finally, in the Appendix we provide the R code used to simulate from this weak limit.

2. Preliminaries

2.1. Aldous's β-model for phylogenetic trees. Birth-death models are popular choices for modelling the evolution of phylogenetic trees. However, in [A96, A01] a different class of models was proposed, the so-called β-model for binary phylogenetic trees.

The main idea behind this model is to consider a (suitable) family $\{q_n\}_{n=2}^{\infty}$ of symmetric, $q_n(i) = q_n(n-i)$, probability distributions on the natural numbers. In particular $q_n : \{1,\dots,n-1\} \to [0,1]$. The tree grows in a natural way. The root node of an n-leaf tree defines a partition of the n nodes into two sets of sizes i and $n-i$ ($i \in \{1,\dots,n-1\}$). We randomly choose the number of leaves of the left subtree, $L_n = i$, according to the distribution q_n and this induces the number of leaves, $n - L_n$, in the right subtree. We then repeat this recursively in the left and right subtrees, i.e. splitting according to the distributions q_{L_n} and q_{n-L_n} respectively. Notice that due to q_n's symmetry the terms left and right do not have any particular meaning attached.

In [A96] a one-parameter, $-2 \leq \beta \leq \infty$, family of probability distributions was proposed:

$$q_n(i) = \frac{1}{a_n(\beta)} \frac{\Gamma(\beta+i)\Gamma(\beta+n-i)}{\Gamma(i)\Gamma(n-i)}, \qquad 1 \leq i \leq n-1, \tag{2.1}$$

where $a_n(\beta)$ is the normalizing constant and $\Gamma(\cdot)$ the Gamma function. We may actually recognize this as the BetaBinomial$(n-2, \beta+1, \beta+1)$ distribution and write

$$q_n(i) = B(\beta+1,\beta+1)^{-1} \int_0^1 \left(\binom{n-2}{i-1} \tau^{i-1}(1-\tau)^{n-i-1} \right) \tau^{\beta}(1-\tau)^{\beta}\, \mathrm{d}\tau, \tag{2.2}$$

where $B(a,b)$ is the Beta function with parameters a and b. Notice that we changed n to $n-2$ and i to $i-1$ on the right side of the equations with respect to [A96] in order to have better correspondence with the rest of our paper. Writing informally, from the probability distribution function (2.2), we can see that if we condition under the integral on τ, then we obtain a binomially distributed random variable. This observation is the key intuition behind the analysis presented here.

Particular values of β correspond to some well known models. The uniform tree model is represented by $\beta = -3/2$, and the pure-birth Yule model by $\beta = 0$. The limit case of $\beta = \infty$ is $q_n(i) \to \binom{n-2}{i-1}2^{-(n-2)}$, i.e. the binomial distribution, with success probability 0.5. This corresponds to the so-called "symmetric binary tree" in the computer science literature (e.g. [M92, Ch. 5.3]) and was mentioned as the "random partition tree" in the evolutionary biology literature [MS91].

Of particular importance to our work is the limiting behaviour of the scaled size of the left (and hence right) subtree, $n^{-1}L_n$. Lemma 3 in [A96] characterizes these asymptotics.

Lemma 2.1 ([A96, Lemma 3 for $\beta > -1$]).

(1) *For $\beta = \infty$, $n^{-1}L_n \xrightarrow{\mathcal{D}} 1/2$.*

(2) *For $-1 < \beta < \infty$, $n^{-1}L_n \xrightarrow{\mathcal{D}} \tau_\beta$, where τ_β has the Beta distribution*

$$f(x) = \frac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)}x^\beta(1-x)^\beta, \qquad 0 < x < 1. \tag{2.3}$$

2.2. Quartet index. In [MCMRV] a new type of balance index was proposed for discrete (i.e. without branch lengths, or in the language of graph theory, without weights assigned to branches) phylogenetic trees—the quartet index. This index is based on considering the number of so-called quartets of each type made up by the leaves of the tree. A (rooted) *quartet* is the induced subtree (a subtree formed by removing all but some given set of leaves and then removing all degree two nodes except the root) from choosing some four leaves. We should make a point here about the nomenclature. Usually in the phylogenetic literature a quartet is an unrooted tree on four leaves (e.g. [SS03]). However, here we consider rooted trees and following [MCMRV] by a (rooted) quartet we mean a rooted tree on four leaves. From now on we will write just "quartet", dropping the "rooted" qualification.

For a given tree T, let $\mathcal{P}_4(T)$ be the set of quartets of T. Then the *quartet index* of T is defined as

$$\mathrm{QI}(T) = \sum_{\mathcal{P}_4(T)} \mathrm{QI}(Q), \tag{2.4}$$

where $\mathrm{QI}(Q)$ assigns a predefined value to a specific quartet (i.e. a given tree

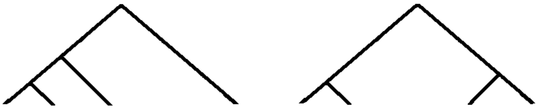

Fig. 1. The two possible rooted quartets for a binary tree. Left: K_4, the four leaf rooted caterpillar tree (also known as a comb or pectinate tree); right: B_4, the fully balanced tree on four leaves (also known as a fork, see e.g. [CS07] for nomenclature).

topology on four leaves). When the tree is a binary one (as here) there are only two possible topologies on four leaves (see Fig. 1). Following [MCMRV, Table 1], we assign the value 0 to K_4 quartets and 1 to B_4 quartets. Therefore, the QI for a binary tree (QIB) will be

$$\mathrm{QIB}(T) = \text{number of } B_4 \text{ quartets in } T. \tag{2.5}$$

Importantly for us, in [MCMRV, Lemma 4] it is shown that for $n > 4$, the quartet index has a recursive representation as

$$\mathrm{QIB}(T_n) = \mathrm{QIB}(T_{L_n}) + \mathrm{QIB}(T_{n-L_n}) + \binom{L_n}{2}\binom{n-L_n}{2}, \tag{2.6}$$

where T_n is the tree on n leaves.

