CORRELATION OF GRAPH-THEORETICAL INDICES.*

STEPHAN G. WAGNER†

Abstract. The correlation of graph characteristics such as the number of independent vertex or edge subsets, the number of connected subsets or the sum of distances, which also play a role in combinatorial chemistry, is studied by a generating function approach and asymptotic analysis. It is shown how an asymptotic formula for the correlation coefficient can be obtained when simply generated families of trees are investigated. For rooted ordered trees, the calculations are done explicitly. Eventually, further feasible correlation measures are discussed.

Key words. Trees, correlation, graph-theoretical index

AMS subject classifications. 05A15, 05A16, 05C05, 05C30

1. Introduction.

In combinatorial chemistry, so-called topological indices are used for the description of the structural properties of molecular graphs. Formally, such an index is a map from the set of graphs into the real numbers (usually integer-valued). Typically, for a fixed number of vertices, the trees of maximal and minimal index are the path and the star respectively (or vice versa). A variety of graph-theoretical indices has been proposed for this purpose, and their connection to the physico-chemical properties of the corresponding molecules has been studied (cf. [19, 23]).

The isomer-discriminating power, a measure for the ability of an index to distinguish between isomeric compounds, has been considered in the paper [15], and there is also a large amount of literature on extremal and asymptotic properties of various indices, we refer to [3, 4, 11, 12, 16, 20, 22].

However, it seems that there is no theoretical result on the correlation between the different indices yet. It should be quite natural to claim some strong correlation between them, since they all reflect the structural properties of graphs in some way. This paper tries to fill this gap a little by proposing measures for the correlation of two indices and discussing them.

The main part of this paper will deal with the asymptotic behavior of the classical correlation coefficient given by

$$r(X_n, Y_n) = \frac{E(X_nY_n) - E(X_n)E(Y_n)}{\sqrt{\operatorname{Var}(X_n)\operatorname{Var}(Y_n)}}. \tag{1.1}$$

Here, $X_n = X(T_n)$ and $Y_n = Y(T_n)$ are the X- and Y- index of a tree T_n on n vertices taken uniformly at random from some family of trees – for simplicity, we will only consider rooted ordered trees in detail; however, the methods can be extended to other families of simply generated trees (such as binary trees, cf. [4, 17]) quite easily.

The asymptotic behavior of the correlation coefficient will give us a measure of the linear correlation of the indices X and Y. Other possibilities to define such a measure are discussed afterwards, but it seems that there is no possibility for a similar asymptotic discussion in these cases.

The indices that will be taken into consideration are the following:

*This work was supported by Austrian Science Fund project no. S-8307-MAT.

†Department of Mathematics, Graz University of Technology, Steyrergasse 30, A-8010 Graz, Austria (wagner@finanz.math.tugraz.at).

(1) The *Merrifield-Simmons-* or σ-index is defined to be the number of independent vertex subsets of a graph, i.e. the number of vertex subsets in which no two vertices are adjacent, including the empty set. Merrifield and Simmons investigated the σ-index in their work [19] and pointed out its correlation to boiling points of molecules.

(2) The *Hosoya-* or Z-index ([8]) is defined as the number of independent edge subsets (also referred to as "matchings"), i.e. the number of edge subsets in which no two edges are adjacent, including the empty set again.

(3) The *number of subtrees* is called ρ-index in [19] and was discussed lately in a paper of Székely and Wang [22].

(4) The *Wiener index* is probably the most popular topological index (s. [3, 4, 26]). It is defined as the sum of all the distances between pairs of vertices, i.e.

$$W(G) = \sum_{v,w \in V(G)} d_G(v,w). \tag{1.2}$$

Section 2 will deal with the correlation of (1), (2) and (3). The Wiener index has another growth structure than the other three, so we need a different approach, which will be presented in section 3. Finally, we will take a look at some other statistical measures in section 4.

2. σ-, Z-, and ρ-index. The method to determine the expected values of these indices for rooted ordered trees on n vertices has been given in several papers [11, 12, 13]. However, for the sake of completeness, it is repeated here. It is well known that the generating function for the number of rooted ordered trees is given by the functional equation

$$T(z) = \frac{z}{1 - T(z)}, \tag{2.1}$$

which is an immediate consequence of the recursive structure of this family of trees. Now, consider the σ-index for instance. We want to determine the function

$$S(z) = \sum_T \sigma(T) z^{|T|},$$

where the sum goes over all trees T and $|T|$ denotes the number of vertices. Now, we distinguish between independent sets containing the root and those not containing it and denote the corresponding quantities by $\sigma_1(T), \sigma_2(T)$. If $T_1, \ldots, T_k$ are the branches of the rooted tree T, it is easy to see that the recursive relations

$$\sigma_1(T) = \prod_{i=1}^{k} \sigma_2(T_i),$$
$$\sigma_2(T) = \prod_{i=1}^{k} (\sigma_1(T_i) + \sigma_2(T_i))$$

hold. These relations can be translated to equations for the corresponding generating functions: if $S_1(z)$ is the generating function for the number of subsets of the first type and $S_2(z)$ the generating function for the number of subsets of the second type, we obtain

$$\begin{aligned}
S_1(z) &= \sum_T \sigma_1(T) z^{|T|} \\
&= \sum_{k\ge 0}\sum_{T_1}\sum_{T_2}\cdots\sum_{T_k}\left(\prod_{i=1}^{k}\sigma_2(T_i)\right) z^{|T_1|+\ldots+|T_k|+1} \\
&= z\sum_{k\ge 0}\left(\sum_T \sigma_2(T) z^{|T|}\right)^k \\
&= z\sum_{k\ge 0} S_2(z)^k = \frac{z}{1-S_2(z)}
\end{aligned} \tag{2.2}$$

