Probability in the Engineering and Informational Sciences, **34**, 2020, 157–171.
doi:10.1017/S0269964819000044

DEGREE-BASED GINI INDEX FOR GRAPHS

CARLY DOMICOLO AND HOSAM MAHMOUD
Department of Statistics, The George Washington University, Washington, D.C. 20052, USA
E-mail: carlydomicolo@gwu.edu; hosam@gwu.edu

In Balaji and Mahmoud [1], the authors introduced a distance-based Gini index for rooted trees. In this paper, we introduce a degree-based Gini index (or just simply degree Gini index) for graphs. The latter index is a topological measure on a graph capturing the proximity to regular graphs. When applied across the random members of a class of graphs, we can identify an average measure of regularity for the class. Whence, we can compare the classes of graphs from the vantage point of closeness to regularity.

We develop a simplified computational formula for the degree Gini index and study its extreme values. We show that the degree Gini index falls in the interval $[0, 1)$. The main focus in our study is the degree Gini index for the class of binary trees. Via a left-packing transformation, we show that, for an arbitrary sequence of binary trees, the Gini index has inferior and superior limits in the interval $[0, 1/4]$. We also show, via the degree Gini index, that uniform rooted binary trees are more regular than binary search trees grown from random permutations.

Keywords: chemical tree, combinatorial probability, Gini index, random tree, topological index, wiener index

1. INTRODUCTION

A topological index of a graph quantifies it by turning its structure into a number. The idea behind capturing structures in numbers is to be able to compare graphs according to certain criteria. Particular values of a topological index may be desirable. For example, the height of a rooted tree, that is, the length of the longest root-to-node path, is a topological index. When the trees are used for data storage, among trees storing the same amount of data, trees of smaller height are preferred for fast data retrieval.

No single topological index adequately describes all the facets of a graph. That is why there is always a need to introduce new indices, hoping that combinations of these indices portray a more accurate picture. Examples for indices that have been introduced for trees include a (distance-based) Gini index [1], Zagreb index [7], Randič index [8], and Wiener index [17], among others. Topological indices of more general graphs appear in [6], which discusses the clustering coefficient of scale-free random graphs, with machine learning applications in [14]. The source [22] deals with the total weight of a random Apollonian network. These are only instances, to name a few in a huge literature on a variety of different families of graphs and indices therein. Many more molecular indices are compiled in [20].

 0269-9648/19 $25.00

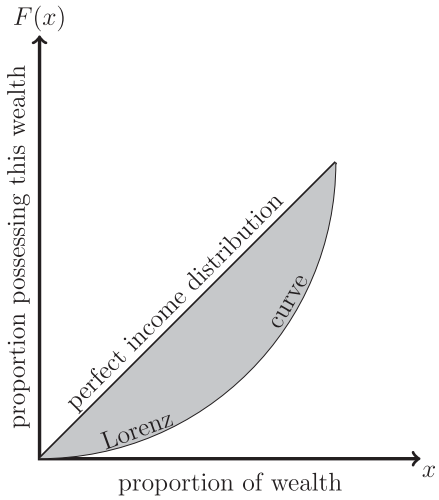

FIGURE **1.** The Gini index as it relates to Lornez curve.

In [1], the authors introduced a distance-based Gini index for rooted trees. The quest for more topological indices is our motivation to propose one more degree-based Gini index. The distance-based Gini index can be a measure to capture the balance within a tree and compare the overall balance of a random class of trees. The degree-based Gini index we propose in this paper is for graphs, not only trees, and serves a similar purpose, but characterizes a different property, namely how close a graph is to "regularity."[1]

2. OVERVIEW OF THE STANDARD GINI INDEX

The Gini index (or coefficient) is a measure of inequality of a distribution, ranging on a scale of 0–1 and relying on the Lorenz curve. The Lorenz curve is an economics plot that represents income or wealth inequality as the cumulative distribution of the income or wealth (or some other economics criterion of interest) of a nation. Viewed as a function $F(x)$, for $x \in [0, 1]$, the Lorenz curve is the proportion $F(x)$ in the population in a nation of individuals possessing proportion at most x of the total wealth of that nation. As a cumulative distribution function, it is graphed as the cumulative percentage of the total income of a nation against the cumulative percentage of the population of the nation possessing this percentage of the wealth, on an increasing scale; see Figure 1. The Gini index is a ratio, with the numerator measuring the area between the Lorenz curve, the actual distribution of wealth in a nation, and the uniform distribution (the 45° line), considered in socialist economics as a perfect distribution (all individuals have the same share of the wealth); see the shaded area in Figure 1. The denominator is the area under the uniform distribution line.

Corrado Gini presented this index in 1912 in his book "Variabilità e Mutabilità" [2,11]. Organizations such as the World Bank and the United Nations use the Gini index to study countries and determine the amount of aid to provide to these areas, which can be found in the records on their websites at

https://data.worldbank.org/indicator/SI.POV.GINI
http://hdr.undp.org/en/content/income-gini-coefficient

[1] A graph is said to be d-regular, if all its vertices are of degree d.

https://doi.org/10.1017/S0269964819000044 Published online by Cambridge University Press

A statistical estimator of Gini index is defined as follows [10]. Suppose $X_1, \ldots, X_n$ are the observations of a sample of size $n \geq 1$ independent, identically distributed (i.i.d.) random variables from a common distribution of known mean $\mu > 0$. The Gini index is estimated by

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

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

3. A DEGREE-BASED GINI INDEX FOR GRAPHS

Suppose we have a graph $H = (V, E)$, where V and E are its sets of vertices and edges, respectively. In this paper, we use $|\mathcal{E}|$ to indicate the cardinality of the set $|\mathcal{E}|$. Let the vertices be arbitrarily labeled with distinct elements of the set $\{1, 2, \ldots, |V|\}$. Let $D_v(H)$ be the degree of node $v \in V$ in H. Often, we shall refer to $|V|$ as n. Toward simpler notation, we shall occasionally refer to $D_v(H)$ as D_v.

