MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **54** (2005) 177-187

ISSN 0340 - 6253

The Covariance of Topological Indices that Depend on the Degree of a Vertex

Boris Hollas

Theoretische Informatik, Universität Ulm, D-89081 Ulm
E-mail: hollas@informatik.uni-ulm.de

Received August 23, 2004

Abstract

We consider topological indices $\mathcal{I}$ that are sums of $f(\deg(u), \deg(v))$, where u, v are adjacent vertices and f is a function. The Randić connectivity index, the 2nd Zagreb index or the Platt number are examples for indices of this kind. In earlier work on topological indices that are sums of independent random variables, we identified the correlation between $\mathcal{I}$ and the number of edges of the molecular graph as the main cause for correlated indices. For random graphs on a Poisson-distributed number of vertices, we show how $\mathcal{I}$ can be transformed to be uncorrelated to the number of edges. More important, we give reason to assume that this should not simply transform a linear to a non-linear dependence. A similar result holds for zero-order indices, i.e. topological indices that are sums of $f(\deg(v))$ as for example the 1st Zagreb index.

1 Introduction

For quite some time it has been known that topological indices (graph invariants on molecular graphs) exhibit considerable mutual correlation [1, 2]. This is a major problem when performing structure-activity studies as the employed statistical methods may fail or give little meaningful results on sets of correlated data. In addition, strong correlations among a set of topological indices raise doubt whether these indices describe different and meaningful biological, chemical or physical properties of molecules. Many topological indices also depend on the number of bonds in a molecule, which is information that should better be coded spearately.

Principal component analysis (PCA) is sometimes used to obtain uncorrelated descriptors from a set of correlated descriptors. By construction, the new set of descriptors is a linear combination of the old ones. Hence, each new descriptor may be a combination of all previous descriptors, which complicates model interpretability. Also, uncorrelated descriptors may still be non-linearly related.

In an attempt to investigate the reasons for these correlations, we used random graphs [3] as a model for chemical graphs and for topological indices of the form

$$\mathcal{I}_{\mathbf{X}}(G) = \sum_{\{u,v\} \in E} X_u X_v$$

where E is the edge set of the molecular graph $G = (V, E)$ and $\mathbf{X} = \{X_v \mid v \in V\}$ is a set of independent random variables with a common expectation $\boldsymbol{E}(X)$ [4, 5, 6]. We proved that $\mathcal{I}_{\mathbf{X}}$, $\mathcal{I}_{\mathbf{Y}}$, and $\mathcal{I}_{\mathbf{1}}$ are necessarily dependent for independent sets of vertex properties $\mathbf{X}, \mathbf{Y}$ with $\boldsymbol{E}(X), \boldsymbol{E}(Y) \neq 0$. For $\boldsymbol{E}(X) = \boldsymbol{E}(Y) = 0$ however these indices are uncorrelated. Here, $\mathcal{I}_{\mathbf{1}}$ denotes a topological index with $X_v = 1$ for all $v \in V$, that is, $\mathcal{I}_{\mathbf{1}} = |E|$. In the first case, these indices are thus correlated as a result of the graph invariant used, not as a result of similar chemical properties.

While the random graph model we used in [6] encompasses graphs of arbitrary structure, including chemical graphs, the notion of vertex (or atom) properties X_v that are *independent* of the molecular graph is a serious abstraction from computational chemistry where atom properties used for topological indices are a *function* of the graph or even the molecule.

In this paper, we use a special case of the random graph model used in [6]. In particular, the number of vertices is a Poisson-distributed random variable N. For a fixed number of vertices $N = n$, edges are chosen independently with a probability $p_n \in \Theta(1/n)$ so that the expected number of edges is a linear function of n. We use this to model an approximately linear relation of bonds to vertices present in molecules. For example, homologous series of aliphatic or aromatic hydrocarbons with n atoms contain $n + c$ bonds for some constant c. Polyphenyls contain $\frac{7}{6}n + c$ bonds as each monomer adds 6 atoms and 7 bonds. On the other hand, there is some variation in the number of bonds for a given number of atoms in a heterogenous set of molecules, which is also true for the random graph model. Note however that this random graph model describes only some specific aspects of chemical graphs, namely the average degree of a vertex and a variation in the number of vertices and edges. Random graphs may not be connected or may contain vertices of degree greater 4.