In [MCMRV] various models of tree growth were considered: Aldous's β-model, Ford's α-model ([F], but see also [MCMR]) and Chen–Ford–Winkel's α-γ-model [CFW09]. In this work we will focus on Aldous's β model of tree growth with $\beta \geq 0$ and characterize the limit distribution of the QI as the number of leaves, n, grows to infinity. We will take advantage of the recursive representation (2.6) that enables the use of the powerful contraction method.

We require knowledge of the mean and variance of the QI for Aldous's β-model [MCMRV, Corollaries 4 and 7]):

$$\begin{aligned} \mathrm{E}[\mathrm{QIB}(T_n)] &= \frac{3\beta+6}{7\beta+18}\binom{n}{4}, \\ \mathrm{Var}[\mathrm{QIB}(T_n)] &= \frac{(\beta+2)(2\beta^2+9\beta+12)}{2(7\beta+18)^2(127\beta^3+1383\beta^2+4958\beta+5880)}n^8 \\ &\quad + O(n^7). \end{aligned} \tag{2.7}$$

3. Contraction method. Consider the space D of distribution functions with finite second moment and first moment zero. On D we define the Wasserstein metric

$$d(F,G) = \inf \|X-Y\|_2$$

where $\|\cdot\|_2$ denotes the L_2 norm and the infimum is over all $X \sim F$, $Y \sim G$. Notice that convergence in d induces convergence in distribution.

Let $\tau \in [0,1]$ be a random variable whose distribution is not a Dirac δ at 0 nor at 1. For $r \in \mathbb{N}_+$ define a transformation $S : D \to D$ by

$$S(F) = \mathcal{L}(\tau^r Y' + (1-\tau)^r Y'' + C(\tau)), \tag{3.1}$$

where $\mathcal{L}(X)$ denotes the law of the random variable X, and Y', Y'', τ are independent with $Y', Y'' \sim F$; moreover we assume that τ satisfies, for all n,

$$2\sum_{i=1}^{n} p_{n,i}\left(\frac{i}{n}\right)^{2r} < 1, \tag{3.2}$$

where $p_{n,i} = P((i-1)/n < \tau \le i/n)$, and the function $C(\cdot)$ is of the form

$$C(\tau) = \sum_{r_1+r_2 \le r} C_{r_1,r_2}\tau^{r_1}(1-\tau)^{r_2} \tag{3.3}$$

for some constants C_{r_1,r_2} and furthermore satisfies $\mathrm{E}[C(\tau)] = 0$. By [R92, Thms. 3 and 4], S is well defined, has a unique fixed point and for any $F \in D$ the sequence $S^n(F)$ converges exponentially fast in the d metric to S's fixed point. Using the exact arguments used to show [R91, Thm. 2.1] one can show that the map S is a contraction. Only the Lipschitz constant of convergence will differ, being $\sqrt{C_\tau}$ with $C_\tau = \max\{\mathrm{E}[\tau^{2r}], \mathrm{E}[(1-\tau)^{2r}]\}$ in our case. Notice that as $\tau \in [0,1]$ and is non-degenerate at the edges, it follows that $C_\tau < 1$ and we have a contraction.

We now state the main result of our work. We show weak convergence, with a characterization of the limit for a class of recursively defined models.

Theorem 3.1 (cf. [R91, Thm. 3.1]). *For $n \ge 2$ and $\beta > 0$ let $L_n \in \{1,\dots,n-1\}$ be such that $L_n - 1$ is* BetaBinomial$(n-2, \beta+1, \beta+1)$ *distributed and let τ be* Beta$(\beta+1, \beta+1) =: F_\tau$ *distributed. Starting from the Dirac δ at 0, i.e. $Y_1 = 0$ and with the convention* BetaBinomial$(0, \beta+1, \beta+1) = \delta_0$, *for $r \in \mathbb{N}_+$ such that the condition* (3.2) *is met with this choice of F_τ, define recursively a sequence of random variables by*

$$Y_n = \left(\frac{L_n}{n}\right)^r Y_{L_n} + \left(1 - \frac{L_n}{n}\right)^r Y_{n-L_n} + C_n(L_n),$$

where

$$C_n(i) = n^{-r}\Big(\sum_{r_1+r_2+r_3 \le r} C_{r_1,r_2,r_3} i^{r_1}(n-i)^{r_2} n^{r_3} + h_n(i)\Big), \tag{3.4}$$

with $\mathrm{E}[C_n(L_n)] = 0$ *and* $\sup_i n^{-r}h_n(i) \to 0$. *If* $\mathrm{E}[Y_n^2]$ *is uniformly bounded then the random variable Y_n converges in the Wasserstein d-metric to a random variable Y_∞ whose distribution is the unique fixed point of the transformation S of* (3.1).

Notice that as $Y_1 = 0$ and by the definition of the recursion we have $\mathrm{E}[Y_n] = 0$ for all n.

The Yule tree case will be the limit of $\beta = 0$, and in this case the proof of the result will be more straightforward (as commented on in the proof of Theorem 3.1).

Notice that $L_n/n \xrightarrow{D} \tau$. It is tempting to suspect that Theorem 3.1 is a corollary of a general result related to the contraction method (as presented in [D09, (8.12), p. 351]). However, to the best of my knowledge, general results assume L_2 convergence of L_n/n (e.g. [D09, Thm. 8.6, p. 354]), while in our phylogenetic balance index case we will only have convergence in distribution. In such a case it seems that convergence has to be proved case by case (see e.g. examples in [RR95]). Here we show the convergence of Theorem 3.1 similarly to [R91].