3.1. A degree-based Gini index as a toplogical measure of a graph

We define a degree-based Gini index as a toplogical measure of a graph. Given a nonempty graph $H = (V, E)$, with $|V| = n$ vertices, let $D^*(H)$ be the degree of a random (uniformly chosen) node in it. Let the average node degree in H be $\mathbb{E}[D^*(H)]$. That is, we have

$$\mathbb{E}[D^*(H)] = \frac{1}{n} \sum_{i=1}^{n} D_i(H).$$

We now define the degree-based Gini index of H as

$$G(H) = \frac{\sum_{1 \leq i < j \leq n} |D_j(H) - D_i(H)|}{n^2 \, \mathbb{E}[D^*(H)]}. \tag{2}$$

In what follows, we simply call this measure the degree Gini index of the graph. We arrived at this definition by replacing X_i in (1) by D_i and replacing μ by an average node degree in the graph.

Example 3.1: Consider the graph H of Figure 2. In this graph, we have

$$D_1 = 2, \quad D_2 = 2, \quad D_3 = 3, \quad D_4 = 7.$$

The average degree in this graph is

$$\mathbb{E}[D^*(H)] = \frac{2 + 2 + 3 + 7}{4} = \frac{7}{2},$$

yielding the degree Gini index

$$G(H) = \frac{|2-2| + |3-2| + |7-2| + |3-2| + |7-2| + |7-3|}{4^2 \times \frac{7}{2}} = \frac{16}{56} \approx 0.2857.$$

https://doi.org/10.1017/S0269964819000044 Published online by Cambridge University Press

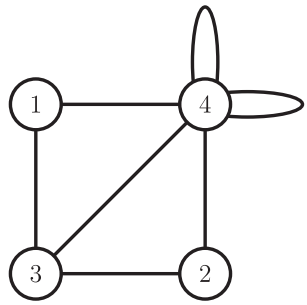

FIGURE **2.** A graph with degree Gini index 16/56.

3.2. A degree-based Gini index for a random class of graphs

Suppose we have a class $\mathcal{H}$ of *random* graphs.[2] We define a relative degree-based Gini index for a random graph $H \in \mathcal{H}$. The *relative degree Gini index* of $H \in \mathcal{H}$ is intended to mimic the standard Gini index. It is like the topological degree Gini index, but measures the sum of the absolute degree differences in H against an average size and an average node degree in the class, instead of comparing these differences with the parameters of H. This relative measure is achieved by replacing the X_i's in (1) with $D_i(H)$'s, the node degrees of H. In general, the degrees in a random class are not i.i.d. random variables. Therefore, we need to adapt the definition to work for random graphs. There is no general n if the class considered comprises graphs of different sizes. However, we have an average size that we can use as a replacement for n. We replace n with $\mathbb{E}|V|$.

We need a replacement for μ, too, as each graph $H \in \mathcal{H}$ has its own average node degree. There is an overall average degree over the entire class, which we propose for the replacement. That is, we consider the degree of a randomly chosen node in a random graph in the class. There is double randomness that can be viewed as a hierarchical model: First generate a graph H from the class $\mathcal{H}$, according to its relative frequency (probability) in $\mathcal{H}$, then choose a node of H uniformly at random. Let $D^*_{\mathcal{H}}$ be the degree of a randomly chosen node in a random graph H according to the hierarchical model.

We can now define the relative degree Gini index of the graph $H = (V, E)$ within the class $\mathcal{H}$ to be

$$G_{\mathcal{H}}(H) = \frac{\sum_{1 \le i \le j \le |V|} |D_j(H) - D_i(H)|}{\mathbb{E}^2|V|\,\mathbb{E}[D^*_{\mathcal{H}}]}.$$

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

Note that if the class $\mathcal{H}$ has only one graph in it (occurring with probability 1), the degree Gini index for the class is then reduced to the topological degree Gini index of that one graph. We illustrate these concepts by an example.

Example 3.2: Consider a family $\mathcal{H}$ of the two random graphs in Figure 3, with corresponding probabilities 1/3 for the graph on the left, and 2/3 for the graph on the right. The graph on the left is that of Example 3.1.

[2] One can define many probability measures on a measurable space. In the sequel, we specify the measure as needed.

https://doi.org/10.1017/S0269964819000044 Published online by Cambridge University Press

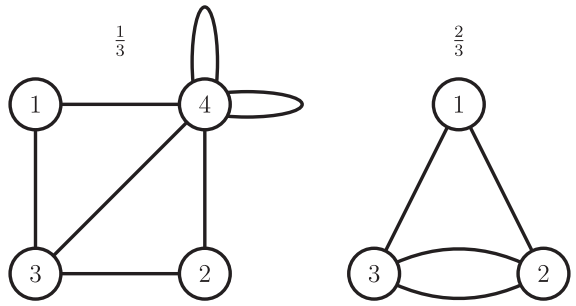

FIGURE **3.** A class of two random graphs. The top row of numbers gives the probabilities of these graphs.

