J. Appl. Prob. **54,** 701–709 (2017)
*doi:*10.1017/jpr.2017.28

THE GINI INDEX OF RANDOM TREES WITH AN APPLICATION TO CATERPILLARS

HRISHIKESH BALAJI,* *Winston Churchill High School*
HOSAM MAHMOUD,** *The George Washington University*

Abstract

We propose two distance-based topological indices (level index and Gini index) as measures of disparity within a single tree and within tree classes. The level index and the Gini index of a single tree are measures of balance within the tree. On the other hand, the Gini index for a class of random trees can be used as a comparative measure of balance between tree classes. We establish a general expression for the level index of a tree. We compute the Gini index for two random classes of caterpillar trees and see that a random multinomial model of trees with finite height has a countable number of limits in $[0, \frac{1}{3}]$, whereas a model with independent level numbers fills the spectrum $(0, \frac{1}{3}]$.

Keywords: Random tree; chemical tree; combinatorial probability; Gini index; Wiener index; topological index

2010 Mathematics Subject Classification: Primary 05C05; 05C12; 92E10
Secondary 60C05

1. Introduction

Many types of graphs arise in important applications, such as networks, data structures, models for algorithms, and models for chemical molecules.

Topological indices numerically quantify aspects of these graphs for multiple purposes, such as ranking and comparing their properties to identify the better selections among them, and predicting properties of molecules. When molecules are viewed as graphs, structural properties such as connectivity and balance are correlated with physical properties such as boiling points.

Several indices have been proposed and analyzed for different classes of graphs. These include the Zagreb index [3], the Randič index [4], and the Wiener index [10], among others. However, no one single index adequately describes all the facets of a tree. Rather, combinations of these indices can coalesce to portray a picture. Our motivation is to propose two other indices, namely, the level index and the Gini index. The standard Gini index (also known as the Gini coefficient) is a measure of discrepancy in a data set and is commonly used in economics to study the distribution of income within a nation. It is sometimes referred to as an inequality index. The World Bank uses it (among other factors) as a basis for the approval of project funding and allocation of aid to countries and maintains a website on Gini index statistics.

The Gini index can be a measure to capture the balance within a tree and compare the overall balance of a random class of trees.

Received 15 March 2016; revision received 15 November 2016.
* Postal address: Winston Churchill High School, Potomac, MD 20854, USA.
Email address: hrishikeshbalaji@gmail.com
** Postal address: Department of Statistics, The George Washington University, Washington, D.C. 20052, USA.
Email address: hosam@gwu.edu

2. Organization

In Section 3 we review the usual Gini index. In Section 4 we introduce some preliminaries about trees, and construct the Gini index of a class in two steps: in Subsection 4.1 we introduce the level index of a tree, and in Subsection 4.2 we build from it a Gini index of a random class of trees. In Subsection 4.3 we provide a computational formula for the level index.

To put these two new indices into perspective, the rest of the paper is devoted to the study of these indices in specific species of random trees. In Section 5 we examine the Gini index of caterpillars, where two random models are considered (the multinomial model in Subsection 5.1 and a model of independent level numbers in Subsection 5.2). We conclude with some remarks in Section 6.

3. The Gini index

The Gini index was first introduced in 1912 by Corrado Gini [7]. The survey [1] provides an excellent exposition to the origin, developments, and uses of this index. This index is the ratio of the area between the Lorenz curve and the 45-degree line to the area under the 45-degree line, where the Lorenz curve is the cumulative distribution function of a nation's wealth, i.e. a graph of the function that gives the percentage of a nation's wealth accounted for by a proportion of the population. The Gini index is equivalent to the relative mean absolute difference of elements of the distribution.

A statistical estimator is defined as follows; see [6]. Suppose that $X_1, \ldots, X_n$ are the observations of a sample of size $n \geq 2$ of independent and identically distributed (i.i.d.) random variables from a common distribution of known mean $\mu > 0$. The Gini index is estimated by

$$G_n = \frac{\sum_{1 \leq i < j \leq n} |X_j - X_i|}{n^2 \mu}. \tag{3.1}$$

If μ is not known, it is replaced by an estimator of it.