and in exactly the same way

$$S_2(z) = \frac{z}{1-S_1(z)-S_2(z)}. \tag{2.3}$$

The asymptotic growth of the coefficients of functions satisfying algebraical equations of this kind can be determined by a standard application of the Flajolet-Odlyzko singularity analysis, which is discussed in several papers such as [1, 2, 5, 18] (sometimes, one can even find exact expressions by means of Lagrange's inversion formula; this is the case for this example (s. [11, 12]), but we won't need the exact solution, which can be given as a hypergeometric sum). However, the details can be intricate, as will be explained in the following. Here, inserting yields

$$S_2(z) = \frac{z}{1-\frac{z}{1-S_2(z)}-S_2(z)}$$

or

$$S_2(z)^3 - 2S_2(z)^2 + S_2(z) - z = 0.$$

Bender [1] gives a general theorem dealing with functional equations of the type $F(z, w(z)) = 0$. His theorem states that, given a minimal solution (with respect to absolute value) (α, β) of the system

$$F(z,w) = 0, F_w(z,w) = 0,$$

which lies within the region of analyticity of F and satisfies $F_z(\alpha,\beta), F_{ww}(\alpha,\beta) \neq 0$, the asymptotic behavior of the coefficients a_n of $w(z)$ is determined by

$$a_n \sim \sqrt{\frac{\alpha F_z(\alpha,\beta)}{2\pi F_{ww}(\alpha,\beta)}} n^{-3/2}\alpha^{-n}.$$

However, there is a slight mistake in this theorem, as was pointed out by Canfield [2], and the method might give erroneous results. The theorem only holds true if α is indeed the radius of convergence of $w(z)$ and the only singularity on the circle of convergence.

In the present case, we know from [7, Th. 12.2.1] (see also [2]) that a singularity of an algebraic function $w(z)$ given by a polynomial equation of the form

$$F(z,w) = \sum_{j=0}^{k} p_{k-j}(z) w^j = 0$$

is either a zero of $p_0(z)$ (here, there is no such zero) or given by a solution of the system $F(z,w)=0$, $F_w(z,w)=0$.

Therefore, the common singularity z_0 of $S_1(z), S_2(z)$ and $S(z)=S_1(z)+S_2(z)$ nearest to the origin is given by the system of equations

$$\begin{aligned} F(s,z) &= s^3-2s^2+s-z=0,\\ \frac{\partial}{\partial s}F(s,z) &= 3s^2-4s+1=0, \end{aligned}$$

yielding $z_0=\frac{4}{27}$. Using the formula for the number of rooted ordered trees on n vertices,

$$t_n=\frac{1}{n}\binom{2n-2}{n-1}\sim\frac{1}{4\sqrt{\pi}}n^{-3/2}4^n,$$

it is easy now to find out the asymptotics for the expected σ-index:

$$E(\sigma_n)\sim\sqrt{3}\left(\frac{27}{16}\right)^{n-1}\approx(1.02640)\cdot(1.6875)^n.$$

Similarly, for the Z-index, we have

$$\begin{aligned} Z_1(T) &= \sum_{j=1}^{k} Z_2(T_j)\prod_{\substack{i=1\\ i\neq j}}^{k}(Z_1(T_i)+Z_2(T_i)),\\ Z_2(T) &= \prod_{i=1}^{k}(Z_1(T_i)+Z_2(T_i)), \end{aligned}$$

where $Z_1(T)$ and $Z_2(T)$ denote the number of independent edge subsets containing resp. not containing an edge incident to the root. From this, we obtain the equations

$$\begin{aligned} Z_1(z) &= \frac{zZ_2(z)}{(1-Z_1(z)-Z_2(z))^2},\\ Z_2(z) &= \frac{z}{1-Z_1(z)-Z_2(z)}, \end{aligned} \tag{2.4}$$

for the respective generating functions. This system gives us the asymptotic expression for the average Z-index:

$$E(Z_n)\sim\sqrt{\frac{65-\sqrt{13}}{78}}\left(\frac{35+13\sqrt{13}}{54}\right)^n\approx(0.88719)\cdot(1.51615)^n.$$

Finally, for the ρ-index,

$$\begin{aligned} \rho_1(T) &= \prod_{i=1}^{k}(1+\rho_1(T_i)),\\ \rho_2(T) &= \sum_{i=1}^{k}(\rho_1(T_i)+\rho_2(T_i)), \end{aligned}$$

where $\rho_1(T)$ and $\rho_2(T)$ denote the number of subtrees containing resp. not containing an edge incident to the root. Here, the system of equations for the corresponding generating functions is

$$\begin{aligned} R_1(z) &= \frac{z}{1 - R_1(z) - T(z)}, \\ R_2(z) &= \frac{z}{(1 - T(z))^2}(R_1(z) + R_2(z)), \end{aligned} \tag{2.5}$$

yielding

$$E(\rho_n) \sim \frac{16}{3\sqrt{15}}\left(\frac{25}{16}\right)^n \approx (1.37706) \cdot (1.5625)^n.$$

All these results have already been given in a paper of Klazar [13]. Now, to find the covariances, one needs four generating functions connected by a system of equations. For the covariance of the σ- and Z-index, for example, we take $\mathrm{SZ}_{11}, \ldots, \mathrm{SZ}_{22}$ to be the generating functions for the product of the number of independent vertex subsets and independent edge subsets such that the root is contained in