The graph on the right has average degree 8/3. The average degree across this class of graphs is

$$\mathbb{E}[D^*_{\mathcal{H}}] = \frac{7}{2} \times \frac{1}{3} + \frac{8}{3} \times \frac{2}{3} = \frac{53}{18}.$$

The average size of a graph in this class is

$$\mathbb{E}|V| = 4 \times \frac{1}{3} + 3 \times \frac{2}{3} = \frac{10}{3}.$$

The graph on the left has relative degree Gini index $16/(10/3)^2 \times 53/18 = 648/1325$, while that of the one on the right has relative index $2/(10/3)^2 \times 53/18 = 81/1325$. The degree Gini index of the class is

$$G^*_{\mathcal{H}} = \mathbb{E}[G_{\mathcal{H}}] = \frac{16 \times 1/3 + 2 \times 2/3}{(10/3)^2 \times 53/18} = \frac{54}{265} \approx 0.20377.$$

Comparing the degree Gini index of the two classes in Examples 3.1 and 3.2, we find that the second class is closer to regularity than the first. The reason is that when we created the random class in Example 3.2, we added to the graph in Example 3.1 a graph that is closer to regularity and assigned to it a higher probability.

3.3. A simplified computational formula

We shall find it convenient in many cases to label the nodes distinctly with elements of the set $\{1, \ldots, |V|\}$ in a manner such that $D_i(H)$ appear in non-decreasing order, for $i = 1, \ldots, |V|$. For a graph of size $|V| = n$, the raw form of the index in (2) requires the computation of $\binom{n}{2}$ differences in $O(n^2)$ time, and the calculation in the denominator requires only $\Theta(n)$ time,[3] with an overall $O(n^2)$ computing cost.

By the convention of labeling the nodes in nondecreasing order according to their degrees, we have

$$D_1 \leq D_2 \leq D_3 \leq \cdots \leq D_n;$$

the nodes in the graphs of Figure 2 are labeled in this canonical order. This yields a simplified degree Gini index formula, as in Thon's axiomatization of the standard Gini index [19].

[3] A function $f(n)$ is said to be $\Theta(g(n))$, if there exist two positive real constants c_1 and c_2 and a positive integer n_0, such that $c_1|g(n)| \leq |f(n)| \leq c_2|g(n)|$, for all $n \geq n_0$.

https://doi.org/10.1017/S0269964819000044 Published online by Cambridge University Press

LEMMA 3.1: *For a graph H of size n with nodes canonically labeled $1, 2, \ldots, n$ and node degrees D_i, for $i = 1, 2, \ldots, n$, the degree Gini index is*

$$G(H) = \frac{\sum_{i=1}^{n}(2i - n - 1)D_i(H)}{n\sum_{i=1}^{n} D_i(H)}.$$

PROOF: In view of the canonical labeling, we can remove the unwieldy absolute value signs, by taking the differences $D_j - D_i$, with $j > i$. The numerator becomes

$$\begin{aligned}\sum_{1\le i\le j\le n} |D_j - D_i| &= \sum_{1\le i\le j\le n} (D_j - D_i)\\ &= \sum_{j=1}^{n}\sum_{i=1}^{j} D_j - \sum_{i=1}^{n}\sum_{j=i}^{n} D_i\\ &= \sum_{j=1}^{n} jD_j - \sum_{i=1}^{n}(n - i + 1)D_i\\ &= \sum_{i=1}^{n}(2i - n - 1)D_i.\end{aligned}$$

It follows that

$$G(H) = \frac{\sum_{i=1}^{n}(2i - n - 1)D_i(H)}{n^2\left(\frac{1}{n}\sum_{i=1}^{n} D_i(H)\right)},$$

which simplifies to the stated formula. ■

By Lemma 3.1, we have an expression for the degree Gini index computable in $O(n)$ time.

4. EXTREMAL DEGREE GINI VALUES

The standard Gini index falls in the interval [0,1]. The extremes are attainable. Suppose we have a population of size $n \ge 1$. The smallest index value 0 corresponds to a population of equal attributes (in the context of economics, all members of the population have the same income, a totally uniform society, ideal socialism). The largest index value 1 corresponds to a population in which one person has all the wealth of the nation, while all others possess nothing, an absolute oligarchy. The degree Gini index has a similar range, as we demonstrate next.

THEOREM 4.1: *Let H_n be a graph of size $n \ge 1$. We then have*

$$0 \le G(H_n) \le 1 - \frac{1}{n},$$

and the bounds are attainable.

Letting $\mathcal{H}$ be the class of all graphs of size $n \ge 1$, and $\widetilde{H}$ be a graph in the class, we have

$$0 \le G(\widetilde{H}) < 1.$$

https://doi.org/10.1017/S0269964819000044 Published online by Cambridge University Press

PROOF: Within the class of graphs of size n, any regular graph has degree Gini index 0, establishing an attainable lower bound.

On the other hand, the computational formula in Lemma 3.1 provides us with an upper bound. Let H_n be a graph of size n. We then have

$$\begin{aligned} G(H_n) &= \frac{\sum_{i=1}^{n}(2i-n-1)D_i(H_n)}{n\sum_{i=1}^{n}D_i(H_n)} \\ &\leq \frac{\sum_{i=1}^{n}(2n-n-1)D_i(H_n)}{n\sum_{i=1}^{n}D_i(H_n)} \\ &= 1-\frac{1}{n}. \end{aligned} \tag{3}$$

A quick computation shows that the upper bound $1-1/n$ (for graphs of size n) is attained by a graph of size n, with $n-1$ isolated nodes (of degree 0 each), and one node with at least one loop on it.[4]

We see from (3) that the degree Gini index of any graph falls in the interval [0, 1), and there exist graphs of degree Gini index arbitrarily close to 1. ■

5. BINARY TREES

As an illustration of the utility of the degree Gini index, we use it to compare two classes of binary trees. A *binary tree* is a structure of a finite number of nodes and edges. One special node is recognized as the *root*. The tree is either empty or has nodes, each having at most two children.