4. The Gini index of trees

We propose a Wiener-like topological index, crafted to capture the degree of balance within the strata of rooted trees. We build from it a (distance-based) model of the Gini index for classes of trees.

We deal with rooted trees, i.e. trees in which a node is distinguished as a center or a root. We first establish some terminology. Suppose that we have a rooted tree T, on $|T|$ nodes (i.e. of *size* $|T|$), and we arbitrarily label its nodes distinctly with elements of the set $\{1, \ldots, |T|\}$. Let $D_i(T)$ be the distance from the root (measured in edges) of the node labelled with i. Some sources call this distance the *level*, *depth*, or *altitude* of the node. The distance from the root to a node at the highest altitude is called the *height* of the tree. We call the number of nodes at a level the *level number* as they are termed in [5] and [12].

4.1. The level index of a tree

Inspired by the Wiener index, we propose an index that is unique to structures with a node distinguished as a root. Let T be a given rooted tree with the parameters described above. The classical Wiener index of the tree T is the sum of all distances between pairs of nodes in it. If we let $d_{i,j}(T)$ be the distance between nodes i and j, the index is defined as

$$\sum_{1 \leq i < j \leq |T|} d_{i,j}(T).$$

https://doi.org/10.1017/jpr.2017.28 Published online by Cambridge University Press

To produce a balance-oriented variation of the Wiener index, we replace the $d_{i,j}$ with the absolute difference between the distance of the two nodes i and j from the root, deriving from the Wiener index a new topological tree index called the *level index*, i.e.

$$L(T) = \sum_{1 \le i < j \le |T|} |D_j(T) - D_i(T)|.$$

4.2. The Gini index of a random class of trees

Suppose that we have a random class $\mathcal{T}$ of trees. (By random we refer to a model of randomness that we leave unspecified at this time.) From the level index for each tree, we build a *relative* Gini index for the trees of the class and consider their average to be a Gini index for the entire class. We take this to be a comparative measure of balance between classes.

We treat a class of random trees as a distribution of depths and we take the Gini index of that distribution. Therefore, the relative Gini index mimics the standard Gini index by replacing the X_i in (3.1) with D_i, the node distances from the root. Thus, the numerator of the relative Gini index is replaced with the level index. The depths would then not represent i.i.d. random variables. Therefore, we need to adapt the definition to work for new types of data. Hence, we also need a replacement for μ, as each tree $T \in \mathcal{T}$ has its own average depth, and there is an overall average depth over the entire class. So, we consider the depth of a randomly chosen node in a random rooted tree (note the double randomness) as a representative average depth in the 'average distribution'. Let $D^*_{\mathcal{T}}$ be the depth of a randomly chosen node in a random tree. We have a well-defined notion of $\mathbb{E}[D^*_{\mathcal{T}}]$, as well as a well-defined average tree size $\mathbb{E}[|T|]$, in the class. We can then define the relative Gini index of the tree T within this class to be

$$G_{\mathcal{T}}(T) = \frac{\sum_{1 \le i < j \le |T|} |D_j(T) - D_i(T)|}{(\mathbb{E}[|T|])^2 \mathbb{E}[D^*_{\mathcal{T}}]} =: \frac{L(T)}{(\mathbb{E}[|T|])^2 \mathbb{E}[D^*_{\mathcal{T}}]}. \tag{4.1}$$

We take $G^*_{\mathcal{T}} = \mathbb{E}[G_{\mathcal{T}}(T)]$ as the Gini index of the class $\mathcal{T}$.

Note that if the class $\mathcal{T}$ has only one tree in it (occurring with probability 1), the Gini index for the class is then reduced to a topological Gini index of the tree. We illustrate these concepts by two examples.