We first derive a lemma that controls the non-homogeneous part of the recursion, i.e. $C_n(\cdot)$ as defined in (3.4).

LEMMA 3.2 (cf. [R91, Prop. 3.2]). *Let $C_n : \{1, \ldots, n-1\} \to \mathbb{R}$ be as in equation* (3.4). *Then*

$$\sup_{x\in[0,1)} \left|C_n(\lfloor (n-1)x\rfloor + 1) - C(x)\right| \le \sup_i n^{-r} h_n(i) + O(n^{-1}). \tag{3.5}$$

Proof. For $1 \le \lfloor (n-1)x\rfloor + 1 \le n-1$ and writing $i = \lfloor (n-1)x\rfloor + 1$ we have, due to (3.3) and (3.4),

$$\begin{aligned}
&|C_n(\lfloor (n-1)x\rfloor + 1) - C(x)| \\
&\qquad \le \max\{C_{r_1,r_2}\}\Bigg(\left|\left(\frac{i}{n}\right)^r - x^r\right| + \left|\left(1-\frac{i}{n}\right)^r - (1-x)^r\right| \\
&\qquad\qquad + \sum_{r_1+r_2\le r}\left|\left(\frac{i}{n}\right)^{r_1}\left(1-\frac{i}{n}\right)^{r_2} - x^{r_1}(1-x)^{r_2}\Bigg)\right| \\
&\qquad\qquad + \sup_i n^{-r} h_n(i).
\end{aligned}$$

Bounding the individual components, using the mean value theorem and the fact that by construction x cannot differ from i/n by more than $1/n$, we have

$$\left|\left(\frac{i}{n}\right)^r - x^r\right| \le r\left|\frac{i}{n} - x\right| \le \frac{r}{n} = O(n^{-1})$$

and

$$\left|\left(1-\frac{i}{n}\right)^r - (1-x)^r\right| \le r\left|\frac{i}{n} - x\right| \le \frac{r}{n} = O(n^{-1}).$$

Furthermore, by the triangle inequality and the above two inequalities,

$$\left|\left(\frac{i}{n}\right)^{r_1}\left(1-\frac{i}{n}\right)^{r_2} - x^{r_1}(1-x)^{r_2}\right| = O(n^{-1}). \blacksquare$$

Lemma 3.3 (cf. [R91, Prop. 3.3]). *Let a_n, b_n, $p_{n,i}$, $n \in \mathbb{N}$, be three sequences such that* $0 \le b_n \to 0$ *as* $n \to \infty$, $0 \le p_{n,i} \le 1$,

$$0 \le a_{n+1} \le 2 \sum_{i=1}^{n} p_{n,i} \left(\frac{i}{n} \right)^R \Big(\sup_{i \in \{1,\dots,n\}} a_i \Big) + b_n. \tag{3.6}$$

and

$$0 < 2 \sum_{i=1}^{n} p_{n,i} \left(\frac{i}{n} \right)^R = C < 1.$$

Then $\lim_{n \to \infty} a_n = 0$.

Proof. The proof is exactly the same as in [R91, proof of Prop. 3.3]. In the last step we will have, with $a := \limsup a_n < \infty$, the sandwiching for all $\epsilon > 0$:

$$0 \le a \le C(a + \epsilon). \ \blacksquare$$

Having Lemmata 3.2 and 3.3 we turn to showing Theorem 3.1.

Proof of Theorem 3.1. Denote the law of Y_n as $\mathcal{L}(Y_n) = G_n$. We take Y_∞ and Y'_∞ independent and distributed as G_∞, the fixed point of S. Then, for $i = 1, \dots, n-1$ we choose independent versions of Y_i and Y'_i. We need to show $d^2(G_n, G_\infty) \to 0$. As the metric is the infimum over all pairs of random variables that have marginal distributions G_n and G_∞, the obvious choice is to take Y_n, Y_∞ such that L_n/n will be close to τ for large n. The Yule model ($\beta = 0$) was considered in [R91] and there $\tau \sim \mathrm{Unif}[0,1]$ and L_n is uniform on $\{1, \dots, n-1\}$. Hence, $\lfloor (n-1)\tau \rfloor + 1$ will be uniform on $\{1, \dots, n-1\}$ (remember $P(\tau = 1) = 0$), and $L_n/n \overset{D}{=} (\lfloor (n-1)\tau \rfloor + 1)/n$.

However, when $\beta > 0$ the situation complicates. For a given n, $L_n - 1$ is $\mathrm{BetaBinomial}(n-2, \beta+1, \beta+1)$ distributed (cf. (2.1) and [A96, (1) and (3)]). Hence, if $\tau \sim \mathrm{Beta}(\beta+1, \beta+1)$ and $L_n - 1 \sim \mathrm{BetaBinomial}(n-2, \beta+1, \beta+1)$ we do not have $L_n/n \overset{D}{=} (\lfloor (n-1)\tau \rfloor + 1)/n$ exactly. We may bound the Wasserstein metric by any coupling that retains the marginal distributions of the two random variables.