In a binary tree, level ℓ can hold at most 2^ℓ nodes. We say level ℓ is *full*, if the tree has all 2^ℓ nodes on level ℓ. The tree is said to be *complete*, if all levels are full, except possibly the highest. Otherwise, the tree is said to be *incomplete*. Note that level ℓ cannot be full, unless all the lower levels (levels $0, 1, \ldots, \ell-1$) are full, too.

5.1. Intensity of left packing in a complete binary tree

The arrangement of the nodes on the highest level of a complete tree plays a role in determining extremal degree Gini index values. We introduce the concept of the intensity of left packing to help us navigate through some proofs.

Let h_n be the height of a complete binary tree of size n, and x_n be the number of leaves on the highest level. Number the 2^{h_n} possible insertion positions on the highest level with $\{1, 2, \ldots, 2^{h_n}\}$ from right to left. We can now describe the arrangement of the positions of the x_n nodes (from right to left on level h_n) with a sequence of increasing numbers from the set $\{1, 2, \ldots, 2^{h_n}\}$ specifying the positions of the x_n nodes; we call that descriptor the *left-packing intensity* of the complete tree. When more of the x_n nodes appear further to the left, we consider this more left packing. So, left packing can be described by viewing the intensity of left packing of the x_n nodes on level h_n lexicographically; the higher (lexicographically) the intensity, the larger the left packing we have. For example, in a complete tree of size 10, we have $h_{10} = 3$, and three nodes appear on level 3 (i.e., $x_{10} = 3$). If they appear in rightmost positions, the left-packing intensity is (3, 2, 1), and if they appear in leftmost positions, the left-packing intensity is (8, 7, 6).

[4] In a graph (V, E), an edge is called a loop, if its two end vertices are the same.

https://doi.org/10.1017/S0269964819000044 Published online by Cambridge University Press

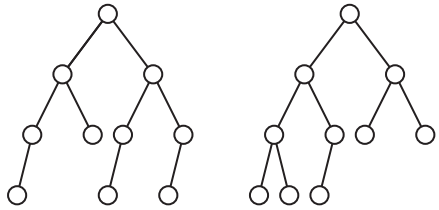

FIGURE **4.** (a) A complete binary tree; (b) Its left-packed counterpart.

We call a complete binary tree *leftmost packed*, when all the leaves on the highest level appear at the leftmost positions, that is, when the tree has the highest left-packing intensity–the tuple $(2^{h_n}, 2^{h_n} - 1, \ldots, 2^{h_n} - x_n + 1)$. Figure 4 shows a complete binary tree (a) with left-packing intensity (8, 4, 2) and a leftmost-packed complete binary tree (b) of the same size with left-packing intensity (8, 7, 6).

5.2. Gini index of binary trees

Suppose T_n is a binary tree on n nodes that has n_1 nodes of degree 1 (i.e., leaves), n_2 nodes of degree 2, and n_3 nodes of degree 3, with $n_1 + n_2 + n_3 = n$. The degree Gini index of T_n is

$$G(T_n) = \frac{n_1 n_2 + n_2 n_3 + 2 n_1 n_3}{n^2 \times (n_1 + 2n_2 + 3n_3)/n}. \tag{4}$$

LEMMA 5.1: *The leftmost-packed complete binary tree has Gini index that is at least that of any other complete binary tree of the same size.*

PROOF: Let T_n be a complete binary tree with n_i nodes of degree i, for $i = 1, 2, 3$. Suppose that the leaves at the highest level are not leftmost packed. There is at least one node x on the highest level with a "gap" to its left, that is, there is a node v on the penultimate level to the left of x with degree 1 or 2 with a vacant insertion position under it. Depending on the degree of v, we move a node or two on the highest level to the left to be a child or children of v.

Let y be the parent of x in T_n. Suppose the degrees of v and y are respectively $D_v \in \{1, 2\}$ and $D_y \in \{2, 3\}$. As we shall see, the same or higher left-packing intensity will be obtained depending on the pair (D_v, D_y) in T_n. Let us call the new complete tree T'_n. Suppose T'_n has n'_i nodes of degree i, for $i = 1, 2, 3$.

In the two cases when the pair (D_v, D_y) is (1, 2) or (1, 3), we transfer all the children (one or two) of y in T_n to become children of v in T'_n. In these two cases the transformation from T_n into T'_n does not change the node counts, and we have $n'_i = n_i$ nodes of degree i, for $i = 1, 2, 3$. Thus, the trees T_n and T'_n have the same degree Gini index.

When the degree pair (D_v, D_y) is (2, 3), the case is similar, we transfer only x to become a second child of v, inducing the same node counts $n'_i = n_i$ nodes of degree i, for $i = 1, 2, 3$, and, the trees T_n and T'_n have the same degree Gini index.

Next, consider the only remaining case, when $(D_v, D_y) = (2, 2)$. Note that $n_2 \geq 2$. The node y has degree 2 in T_n and is now only a leaf of degree 1 in T'_n, inducing the change $n'_1 = n_1 + 1$. Two nodes of degree 2 in T_n are now of different degrees, and $n'_2 = n_2 - 2$,

https://doi.org/10.1017/S0269964819000044 Published online by Cambridge University Press

$n_3' = n_3 + 1$. We compute

$$\begin{aligned} G(T_n') &= \frac{n_1'n_2' + n_2'n_3' + 2n_1'n_3'}{n(n_1' + 2n_2' + 3n_3')} \\ &= \frac{(n_1+1)(n_2-2) + (n_2-2)(n_3+1) + 2(n_1+1)(n_3+1)}{n((n_1+1) + 2(n_2-2) + 3(n_3+1))} \\ &= \frac{(n_1n_2 + n_2n_3 + 2n_1n_3) + 2(n_2-1)}{n(n_1 + 2n_2 + 3n_3)} \\ &= G(T_n) + \frac{2(n_2-1)}{n(n_1 + 2n_2 + 3n_3)} \\ &> G(T_n). \end{aligned}$$

In all four possible pair degree cases of (D_v, D_y), we have a possible transformation from a complete binary tree T_n into a complete binary tree T_n' with a higher left-packing intensity and a degree Gini index for T_n' that is at least as large as that of T_n.