As a more significant difference we consider the vertex properties to be a function of the vertex degree instead of being independent. Thus, our results are valid for topological indices such as the Randić connectivity index, the second Zagreb index or the Platt number [7].

Since we discovered that topological indices that are correlated with $\mathcal{I}_{\mathbf{1}}$ are also mutually correlated within the setting of [4, 5, 6], we focus on the covariance between $\mathcal{I}_{\mathbf{X}}$ and $\mathcal{I}_{\mathbf{1}}$. We show how to make $\mathcal{I}_{\mathbf{X}}$ uncorrelated with $\mathcal{I}_{\mathbf{1}}$ in a way that should not result in a non-linear dependence. In section 7 we also consider zero-order indices, which include the first Zagreb index and the zero-order Kier & Hall index.

2 Preliminaries

The random graph model is constructed in two steps. We shall first consider a fixed, then a variable number of vertices. In the first step, we obtain random graph model $\mathscr{G}(n, p_n)$. From here, we obtain model $\mathscr{G}(N, p_N)$ by the use of expectation.

For a graph (V, E) on the vertex set $V = \{1, \ldots, N\}$ where N is a random variable let

$$1_{uv} = 1_{\{\{u,v\}\in E\}} = \begin{cases} 1 & \text{if } \{u, v\} \in E \\ 0 & \text{else} \end{cases}$$

be the indicator function for $\{\{u, v\} \in E\}$. For N fixed, let 1_{uv} $(u, v \in V)$ be independent random variables with $P(1_{uv} = 1 \mid N = n) = p_n$ for some $p_n \in (0, 1)$. The space of random graphs $\mathscr{G}(n, p_n)$ can be identified with the distribution of $(1_{uv})_{u,v\in V}$ with respect to $P(\cdot \mid N = n)$, that is, all edges are chosen independently with probability p_n. We choose the edge-probabilities p_n in a way such that $\boldsymbol{E}|E| = \alpha(n - 2)$ for a fixed parameter $\alpha > 0$ as motivated in the introduction. For the expected degree of a vertex follows

$$\boldsymbol{E}(\deg(v)) = \frac{2\boldsymbol{E}|E|}{n} \sim 2\alpha$$

The parameter α thus describes the branching of the graphs.

To let the number of vertices vary, let N be a Poisson-distributed random variable and $\mathscr{G}(N, p_N)$ a space of random graphs with distribution

$$P(G) = \boldsymbol{E}(P(G \mid N)) = \sum_n P(G|N = n)\, P(N = n)$$

To describe the vertex properties, let $f : \mathbb{N}^2 \to \mathbb{R}$ be a function and let

$$X_{uv} = f(\deg(u), \deg(v)) \tag{2.1}$$

be the properties of vertices u, v. We consider the topological index

$$\mathcal{I}_{\mathbf{X}} = \mathcal{I}_{\mathbf{X}}(G) = \sum_{\{u,v\}\in E(G)} X_{uv} \tag{2.2}$$

where G is a random graph. In section 3 we will write this as

$$\mathcal{I}_{\mathbf{X}} = \sum_{u<v} X_{uv} 1_{uv} \tag{2.3}$$

which is better suited to employ the expectation operator.

In section 7, we also consider topological indices of the form

$$\mathcal{I}_{\mathbf{X}} = \sum_{v=1}^{N} X_v$$

with $X_v = f(\deg(v))$ for a function f.