Therefore, from now on we will be considering a version where conditional on τ, the random variable $L_n - 1$ is $\mathrm{Binomial}(n-2, \tau)$ distributed. Let r_n be any sequence such that $r_n/n \to 0$ and $n/r_n^2 \to 0$, e.g. $r_n = n \ln^{-1} n$. Then, by Chebyshev's inequality,

$$P\big(|L_n - \mathrm{E}[L_n | \tau]| \ge r_n \bigm| \tau \big) \le \frac{n\tau(1-\tau)}{r_n^2} \le \frac{n}{4r_n^2} \to 0.$$

We now want to show $d^2(G_n, G_\infty) \to 0$ and we will exploit the above coupling in the bound:

$$
\begin{aligned}
&d^2(G_n, G_\infty)\\
&\quad\le \mathrm{E}\Bigg[\Bigg(\Bigg(\Big(\frac{L_n}{n}\Big)^r Y_{L_n} - \tau^r Y_\infty\Bigg) + \Bigg(\Big(\frac{n-L_n}{n}\Big)^r Y_{n-L_n} - (1-\tau)^r Y'_\infty\Bigg)\\
&\qquad + (C_n(L_n) - C(\tau))\Bigg)^2\Bigg]\\
&\quad= \mathrm{E}\Bigg[\Bigg(\Big(\frac{L_n}{n}\Big)^r Y_{L_n} - \tau^r Y_\infty\Bigg)^2\Bigg]\\
&\quad+ \mathrm{E}\Bigg[\Bigg(\Big(\frac{n-L_n}{n}\Big)^r Y_{n-L_n} - (1-\tau)^r Y'_\infty\Bigg)^2\Bigg] + \mathrm{E}[(C_n(L_n) - C(\tau))^2],
\end{aligned}
$$

where $Y_\infty, Y'_\infty \sim G_\infty$ are independent. Remember that $\mathrm{E}[Y_i] = \mathrm{E}[Y_\infty] = 0$ so that the expectation of the cross products disappears.

Our main step is to give a bound where the L_n/n term is replaced by some transformation of τ. Let $\tilde{r}_n$ be an appropriate random integer in $\{\pm 1, \ldots, \pm\lceil r_n \rceil\}$ and we may write (with the chosen coupling of L_n and τ)

$$
\begin{aligned}
&\mathrm{E}\Bigg[\Bigg(\Big(\frac{L_n}{n}\Big)^r Y_{L_n} - \tau^r Y_\infty\Bigg)^2\Bigg] = \mathrm{E}\Bigg[\mathrm{E}\Bigg[\Bigg(\Big(\frac{L_n}{n}\Big)^r Y_{L_n} - \tau^r Y_\infty\Bigg)^2 \;\Bigg|\; \tau\Bigg]\Bigg]\\
&= \mathrm{E}\Bigg[\mathrm{E}\Bigg[\Bigg(\Big(\frac{\lfloor (n-1)\tau\rfloor + 1 + \tilde{r}_n}{n}\Big)^r Y_{L_n} - \tau^r Y_\infty\Bigg)^2 \;\Bigg|\; |L_n - \mathrm{E}[L_n]| \le r_n, \tau\Bigg]\\
&\qquad\qquad\qquad\qquad\qquad\qquad \cdot P(|L_n - \mathrm{E}[L_n]| \le r_n \mid \tau)\Bigg]\\
&+ \mathrm{E}\Bigg[\mathrm{E}\Bigg[\Bigg(\Big(\frac{L_n}{n}\Big)^r Y_{L_n} - \tau^r Y_\infty\Bigg)^2 \;\Bigg|\; |L_n - \mathrm{E}[L_n]| \ge r_n, \tau\Bigg] P(|L_n - \mathrm{E}[L_n]| \ge r_n \mid \tau)\Bigg]\\
&\le \mathrm{E}\Bigg[\Bigg(\Big(\frac{\lfloor (n-1)\tau\rfloor + 1 + \tilde{r}_n}{n}\Big)^r Y_{L_n} - \tau^r Y_\infty\Bigg)^2 + \frac{n}{4r_n^2}\,\mathrm{E}\Bigg(\Big(\frac{L_n}{n}\Big)^r Y_{L_n} - \tau^r Y_\infty\Bigg)^2\\
&= \mathrm{E}\Bigg[\Bigg(\Bigg(\Big(\frac{\lfloor (n-1)\tau\rfloor + 1}{n}\Big)^r + r\frac{\tilde{r}_n}{n}\Big(\frac{\lfloor (n-1)\tau\rfloor + 1 + \xi_{\tilde{r}_n}}{n}\Big)^{r-1}\Bigg) Y_{L_n} - \tau^r Y_\infty\Bigg)^2\Bigg]\\
&\quad + \frac{n}{4r_n^2}\,\mathrm{E}\Bigg[\Bigg(\Big(\frac{L_n}{n}\Big)^r Y_{L_n} - \tau^r Y_\infty\Bigg)^2\Bigg]\\
&= \mathrm{E}\Bigg[\Bigg(\Big(\frac{\lfloor (n-1)\tau\rfloor + 1}{n}\Big)^r Y_{L_n} - \tau^r Y_\infty\Bigg)^2\\
&\qquad\qquad\qquad + r^2 n^{-2}\,\mathrm{E}\Bigg[\tilde{r}_n^2\Big(\frac{\lfloor (n-1)\tau\rfloor + 1 + \xi_{\tilde{r}_n}}{n}\Big)^{2(r-1)} Y_{L_n}^2\Bigg]\\
&+ 2rn^{-1}\,\mathrm{E}\Bigg[\tilde{r}_n\Big(\frac{\lfloor (n-1)\tau\rfloor + 1 + \xi_{\tilde{r}_n}}{n}\Big)^{r-1} Y_{L_n}\cdot\Bigg(\Big(\frac{\lfloor (n-1)\tau\rfloor + 1}{n}\Big)^r Y_{L_n} - \tau^r Y_\infty\Bigg)\Bigg]\\
&+ \frac{n}{4r_n^2}\,\mathrm{E}\Bigg[\Bigg(\Big(\frac{L_n}{n}\Big)^r Y_{L_n} - \tau^r Y_\infty\Bigg)^2\Bigg],
\end{aligned}
$$