If T_n' is not the leftmost-packed tree, perform another similar transformation on T_n' to obtain complete binary trees T_n'' of higher left-packing intensity and of the same or higher degree Gini index. Keep making these transformations until it is no longer possible. At this point, the complete tree has $\lfloor x_n/2 \rfloor$ nodes at positions $2^{h_n}, \ldots, 2^{h_n} - 2\lfloor x_n/2 \rfloor + 1$, and if x_n is odd, one additional node appears at either position $2^{h_n} - x_n + 1$ or $2^{h_n} - x_n$. If the position of that lone node is $2^{h_n} - x_n$, the node is a right child of its parent; we make it a left child. We have reached the leftmost-packed tree and left-packing moves are no longer possible. This leftmost-packed tree has the highest Gini index among all trees of the same size. ■

THEOREM 5.1: *Within the class of binary trees of size n, the tree with the least degree Gini index is a path*[5] *and the tree with the largest degree Gini index is the leftmost-packed complete tree.*

PROOF: Suppose T_n is a binary tree of size n, with a node v of degree 3 in it. Cut the edge leading to the right child r of v and move the entire subtree rooted at r to be a subtree of one of the leaves, say x, of T_n.[6] This transformation produces a binary tree T_n'. Figure 5 shows the transformation from T_n to a tree T_n' with a lower degree Gini index. A second transformation on T_n' produces T_n'', which is a path.

We next show that $G(T_n) > G(T_n')$. All the nodes of T_n preserve their degrees in T_n', except v and x: The degree of v is reduced by 1; that of x is increased by 1. The new tree T_n' thus has $n_1' = n_1 - 1$ nodes of degree 1, $n_2' = n_2 + 2$ nodes of degree 2, and $n_3' = n_3 - 1$ nodes of degree 3. Using (4), we find the new degree Gini index to be

$$\begin{aligned} G(T_n') &= \frac{n_1'n_2' + n_2'n_3' + 2n_1'n_3'}{n(n_1' + 2n_2' + 3n_3')} \\ &= \frac{(n_1-1)(n_2+2) + (n_2+2)(n_3-1) + 2(n_1-1)(n_3-1)}{n((n_1-1) + 2(n_2+2) + 3(n_3-1))} \\ &= \frac{(n_1n_2 + n_2n_3 + 2n_1n_3) - 2(n_2+1)}{n(n_1 + 2n_2 + 3n_3)} \end{aligned}$$

[5] In this context, we mean when the tree is viewed as unrooted, it is a path.
[6] It does not matter whether we make the moved subtree a left or a right subtree of x.

https://doi.org/10.1017/S0269964819000044 Published online by Cambridge University Press

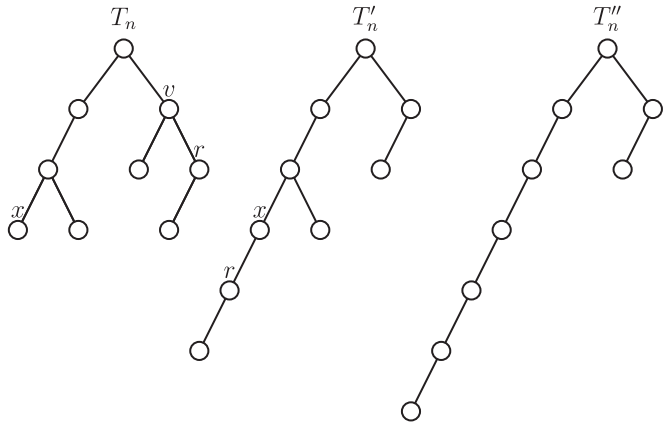

FIGURE **5.** Transformations that lower the degree Gini index.

$$= G(T_n) - \frac{2(n_2+1)}{n(n_1+2n_2+3n_3)}$$
$$< G(T_n).$$

The degree Gini index is reduced by the transformation of T_n into T'_n. If T'_n still has nodes of degree 3, we transform it in the same way to get yet another tree with lower degree Gini index. We keep going through a series of transformations, each lowering the degree Gini index, till it is no longer possible, that is, when there are no nodes left of degree 3. The only possible such tree is a binary tree in the form of a path.

At the other end of the extreme, we obtain trees with higher degree Gini index by compressing them via an inverse transformation: If there is any internal node with degree 2, we remove (one of) the highest leaves and adjoin it as a child of that node, producing a tree of a higher degree Gini index. We continue these transformations until it is no longer possible to transfer nodes to lower levels. The resulting tree is complete. We then go through a series of left-packing moves obtaining trees of the same or higher Gini index. When it is no longer possible to proceed with left packing, we have a leftmost-packed tree of the highest Gini index among all trees of size n. ■

THEOREM 5.2: *Let $\{T_n\}_{n=1}^{\infty}$ be a sequence of binary trees, and T_n has n vertices. We then have*

$$0 \le \liminf_{n\to\infty} G(T_n) \le \limsup_{n\to\infty} G(T_n) \le \frac{1}{4},$$

and the bounds are attainable by certain such sequences.