Example 4.1. Consider the tree T of Figure 1 as a deterministic family $\mathcal{T}$ of trees in which T occurs with probability 1. The average depth of a node in this tree is

$$\mathbb{E}[D^*_{\mathcal{T}}] = \tfrac{1}{6}(0 + (1 + 1) + (2 + 2 + 2)) = \tfrac{4}{3},$$

and the level index of the tree is

$$L(T) = 1 + 1 + 2 + 2 + 2 + 1 + 1 + 1 + 1 + 1 + 1 = 14,$$

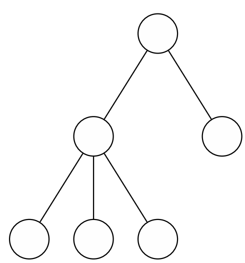

FIGURE 1: A rooted tree with Gini index $\frac{7}{24}$.

https://doi.org/10.1017/jpr.2017.28 Published online by Cambridge University Press

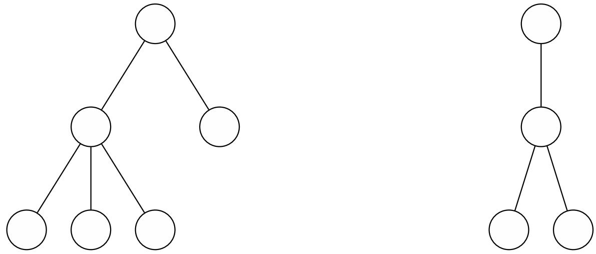

FIGURE 2: A class of random trees: the tree on the left occurs with probability $\frac{1}{3}$, and the tree on the right occurs with probability $\frac{2}{3}$. The Gini index of the class is $\frac{54}{161}$.

yielding the Gini index of the tree

$$G^*_{\mathcal{T}} = G_{\mathcal{T}}(T) = \frac{14}{6^2 \times 4/3} = \frac{7}{24} \approx 0.292.$$

Example 4.2. Consider a family $\mathcal{T}$ of the two random trees of Figure 2 with corresponding probabilities $\frac{1}{3}$ for the tree on the left and $\frac{2}{3}$ for the tree on the right. The tree on the left is that of Example 4.1 with average depth $\frac{4}{3}$ and level index 14. The tree on the right has average depth $\frac{5}{4}$ and level index 7. The average depth across the trees of this class is

$$\mathbb{E}[D^*_{\mathcal{T}}] = \tfrac{4}{3} \times \tfrac{1}{3} + \tfrac{5}{4} \times \tfrac{2}{3} = \tfrac{23}{18}.$$

The average size of trees in this class is

$$\mathbb{E}[|T|] = 6 \times \tfrac{1}{3} + 4 \times \tfrac{2}{3} = \tfrac{14}{3}.$$

The tree on the left has relative Gini index $14/((14/3)^2 \times 23/18) = \frac{81}{161}$, while that of the one on the right is $7/((14/3)^2 \times 23/18) = \frac{81}{322}$. The Gini index of the class is

$$G^*_{\mathcal{T}} = \mathbb{E}[G_{\mathcal{T}}] = \frac{14 \times 1/3 + 7 \times 2/3}{(14/3)^2 \times 23/18} = \frac{54}{161} \approx 0.335.$$

Comparing the Gini index of the two classes in Examples 4.1 and 4.2, we find that the first class is better balanced than the second.

4.3. Computing the level index

Let T be a tree of height h and level numbers $1 = N_0, \dots, N_h$. Nodes at levels i and $i + j$ have a difference in depth of j. Therefore, the level index is

$$L = \sum_{i=0}^{h} \sum_{j=0}^{h-i} j N_i N_{i+j}. \tag{4.2}$$

Technically speaking, L is $L(T)$, a function of T, and so are $h = h(T)$ and $N_i = N_i(T)$ for each i, and so forth. We drop the T for simplicity. In what follows, we use a variety of notation depending on what we believe enhances readability. For example, the level index for a tree T_n from a class of trees all of the same size, say n, can be represented as L_n or $L(T_n)$. Some of the classes we consider have two parameters, such as tree size and height, and it is natural there to use double subscripts for tree parameters.

If we are computing the level index or Gini index of a given tree (i.e. a deterministic class of trees comprised of one tree) we have all we need: all the components in (4.1) and (4.2) are

https://doi.org/10.1017/jpr.2017.28 Published online by Cambridge University Press

determined by the given tree. However, if the tree belongs to a random class, we need to compute $\mathbb{E}[|T|]$ and $\mathbb{E}[D^*_{\mathcal{T}}]$, the average tree size and depth of nodes in the class, for the denominator of the Gini index in (4.1). This cannot be determined without imposing a probability measure $\mathbb{P}$ on a space of trees. For the same class of trees $\mathcal{T}$, the average depth $\mathbb{E}[D^*_{\mathcal{T}}]$ will vary with $\mathbb{P}$. For instance, if the class of trees we are considering consists of one given tree, occurring then with probability 1, $\mathbb{E}[D^*_{\mathcal{T}}]$ is the average depth of a node in *that* tree, as in Example 4.1. However, if the class has several trees in it, each having its own positive probability, $\mathbb{E}[D^*_{\mathcal{T}}]$ is the average depth of a node *across the entire family* of trees, each contributing according to its assigned probability, as in Example 4.2.