where $\xi_{\tilde{r}_n} \in (0, \tilde{r}_n)$ is (a random variable) such that the mean value theorem holds (for the function $(\cdot)^r$). As Y_n, Y_∞ have uniformly bounded second moments and $0 \le \xi_{\tilde{r}_n} \le \tilde{r}_n \le r_n \le n$, we have, by the assumptions $r_n/n \to 0$ and $n/r_n^2 \to 0$,

$$\left(r\frac{r_n}{n}\right)^2 \mathrm{E}\left[\left(\frac{\lfloor(n-1)\tau\rfloor+1+\xi_{\tilde{r}_n}}{n}\right)^{2(r-1)} Y_{L_n}^2\right]+\frac{n}{4r_n^2}\mathrm{E}\left[\left(\left(\frac{L_n}{n}\right)^r Y_{L_n}-\tau^r Y_\infty\right)^2\right]$$
$$+2r\frac{r_n}{n}\mathrm{E}\left[\left(\frac{\lfloor(n-1)\tau\rfloor+1+\xi_{\tilde{r}_n}}{n}\right)^{r-1} Y_{L_n}\cdot\left(\left(\frac{\lfloor(n-1)\tau\rfloor+1}{n}\right)^r Y_{L_n}-\tau^r Y_\infty\right)\right]\to 0,$$

and hence for some sequence $u_n \to 0$,

$$\mathrm{E}\left[\left(\left(\frac{L_n}{n}\right)^r Y_{L_n}-\tau^r Y_\infty\right)^2\right] \le \mathrm{E}\left[\left(\left(\frac{\lfloor(n-1)\tau\rfloor+1}{n}\right)^r Y_{L_n}-\tau^r Y_\infty\right)^2\right]+u_n.$$

Remembering the assumption $\sup_i n^{-r}h_n(i) \to 0$, the other component can be treated in the same way as $\mathrm{E}[((L_n/n)^r Y_{L_n} - \tau^r Y_\infty)^2]$ with conditioning on τ and then controlling L_n's deviation from its expected value by r_n and Chebyshev's inequality. Therefore, for some sequence $v_n \to 0$,

$$\begin{aligned} d^2(G_n, G_\infty) \le{}& \mathrm{E}\left[\left(\left(\frac{\lfloor(n-1)\tau\rfloor+1}{n}\right)^r Y_{L_n}-\tau^r Y_\infty\right)^2\right] \\ &+\mathrm{E}\left[\left(C_n(\lfloor(n-1)\tau\rfloor+1)-C(\tau)\right)^2\right] \\ &+\mathrm{E}\left[\left(\left(\frac{n-\lfloor(n-1)\tau\rfloor-1}{n}\right)^r Y_{n-L_n}-(1-\tau)^r Y'_\infty\right)^2\right]+v_n. \end{aligned}$$

In order to estimate the first term of the right-hand side, let us denote $d_{n-1}^2 := \sup_{i\in\{1,\dots,n-1\}} d^2(G_i, G_\infty)$. Then

$$\begin{aligned} &\mathrm{E}\left[\left(\left(\frac{\lfloor(n-1)\tau\rfloor+1}{n}\right)^r Y_{L_n}-\tau^r Y_\infty\right)^2\right] \\ &\qquad= \mathrm{E}\left[\sum_{i=1}^{n-1} 1_{(i-1)/(n-1)<\tau\le i/(n-1)}\left(\left(\frac{i}{n}\right)^r Y_{L_n}-\tau^r Y_\infty\right)^2\right] \\ &\qquad\le \sum_{i=1}^{n-1} p_{n-1,i}\left(\frac{i}{n}\right)^{2r} \mathrm{E}[(Y_{L_n}-Y_\infty)^2] \\ &\qquad= \sum_{i=1}^{n-1} p_{n-1,i}\left(\frac{i}{n}\right)^{2r} d_{n-1}^2, \end{aligned}$$

where $p_{n,i} = P((i-1)/(n-1) < \tau \le i/(n-1))$. Invoking Lemmata 3.2, 3.3 and using the assumption (3.2) with $R = 2r$ we have

$$d^2(G_n, G_\infty) \leq 2\sum_{i=1}^{n-1} p_{n-1,i}\left(\frac{i}{n}\right)^{2r} d_{n-1}^2 + \Big(n^{-r}\sup_i h_n(i)\Big)^2 + v_n + O(n^{-2}),$$

which converges to 0. ∎

4. Limit distribution of the quartet index for Aldous's ($\beta \geq 0$)**-model trees.** We show here that the QIB of Aldous's ($\beta \geq 0$)-model trees satisfies the conditions of Theorem 3.1 with $r = 4$ and hence the QIB has a well characterized limit distribution. We define a centred and scaled version of the QIB for Aldous's ($\beta \geq 0$)-model tree on $n \geq 4$ leaves:

$$Y_n^Q = n^{-4}\left(\mathrm{QIB}(T_n) - \frac{3\beta+6}{7\beta+18}\binom{n}{4}\right). \tag{4.1}$$

We now specialize Theorem 3.1 to the QIB case and assume $Y_1 = Y_2 = Y_3 = 0$ for completeness.