PROOF: By Theorem 5.1, a path on n nodes, P_n, is a binary graph with the least possible degree Gini index value among all trees of size n. Such a path has degree Gini value

$$G(P_n) = \frac{2(n-2)}{n(2+2(n-2))} = \frac{n-2}{n(n-1)} \to 0, \quad \text{as } n \to \infty.$$

The degree Gini index of leftmost-packed complete trees of size 1 and 2 is 0.

To compute the highest degree Gini index for trees of size $n \ge 3$, we need to dissect the leftmost-packed complete binary tree on n nodes. Assume $n \ge 3$. The degree Gini index of

https://doi.org/10.1017/S0269964819000044 Published online by Cambridge University Press

such a leftmost-packed complete binary tree C_n, can be computed from some basic elements in the tree. Let x_n be the number of nodes at the highest level of C_n. If x_n is even, all the parents of these nodes on the highest level are of degree 3, otherwise, one of them is of degree 2. Let $\mathbb{I}_n$ be an indicator that assumes the value 1 if x_n is odd, otherwise the indicator is 0. The only other node of degree 2 is the root. We thus have

$$n_2 = 1 + \mathbb{I}_n.$$

It is well known that a complete binary tree on n nodes has height $h_n = \lfloor \log_2 n \rfloor$; see [13], p. 400. There are $h_n - 1$ full levels containing $1 + 2 + \cdots + 2^{h_n - 1} = 2^{h_n} - 1$ nodes. Therefore, we have

$$x_n = n - 2^{h_n} + 1.$$

Each of the nodes on levels $1, 2, \ldots, h_{n-2}$, has degree 3, and only $\lfloor x_n/2 \rfloor$ nodes on level $h_n - 1$ are of degree 3, too, yielding the count

$$\begin{aligned} n_3 &= 2 + 2^2 + \cdots + 2^{h_n - 2} + \left\lfloor \frac{x_n}{2} \right\rfloor \\ &= 2^{h_n - 1} - 2 + \left\lfloor \frac{x_n}{2} \right\rfloor \\ &= 2^{\lfloor \log_2 n \rfloor - 1} - 2 + \left\lfloor \frac{n - 2^{\lfloor \log_2 n \rfloor} + 1}{2} \right\rfloor. \end{aligned}$$

The degree Gini index of the complete tree on n nodes can now be computed via the formula

$$G(C_n) = \frac{n_1 n_2 + n_2 n_3 + 2 n_3 n_1}{n(n_1 + 2n_2 + 3n_3)}.$$

To assess this degree Gini index, we note that the denominator is of the order $\Theta(n)$, and

$$n_2 = O(1), \quad n_3 = \frac{n}{2} + O(1), \quad n_1 = n - n_2 - n_3 = \frac{n}{2} + O(1).$$

So, we have

$$\begin{aligned} G(C_n) &= \frac{n_1 n_2}{n(n_1 + 2n_2 + 3n_3)} + \frac{n_2 n_3}{n(n_1 + 2n_2 + 3n_3)} + \frac{2 n_3 n_1}{n(n_1 + 2n_2 + 3n_3)} \\ &= O\left(\frac{1}{n}\right) + O\left(\frac{1}{n}\right) + \frac{2(n/2 + O(1))^2}{n((n/2 + O(1)) + O(1) + 3(n/2 + O(1)))} \\ &\to \frac{1}{4}, \quad \text{as } n \to \infty. \end{aligned}$$

The claimed attainability has been established.

For any general binary tree T_n, Theorem 5.1 tells us that

$$G(P_n) \le G(T_n) \le G(C_n).$$

The stated bounds follow. ■

Remark 5.1: Note that the remainder $O(1)$ functions are oscillating, creating rather turbulent but damped oscillations in the degree Gini index of leftmost-packed complete binary trees. The oscillations are pronounced for small n, but die out for large n. Figure 6 shows the fluctuations in the Gini index for the leftmost-packed complete binary trees of sizes $30, 31, \ldots, 50$. The plot indicates that the limit is approached from above.

https://doi.org/10.1017/S0269964819000044 Published online by Cambridge University Press

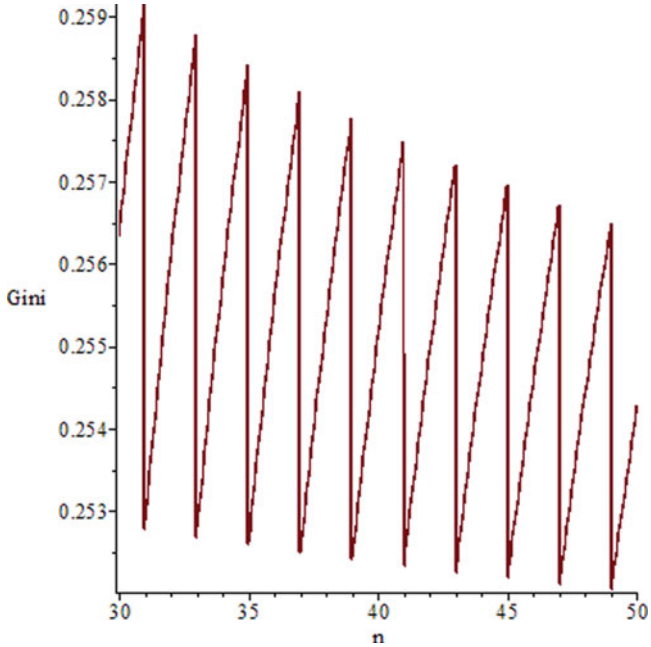

FIGURE **6.** Oscillations in the Gini index of leftmost-packed complete binary trees.

