

ON THE ZAGREB INDEX OF RANDOM RECURSIVE TREES

QUNQIANG FENG * ** AND

ZHISHUI HU, * *** *University of Science and Technology of China*

Abstract

We investigate the Zagreb index, one of the topological indices, of random recursive trees in this paper. Through a recurrence equation, the first two moments of Z_n , the Zagreb index of a random recursive tree of size n , are obtained. We also show that the random process $\{Z_n - E[Z_n], n \geq 1\}$ is a martingale. Then the asymptotic normality of the Zagreb index of a random recursive tree is given by an application of the martingale central limit theorem. Finally, two other topological indices are also discussed in passing.

Keywords: Random tree; Zagreb index; recursive tree; martingale central limit theorem

2010 Mathematics Subject Classification: Primary 05C05; 60C05
Secondary 60F05

1. Introduction

Many chemists who derive and use molecular complexity indices employ concepts and terminology from (chemical) graph theory in their work (see, for example, Harary (1972) and Trinajstić (1992)), where graphs are generated from molecules by replacing atoms with vertices and bonds with edges, or represent only bare molecular skeletons, that is, molecular skeletons without hydrogen atoms.

The *Zagreb index* of a (molecular) graph is defined as the sum of the squares of the degrees of all vertices in the graph. This index, one of the topological indices of a graph, was introduced by chemists Gutman and Trinajstić (1972). Nikolić *et al.* (2000) studied the mathematical properties of the Zagreb index. Recently, the Zagreb index and its variants have been used to study molecular complexity, chirality, ZE isomerism, and heterosystems, whilst the overall Zagreb indices exhibited a potential applicability for deriving multilinear regression models (see, for example, Nikolić *et al.* (2003b)). The Zagreb index has also been used in the studies of quantitative structure-property/activity relationships (QSPR/QSAR) (see, for example, Devillers and Balaban (1999, p. 28)). For its connection with several topological indices and more background, we refer the reader to Nikolić *et al.* (2003a) and the references therein.

Since the structures of many molecules are tree like, our interest here is to study the Zagreb index of random trees. For a tree T of size n , Li *et al.* (2003) studied the extreme values of the Zagreb index of T : it reaches the minimum value $4n - 6$ if and only if T is a path on n nodes, and reaches the maximum value $n(n - 1)$ if and only if T is a star on n nodes.

Received 14 June 2011; revision received 22 August 2011.

* Postal address: Department of Statistics and Finance, University of Science and Technology of China, Hefei, Anhui 230026, P. R. China.

** Email address: fengqq@ustc.edu.cn

*** Email address: huzs@ustc.edu.cn

Several other topological indices of random trees have been studied by many authors. We refer the reader to Ali Khan and Neininger (2007), Janson (2003), Janson and Chassaing (2004), and Neininger (2002) for the Wiener index, and to Clark and Moon (2000), Feng *et al.* (2008), and Hollas (2005) for the Randić index. Our aim in this paper is to consider the probabilistic behaviors of the Zagreb index for random recursive trees.

A *random recursive tree* is an outgrowth from a single node labeled 1. Progressively, nodes are added in stages: at the n th stage a node in the existing tree is chosen at random as the parent of the n th entrant (labeled n). In this context *random* means that all nodes in the tree of size $n - 1$ are equally likely to be the parent of n . The model of random recursive trees induces a uniform distribution on the trees of order n : all $(n - 1)!$ recursive trees of size n are generated with equal probability. The random recursive tree is a naturally growing structure that underlies many stochastic phenomena and algorithms. For numerous applications of random recursive trees, we refer the reader to Smythe and Mahmoud (1994).

The paper is organized as follows. We first show a recurrence for the Zagreb indices of random recursive trees, from which a martingale is constructed and the mean and variance of the Zagreb index are given. Then the asymptotic normality of the Zagreb index of a random recursive tree is proved as the size of the tree grows to infinity. Finally, two topological indices which are closely related to the Zagreb index are mentioned in passing.

2. The martingale and the moments

For convenience, we denote by H_n the n th harmonic number, i.e.

$$H_n = \sum_{j=1}^n \frac{1}{j}, \quad n \geq 1,$$

with $H_0 = 0$.