Theorem 4.1. *The sequence of random variables Y_n^Q for trees generated by Aldous's β-model with $\beta \geq 0$ converges as $n \to \infty$ in the Wasserstein d-metric* (*and hence in distribution*) *to a random variable $Y_Q \sim \mathcal{Q} \equiv G_\infty$ satisfying the following equality in distribution:*

$$\begin{aligned} Y_Q \overset{\mathcal{D}}{=} \tau^4 Y_Q' + (1-\tau)^4 Y_Q'' &+ \frac{3\beta+6}{24(7\beta+18)}(\tau^4 + (1-\tau)^4) \\ &- \frac{3\beta+6}{24(7\beta+18)} + \frac{1}{4}\tau^2(1-\tau)^2, \end{aligned} \tag{4.2}$$

where $\tau \sim F_\tau$ has the Beta distribution of equation (2.3), *$Y_Q, Y_Q', Y_Q'' \sim \mathcal{Q}$ and Y_Q', Y_Q'', τ are all independent.*

Proof. Denote by $P_3(x,y)$ a polynomial of degree at most three in the variables x, y. From the recursive representation (2.6), for $n > 4$,

$$\begin{aligned} Y_n^Q &= n^{-4}\bigg(\mathrm{QIB}(T_{L_n}) - \frac{3\beta+6}{(7\beta+18)}\binom{L_n}{4} + \mathrm{QIB}(T_{n-L_n}) \\ &\quad - \frac{3\beta+6}{(7\beta+18)}\binom{n-L_n}{4} + \binom{L_n}{2}\binom{n-L_n}{2} + \frac{3\beta+6}{(7\beta+18)}\binom{L_n}{4} \\ &\quad + \frac{3\beta+6}{(7\beta+18)}\binom{n-L_n}{4} - \frac{3\beta+6}{(7\beta+18)}\binom{n}{4}\bigg) \\ &= \left(\frac{L_n}{n}\right)^4 Y_{L_n}^Q + \left(1 - \frac{L_n}{n}\right)^4 Y_{n-L_n}^Q + \frac{1}{4}\left(\frac{L_n}{n}\right)^2\left(1-\frac{L_n}{n}\right)^2 \\ &\quad + \frac{3\beta+6}{24(7\beta+18)}\left(\frac{L_n}{n}\right)^4 + \frac{3\beta+6}{24(7\beta+18)}\left(1-\frac{L_n}{n}\right)^4 - \frac{3\beta+6}{24(7\beta+18)} \\ &\quad + n^{-4}P_3(n, L_n). \end{aligned}$$

We therefore have $r = 4$ and

$$\begin{aligned}C_n(i) = \frac{1}{4}\left(\frac{i}{n}\right)^2\left(1-\frac{i}{n}\right)^2 + \frac{3\beta+6}{24(7\beta+18)}\left(\left(\frac{i}{n}\right)^4 + \left(1-\frac{i}{n}\right)^4\right)\\ - \frac{3\beta+6}{24(7\beta+18)} + n^{-4}P_3(n,i).\end{aligned}$$

By scaling and centring we know that $EY_n^Q = 0$ and $E(Y_n^Q)^2$ is uniformly bounded by (2.7). Because of the Beta law of τ we need to examine, for all i,

$$p_{n,i} := P\left(\frac{i-1}{n} \le \tau < \frac{i}{n}\right) = \frac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)} \int_{(i-1)/n}^{i/n} x^\beta(1-x)^\beta\,\mathrm{d}x.$$

We consider two cases.

If $\beta > 0$, we have to check whether (3.2) is satisfied. Let

$$B_x(\beta+1,\beta+1) = \int_0^x u^\beta(1-u)^\beta\,\mathrm{d}u$$

be the incomplete Beta function. Then

$$\begin{aligned}p_{n,i} &= \frac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)}\big(B_{i/n}(\beta+1,\beta+1) - B_{(i-1)/n}(\beta+1,\beta+1)\big)\\ &= n^{-1}\frac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)}B'_\xi(\beta+1,\beta+1)\end{aligned}$$

for some $\xi \in ((i-1)/n, i/n)$ by the mean value theorem. Obviously

$$B'_\xi(\beta+1,\beta+1) = \xi^\beta(1-\xi)^\beta \le \left(\frac{i}{n}\right)^\beta\left(1-\frac{i-1}{n}\right)^\beta$$

and now

$$\begin{aligned}\sum_{i=1}^n p_{n,i}\left(\frac{i}{n}\right)^R &\le \frac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)}n^{-1}\sum_{i=1}^n\left(\frac{i}{n}\right)^R\left(\frac{i}{n}\right)^\beta\left(1-\frac{i-1}{n}\right)^\beta\\ &\to \frac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)}\int_0^1 u^{\beta+R}(1-u)^\beta\,\mathrm{d}u\\ &= \frac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)}\,\frac{\Gamma(\beta+R+1)\Gamma(\beta+1)}{\Gamma(2\beta+R+2)}\\ &= \frac{\Gamma(2\beta+2)}{\Gamma(\beta+1)}\,\frac{\Gamma(\beta+R+1)}{\Gamma(2\beta+R+2)}.\end{aligned} \tag{4.3}$$

Take $1 < R_1 < R_2$ and consider the ratio

$$\begin{aligned}A &= \frac{\Gamma(2\beta+2)}{\Gamma(\beta+1)}\,\frac{\Gamma(\beta+R_2+1)}{\Gamma(2\beta+R_2+2)}\left(\frac{\Gamma(2\beta+2)}{\Gamma(\beta+1)}\,\frac{\Gamma(\beta+R_1+1)}{\Gamma(2\beta+R_1+2)}\right)^{-1}\\ &= \frac{\Gamma(\beta+R_2+1)}{\Gamma(2\beta+R_2+2)}\,\frac{\Gamma(\beta)}{\Gamma(\beta)}\,\frac{\Gamma(2\beta+R_1+1)}{\Gamma(\beta+R_1+1)}\\ &= \frac{B(\beta+1+R_2,\beta)}{B(\beta+1+R_1,\beta)}.\end{aligned}$$