6. COMPARING THE REGULARITY OF TWO CLASSES OF BINARY TREES

Many probability measures can be assigned to the space of binary trees of size n. Two popular measures are the uniform and the one induced by random permutations, giving rise to binary search trees.

6.1. Uniform binary trees

Under a uniform model, all binary trees of order n are taken to be equally likely. The model is sometimes called *Catalan trees*, and is deemed relevant to compilers. It is sometimes thought of as the tree model representing the parsing of random arithmetic expressions that appear in computer algebra systems [12].

For $i = 1, 2, 3$, let N_i be the number of nodes of degree i in a uniformly random binary tree T_n. These numbers are random variables, dependent (of course) in view of the constraint

$$N_1 + N_2 + N_3 = n.$$

The components of the row vector (N_1, N_2, N_3) have a known joint exact distribution; see [15] and the more explicit results in[18]. The source [15] gives a multivariate normal limit distribution and weak limit laws. Pertinent to the current investigation is the weak laws

$$\frac{N_1}{n} \xrightarrow{P} \frac{1}{4}, \quad \frac{N_2}{n} \xrightarrow{P} \frac{1}{2}, \quad \frac{N_3}{n} \xrightarrow{P} \frac{1}{4}.$$

https://doi.org/10.1017/S0269964819000044 Published online by Cambridge University Press

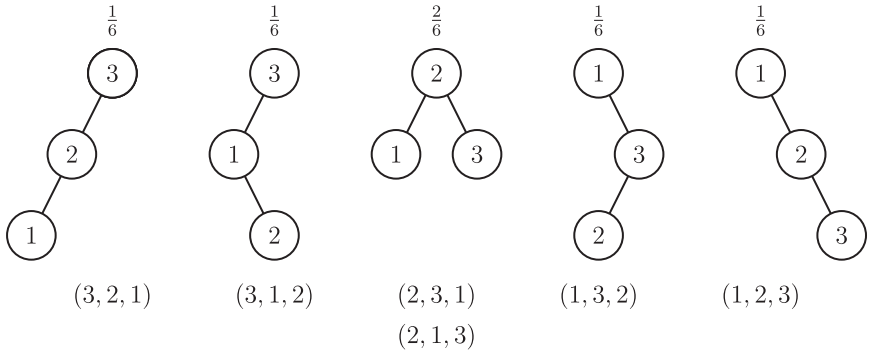

FIGURE **7.** Binary search trees of order 3 arising from permutations of $\{1, 2, 3\}$. Below each tree are the permutations associated with it, and above each is its probability.

We can utilize this in the definition of the relative degree Gini index of a random tree T_n in the class $\mathcal{B}_n$ of uniform binary trees, and obtain

$$\begin{aligned} G_{\mathcal{B}_n}(T_n) &= \frac{N_1N_2 + N_2N_3 + 2N_1N_3}{n(N_1 + 2N_2 + 3N_3)} \\ &= \frac{N_1/n \times N_2/n + N_2/n \times N_3/n + 2\,N_3/n \times N_1/n}{N_1/n + 2N_2/n + 3N_3/n} \\ &\xrightarrow{P} \frac{1/4 \times 1/2 + 1/2 \times 1/4 + 2 \times 1/4 \times 1/4}{1/4 + 2 \times 1/2 + 3 \times 1/4} \\ &= \frac{3}{16}. \end{aligned}$$

With the degree Gini index being bounded by 1, this convergence in probability is reflected in convergence on average. This yields the degree Gini index for the class of uniform binary search trees of size n as

$$G^*_{\mathcal{B}_n} = \mathbb{E}[G_{\mathcal{B}_n}] \to \frac{3}{16} = 0.1875.$$

6.2. Binary search trees

A model of relevance to the use of binary trees as data storage devices is their growth from a random permutation. Suppose $K_1, \ldots, K_n$ are keys sampled from a continuous distribution. Their ranks then almost surely are a random permutation of $\{1, \ldots, n\}$. A binary tree is created for the storage of these keys in the following way. Initially, the key K_1 is retained in a root node, with empty left and right subtrees. When a subsequent key appears, it is guided to the left or right subtree according to whether it is less than K_1 or is at least as large. With the ranks being the only facet of the data that drives the shape of the resulting tree and being almost surely a random permutation of $\{1, \ldots, n\}$, the process can be assimilated by insertions of the n distinct keys in a random permutation of $\{1, \ldots, n\}$. The resulting tree is called a *binary search tree* (BST). Figure 7 shows the five BST's of size 3 and the corresponding permutations. The top row of numbers shows the (nonuniform) probabilities of these trees.

Let N'_1 be the number of nodes of degree 1 (i.e., leaves) in a BST T_n on n vertices, N'_2 be the number of nodes of degree 2 in it, and N'_3 be the number of nodes of degree 3 in it.

https://doi.org/10.1017/S0269964819000044 Published online by Cambridge University Press

These random variables follow the constraint

$$N_1' + N_2' + N_3' = n.$$

The three components of the row vector (N_1', N_2', N_3') asymptotically have a multivariate normal distribution, as shown in [4]; see also Section 8 of [16]. These components follow a weak law:

$$\frac{N_1'}{n} \xrightarrow{P} \frac{1}{3}, \quad \frac{N_2'}{n} \xrightarrow{P} \frac{1}{3}, \quad \frac{N_3'}{n} \xrightarrow{P} \frac{1}{3}.$$