Let Z_n be the Zagreb index of a random recursive tree of size n . Let $D_{k,n}$ denote the degree of node k in the random recursive tree of size n . We also define \mathcal{F}_n to be the σ -field generated by the first n stages of the random recursive trees. Considering the insertion of node n at the n th stage, we obtain

$$\begin{aligned} Z_n &= Z_{n-1} + (D_{U_n,n-1} + 1)^2 - D_{U_n,n-1}^2 + 1 \\ &= Z_{n-1} + 2D_{U_n,n-1} + 2, \end{aligned} \tag{1}$$

where $U_n = \sum_{k=1}^{n-1} k \mathbf{1}$ (node k is the parent of node n) is uniformly distributed on the set $\{1, 2, \dots, n - 1\}$ and independent of \mathcal{F}_{n-1} .

We can compute the moments of Z_n by (1). Clearly,

$$\begin{aligned} E[Z_n \mid \mathcal{F}_{n-1}] &= E[Z_n \mid D_{k,n-1}, k = 1, 2, \dots, n - 1] \\ &= Z_{n-1} + \frac{2}{n - 1} \sum_{k=1}^{n-1} D_{k,n-1} + 2. \end{aligned}$$

Note that the sum of the degrees of all nodes in a tree of size n is $2(n - 1)$. Then we have

$$E[Z_n \mid \mathcal{F}_{n-1}] = Z_{n-1} + 6 - \frac{4}{n - 1}. \tag{2}$$

Thus,

$$E[Z_n] = E[Z_{n-1}] + 6 - \frac{4}{n-1}.$$

With the initial values $Z_1 = 0$ and $Z_2 = 2$, the above recurrence equation has the solution

$$E[Z_n] = 6n - 4H_{n-1} - 6, \quad n \geq 1. \tag{3}$$

Relation (2) also implies that

$$E[Z_n - E[Z_n] \mid \mathcal{F}_{n-1}] = Z_{n-1} - E[Z_{n-1}].$$

That is, the process $\{Z_n - E[Z_n], n \geq 1\}$ is a martingale. Let $M_1 = 0$ and

$$M_n := Z_n - Z_{n-1} - 6 + \frac{4}{n-1}, \quad n \geq 2.$$

Then $\{M_n, n \geq 1\}$ is the martingale difference sequence of $\{Z_n - E[Z_n], n \geq 1\}$. Hence, $E[M_n \mid \mathcal{F}_{n-1}] = 0$.

For the second moment of Z_n , also computed by (1), we have

$$E[(Z_n - Z_{n-1} - 2)^2 \mid \mathcal{F}_{n-1}] = 4E[D_{U_n, n-1}^2 \mid \mathcal{F}_{n-1}] = \frac{4}{n-1}Z_{n-1}. \tag{4}$$

By noting that, for $n \geq 2$,

$$\begin{aligned} E[(Z_n - Z_{n-1} - 2)^2 \mid \mathcal{F}_{n-1}] &= E\left[\left(M_n + 4 - \frac{4}{n-1}\right)^2 \mid \mathcal{F}_{n-1}\right] \\ &= E[M_n^2 \mid \mathcal{F}_{n-1}] + \left(4 - \frac{4}{n-1}\right)^2, \end{aligned} \tag{5}$$

we have

$$\begin{aligned} E[M_n^2] &= \frac{4}{n-1}E[Z_{n-1}] - \left(4 - \frac{4}{n-1}\right)^2 \\ &= 8 - \frac{16H_{n-2} - 8}{n-1} + \frac{16}{(n-1)^2} \\ &= 8 + O\left(\frac{\log n}{n}\right), \quad n \geq 2. \end{aligned} \tag{6}$$

Since $E[M_i M_j] = 0$ for any $1 \leq i \neq j \leq n$, we obtain

$$\begin{aligned} \text{var}[Z_n] &= E(Z_n - E[Z_n])^2 \\ &= E\left[\sum_{j=1}^n M_j\right]^2 \\ &= \sum_{j=1}^n E[M_j^2] \\ &= 8n + O(\log^2 n), \quad n \geq 2. \end{aligned}$$

We have thus proved the following result.