- the vertex and the edge subset,
- the vertex, but not the edge subset,
- the edge, but not the vertex subset,
- neither,

respectively. The functional equations can be seen to be a combination of those for S_1 and S_2 resp. Z_1 and Z_2:

$$\begin{aligned} \mathrm{SZ}_{11}(z) &= \frac{z\,\mathrm{SZ}_{22}(z)}{(1 - \mathrm{SZ}_{21}(z) - \mathrm{SZ}_{22}(z))^2}, \\ \mathrm{SZ}_{12}(z) &= \frac{z}{1 - \mathrm{SZ}_{21}(z) - \mathrm{SZ}_{22}(z)}, \\ \mathrm{SZ}_{21}(z) &= \frac{z(\mathrm{SZ}_{12}(z) + \mathrm{SZ}_{22}(z))}{(1 - \mathrm{SZ}_{11}(z) - \mathrm{SZ}_{12}(z) - \mathrm{SZ}_{21}(z) - \mathrm{SZ}_{22}(z))^2}, \\ \mathrm{SZ}_{22}(z) &= \frac{z}{1 - \mathrm{SZ}_{11}(z) - \mathrm{SZ}_{12}(z) - \mathrm{SZ}_{21}(z) - \mathrm{SZ}_{22}(z)}. \end{aligned} \tag{2.6}$$

For instance, the functional equation for SZ_{11} is derived as follows:

$$\begin{aligned} \mathrm{SZ}_{11}(z) &= \sum_T \sigma_1(T) Z_1(T) z^{|T|} \\ &= \sum_{k\geq 0}\sum_{j=1}^{k}\sum_{T_1}\sum_{T_2}\cdots\sum_{T_k}\left(\sigma_2(T_j)Z_2(T_j)\prod_{i\neq j}\sigma_2(T_i)(Z_1(T_i) + Z_2(T_i))\right) \\ &\quad \cdot z^{|T_1|+\ldots+|T_k|+1} \\ &= z\sum_{k\geq 0} k\,\mathrm{SZ}_{22}(z)(\mathrm{SZ}_{21}(z) + \mathrm{SZ}_{22}(z))^{k-1} \\ &= \frac{z\,\mathrm{SZ}_{22}(z)}{(1 - SZ_{21}(z) - \mathrm{SZ}_{22}(z))^2}. \end{aligned}$$

Since all the functional equations can be written in polynomial form, it is possible to employ the method of Gröbner bases (cf. [6]) and a computer algebra package such as Mathematica® (for details, see [24]) to obtain a single polynomial equation from the system. In this case, we find that $s = \mathrm{SZ}_{22}(z)$ satisfies the polynomial equation

$$F(z,s) = s^{10} + 2zs^8 - 3zs^7 + z^2s^6 - 4z^2s^5 + 3z^2s^4 - z^3s^3 + 2z^3s^2 - z^3s + z^4 = 0.$$

Since $\mathrm{SZ}(z) = \mathrm{SZ}_{11}(z) + \mathrm{SZ}_{12}(z) + \mathrm{SZ}_{21}(z) + \mathrm{SZ}_{22}(z) = 1 - \frac{z}{\mathrm{SZ}_{22}(z)}$, the smallest singularity of SZ is either a singularity of SZ_{22} or a zero of SZ_{22}. However, from the functional equation we know that SZ_{22} has only one zero at $z = 0$, where the zero cancels out with the numerator. Therefore, we only have to find the smallest singularity of SZ_{22} to apply Bender's theorem. Fortunately, things are still comparatively simple since we can bound the range of the singularity by an a-priori estimate.

Again, the leading coefficient of the polynomial equation is 1, so it has no zeroes. Therefore, the dominating singularity is a solution of the system $F(z,w) = 0$, $F_w(z,w) = 0$ again. The solutions of this system can be found by the method of Gröbner bases as well – it turns out that a singularity z_0 of SZ must be a solution of

$$5038848z^4 - 221833728z^3 + 5017360096z^2 + 3451610880z - 387420489 = 0.$$

Now we note that, for trivial reasons, $1 \leq \sigma(T), Z(T), \rho(T) \leq 2^{|T|}$ for all trees T. This shows that the coefficients c_n of SZ are bounded by

$$\frac{1}{n}\binom{2n-2}{n-1} \leq c_n \leq \frac{1}{n}\binom{2n-2}{n-1} \cdot 4^n,$$

so the radius of convergence of SZ lies in the interval $\left[\frac{1}{16}, \frac{1}{4}\right]$. Thus we only have to search for a solution whose absolute value lies within this interval. There is only one such solution in this case, which is given by $z_0 \approx 0.0982673$. Expanding SZ_{22} and SZ around this singularity and applying Bender's formula yields an asymptotic expression for the expected product of σ- and Z-index:

$$E(\sigma_n Z_n) \sim (0.92565) \cdot (2.54408)^n.$$

Of course, the same way of reasoning can also be used to determine the other expected values $E(\sigma_n \rho_n)$ and $E(Z_n \rho_n)$ as well as the variances of all our random variables. All details (which are mostly analogous to the example) are given in [24]. Therefore, we only list all the asymptotics in Table 2.1.