We have $A < 1$ as the Beta function is decreasing in each of its arguments, hence the upper bound in (4.3) is decreasing in R. For $R = 1$ the bound equals

$$\frac{\Gamma(2\beta+2)}{\Gamma(\beta+1)}\,\frac{\Gamma(\beta+2)}{\Gamma(2\beta+3)} = \frac{\Gamma(2\beta+2)}{\Gamma(\beta+1)}\,\frac{(\beta+1)\Gamma(\beta+1)}{(2\beta+2)\Gamma(2\beta+2)} = \frac{\beta+1}{2(\beta+1)} = \frac{1}{2},$$

and hence for all $R > 1$ and all $\beta > 0$,

$$\sum_{i=1}^{n} p_{n,i}\left(\frac{i}{n}\right)^{R} < \frac{1}{2}.$$

As in our case we have $r \geq 1$, for $R = 2r \geq 2$ the assumptions of Lemma 3.3 are satisfied, and the statement of the theorem follows.

If $\beta = 0$, then directly $p_{n,i} = n^{-1}$, (3.2) and the assumptions of Lemma 3.3 are immediately satisfied, and the statement of the theorem follows. This is the Yule model case, in which the proof of the counterpart of Theorem 3.1 is much more straightforward, as mentioned before. ■

Remark 4.2. When $\beta < 0$ the process L_n/n seems to have a more involved asymptotic behaviour (cf. [A96, Lemma 3] in the $\beta \leq -1$ case). Furthermore, the bounds applied here do not hold for $\beta < 0$. Therefore, this family of tree models (including the important uniform model, $\beta = -3/2$) deserves a separate study with respect to its quartet index.

5. Comparing with simulations. To verify the results we compared the simulated values from the limiting theoretical distribution of Y_Q with scaled and centred values of Yule tree QI values. The 500-leaf Yule trees were simulated using the `rtreeshape()` function of the apTreeshape [BDBF12] R [R17] package and Tomás Martínez-Coronado's in-house Python code. Then, for each tree the QI value was calculated by Gabriel Valiente's and Tomás Martínez-Coronado's in-house programs. The raw values $\mathrm{QIB}(\mathrm{Yule}_{500})$ were scaled and centred as

$$Y_n^Q = 500^{-4}\left(\mathrm{QIB}(\mathrm{Yule}_{500}) - \frac{1}{3}\binom{500}{4}\right).$$

The Y_Q values were simulated using the heuristic algorithm proposed in [B18, Algorithm 3] (see R code in Appendix). The results of the simulation are presented in Fig. 2.

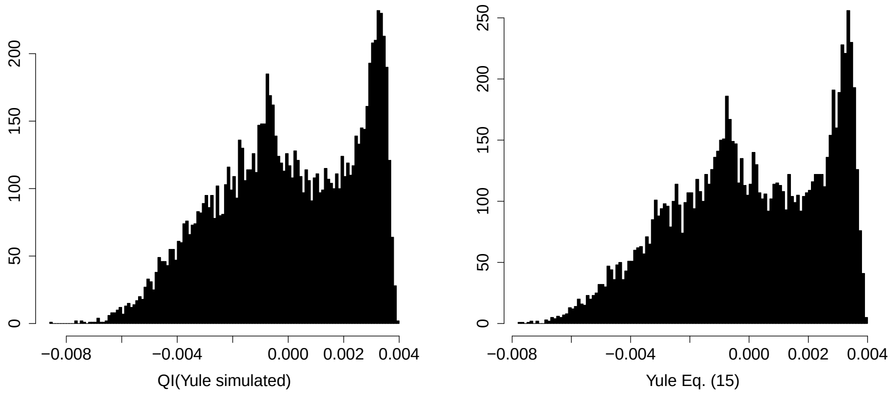

Fig. 2. Left: histogram of scaled and centred simulated values of the QIB for the Yule tree, Y_n^Q; right: histogram of Y_Q for the Yule model, $\beta = 0$. The mean, variance, skewness and excess kurtosis of the simulated values are $-3.177 \cdot 10^{-6}$, $6.321 \cdot 10^{-6}$, -0.308, -0.852 (left, simulated values) and $1.682 \cdot 10^{-5}$, $6.38 \cdot 10^{-6}$, -0.317, -0.834 (right, theoretical values of the heuristic Algorithm 3 in [B18] with recursion depth 15). For $\beta = 0$ the leading constant of the variance in (2.7) is $5/(24 \cdot 33075) \approx 6.299 \cdot 10^{-6}$.

Appendix: R code for simulating from the limit distribution of the normalized quartet index

```
fCtau_QIB<-function(x,beta=0){
    PB4beta<-(3*beta+6)/(7*beta+18);
    PB4beta*(x^4)/24+PB4beta*((1-x)^4)/24-PB4beta/24+0.25*x*x*(1-x)^2
}

fdistribution_limitQIB<-function(num.iter=10,popsize=10000,Y0=0){
    replicate(popsize,fdraw_limitQIB(num.iter,Y0))
}

fdraw_limitQIB<-function(num.iter=15,Y0=0){
    res<-0
    if (num.iter==0){
        Y1<-Y0
        Y2<-Y0
    }
    else{
        Y1<-fdraw_limitQIB(num.iter-1,Y0)
        Y2<-fdraw_limitQIB(num.iter-1,Y0)
    }
    tau<-runif(1)
    res<-((tau)^4)*Y1+((1-tau)^4)*Y2+fCtau_QIB(tau)
    res
}

popsize<-10000 ## size of sample for histogram
num.iter<-15 ## depth of the recursion
Y0 <- 0 ## initial value

vlimitQIB<-fdistribution_limitQIB(num.iter=num.iter,popsize=popsize)
```