5. The Gini index of random caterpillars

Let P_h be a rooted path made up of the h nodes $v_0, \dots, v_{h-1}$, with the root v_0 being an endpoint, and v_{h-1} being the only leaf at the other end. We call that path the *spine*. Suppose that X_i nodes for $i = 1, \dots, h$ join the structure by attaching themselves as children of v_{i-1}. Thus, for $i = 1, \dots, h-1$, level i has $N_i = X_i + 1$ nodes (the 1 accounts for one sibling on the spine), but level h has X_h nodes only. Such a tree is known in the literature as a caterpillar [9]. This source poetically describes a tree from this species as 'a tree which metamorphoses into a path, when its cocoon of endpoints is removed'. The chemistry literature refers to caterpillars as Gutman trees or benzenoid trees [2]. A topological perspective of caterpillars was given in [8] and [11].

It turns out that the Gini index of a caterpillar depends critically on the interplay of the level numbers and their rates of growth. We shall discuss a case in which the level numbers are identically distributed but dependent, and several cases where they are independent.

5.1. A multinomial model

Let $\operatorname{bin}(n, p)$ represent a binomial random variable that counts the number of successes in n i.i.d. experiments, with probability of success p, and let $\operatorname{multinom}(n, p_1, \dots, p_r)$ be the multinomial vector of shares obtained by dropping n distinguishable balls in r distinguishable boxes, with the ith box attracting the balls with probability p_i for $i = 1, \dots, r$. Suppose that n organisms attack the spine and each predator chooses any of the h nodes of the spine with equal probability. Thus, X_i is distributed like $\operatorname{bin}(n, 1/h)$, and collectively $(X_1, \dots, X_h)$ is a

$$\operatorname{multinom}\Big(n, \underbrace{\frac{1}{h}, \dots, \frac{1}{h}}_{h \text{ times}}\Big)$$

vector.

We can think of the emerging caterpillar as an algae as it represents colonies of biological organisms attacking a natural resource. Note that here the Gini index (of the class) is a function of h and n and so are the other parameters such as the level index (of a tree in the class); it is natural to use double subscripts for such parameters. To keep the notation simple, we do not use additional parameters for the N_i. We then have

$$\begin{aligned}\mathbb{E}[L_{n,h}] &= \sum_{i=0}^{h-1} \sum_{j=0}^{h-i-1} j\mathbb{E}[N_i N_{i+j}] + \sum_{i=0}^{h-1} (h-i)\mathbb{E}[N_i N_h] \\ &= \sum_{j=1}^{h-1} j\mathbb{E}[N_0(X_j+1)] + \sum_{i=1}^{h-1} \sum_{j=1}^{h-i-1} j\mathbb{E}[(X_i+1)(X_{i+j}+1)] + h\mathbb{E}[N_0 X_h]\end{aligned}$$

https://doi.org/10.1017/jpr.2017.28 Published online by Cambridge University Press

$$+\sum_{i=1}^{h-1}(h-i)\mathbb{E}[(X_i+1)X_h]$$

$$= (\mathbb{E}[X_1]+1)\sum_{j=1}^{h-1} j + (\mathbb{E}[X_1X_2]+2\mathbb{E}[X_1]+1)\sum_{i=1}^{h-1}\sum_{j=1}^{h-i-1} j + h\mathbb{E}[X_1]$$

$$+ (\mathbb{E}[X_1X_2]+\mathbb{E}[X_1])\sum_{j=1}^{h-1} j,$$

where in the last step we used the identical distribution of the X_i and pairs of them. The mean and covariance structure of a multinom$(n, p_1, \dots, p_r)$ vector is well known; we have

$$\mathbb{E}[X_1X_2] = \frac{n(n-1)}{h^2}, \qquad \mathbb{E}[X_1] = \frac{n}{h}.$$