Now we can turn to the correlation coefficients. We see that

$$\begin{aligned} r(\sigma_n, Z_n) &\sim (-1.01706) \cdot (0.99405)^n, \\ r(\sigma_n, \rho_n) &\sim (1.05088) \cdot (0.99023)^n, \\ r(Z_n, \rho_n) &\sim (-1.08924) \cdot (0.97853)^n, \end{aligned}$$

and conclude that the σ and ρ-index are positively correlated, whereas they are both negatively correlated to the Z-index. The correlation coefficient tends to zero as $n \to \infty$, but very slowly. The constant factor as well as the basis of the exponential term can be used as a measure for the correlation. So we may claim that the closest correlation of the three is between the σ- and the Z-index.

3. Correlation to the Wiener index. The Wiener index has a different recursive structure than the indices discussed in the preceding chapter, and its growth is not exponential. Entringer et al. [4] were able to show that the average Wiener

$E(\sigma_n)$	$\sqrt{3}\left(\frac{27}{16}\right)^{n-1} \sim (1.02640)\cdot(1.6875)^n$
$E(Z_n)$	$\sqrt{\frac{65-\sqrt{13}}{78}}\left(\frac{35+13\sqrt{13}}{54}\right)^n \sim (0.88719)\cdot(1.51615)^n$
$E(\rho_n)$	$\frac{16}{3\sqrt{15}}\left(\frac{25}{16}\right)^n \sim (1.37706)\cdot(1.5625)^n$
$E(\sigma_n Z_n)$	$(0.92565)\cdot(2.54408)^n$
$E(\sigma_n \rho_n)$	$(1.36653)\cdot(2.66477)^n$
$E(Z_n \rho_n)$	$\frac{1}{116}\sqrt{\frac{5(128985+57683\sqrt{5})}{58}}\cdot\left(8(7-3\sqrt{5})\right)^n \sim (1.28557)\cdot(2.33437)^n$
$\mathrm{Var}(\sigma_n)$	$(1.03802)\cdot(2.86096)^n$
$\mathrm{Var}(Z_n)$	$(0.77227)\cdot(2.31549)^n$
$\mathrm{Var}(\rho_n)$	$\frac{64\sqrt{14}}{147}\cdot\left(\frac{81}{32}\right)^n \sim (1.79509)\cdot(2.53125)^n$

TABLE 2.1
Asymptotic formulas for expected values and variances.

index is asymptotically $K \cdot n^{5/2}$ for a simply generated family of trees, where K is a constant depending on the specific family. For rooted ordered trees, the constant K is $\frac{\sqrt{\pi}}{4}$. We repeat their argument here since it will be needed for the computation of the covariances.

We are first going to consider an auxiliary value, $D(T)$, denoting the sum of the distances of all vertices from the root. This is also known as the *total height* [21] or *internal path length* [9] of the tree T. Then, we set

$$D(z) := \sum_T D(T) z^{|T|},$$

where the sum runs over all rooted ordered trees T again. The value $D(T)$ can be calculated recursively from the branches of T: in fact, if $T_1, \ldots, T_k$ are the branches of T, we have

$$D(T) = \sum_{i=1}^{k} D(T_i) + |T| - 1, \tag{3.1}$$

where $|T|$ is the size (number of vertices) of T. In terms of $D(z)$, this gives

$$\begin{aligned} D(z) &= \sum_T D(T) z^{|T|} \\ &= \sum_{k\ge 0}\sum_{i=1}^{k}\sum_{T_1}\sum_{T_2}\cdots\sum_{T_k} D(T_i) z^{|T_1|+\ldots|T_k|+1} + \sum_T (|T|-1) z^{|T|} \\ &= z\sum_{k\ge 0} kD(z)T(z)^{k-1} + zT'(z) - T(z) \\ &= \frac{zD(z)}{(1-T(z))^2} + zT'(z) - T(z). \end{aligned} \tag{3.2}$$

Now, the Wiener index of a tree can also be determined recursively from its branches:

$$W(T) = D(T) + \sum_{i=1}^{k} W(T_i) + \sum_{i\ne j}\Big(D(T_i) + |T_i|\Big)|T_j|, \tag{3.3}$$

where the last sum goes over all $k(k-1)$ pairs of different branches. Thus, if

$$W(z) := \sum_T W(T) z^{|T|},$$

we have

$$W(z) = D(z) + \frac{zW(z)}{(1-T(z))^2} + \frac{2z^2 T'(z)(D(z)+zT'(z))}{(1-T(z))^3}. \tag{3.4}$$

It turns out that $W(z) = \frac{z^2}{(1-4z)^2}$, giving an average Wiener index of asymptotically $\frac{\sqrt{\pi}}{4} n^{5/2}$. Now, we introduce various generating functions for the correlation of $D(T), W(T)$ and $\sigma(T)$: let $\mathrm{DS}_1, \mathrm{DS}_2, \mathrm{WS}_1$ and WS_2 be the generating functions for the product of $D(T)$ resp. $W(T)$ with the number of independent vertex subsets containing resp. not containing the root. In analogy to the functional equations for $D(z)$ and $W(z)$ we obtain a system of linear equations – for example, we have

$$\begin{aligned}
\mathrm{DS}_1(z) &= \sum_T D(T)\sigma_1(T) z^{|T|} \\
&= \sum_{k\geq 0} \sum_{T_1} \cdots \sum_{T_k} \left(\sum_{i=1}^k D(T_i) \prod_{j=1}^k \sigma_2(T_j) \right) z^{|T_1|+\ldots+|T_k|+1} \\
&\quad + \sum_T (|T|-1)\sigma_1(T) z^{|T|} \\
&= \sum_{k\geq 0} \sum_{T_1} \cdots \sum_{T_k} \left(\sum_{i=1}^k D(T_i)\sigma_2(T_i) \prod_{j\neq i} \sigma_2(T_j) \right) z^{|T_1|+\ldots+|T_k|+1} \\
&\quad + zS_1'(z) - S_1(z) \\
&= z \sum_{k\geq 0} k\, \mathrm{DS}_2(z) S_2(z)^{k-1} + zS_1'(z) - S_1(z) \\
&= \frac{z\,\mathrm{DS}_2(z)}{(1-S_2(z))^2} + zS_1'(z) - S_1(z).
\end{aligned}$$