We use the following notations throughout the text:

$O(f)$	denotes	a function g with $g(x) \le cf(x)$ for all large x and a constant $c > 0$
$X_n \xrightarrow{\mathscr{D}} X$	denotes	that random element X_n converges to X in distribution
$\mathscr{P}_\alpha$	denotes	the Poisson distribution with parameter α.

3 Expectations for n fixed

In this section, we use the random graph model $\mathscr{G}(n, p_n)$, that is, all expectation values are defined via $P(\cdot \mid N = n)$.

To determine expectation values, we have to eliminate the dependence among X_{uv} and $(1_{uv})_{u,v\in V}$ in (2.1). Therefore, we define the conditional expectation

$$\delta_{f,n}^{(i,j)} = \boldsymbol{E}\left(X_{12} \mid 1_{12} \cdot \prod_{k=3}^{i+1} 1_{1k} \cdot \prod_{l=3}^{j+1} 1_{2l} = 1 \right) \tag{3.1}$$

Since 1_{uv} $(u, v \in V)$ are independent and identically distributed, we have $\delta_{f,n}^{(i,j)} = \delta_{f,n}^{(j,i)}$.

Note that this definition does not depend on the choice of $(u, v) = (1, 2)$ since all X_{uv} $(u, v \in V)$ have the same distribution. As we shall see in section 4, $\lim_{n\to\infty} \delta_{f,n}^{(i,j)}$ exists and is a function of α if f does not grow too steeply. Thus, we may regard $\delta_{f,n}^{(i,j)}$ as almost constant for large n.

Lemma 1.
$\boldsymbol{E}(\mathcal{I}_{\mathbf{X}}) = \delta_{f,n}^{(1,1)} \boldsymbol{E}|E|$

Proof.

$$\begin{aligned} \boldsymbol{E}(\mathcal{I}_{\mathbf{X}}) &= \sum_{u<v} \boldsymbol{E}(X_{uv} \mid 1_{uv} = 1)\, p_n && \text{by (2.3)} \\ &= \delta_{f,n}^{(1,1)} \boldsymbol{E}|E| && \text{by (3.1)} \end{aligned}$$

□

Lemma 2.

$$\boldsymbol{E}(\mathcal{I}_{\mathbf{X}}\mathcal{I}_{\mathbf{1}}) = \left[\delta_{f,n}^{(1,1)} \binom{n-2}{2} p_n + 2\delta_{f,n}^{(1,2)} (n-2) p_n + \delta_{f,n}^{(1,1)} \right] \boldsymbol{E}|E|$$

Proof. To dissect the sum

$$\boldsymbol{E}(\mathcal{I}_{\mathbf{X}}\mathcal{I}_{\mathbf{1}}) = \sum_{u<v} \sum_{u'<v'} \boldsymbol{E}(X_{uv}1_{uv}1_{u'v'})$$

according to $|\{u,v\} \cap \{u',v'\}|$, consider

$$S_k = \{(u,v,u',v') \mid u<v \wedge u'<v' \wedge |\{u,v\} \cap \{u',v'\}| = k\}, \quad 0 \le k \le 2$$

Then

$$|S_0| = \binom{n}{2}\binom{n-2}{2} \tag{3.2}$$

$$|S_1| = 6\binom{n}{3} \tag{3.3}$$

$$|S_2| = \binom{n}{2} \tag{3.4}$$

(3.2) and (3.4) are obvious. To verify (3.3) let $(u,v,u',v') \in S_1$. Exactly two numbers are equal as indicated in figure 1. Cases (a), (b) allow just one way to distribute three distinct numbers on u,v,u',v' while there are two ways for cases (c), (d). For symmetry reasons, $\boldsymbol{E}(X_{uv}1_{uv}1_{u'v'}) =$