Proposition 1. *Let Z_n be the Zagreb index of a random recursive tree of size n . Then the process $\{Z_n - E[Z_n], n \geq 1\}$ is a martingale, and*

$$E[Z_n] = 6n - 4H_{n-1} - 6 = 6n + O(\log n),$$

$$\text{var}[Z_n] = 8n + O(\log^2 n).$$

By Proposition 1 and Chebyshev’s inequality, it follows that $Z_n/E[Z_n] \xrightarrow{P} 1$. This, together with (3), yields the following weak law of large numbers of Z_n .

Proposition 2. *For Z_n , the Zagreb index of a random recursive tree of size n , as $n \rightarrow \infty$, we have*

$$\frac{Z_n}{n} \xrightarrow{P} 6.$$

Remark. A random recursive tree of size n is much ‘closer’ to a path of length n (with the Zagreb index $4n - 6$) than a star of size n (with the Zagreb index $n(n - 1)$), when n is large. However, as expected, a random binary tree is a bit closer to a path than a random recursive tree when their sizes grow (see Feng (2011)). The expected Zagreb index of a random binary search tree of size n is $14n/3 + O(1)$, while that of a binary Catalan tree of size n is $9n/2 + O(1)$.

3. Asymptotic normality

For the asymptotic normality of Z_n , we need the following two lemmas.

Lemma 1. *We have*

$$\frac{1}{n} \sum_{j=1}^n \frac{Z_j - E[Z_j]}{j} \xrightarrow{P} 0. \tag{7}$$

Proof. Since

$$\sum_{j=1}^n \frac{Z_j - E[Z_j]}{j} = \sum_{j=1}^n \frac{1}{j} \sum_{i=1}^j M_i = \sum_{j=1}^n (H_n - H_{j-1})M_j,$$

by (6) we have, for $n \geq 2$,

$$E \left[\sum_{j=1}^n \frac{Z_j - E[Z_j]}{j} \right]^2 = \sum_{j=1}^n (H_n - H_{j-1})^2 E[M_j^2]$$

$$\leq (H_n - 1)^2 \sum_{j=1}^n E[M_j^2]$$

$$= 8n \log^2 n + O(\log^4 n).$$

Then (7) follows by Chebyshev’s inequality.

Lemma 2. *For the maximum degree of the nodes in a random recursive tree of size n , as $n \rightarrow \infty$,*

$$\frac{1}{\sqrt{n}} \max_{1 \leq j \leq n} D_{j,n} \xrightarrow{P} 0.$$

Proof. Define $I_{ji} = \mathbf{1}(\text{node } j \text{ is the parent of node } i)$ for $1 \leq j < i \leq n$. Then $I_{ji} \sim \text{Ber}(1/(i - 1))$, and the degree of node j in a random recursive tree of size n can be expressed as

$$D_{j,n} = \begin{cases} \sum_{i=j+1}^n I_{ji} & \text{if } j = 1, \\ 1 + \sum_{i=j+1}^n I_{ji} & \text{if } 2 \leq j \leq n. \end{cases}$$

Note that, for fixed j , the I_{ji} , $i = j + 1, \dots, n$, are independent. Then by the inequality $(a + b)^3 \leq 4(a^3 + b^3)$ for any $a, b \geq 0$ we have, for $n \geq 2$,

$$\begin{aligned} E[D_{j,n}^3] &\leq E\left(1 + \sum_{i=j+1}^n I_{ji}\right)^3 \\ &\leq 4 + 4E\left(\sum_{i=j+1}^n I_{ji}\right)^3 \\ &\leq 4 + 4E\left[\sum_{i=j+1}^n \left(I_{ji} - \frac{1}{i-1}\right)\right]^3 + 4\left[\sum_{i=j+1}^n \frac{1}{i-1}\right]^3 \\ &\quad + 12\left[\sum_{i=j+1}^n \frac{1}{i-1}\right]E\left[\sum_{i=j+1}^n \left(I_{ji} - \frac{1}{i-1}\right)\right]^2 \\ &\quad + 12\left[\sum_{i=j+1}^n \frac{1}{i-1}\right]^2E\left[\sum_{i=j+1}^n \left(I_{ji} - \frac{1}{i-1}\right)\right] \\ &= 4 + 4\sum_{i=j+1}^n E\left[I_{ji} - \frac{1}{i-1}\right]^3 + O(\log^3 n) \\ &\quad + O(\log n)\sum_{i=j+1}^n E\left[I_{ji} - \frac{1}{i-1}\right]^2 \\ &= O(\log^3 n). \end{aligned}$$