Altogether, we obtain

$$\begin{aligned}
\mathrm{DS}_1(z) &= \frac{z\,\mathrm{DS}_2(z)}{(1-S_2(z))^2} + zS_1'(z) - S_1(z), \\
\mathrm{DS}_2(z) &= \frac{z(\mathrm{DS}_1(z)+\mathrm{DS}_2(z))}{(1-S_1(z)-S_2(z))^2} + zS_2'(z) - S_2(z), \\
\mathrm{WS}_1(z) &= \mathrm{DS}_1(z) + \frac{z\,\mathrm{WS}_2(z)}{(1-S_2(z))^2} + \frac{2z^2 S_2'(z)(\mathrm{DS}_2(z)+zS_2'(z))}{(1-S_2(z))^3}, \\
\mathrm{WS}_2(z) &= \mathrm{DS}_2(z) + \frac{z(\mathrm{WS}_1(z)+\mathrm{WS}_2(z))}{(1-S_1(z)-S_2(z))^2} \\
&\quad + \frac{2z(zS_1'(z)+zS_2'(z))(\mathrm{DS}_1(z)+\mathrm{DS}_2(z)+zS_1'(z)+zS_2'(z))}{(1-S_1(z)-S_2(z))^3}.
\end{aligned} \tag{3.5}$$

We solve this system for WS_1 and WS_2 (which can be done explicitly in terms of S_1 and S_2 since the system is linear) and write the total generating function $\mathrm{WS}(z) =$

$\mathrm{WS}_1(z) + \mathrm{WS}_2(z)$ in terms of S_1, S_2, S_1', S_2'. Then we make use of the functional equations for S_1 and S_2 and replace $S_1(z)$ by $\frac{z}{1-S_2(z)}$. Implicit differentiation of the equation $S_2(z)^3 - 2S_2(z)^2 + S_2(z) - z = 0$ yields

$$S_2'(z) = \frac{1}{3S_2(z)^2 - 4S_2(z) + 1},$$

so WS can be written in terms of S_2 and z only. In fact, we have

$$\mathrm{WS}(z) = \frac{N}{(1-3S_2(z))^2(1-S_2(z))^3(S_2(z)^2 + S_2(z)^3 - z)^2},$$

where N is a polynomial in S_2 and z. The denominator only vanishes at 0 and at the dominating singularity $\frac{4}{27}$ of S_2. Therefore, we only have to expand WS around $\frac{4}{27}$:

$$\mathrm{WS}(z) \sim \frac{5}{81\left(1 - \frac{27z}{4}\right)^2},$$

which gives us the expected value $E(W_n \sigma_n)$ by means of the Flajolet-Odlyzko singularity analysis [5] once again:

$$E(W_n \sigma_n) \sim \frac{20\sqrt{\pi}}{81} n^{5/2} \left(\frac{27}{16}\right)^n.$$

It was shown by Janson [9] that the variance of the Wiener index for rooted ordered trees is given asymptotically by

$$\mathrm{Var}(W_n) \sim \frac{16 - 5\pi}{80} n^5,$$

and thus the correlation coefficient of W_n and σ_n is

$$r(W_n, \sigma_n) \sim (-0.27891) \cdot (0.99767)^n.$$

Similarly, we obtain

$$r(W_n, Z_n) \sim (0.40351) \cdot (0.99637)^n,$$
$$r(W_n, \rho_n) \sim (-1.78357) \cdot (0.98209)^n.$$

Again, the calculational details are given in [24].

4. Some numerical values and their interpretation. We have seen that in all the considered cases, the correlation coefficient was asymptotically of the form

$$\alpha \cdot \beta^n$$

for some constants α and β. The significance of these constants can be roughly described as follows:

- A large value of α usually means a higher correlation for trees with few vertices.
- A large value of β means that the correlation decreases very slowly – thus, it is a measure for the correlation of the indices when the number of vertices is large.

When the correlation of σ, Z and ρ was considered, β depended on the growth of both indices. If the correlation was negative in these cases (which it was except for the correlation of σ- and ρ-index), the exact asymptotics of the expected value of their product were redundant for the asymptotics of the correlation coefficient. So, in order to exploit this piece of information as well, one should separately consider normalized values of the form

$$\frac{E(X_nY_n)}{\sqrt{\operatorname{Var}(X_n)\operatorname{Var}(Y_n)}} \text{ and } \frac{E(X_n)E(Y_n)}{\sqrt{\operatorname{Var}(X_n)\operatorname{Var}(Y_n)}},$$

where X_n and Y_n are X- and Y-indices of random trees.
Further problems arise in the study of the Wiener index. Since the Wiener index only grows polynomially, β only depends on the expected value and variance of the second index. Again, one should also consider the coefficients given above separately. We have seen that they are of the same asymptotic order except from the constant factors, so one might use their quotient as a correlation measure as well. The following table gives the asymptotic behavior of these coefficients and their quotient:

Indices	$\frac{E(X_nY_n)}{\sqrt{\operatorname{Var}(X_n)\operatorname{Var}(Y_n)}}$	$\frac{E(X_n)E(Y_n)}{\sqrt{\operatorname{Var}(X_n)\operatorname{Var}(Y_n)}}$	$\frac{E(X_nY_n)}{E(X_n)E(Y_n)}$
$\sigma - Z$	$(1.03386) \cdot (0.988448)^n$	$(1.01706) \cdot (0.99405)^n$	$(1.01652) \cdot (0.99436)^n$
$\sigma - \rho$	$(1.05088) \cdot (0.99023)^n$	$(1.08694) \cdot (0.97981)^n$	$(0.96683) \cdot (1.01064)^n$
$Z - \rho$	$(1.14617) \cdot (0.96423)^n$	$(1.08924) \cdot (0.97853)^n$	$(1.05227) \cdot (0.98539)^n$
$\sigma - W$	$(7.10957) \cdot (0.99767)^n$	$(7.38848) \cdot (0.99767)^n$	0.96225
$Z - W$	$(7.80764) \cdot (0.99637)^n$	$(7.40413) \cdot (0.99637)^n$	1.05450
$\rho - W$	$(6.12924) \cdot (0.98209)^n$	$(7.91281) \cdot (0.98209)^n$	0.77460

Table 4.1
$E(X_nY_n)$ and $E(X_n)E(Y_n)$ separated.

In any case, our approach will only yield us quantitative correlation measures; qualitative information on the correlation structure is not provided.

One can calculate the exact correlation coefficients for small values of n quite easily from the functional equations. In Table 4.2, some numerical examples are given – note that the correlation coefficient only makes sense for $n \geq 4$: for $n \leq 3$, all trees are isomorphic.

We see that the correlation coefficient between σ- and Z-index is largest among those investigated in section 2. Likewise, the correlation to the Wiener index is highest for the ρ-index. This observation agrees with the asymptotic results of the preceding sections. The following plots (Fig. 4.1) suggest that the correlation is in fact very strong in both cases (much stronger than for the other pairs, which is quite remarkable), but not entirely linear, which is clear from the exponential growth of σ-, Z- and ρ-index (this phenomenon will be discussed in detail in the following section). The plots show the values of all trees with 12 vertices.

5. Other correlation measures. Unfortunately, there are some drawbacks in our approach. Apart from the obvious fact that asymptotic correlations might only hold for a considerably large number of vertices, the correlation coefficient principally measures linear dependence. But since the $\sigma-$, $Z-$ and $\rho-$ indices grow exponentially

n	$r(\sigma_n, Z_n)$	$r(\sigma_n, \rho_n)$	$r(Z_n, \rho_n)$	$r(\sigma_n, W_n)$	$r(Z_n, W_n)$	$r(\rho_n, W_n)$
4	-1.000000	1.000000	-1.000000	-1.000000	1.000000	-1.000000
5	-0.991189	0.971494	-0.994334	-0.923381	0.966092	-0.988064
6	-0.970054	0.947369	-0.955649	-0.870581	0.918482	-0.977131
7	-0.959741	0.926080	-0.926321	-0.829908	0.883867	-0.966673
8	-0.950801	0.907123	-0.898558	-0.796570	0.853248	-0.956356
9	-0.943296	0.890225	-0.873371	-0.768197	0.826459	-0.945962
10	-0.936479	0.875159	-0.850213	-0.743446	0.802492	-0.935353
11	-0.930116	0.861703	-0.828817	-0.721477	0.780828	-0.924449
12	-0.924048	0.849641	-0.808906	-0.701723	0.761060	-0.913214
13	-0.918187	0.838772	-0.790246	-0.683782	0.742891	-0.901641
14	-0.912479	0.828909	-0.772640	-0.667357	0.726088	-0.889750
15	-0.906888	0.819890	-0.755923	-0.652218	0.710467	-0.877574
20	-0.880077	0.783214	-0.681768	-0.590624	0.645700	-0.814057
25	-0.854498	0.753917	-0.617683	-0.544547	0.596088	-0.750155

TABLE 4.2
Correlation coefficients for rooted ordered trees, $n \leq 25$.

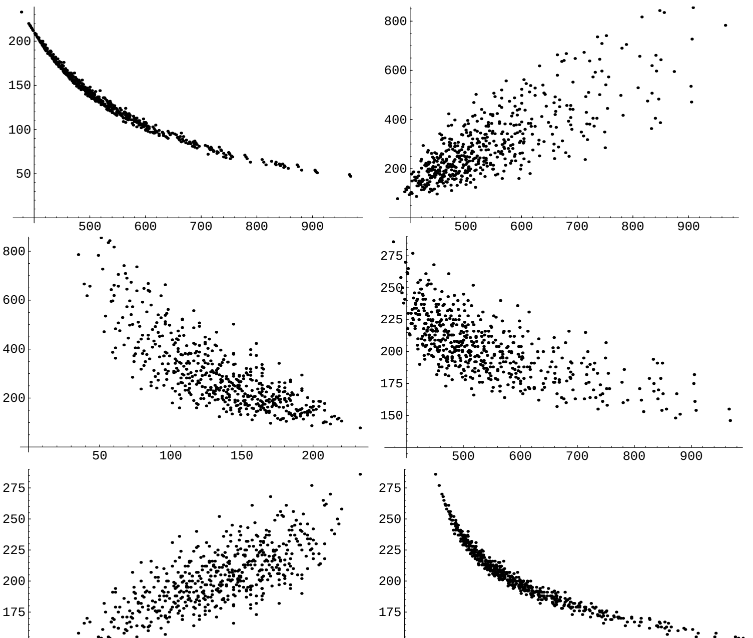

FIG. 4.1. *From top to bottom: σ- and Z-index, σ- and ρ-index, Z- and ρ-index, σ- and Wiener index, Z- and Wiener index, ρ- and Wiener index.*