Using these weak laws in the relative degree Gini index of a random tree T_n in the class $\mathcal{BST}(n)$ of binary search trees of size n, we obtain

$$\begin{aligned} G_{\mathcal{BST}(n)}(T_n) &= \frac{N_1'N_2' + N_2'N_3' + 2N_1'N_3'}{n(N_1' + 2N_2' + 3N_3')} \\ &= \frac{N_1'/n \times N_2'/n + N_2'/n \times N_3'/n + 2\,N_3'/n \times N_1'/n}{N_1'/n + 2N_2'/n + 3N_3'/n} \\ &\xrightarrow{P} \frac{1/3 \times 1/3 + 1/3 \times 1/3 + 2 \times 1/3 \times 1/3}{1/3 + 2 \times 1/3 + 3 \times 1/3} \\ &= \frac{2}{9}. \end{aligned}$$

However, the Gini index is bounded by 1. Under this condition, convergence in probability implies convergence in L_1, yielding the degree Gini index for the class of binary search trees:

$$G^*_{\mathcal{BST}(n)} = \mathbb{E}[G_{\mathcal{B}(n)}] \to \frac{2}{9} \approx 0.2222.$$

Remark 6.1: Tall trees must have long paths, with a dominant number of nodes of degree 2, which increases the regularity, ultimately impacting a reduction in the degree Gini index. This is manifested in the contrast between the tall and scrawny uniform binary trees when compared with the short and shrubby binary search trees.

The class $\mathcal{B}_n$ of uniform binary trees has degree Gini index 0.1875, whereas that of the class $\mathcal{BST}(n)$ is $0.2222\ldots$. The class $\mathcal{B}_n$ is more regular than the class $\mathcal{BST}(n)$. This is consistent with the fact that the average height of the former is about $2\sqrt{\pi n}$ (see [9]), which is higher than the that of the latter, which is about $4.31107 \ln n$ (see [3,5]). Thus, the uniform random binary tree is more "spread out" than the random BST of the same size, rendering the uniform random binary tree closer to a path, and the BST closer to a complete tree, and the effect of Theorem 5.1 kicks in.

Acknowledgments

The authors are indebted to Dr. Panpan Zhang for a discussion on the subject, and for bringing to their attention a closely related manuscript [21]. Part of this research was completed while the second author was on sabbatical leave visiting University of Southern California, where he finds an excellent academic environment.

References

1. Balaji, H. & Mahmoud, H. (2017). The Gini index of random trees with an application to caterpillars. *Journal of Applied Probability* 54: 701–709.
2. Ceriani, L. & Verme, P. (2012). The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *Journal of Economic Inequality* 10: 421–443.

3. Devroye, L. (1986). A note on the height of binary search trees. *Journal of the ACM* 33: 489–498.
4. Devroye, L. (1991). Limit laws for local counters in random binary search trees. *Random Structures & Algorithms* 2: 303–315.
5. Drmota, M. (2003). On the variance of the height of random binary search trees. *Journal of the ACM* 50: 333–374.
6. Eggemann, N. & Noble, S. (2011). The clustering coefficient of a scale-free random graph. *Discrete Applied Mathematics* 159: 953–965.
7. Feng, Q. & Hu, Z. (2011). On the Zagreb index of random recursive trees. *Journal of Applied Probability* 48: 1189–1196.
8. Feng, Q., Mahmoud, H., & Panholzer, A. (2008). Limit laws for the Randič index of random binary tree models. *The Annals of the Institute of Statistical Mathematics* 60: 319–343.
9. Flajolet, P. & Odlyzko, A. (1982). The height of binary trees and other families of simple trees. *Journal of Computer and System Sciences* 25: 171–213.
10. Gastwirth, J. (1972). The estimation of the Lorenz curve and Gini index. *Review of Economics and Statistics* 54: 306–316.
11. Gini, C. (1912). *Veriabilità e Mutabilità*, C. Cuppini, Bologna, Italy.
12. Kemp, R. (1984). *Fundamentals of the Average Case Analysis of Particular Algorithms*. Stuttgart, Germany: Wiley-Teubner.
13. Knuth, D. (1997). *The Art of Computer Programming*, Vol. 1, Fundamental Algorithms, 3rd ed., Reading, Massachusetts, USA: Addison-Wesley Longman.
14. Lee, J., Manzil, Z., Günnemann, S., & Smola, A. (2015). The clustering coefficient of preferential attachment networks with affinities. *Proceedings of Machine learning Research* 38: 571–580.
15. Mahmoud, H. (1995). The joint distribution of the three types of nodes in uniform binary trees. *Algorithmica* 13: 313–323.
16. Mahmoud, H. (2008). *Pólya Urn Models*. Boca Raton, Florida, USA: Chapman-Hall.
17. Neininger, R. (2002). The Wiener index of random trees. *Combinatorics, Probability & Computing* 11: 587–597.
18. Prodinger, H. (1996). A note on the distribution of the three types of nodes in uniform binary trees. *Seminaire Lotharingien de Combinatoire* 38: 5.
19. Thon, D. (1982). An Axiomatization of the Gini Coefficient. *Mathematical Social Sciences* 2: 131–143.
20. Timmerman, H., Todeschini, R., Consonni, V., Mannhold, R., & Kubinyi, H. (2002). *Handbook of Molecular Descriptors*. Weinheim: Wiley-VCH.
21. Zhang, P. & Dey, D. (2019+). The degree profile and Gini index of random caterpillar trees. *Probability in the Engineering and Informational Sciences* (to appear). doi:10.1017/S0269964818000475
22. Zhang, P. & Mahmoud, H. (2016). The degree profile and weight in Apollonian networks and k-trees. *Advances in Applied Probability* 48: 163–175.