Thus, for any $\varepsilon > 0$,

$$P\left(\max_{1 \leq j \leq n} D_{j,n} > \varepsilon\sqrt{n}\right) \leq \sum_{j=1}^n P(D_{j,n} > \varepsilon\sqrt{n}) \leq \frac{1}{\varepsilon^3 n^{3/2}} \sum_{j=1}^n E[D_{j,n}^3] \rightarrow 0,$$

which completes the proof.

The result of the above lemma can be strengthened. Devroye and Lu (1995) showed that $\max_{1 \leq j \leq n} D_{j,n} / \log_2 n \xrightarrow{P} 1$. For more results concerning the maximum degree of the nodes in a random recursive tree, we refer the reader to Goh and Schmutz (2002).

The following theorem is our main result.

Theorem 1. For Z_n , the Zagreb index of a random recursive tree of size n , as $n \rightarrow \infty$, we have

$$\frac{Z_n - 6n}{\sqrt{8n}} \xrightarrow{D} N(0, 1).$$

Proof. By (3) we only need to prove that

$$\frac{Z_n - E[Z_n]}{\sqrt{8n}} \xrightarrow{D} N(0, 1).$$

Let

$$X_{ni} = \frac{M_i}{\sqrt{8n}}, \quad i = 1, 2, \dots, n, n = 1, 2, \dots$$

By Corollary 3.1 of Hall and Heyde (1980), it is sufficient to show that

$$\sum_{i=1}^n E[X_{ni}^2 \mid \mathcal{F}_{i-1}] \xrightarrow{P} 1, \tag{8}$$

and, for any $\varepsilon > 0$,

$$\sum_{i=1}^n E[X_{ni}^2 \mathbf{1}(|X_{ni}| > \varepsilon) \mid \mathcal{F}_{i-1}] \xrightarrow{P} 0. \tag{9}$$

We check (8) first. By (4) and (5),

$$\begin{aligned} \sum_{i=1}^n E[X_{ni}^2 \mid \mathcal{F}_{i-1}] &= \frac{1}{8n} \sum_{i=1}^n E[M_i^2 \mid \mathcal{F}_{i-1}] \\ &= \frac{1}{2n} \sum_{j=1}^{n-1} \left[\frac{Z_j}{j} - \left(2 - \frac{2}{j}\right)^2 \right] \\ &= \frac{1}{2n} \sum_{j=1}^{n-1} \frac{Z_j - E[Z_j]}{j} + \frac{1}{2n} \sum_{j=1}^{n-1} \left[\frac{E[Z_j]}{j} - \left(2 - \frac{2}{j}\right)^2 \right]. \end{aligned}$$

Then (8) follows by Lemma 1 and (3).

Note that, by (1), for any $\varepsilon > 0$, we have, for sufficiently large n ,

$$\begin{aligned} &\sum_{i=1}^n E[X_{ni}^2 \mathbf{1}(|X_{ni}| > \varepsilon) \mid \mathcal{F}_{i-1}] \\ &= \frac{1}{8n} \sum_{i=1}^n E[M_i^2 \mathbf{1}(|M_i| > \varepsilon\sqrt{8n}) \mid \mathcal{F}_{i-1}] \\ &= \frac{1}{2n} \sum_{i=2}^n E \left[\left(D_{U_i, i-1} - \frac{2(i-2)}{i-1} \right)^2 \mathbf{1} \left(\left| 2D_{U_i, i-1} - \frac{4(i-2)}{i-1} \right| > \varepsilon\sqrt{8n} \right) \mid \mathcal{F}_{i-1} \right] \\ &\leq \frac{1}{2n} \sum_{i=2}^n E[D_{U_i, i-1}^2 \mathbf{1}(D_{U_i, i-1} > \varepsilon\sqrt{n}) \mid \mathcal{F}_{i-1}] \\ &\quad + \frac{2}{n} \sum_{i=2}^n E[\mathbf{1}(D_{U_i, i-1} > \varepsilon\sqrt{n}) \mid \mathcal{F}_{i-1}]. \end{aligned}$$