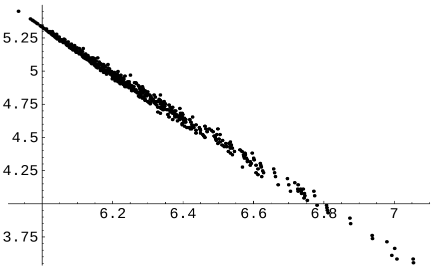

FIG. 5.1. *σ- and Z-index after logarithmic transformation.*

with different growth rates, the dependence cannot be completely linear. Thus, it might be reasonable to study the correlation of their logarithms instead. The problem with that approach is the fact that generating function methods as presented in this paper will not be applicable any longer. The corresponding plot for the correlation of $\log \sigma_n$ and $\log Z_n$ (the random variables are rescaled in such a way that they are of equal order now!) suggests that it is reasonable to use a logarithmic transformation – it shows an almost linear correspondence (Fig. 5.1). This suggests that a sharp inequality of the form

$$f_n(\sigma(T)) \leq Z(T) \leq g_n(\sigma(T)) \tag{5.1}$$

should hold for all trees T on n vertices, where $f_n(x), g_n(x)$ behave like negative powers of x, i.e. $f_n(x) \sim a_1(n)x^{-c_1}$, $g_n(x) \sim a_2(n)x^{-c_2}$. However, it is not difficult to construct discordant pairs of trees, i.e. two trees T_1, T_2 such that $Z(T_1) > Z(T_2)$ and $\sigma(T_1) > \sigma(T_2)$.

This leads us to an alternative method of measuring correlation – the use of rank statistics (cf. [10, 14]): given two indices X and Y, we assign ranks x_i and y_i to all trees $T_1, \ldots, T_s$ on n vertices such that x_i and y_i range from 1 to s and $x_i < x_j$ if $X(T_i) < X(T_j)$ resp. $y_i < y_j$ if $Y(T_i) < Y(T_j)$. Then, a correlation measure is given by Spearman's ρ:

$$\rho_S(X_n, Y_n) = 1 - \frac{6\sum_{i=1}^{s}(x_i - y_i)^2}{s^3 - s} \tag{5.2}$$

which ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation). Unfortunately, even though rank statistics are an interesting means of measuring the statistical dependence of random variables, it seems virtually impossible to apply them to our problem, since generating function methods are not apt to the treatment of ranks. It seems that rank statistics can only be applied to our problem if the number of vertices is considerably small, so that everything can be calculated explicitly.

Another problem with them is the occurrence of ties – all the random variables under consideration are discrete, and the number of trees grows larger than the maximal index in all our cases, so ties (i.e. several non-isomorphic trees of the same index) are inevitable. There are statistical methods to cope with this problem (cf. [10, 14]) – usually, if ties occur, the average rank is allotted to all tied elements. This method is used in the examples at the end of this section.

The problem of ties leads us to our final remark. The methods of this paper easily generalize to all simply generated families of trees. However, one would like to apply them to unordered rooted trees or trees (so one can take isomorphisms into account).

This should be doable (in essentially the same way as in [25]), but is certainly requires very lengthy calculations.

In the following table, correlation coefficients for trees with ≤ 14 vertices are given. If we compare them to the values of Table 4.2, we see that the correlation coefficients for rooted ordered trees provide suitable estimates.

n	$r(\sigma_n, Z_n)$	$r(\sigma_n, \rho_n)$	$r(Z_n, \rho_n)$	$r(\sigma_n, W_n)$	$r(Z_n, W_n)$	$r(\rho_n, W_n)$
4	-1.000000	1.000000	-1.000000	-1.000000	1.000000	-1.000000
5	-0.995871	0.986241	-0.997176	-0.960769	0.981981	-0.993399
6	-0.977051	0.969611	-0.982970	-0.901473	0.953231	-0.977255
7	-0.955329	0.959254	-0.943865	-0.863896	0.911843	-0.959471
8	-0.930868	0.947142	-0.918181	-0.819996	0.886845	-0.940935
9	-0.908594	0.932074	-0.869200	-0.778345	0.841803	-0.91815
10	-0.890714	0.920543	-0.836300	-0.748034	0.816189	-0.899454
11	-0.877343	0.903475	-0.797497	-0.714065	0.782806	-0.879018
12	-0.869047	0.889422	-0.767693	-0.689129	0.758290	-0.860836
13	-0.862946	0.872456	-0.739304	-0.663493	0.732342	-0.843721
14	-0.859211	0.857532	-0.715078	-0.642464	0.710476	-0.827013

TABLE 5.1
Correlation coefficients for trees, $n \leq 14$.