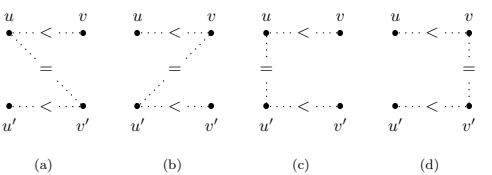

Figure 1: Possibilities for $(u,v,u',v') \in S_1$

$\boldsymbol{E}(X_{12}1_{12}1_{13})$ for all $(u,v,u',v') \in S_1$. Hence, we get

$$\begin{aligned}
\boldsymbol{E}(\mathcal{I}_{\mathbf{X}}\mathcal{I}_{\mathbf{1}}) &= |S_0|\boldsymbol{E}(X_{12}1_{12}1_{34}) \\
&\quad + |S_1|\boldsymbol{E}(X_{12}1_{12}1_{13}) \\
&\quad + |S_2|\boldsymbol{E}\big(X_{12}1_{12}^2\big) \\
&= |S_0|\boldsymbol{E}(X_{12} \mid 1_{12}=1)\, p_n^2 \\
&\quad + |S_1|\boldsymbol{E}(X_{12} \mid 1_{12}1_{13}=1)\, p_n^2 \\
&\quad + |S_2|\boldsymbol{E}(X_{12} \mid 1_{12}=1)\, p_n
\end{aligned} \tag{3.5}$$

With

$$\binom{n}{3} = \binom{n}{2}\frac{n-2}{3}$$

and (3.2)-(3.4), (3.5), we have

$$\begin{aligned}
\boldsymbol{E}(\mathcal{I}_{\mathbf{X}}\mathcal{I}_{\mathbf{1}}) &= \delta_{f,n}^{(1,1)}\boldsymbol{E}|E|\binom{n-2}{2}p_n \\
&\quad + 2\delta_{f,n}^{(1,2)}\boldsymbol{E}|E|(n-2)p_n \\
&\quad + \delta_{f,n}^{(1,1)}\boldsymbol{E}|E|
\end{aligned}$$

□

Remark. With $f \equiv 1$, lemma 1 and the help of Mathematica it follows $Var(\mathcal{I}_{\mathbf{1}}) = \boldsymbol{E}|E|(1-p_n)$, as it should be.

4 Convergence of $\delta_{f,n}^{(i,j)}$

In this section, we prove a result on the convergence of $\delta_{f,n}^{(i,j)}$ for $n \to \infty$. We need this to generalize the results in section 3 for a Poisson-distributed number of vertices and to justify the main conclusion of this paper.

Recall that for random vectors X_n, X in $\mathbb{R}^d$ holds

$$X_n \xrightarrow{\mathscr{D}} X$$

iff

$$\boldsymbol{E}(f(X_n)) \to \boldsymbol{E}(f(X))$$

for all bounded and continuous functions $f : \mathbb{R}^d \to \mathbb{R}^d$. This does not hold for arbitrary unbounded functions f. Therefore, we require that f is bounded by an exponential function in theorem 3. This does not have to be the most general restriction but it is sufficient to treat currently used topological indices.

Theorem 3.
If $|f(x,y)| \leq b^{x+y}$ for a constant $b > 0$ and $lim_{n\to\infty} p_n n = \alpha$, then for all $i, j \in \mathbb{N}$

$$\lim_{n\to\infty} \delta_{f,n}^{(i,j)} = \boldsymbol{E}(f(i + P_1, j + P_2))$$

where P_1, P_2 are independent and $\mathscr{P}_\alpha$-distributed random variables.

Proof. First, we show that the claims holds for bounded functions. Let

$$S_{n,i}^{(1)} = i + \sum_{k=i+2}^{n} 1_{1k}$$

and

$$S_{n,j}^{(2)} = j + \sum_{k=j+2}^{n} 1_{2k}$$

By definition (3.1),

$$\begin{aligned} \delta_{f,n}^{(i,j)} &= \boldsymbol{E}\left(X_{12} \mid 1_{12} \cdot \prod_{k=3}^{i+1} 1_{1k} \cdot \prod_{l=3}^{j+1} 1_{2l} = 1 \right) \\ &= \boldsymbol{E}\Big(f\left(S_{n,i}^{(1)}, S_{n,j}^{(2)} \right) \Big) \\ &= \boldsymbol{E}(f(\mathbf{S}_n)) \end{aligned} \tag{4.1}$$