We can now compute the average level index:

$$\mathbb{E}[L_{n,h}] = \frac{1}{6h}((h^2-1)n^2 + (2h^3-h^2+4h+1)n + h^4 - h^2).$$

Further, the average depth of a randomly selected node across this class of caterpillars is

$$\mathbb{E}[D^*_{n,h}] = \frac{1}{n+h}\sum_{i=0}^{h} i\mathbb{E}[N_i] = \frac{1}{n+h}\bigg(\sum_{i=1}^{h-1} i(\mathbb{E}[X_i]+1) + h\mathbb{E}[X_h]\bigg) = \frac{n}{n+h} + \frac{h-1}{2}.$$

The Gini index of the class follows as the average across all multinomial caterpillars, i.e.

$$G^*_{n,h} = \mathbb{E}[G_{n,h}] = \frac{((h^2-1)n^2 + (2h^3-h^2+4h+1)n + h^4 - h^2)/6h}{(n+h)^2(n/(n+h)+(h-1)/2)}.$$

After asymptotic simplification, we obtain (as $n \to \infty$)

$$\lim_{n\to\infty} G^*_{n,h} = \frac{h-1}{3h}.$$

In addition, if $h \to \infty$, we obtain

$$\lim_{h\to\infty}\lim_{n\to\infty} G^*_{n,h} \to \tfrac{1}{3}.$$

We note that the asymptotics of the average Gini index for multinomial caterpillars do not fill the spectrum of possible values. The first few values for $\lim_{n\to\infty} G^*_{n,h}$ are $0, \frac{1}{6}, \frac{2}{9}, \frac{1}{4}, \frac{4}{15}$ for $h = 1, 2, 3, 4, 5$, and at $h = 300$ this limit is $\frac{299}{900} \approx 0.332\,222$, already at 99.6666% of the limit $\frac{1}{3}$, when h itself goes to ∞. For multinomial caterpillars of finite height, we see that the limit of the class Gini index falls in a discrete set of values.

5.2. A model of independent level numbers

In the multinomial model of caterpillars, we assumed a total of n organisms attacking the spine. This imposition on the total number of organisms introduces dependency among the level numbers. In this section we employ a model in which the level numbers are independent. The model is pertinent to the formation of computer networks, where there is a common ‘bus’

https://doi.org/10.1017/jpr.2017.28 Published online by Cambridge University Press

or 'hub' of servers to which users subscribe to receive internet services. Suppose that there is also a main server at one endpoint of the hub (root of the path) considered as the master server.

The Gini index of such a network may be correlated with the speed of communication within the network. It is natural to use h as a subscript for the level index, the depth of a randomly chosen node and the Gini index. We do not use additional parameters for the N_i. Let X_i for $i = 1, \dots, h$ be independent random variables. Suppose that level i attracts X_i users. Thus, the level numbers in the network are

$$N_0 = 1, \qquad N_i = X_i + 1 \quad \text{for } i = 1, \dots, h-1, \qquad N_h = X_h.$$

To ease some of the forthcoming computations, we impose a regularity condition. Namely, we assume that

$$\sum_{i=1}^{h} \mathrm{var}[N_i] = o(h^{2p+2}). \tag{5.1}$$

One main result of this paper is the following.

Theorem 5.1. *Consider a class of caterpillars with independent random level numbers. We assume that the distribution of the level number N_i has average value $\mu_i \sim K i^p$ for some power $p \ge 0$ and constant $K > 0$, and satisfies the regularity condition (5.1). This class of random caterpillars has a Gini index $G_h^* \to 1/(2p+3)$ as $h \to \infty$.*

Proof. All asymptotics in this proof are taken as $h \to \infty$. Let $Y_h = \sum_{i=0}^{h} N_i / h^{p+1}$. By independence and the regularity condition (5.1), we have

$$\mathrm{var}[Y_h] = \sum_{i=0}^{h} \frac{\mathrm{var}[N_i]}{h^{2p+2}} = \frac{1}{h^{2p+2}} \times o(h^{2p+2}) \to 0.$$