Thus, in order to prove (9), we only need to prove that, for any $\varepsilon > 0$,

$$\frac{1}{n} \sum_{i=2}^n E[D_{U_i, i-1}^2 \mathbf{1}(D_{U_i, i-1} > \varepsilon\sqrt{n}) \mid \mathcal{F}_{i-1}] \xrightarrow{P} 0 \tag{10}$$

and

$$\frac{1}{n} \sum_{i=2}^n E[\mathbf{1}(D_{U_i, i-1} > \varepsilon\sqrt{n}) \mid \mathcal{F}_{i-1}] \xrightarrow{P} 0. \tag{11}$$

Since

$$\begin{aligned} & \frac{1}{n} \sum_{i=2}^n E[D_{U_i, i-1}^2 \mathbf{1}(D_{U_i, i-1} > \varepsilon\sqrt{n}) \mid \mathcal{F}_{i-1}] \\ &= \frac{1}{n} \sum_{i=2}^n \frac{1}{i-1} \sum_{j=1}^{i-1} D_{j, i-1}^2 \mathbf{1}(D_{j, i-1} > \varepsilon\sqrt{n}) \\ &\leq \left(\frac{1}{n} \sum_{i=2}^n \frac{1}{i-1} \sum_{j=1}^{i-1} D_{j, i-1}^2 \right) \mathbf{1}\left(\max_{1 \leq j \leq n} D_{j, n} > \varepsilon\sqrt{n}\right), \end{aligned}$$

by Lemma 2, for any $\delta > 0$, we have

$$\begin{aligned} & P\left(\frac{1}{n} \sum_{i=2}^n E[D_{U_i, i-1}^2 \mathbf{1}(D_{U_i, i-1} > \varepsilon\sqrt{n}) \mid \mathcal{F}_{i-1}] > \delta\right) \\ &\leq P\left(\left(\frac{1}{n} \sum_{i=2}^n \frac{1}{i-1} \sum_{j=1}^{i-1} D_{j, i-1}^2\right) \mathbf{1}\left(\max_{1 \leq j \leq n} D_{j, n} > \varepsilon\sqrt{n}\right) > \delta\right) \\ &\leq P\left(\max_{1 \leq j \leq n} D_{j, n} > \varepsilon\sqrt{n}\right) \\ &\rightarrow 0. \end{aligned}$$

Hence, (10) holds. Similarly,

$$\frac{1}{n} \sum_{i=2}^n E[\mathbf{1}(D_{U_i, i-1} > \varepsilon\sqrt{n}) \mid \mathcal{F}_{i-1}] \leq \mathbf{1}\left(\max_{1 \leq j \leq n} D_{j, n} > \varepsilon\sqrt{n}\right),$$

which, together with Lemma 2, implies (11). The proof of Theorem 1 is complete.

4. The Gordon–Scantlebury and Platt indices

A path in a graph is a sequence of adjacent edges, which do not pass through the same vertex more than once, and the length of the path is the number of edges in it. The degree of an edge is equal to the number of its adjacent edges. For a simple graph G , the *Gordon–Scantlebury index* of G is equal to the number of paths of length two in G (see Gordon and Scantlebury (1964)), and the *Platt index* is equal to the total sum of the degrees of all edges in G (see Platt (1947)).

The Zagreb index is related to the Gordon–Scantlebury and Platt indices (see, for example, Barysz *et al.* (1986)). Let Z_G , S_G , and P_G be the Zagreb index, the Gordon–Scantlebury index, and the Platt index of the graph G , respectively. Nikolić *et al.* (2003a) showed that

$$Z_G = 2(S_G + E(G)) \quad \text{and} \quad P_G = 2S_G,$$

where $E(G)$ denotes the number of edges in G . Since the number of edges in a tree of size n is obviously $n - 1$, we can directly establish parallel results for the Gordon–Scantlebury and the Platt indices of a random recursive tree. We omit the details.