if we write $\mathbf{S}_n = \left(S_{n,i}^{(1)}, S_{n,j}^{(2)} \right)$. Since $p_n n \to \alpha$, Poisson's limit theorem gives

$$\sum_{k=i+2}^{n} 1_{1k} \xrightarrow{\mathscr{D}} \mathscr{P}_\alpha \quad (n \to \infty)$$

The function $f : \mathbb{N}^2 \to \mathbb{R}$ can be extended to a continuous function $f : \mathbb{R}^2 \to \mathbb{R}$ in an arbitrary way. Hence, the continuous mapping theorem gives

$$f(\mathbf{S}_n) \xrightarrow{\mathscr{D}} f(i + P_1, j + P_2) \quad (n \to \infty)$$

For all bounded and continuous functions $f^* : \mathbb{R}^2 \to \mathbb{R}$ follows by (4.1)

$$\delta_{f^*,n}^{(i,j)} \to \boldsymbol{E}(f^*(i + P_1, j + P_1)) \quad (n \to \infty) \tag{4.2}$$

To prove (4.2) for arbitrary functions f with $|f(x,y)| \leq b^{x+y}$ we cut f off above a limit to divide f into a bounded and an unbounded part. We show that the unbounded part tends to zero as the limit tends to infinity.

To do so, let

$$c_m(x) = \begin{cases} x & \text{if } |x| < m \\ 0 & \text{else} \end{cases}$$

and

$$\tilde{c}_m(x) = \begin{cases} 0 & \text{if } |x| < m \\ x & \text{else} \end{cases}$$

Then

$$\begin{aligned} |\boldsymbol{E}((\tilde{c}_m \circ f)(\mathbf{S}_n))| &= |\boldsymbol{E}\big(f(\mathbf{S}_n) 1_{\{f(\mathbf{S}_n) \geq m\}}\big)| \\ &\leq \boldsymbol{E}\left(f(\mathbf{S}_n)\frac{f(\mathbf{S}_n)}{m}\right) \\ &\leq \frac{1}{m}\boldsymbol{E}\left(b^{2S_{n,i}^{(1)}+2S_{n,j}^{(2)}}\right) \\ &= \frac{1}{m} b^{2i}\boldsymbol{E}\big(b^{2\cdot 1_{12}}\big)^{n-(i-1)}\, b^{2j}\boldsymbol{E}\big(b^{2\cdot 1_{12}}\big)^{n-(j-1)} \end{aligned}$$

since $S_{n,i}^{(1)}$ and $S_{n,j}^{(2)}$ are independent

$$\begin{aligned} &= \frac{b^{2(i+j)}}{m} \cdot \frac{(1+b^2 p_n)^{2n}}{(1+b^2 p_n)^{i+j-2}} \\ &= O(1/m) \end{aligned} \tag{4.3}$$

since $p_n n \to \alpha$. Thus,

$$\lim_{n\to\infty} \delta_{f,n}^{(i,j)} = \lim_{m\to\infty} \lim_{n\to\infty} \boldsymbol{E}(f(\mathbf{S}_n))$$

by (4.1)

$$\begin{aligned} &= \lim_{m\to\infty} \lim_{n\to\infty} [\boldsymbol{E}((c_m \circ f)(\mathbf{S}_n)) + \boldsymbol{E}((\tilde{c}_m \circ f)(\mathbf{S}_n))] \\ &= \lim_{m\to\infty} [\boldsymbol{E}((c_m \circ f)(i+P_1, j+P_2)) + O(1/m)] \end{aligned}$$

by (4.2) and (4.3)

$$= \boldsymbol{E}(f(i+P_2, j+P_1))$$

by the convergence theorem of Lebesgue. □

5 Expectations for Poisson-distributed N

In this section, we use the random graph model $\mathscr{G}(N, p_N)$. Hence, the number of vertices N is not constant anymore. As we need the results of section 3 here, we denote by E_n the edge set of a random graph in $\mathscr{G}(n, p_n)$, i. e. on a *fixed* number of vertices, whereas E is the edge set of a graph in $\mathscr{G}(N, p_N)$.