Subsequently, by Chebyshev's inequality, for any $\varepsilon > 0$, we have

$$\mathbb{P}(|Y_h - \mathbb{E}[Y_h]| > \varepsilon) \le \frac{\mathrm{var}[Y_h]}{\varepsilon^2} \to 0.$$

Therefore, we have

$$Y_h - \mathbb{E}[Y_h] \xrightarrow{\mathbb{P}} 0. \tag{5.2}$$

We further have

$$\mathbb{E}[Y_h] \sim \sum_{i=0}^{h} \frac{K i^p}{h^{p+1}} = \frac{K h^{p+1} + O(h^p)}{(p+1)h^{p+1}} \to \frac{K}{p+1}.$$

Adding this latter convergence relation to (5.2), we obtain

$$\sum_{i=0}^{h} \frac{N_i}{h^{p+1}} \xrightarrow{\mathbb{P}} \frac{K}{p+1}. \tag{5.3}$$

Similarly, we have

$$\mathrm{var}\Bigg[\sum_{i=0}^{h} \frac{i N_i}{h^{p+2}}\Bigg] = \sum_{i=0}^{h} i^2 \,\mathrm{var}\Bigg[\frac{N_i}{h^{p+2}}\Bigg] \le h^2 \sum_{i=0}^{h} \frac{1}{h^2} \,\mathrm{var}\Bigg[\frac{N_i}{h^{p+1}}\Bigg] \to 0,$$

https://doi.org/10.1017/jpr.2017.28 Published online by Cambridge University Press

and, by almost identical steps, we have

$$\sum_{i=0}^{h} \frac{iN_i}{h^{p+2}} \xrightarrow{\mathbb{P}} \frac{K}{p+2}. \tag{5.4}$$

Using (5.3) and (5.4), we obtain

$$\frac{\sum_{i=0}^{h} iN_i}{h\sum_{i=0}^{h} N_i} \xrightarrow{\mathbb{P}} \frac{p+1}{p+2}.$$

But, then we have

$$0 \le \frac{\sum_{i=0}^{h} iN_i}{h\sum_{i=0}^{h} N_i} \le 1,$$

and the boundedness ensures that convergence in probability implies convergence in moments. In other words, we have

$$\frac{1}{h}\mathbb{E}[D_h^*] = \mathbb{E}\bigg[\frac{\sum_{i=0}^{h} iN_i}{h\sum_{i=0}^{h} N_i}\bigg] \to \frac{p+1}{p+2}.$$

We have calculated an important component in the denominator of the Gini index. The other one is

$$\mathbb{E}[S_h] = \sum_{i=0}^{h} \mathbb{E}[N_i] = 1 + \sum_{i=1}^{h} \mu_i \sim \sum_{i=0}^{h} Ki^p \sim \frac{Kh^{p+1}}{p+1}.$$

For the numerator of the Gini index, we need the average level index, i.e.

$$\begin{aligned}
\mathbb{E}[L_h] &= \sum_{i=0}^{h}\sum_{j=0}^{h-i} j\mathbb{E}[N_i N_{i+j}] \\
&= \sum_{j=0}^{h} j\mathbb{E}[N_j] + \sum_{i=1}^{h-1}\sum_{j=0}^{h-i} j\mathbb{E}[N_i N_{i+j}] \\
&= \sum_{j=1}^{h} j\mu_j + \sum_{i=1}^{h-1}\sum_{j=0}^{h-i} j\mu_i\mu_{i+j} \quad \text{(by independence)} \\
&\sim \sum_{j=1}^{h-1} Kj^{p+1} + \sum_{i=1}^{h-1}\sum_{j=0}^{h-i} K^2 j i^p (i+j)^p \\
&\sim \frac{K^2 h^{2p+3}}{(p+1)(p+2)(2p+3)},
\end{aligned}$$

and we obtain the last asymptotic relation by comparing the sums to integrals.