Finally, we examine the rank correlation. The table shows the numerical values of Spearman's ρ for all trees with ≤ 14 vertices.

n	$\rho_S(\sigma_n, Z_n)$	$\rho_S(\sigma_n, \rho_n)$	$\rho_S(Z_n, \rho_n)$	$\rho_S(\sigma_n, W_n)$	$\rho_S(Z_n, W_n)$	$\rho_S(\rho_n, W_n)$
4	-1.000000	1.000000	-1.000000	-1.000000	1.000000	-1.000000
5	-1.000000	1.000000	-1.000000	-1.000000	1.000000	-1.000000
6	-1.000000	0.942857	-0.942857	-0.942857	0.942857	-1.000000
7	-1.000000	0.918182	-0.918182	-0.877273	0.886364	-0.986364
8	-0.994071	0.881670	-0.876729	-0.867836	0.870800	-0.996789
9	-0.996126	0.854591	-0.852798	-0.805273	0.809349	-0.990171
10	-0.997048	0.832577	-0.834320	-0.774514	0.777381	-0.992314
11	-0.997392	0.811737	-0.814267	-0.746093	0.749423	-0.990921
12	-0.997471	0.796388	-0.801514	-0.724382	0.729450	-0.990146
13	-0.997421	0.781437	-0.787808	-0.697123	0.703244	-0.987169
14	-0.997383	0.770002	-0.777472	-0.675956	0.682617	-0.984820

TABLE 5.2
Spearman's ρ for $n \leq 14$.

Again, we observe the striking correspondence between σ- and Z-index resp. ρ- and Wiener index. It seems to be a challenging graph-theoretical problem to explain this phenomenon.

Acknowledgement. The author thanks the referees for their valuable help in improving this paper and in particular for bringing references [2, 18] to his attention.

REFERENCES

[1] Edward A. Bender, *Asymptotic methods in enumeration*, SIAM Rev., 16 (1974), pp. 485–515.

[2] E. R. Canfield, *Remarks on an asymptotic method in combinatorics*, J. Combin. Theory Ser. A, 37 (1984), pp. 348–352.

[3] A. A. Dobrynin, R. Entringer, and I. Gutman, *Wiener index of trees: theory and applications*, Acta Appl. Math., 66 (2001), pp. 211–249.

[4] R. C. Entringer, A. Meir, J. W. Moon, and L. A. Székely, *The Wiener index of trees from certain families*, Australas. J. Combin., 10 (1994), pp. 211–224.

[5] P. Flajolet and A. Odlyzko, *Singularity analysis of generating functions*, SIAM J. Discrete Math., 3 (1990), pp. 216–240.

[6] R. Fröberg, *An introduction to Gröbner bases*, Pure and Applied Mathematics (New York), John Wiley & Sons Ltd., Chichester, 1997.

[7] E. Hille, *Analytic function theory. Vol. II*, Introductions to Higher Mathematics, Ginn and Co., Boston, Mass.-New York-Toronto, Ont., 1962.

[8] H. Hosoya, *Topological index as a common tool for quantum chemistry, statistical mechanics, and graph theory*, in Mathematical and computational concepts in chemistry (Dubrovnik, 1985), Ellis Horwood Ser. Math. Appl., Horwood, Chichester, 1986, pp. 110–123.

[9] S. Janson, *The Wiener index of simply generated random trees*, Random Structures Algorithms, 22 (2003), pp. 337–358.

[10] M. Kendall and J. D. Gibbons, *Rank correlation methods*, A Charles Griffin Title, Edward Arnold, London, fifth ed., 1990.

[11] P. Kirschenhofer, H. Prodinger, and R. F. Tichy, *Fibonacci numbers of graphs. II*, Fibonacci Quart., 21 (1983), pp. 219–229.

[12] ———, *Fibonacci numbers of graphs. III. Planted plane trees*, in Fibonacci numbers and their applications (Patras, 1984), vol. 28 of Math. Appl., Reidel, Dordrecht, 1986, pp. 105–120.

[13] M. Klazar, *Twelve countings with rooted plane trees*, European J. Combin., 18 (1997), pp. 195–210.

[14] E. L. Lehmann, *Nonparametrics: statistical methods based on ranks*, Holden-Day Inc., San Francisco, Calif., 1975. With the special assistance of H. J. M. d'Abrera, Holden-Day Series in Probability and Statistics.

[15] M. Lepović and I. Gutman, *A Collective Property of Trees and Chemical Trees*, J. Chem. Inf. Comput. Sci., 38 (1998), pp. 823–826.

[16] X. Li, Z. Li, and L. Wang, *The inverse problems for some topological indices in combinatorial chemistry*, J. Computational Biology, 10 (2003), pp. 47–55.

[17] A. Meir and J. W. Moon, *On the altitude of nodes in random trees*, Canad. J. Math., 30 (1978), pp. 997–1015.

[18] ———, *On an asymptotic method in enumeration*, J. Combin. Theory Ser. A, 51 (1989), pp. 77–89.

[19] R. E. Merrifield and H. E. Simmons, *Topological Methods in Chemistry*, Wiley, New York, 1989.

[20] H. Prodinger and R. F. Tichy, *Fibonacci numbers of graphs*, Fibonacci Quart., 20 (1982), pp. 16–21.

[21] J. Riordan and N. J. A. Sloane, *The enumeration of rooted trees by total height*, J. Austral. Math. Soc., 10 (1969), pp. 278–282.

[22] L. A. Székely and H. Wang, *On subtrees of trees*, Adv. in Appl. Math., 34 (2005), pp. 138–155.

[23] N. Trinajstić, *Chemical graph theory*, CRC Press, Boca Raton, FL., 1992.

[24] S. Wagner, *Calculating the correlation of graph-theoretical indices.* arXiv:math.CO/0608753; also available at `http://finanz.math.tugraz.at/~wagner/Correlation`, 2006.

[25] ———, *Subset counting in trees.* To appear in Ars Combinatoria, 2006.

[26] H. Wiener, *Structural determination of paraffin boiling points*, J. Amer. Chem. Soc., 69 (1947), pp. 17–20.