To circumvent dividing by zero and other technical problems we require that $N \geq 2$ in this section. Therefore, N is a random variable with

$$P(N=n) = \begin{cases} \frac{\beta^{n-2}}{(n-2)!} e^{-\beta} & \text{if } n \geq 2 \\ 0 & \text{else} \end{cases}$$

for $\beta > 0$. In the strict sense, N is thus not Poisson-distributed, but $N-2$ is. Since $\boldsymbol{E}(N) = \boldsymbol{E}(N-2)+2 = \beta+2$, the parameter β describes the expectation of N.

For $N = n$ and $\alpha \in (0,3)$ we define the edge-probabilities p_n as

$$p_n = \alpha \frac{n-2}{\binom{n}{2}}$$

so that $\boldsymbol{E}|E_n| = \alpha(n-2)$. This results in the useful properties

$$\begin{aligned} \boldsymbol{E}|E| &= \sum_{n\geq 2} \boldsymbol{E}|E_n| P(N=n) \\ &= \sum_{n\geq 3} \alpha\beta \frac{\beta^{n-3}}{(n-3)!} e^{-\beta} \\ &= \alpha\beta \end{aligned} \tag{5.1}$$

and

$$\boldsymbol{E}|E_n| P(N=n) = \boldsymbol{E}|E| P(N=n-1) \tag{5.2}$$

Thus, $\boldsymbol{E}|E|$ is the product of a parameter that is related to the average degree of a vertex (see section 2) and a parameter related to the number of vertices in the graph.

With lemma 1 and (5.2) we get

Lemma 4.
$\boldsymbol{E}(\mathcal{I}_{\mathbf{X}}) = \boldsymbol{E}\left(\delta^{(1,1)}_{f,N+1}\right) \boldsymbol{E}|E|$

The generalization of lemma 2 requires more effort.

Lemma 5.

$$\begin{aligned} \boldsymbol{E}(\mathcal{I}_{\mathbf{X}}\mathcal{I}_{\mathbf{1}}) = \boldsymbol{E}|E| \Big[\boldsymbol{E}|E|\, \boldsymbol{E}\left(\delta^{(1,1)}_{f,N+2}\right) + (1-3\alpha)\boldsymbol{E}\left(\delta^{(1,1)}_{f,N+1}\right) \\ + 4\alpha \boldsymbol{E}\left(\delta^{(1,2)}_{f,N+1}\right) + O(1/\beta) \Big] \end{aligned}$$

Proof. By lemma 2 and (5.2), we have

$$\begin{aligned} \boldsymbol{E}(\mathcal{I}_{\mathbf{X}}\mathcal{I}_{\mathbf{1}}) &= \boldsymbol{E}(\boldsymbol{E}(\mathcal{I}_{\mathbf{X}}\mathcal{I}_{\mathbf{1}} \mid N)) \\ &= \boldsymbol{E}|E| \sum_{n\geq 3} \left[\delta^{(1,1)}_{f,n} \binom{n-2}{2} p_n + 2\delta^{(1,2)}_{f,n}(n-2)p_n + \delta^{(1,1)}_{f,n} \right] \\ &\qquad P(N=n-1) \end{aligned}$$