We can compose the asymptotic Gini index from the following components:

$$G_h^* \sim \bigg[\frac{K^2 h^{2p+3}}{(p+1)(p+2)(2p+3)}\bigg]\bigg[\bigg(\frac{Kh^{p+1}}{p+1}\bigg)^2\bigg(\frac{p+1}{p+2}h\bigg)\bigg]^{-1} \to \frac{1}{2p+3}. \qquad \square$$

5.3. An illustrative example

We discuss an instance of Theorem 5.1. Note that p can be any nonnegative real number, rendering the possible limiting Gini index values fill the range $(0, \frac{1}{3}]$, as opposed to the usual economics Gini index which fills the entire [0, 1] interval.

Example 5.1. Take the case

$$X_i = \left\lfloor \left(3 + \frac{(-1)^i}{i}\right) \mathrm{Poi}(i^{5/2}) + \mathrm{bin}(i^2, p) + 30 \ln i + 82 \right\rfloor,$$

where $\mathrm{Poi}(i^{5/2})$ is a Poisson random variable with parameter $i^{5/2}$. Here, we have $\mu_i \sim 3i^{5/2}$. The limiting Gini index is $\frac{1}{8}$.

6. Concluding remarks

We introduced a notion of the Gini index for a class of rooted trees. The concept extends to all connected graphs with a distinguished vertex, which is a topic for future research. The usual Gini index is known to have range [0, 1]. In contrast, we found that the limiting Gini index of certain random caterpillars falls in the reduced range $(0, \frac{1}{3}]$.

The independence of the data in the usual Gini index makes it possible in a poor country for a few outliers of very wealthy individuals to catapult the index to values close to 1. By contrast, in trees if there are outliers (nodes at very high altitudes), they cannot be there alone, there must be other nodes nearby to hold the tree together. This dependence reduces the influence of outliers and results in a smaller range of Gini indices for trees.

The Gini index proposed in this paper is distance-based. One can think of a Gini index comparing node properties in a tree other than their depths, such as their degrees. This is the subject of a future research.

Acknowledgements

We are indebted to a dear father and colleague, Dr Srinivasan Balaji, for many discussions and car rides that helped facilitate and enhance our thinking. We also thank Dr Hossam Elgindy, Dr Ralph Neininger, and Dr Stephan Wagner for critique that helped us tighten loose ends and improve the exposition. We are thankful to the anonymous referees for suggesting useful reorganization of the manuscript. One referee has graciously suggested an alternative elegant shortcut to the submitted proof of Theorem 5.1, which simplified the proof and the regularity conditions. The present proof is based on the referee's ideas. We include it with thanks.

References

[1] Ceriani, L. and Verme, P. (2012). The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *J. Econom. Inequality* **10,** 421–443.

[2] El-Basil, S. (1990). Caterpillar (Gutman) trees in chemical graph theory. In *Advances in the Theory of Benzenoid Hydrocarbons* (Topics Current Chem. **153**), Springer, Berlin, pp. 273–289.

[3] Feng, Q. and Hu, Z. (2011). On the Zagreb index of random recursive trees. *J. Appl. Prob.* **48,** 1189–1196.

[4] Feng, Q., Mahmoud, H. M. and Panholzer, A. (2008). Limit laws for the Randić index of random binary tree models. *Ann. Inst. Statist. Math.* **60,** 319–343.

[5] Flajolet, P. and Prodinger, H. (1987). Level number sequences for trees. *Discrete Math.* **65,** 149–156.

[6] Gastwirth, J. L. (1972). The estimation of the Lorenz curve and Gini index. *Rev. Econom. Statist.* **54,** 306–316.

[7] Gini, C. (1912). *Veriabilità e Mutabilità*. Cuppini, Bologna.

[8] Gutman, I. (1977). Topological properties of benzenoid systems. *Theoretica Chimica Acta* **45,** 309–315.

[9] Harary, F. and Schwenk, A. J. (1973). The number of caterpillars. *Discrete Math.* **6,** 359–365.

[10] Neininger, R. (2002). The Wiener index of random trees. *Combinatorics Prob. Comput.* **11,** 587–597.

[11] Schmuck N. S., Wagner, S. G. and Wang, H. (2012). Greedy trees, caterpillars, and Wiener-type graph invariants. *MATCH Commun. Math. Comput. Chem.* **68,** 273–292.

[12] Tangora, M. C. (1991). Level number sequences of trees and the lambda algebra. *Europ. J. Combinatorics* **12,** 433–443.