Acknowledgements. I would like to thank the whole Computational Biology and Bioinformatics Research Group of the Balearic Islands University for hosting me on multiple occasions, introducing me to problems related to the quartet index and for valuable comments on this article. The simulated values of the quartet index for the Yule tree were provided by Gabriel Valiente and Tomás Martínez-Coronado. I was supported by the Knut and Alice Wallenberg Foundation and am by the Swedish Research Council (Vetenskapsrådet) grant no. 2017-04951. My collaboration with the Balearic Islands University was partially supported by the the G S Magnuson Foundation of the Royal Swedish Academy of Sciences (grants no. MG2015-0055, MG2017-0066) and The Foundation for Scientific Research and Education in Mathematics (SVeFUM). I am grateful to anonymous reviewers for comments that significantly improved this work.

References

[A91] D. Aldous, *The continuum random tree II: an overview*, in: Stochastic Analysis (Durham, 1990), M. T. Barlow and N. H. Bingham (eds.), London Math. Soc. Lecture Note Ser. 167, Cambridge Univ. Press, 1991, 23–70.

[A96] D. Aldous, *Probability distributions on cladograms*, in: Random Discrete Structures, D. Aldous and R. Pemantle (eds.), Springer, 1996, 1–18.

[A01] D. Aldous, *Stochastic models and descriptive statistics for phylogenetic trees*, Statist. Sci. 16 (2001), 23–34.

[B18] K. Bartoszek, *Exact and approximate limit behaviour of the Yule tree's cophenetic index*, Math. Biosci. 303 (2018), 26–45.

[BF06] M. G. B. Blum and O. François, *On statistical tests of phylogeny imbalance: The Sackin and other indices revisited*, Math. Biosci. 195 (2006), 141–153.

[BFJ06] M. G. B. Blum, O. François and S. Janson, *The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance*, Ann. Appl. Probab. 16 (2006), 2195–2214.

[BDBF12] N. Bortolussi, E. Durand, M. Blum and O. François, *apTreeshape: Analyses of phylogenetic treeshape. R package version 1.4-5*, 2012; https://CRAN.R-project.org/package=apTreeshape.

[CF10] H. Chang and M. Fuchs, *Limit theorems for patterns in phylogenetic trees*, J. Math. Biol. 60 (2010), 481–512.

[CFW09] B. Chen, D. Ford and M. Winkel, *A new family of Markov branching trees: the alpha-gamma model*, Electron. J. Probab. 14 (2009), 400–430.

[CS07] B. Chor and S. Snir, *Analytic solutions of maximum likelihood on forks of four taxa*, Math. Biosci. 208 (2007) 347–358.

[C82] D. H. Colless, *Review of "Phylogenetics: the theory and practise of phylogenetic systematics"*, Systematic Zoology 31 (1982), 100–104.

[D09] M. Drmota, *Random Trees: an Interplay between Combinatorics and Probability*, Springer, 2009.

[F04] J. Felsenstein, *Inferring Phylogenies*, Sinauer Associates, Sunderland, MA, 2004.

[F] D. J. Ford, *Probabilities on cladograms: introduction to the alpha model*, arXiv: math/0511246 (2005).

[H62] C. A. R. Hoare, *Quicksort*, Computer J. 5 (1962), 10–15.

[MS91] W. P. Maddison and M. Slatkin, *Null models for the number of evolutionary steps in a character on a phylogenetic tree*, Evolution 45 (1991) 1184–1197.

[M92] H. M. Mahmoud, *Evolution of Random Search Trees*, Wiley, 1992.

[MCMR] T. Martínez-Coronado, A. Mir and F. Rosselló, *The probabilities of trees and cladograms under Ford's α-model*, arXiv:1801.03843 (2018).

[MCMRV] T. Martínez-Coronado, A. Mir, F. Rosselló, and G. Valiente, *A balance index for phylogenetic trees based on quartets*, arXiv:1803.01651 (2018).

[MRR] A. Mir, F. Rosselló and L. Rotger, *A new balance index for phylogenetic trees*, Math. Biosci. 241 (2013), 125–136.

[R17] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Wien, 2017; https://www.R-project.org/.

[RR95] S. T. Rachev and L. Rüschendorf, *Probability metrics and recursive algorithms*, Adv. Appl. Probab. 27 (1995), 770–779.

[RS13] S. Roch and S. Snir, *Recovering the treelike trend of evolution despite extensive lateral gene transfer: a probabilistic analysis*, J. Comput. Biol. 20 (2013) 93–112.

[R91] U. Rösler, *A limit theorem for "Quicksort"*, RAIRO Inform. Théor. Appl. 25 (1991) 85–100.

[R92] U. Rösler, *A fixed point theorem for distributions*, Stoch. Process. Appl. 42 (1992) 195–214.

[S72] M. J. Sackin, *"Good" and "bad" phenograms*, Systematic Zoology 21 (1972) 225–226.

[SS03] C. Semple and M. Steel, *Phylogenetics*, Oxford Univ. Press, 2003.

[SM01] M. Steel and A. McKenzie, *Properties of phylogenetic trees generated by Yule-type speciation models*, Math. Biosci. 170 (2001), 91–112.

[Y24] G. U. Yule, *A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis F.R.S.*, Philos. Trans. Roy. Soc. B 213 (1924), 21–87.

Krzysztof Bartoszek
Department of Computer and Information Science
Linköping University
Linköping 581 83, Sweden
ORCID: 0000-0002-5816-4345
E-mail: krzysztof.bartoszek@liu.se
krzbar@protonmail.ch