Acknowledgements

The work of the first author was supported by the RFDP (grant number 20093402120014) and the Fundamental Research Funds for the Central Universities. The work of the second author was supported by the NSFC (grant number 10801122), the RFDP (grant number 200803581009), and the Fundamental Research Funds for the Central Universities.

References

- ALI KHAN, T. AND NEININGER, R. (2007). Tail bounds for the Wiener index of random trees. In *2007 Conference on Analysis of Algorithms, AofA 07* (Discrete Math. Theoret. Comput. Sci. Proc. **AH**), Association of Discrete Mathematics and Theoretical Computer Science, Nancy, pp. 279–289.
- BARYSZ, M., PLAVŠIĆ, D. AND TRINAJSTIĆ, N. (1986). A note on topological indices. *MATCH Commun. Math. Comput. Chem.* **19**, 89–116.
- CLARK, L. H. AND MOON, J. W. (2000). On the general Randić index for certain families of trees. *Ars Combinatoria* **54**, 223–235.
- DEVILLERSAND, J. AND BALABAN, A. T. (1999). *Topological Indices and Related Descriptors in QSAR and QSPR*. Gordon and Breach, Amsterdam.
- DEVROYE, L. AND LU, J. (1995). The strong convergence of maximal degrees in uniform random recursive trees and dags. *Random Structures Algorithms* **7**, 1–14.
- FENG, Q. (2011). The Zagreb index of random binary trees. Unpublished manuscript.
- 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.
- GOH, W. AND SCHMUTZ, E. (2002). Limit distribution for the maximum degree of a random recursive tree. *J. Comput. Appl. Math.* **142**, 61–82.
- GORDON, M. AND SCANTLEBURY, G. R. (1964). Non-random polycondensation: statistical theory of the substitution effect. *Trans. Faraday Soc.* **60**, 604–621.
- GUTMAN, I. AND TRINAJSTIĆ, N. (1972). Graph theory and molecular orbitals. Total φ -electron energy of alternant hydrocarbons. *Chem. Phys. Lett.* **17**, 535–538.
- HALL, P. AND HEYDE, C. C. (1980). *Martingale Limit Theory and Its Application*. Academic Press, New York.
- HARARY, F. (1972). *Graph Theory*. Addison-Wesley, Reading, MA.
- HOLLAS, B. (2005). Asymptotically independent topological indices on random trees. *J. Math. Chem.* **38**, 379–387.
- JANSON, S. (2003). The Wiener index of simply generated random trees. *Random Structures Algorithms* **22**, 337–358.
- JANSON, S. AND CHASSAING, P. (2004). The center of mass of the ISE and the Wiener index of trees. *Electron. Commun. Probab.* **9**, 178–187.
- LI, X., LI, Z. AND WANG, L. (2003). The inverse problems for some topological indices in combinatorial chemistry. *J. Comput. Biol.* **10**, 47–55.
- NEININGER, R. (2002). The Wiener index of random trees. *Combinatorics Probab. Comput.* **11**, 587–597.
- NIKOLIĆ, S., KOVAČEVIĆ, G., MILIČEVIĆ, A. AND TRINAJSTIĆ, N. (2003a). The Zagreb indices 30 years after. *Croatica Chemica Acta* **76**, 113–124.
- NIKOLIĆ, S., TOLIĆ, I. M., TRINAJSTIĆ, N. AND BAUČIĆ, I. (2000). On the Zagreb indices as complexity indices. *Croatica Chemica Acta* **73**, 909–921.
- NIKOLIĆ, S. *et al.* (2003b). On molecular complexity indices. In *Complexity in Chemistry: Introduction and Fundamentals*, eds D. Bonchev and D. H. Rouvray, Taylor and Francis, London, pp. 29–89.
- PLATT, J. R. (1947). Influence of neighbor bonds on additive bond properties in paraffins. *J. Chem. Phys.* **15**, 419–420.
- SMYTHE, R. T. AND MAHMOUD, H. M. (1994). A survey of recursive trees. *Teor. Imovir. Mat. Stat.* **51**, 1–29 (in Ukrainian). English translation: *Theory Prob. Math. Statist.* **51** (1996), 1–27.
- TRINAJSTIĆ, N. (1992). *Chemical Graph Theory*, 2nd edn. CRC Press, Boca Raton, FL.