We deal with the three summands separately using

$$\begin{aligned} \alpha(n-3)P(N=n-1) &= \boldsymbol{E}|E| P(N=n-2) \\ \binom{n-2}{2} p_n &= \alpha(n-3) - 3\alpha + O\left(\frac{1}{n-2}\right) \\ (n-2)p_n &= 2\alpha + O\left(\frac{1}{n-2}\right) \end{aligned}$$

and

$$\sum_{n\geq 3} \delta^{(i,j)}_{f,n} O\left(\frac{1}{n-2}\right) P(N=n-1) = O(1/\beta)$$

which holds since $\delta_{f,n}^{(i,j)}$ converges by theorem 3. Hence, we get

$$\begin{aligned}\boldsymbol{E}(\mathcal{I}_{\mathbf{X}}\mathcal{I}_{\mathbf{1}}) = \boldsymbol{E}|E|\Bigg[&\boldsymbol{E}|E|\sum_{n\geq 4}\delta_{f,n}^{(1,1)}P(N=n-2)\\ &-3\alpha\sum_{n\geq 3}\delta_{f,n}^{(1,1)}P(N=n-1)\;+O(1/\beta)\\ &+4\alpha\sum_{n\geq 3}\delta_{f,n}^{(1,2)}P(N=n-1)\;+O(1/\beta)\\ &+\sum_{n\geq 3}\delta_{f,n}^{(1,1)}P(N=n-1)\Bigg]\\ =\boldsymbol{E}|E|\Big[&\boldsymbol{E}|E|\,\boldsymbol{E}\Big(\delta_{f,N+2}^{(1,1)}\Big)+(1-3\alpha)\boldsymbol{E}\Big(\delta_{f,N+1}^{(1,1)}\Big)\\ &+4\alpha\boldsymbol{E}\Big(\delta_{f,N+1}^{(1,2)}\Big)\;+O(1/\beta)\Big]\end{aligned}$$

□

With lemma 4 follows

Corollary 6.
$Var(\mathcal{I}_{\mathbf{1}}) = \boldsymbol{E}|E|(1+\alpha+O(1/\beta))$

6 Covariance with $\mathcal{I}_{\mathbf{1}}$

Any topological index $\mathcal{I}_{\mathbf{X}}$ of the form (2.2) can be made uncorrelated to $\boldsymbol{E}|E| = \mathcal{I}_{\mathbf{1}}$ by adding a suitable constant: If we write

$$\mathcal{I}_{\tilde{\mathbf{X}}} = \sum_{\{u,v\}\in E(G)}(X_{uv}+c) \tag{6.1}$$

we have

$$0 = Cov\big(\mathcal{I}_{\tilde{\mathbf{X}}},\mathcal{I}_{\mathbf{1}}\big) = Cov(\mathcal{I}_{\mathbf{X}},\mathcal{I}_{\mathbf{1}}) + c\,Var(\mathcal{I}_{\mathbf{1}})$$

iff

$$c = -\frac{Cov(\mathcal{I}_{\mathbf{X}},\mathcal{I}_{\mathbf{1}})}{Var(\mathcal{I}_{\mathbf{1}})} \tag{6.2}$$

However, this does not make any sense if (6.1), (6.2) merely transform a linear into a non-linear dependence. Figure 2 illustrates how a non-linear dependence may appear. In this example however $Cov(\mathcal{I}_{\mathbf{X}},\mathcal{I}_{\mathbf{1}})$ is not zero if $\boldsymbol{E}(\mathcal{I}_{\mathbf{1}})$ is in a region where the curve rises or falls. We therefore require that $Cov\big(\mathcal{I}_{\tilde{\mathbf{X}}},\mathcal{I}_{\mathbf{1}}\big) = 0$ for different values of β, or equivalently, that c does not depend on β. It will turn out that this can be achieved approximately, which will precisely be stated in terms of a limit of c for $\beta\to\infty$.

With lemma 1, 2 can be shown that

$$\frac{Cov(\mathcal{I}_{\mathbf{X}},\mathcal{I}_{\mathbf{1}})}{Var(\mathcal{I}_{\mathbf{1}})} = \delta_{f,n}^{(1,1)}(1-2\alpha)+2\delta_{f,n}^{(1,2)}\alpha+O(1/n)$$

holds for every fixed n, hence (6.2) does not much depend on the number of vertices in this case. On the other hand, if the number of vertices is a Poisson-distributed random variable N, the results of section 5 show that $\frac{Cov(\mathcal{I}_{\mathbf{X}},\mathcal{I}_{\mathbf{1}})}{Var(\mathcal{I}_{\mathbf{1}})}$ contains the product

$$\boldsymbol{E}|E|\boldsymbol{E}\Big(\delta_{f,N+2}^{(1,1)}-\delta_{f,N+1}^{(1,1)}\Big) \tag{6.3}$$

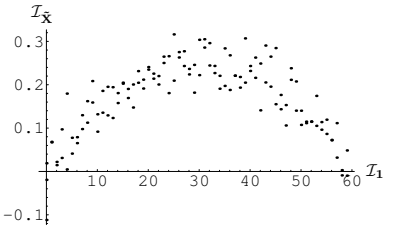

Figure 2: A non-linear dependence

If the convergence speed of $\delta_{f,n}^{(1,1)}$ is not known, nothing can be proved about the behavior of (6.3) for different values of β.

Therefore, we define

$$\epsilon_n = \delta_{f,\infty}^{(1,1)} - \delta_{f,n}^{(1,1)} \tag{6.4}$$

with $\delta_{f,\infty}^{(1,1)} = \lim_{n\to\infty} \delta_{f,n}^{(1,1)}$ and

$$\hat{X}_{uv} = X_{uv} + \epsilon_n \tag{6.5}$$

so that $\delta_{\hat{f},n}^{(1,1)} = \delta_{f,\infty}^{(1,1)}$ for all n and thus $\boldsymbol{E}\left(\delta_{\hat{f},N+k}^{(1,1)}\right) = \delta_{f,\infty}^{(1,1)}$ for all k. Then, product (6.3) is always zero and

$$\frac{Cov(\mathcal{I}_{\tilde{\mathbf{X}}}, \mathcal{I}_{\mathbf{1}})}{Var(\mathcal{I}_{\mathbf{1}})} = \frac{(1-3\alpha)\delta_{f,\infty}^{(1,1)} + 4\alpha \boldsymbol{E}\left(\delta_{\hat{f},N+1}^{(1,2)}\right) + O(1/\beta)}{1+\alpha+O(1/\beta)} \tag{6.6}$$

by lemma 4, 5 and corollary 6. This still depends on β, but only to a small extent. More precisely: $\lim_{\beta\to\infty} c$ exists and is finite. Note that $\beta \to \infty$ iff $\boldsymbol{E}(N) \to \infty$.

As a result, we get

Theorem 7.
For a random graph in $\mathscr{G}(N, p_N)$, let

$$\hat{X}_{uv} = X_{uv} + \epsilon_N$$
$$\tilde{X}_{uv} = \hat{X}_{uv} + c$$

with

$$c = -\frac{Cov(\mathcal{I}_{\tilde{\mathbf{X}}}, \mathcal{I}_{\mathbf{1}})}{Var(\mathcal{I}_{\mathbf{1}})}$$

and let ϵ_n be defined according to (6.4)*. Then we have:*

1. *$\mathcal{I}_{\tilde{\mathbf{X}}}$ is uncorrelated to $\mathcal{I}_{\mathbf{1}}$*
2. *$\lim_{\beta\to\infty} c$ exists and is finite*

Proof. The first claim follows from (6.1), (6.2). The only assertion left to show is that $\boldsymbol{E}\left(\delta_{\hat{f},N+1}^{(1,2)}\right)$ converges for $\beta \to \infty$. By theorem 3 and (6.3), (6.5), $\delta_{\hat{f},n}^{(1,2)}$ converges. Hence, there is for any

$\varepsilon > 0$ an n_ε such that $\left|\delta^{(1,2)}_{\hat{f},n+1} - \delta^{(1,2)}_{\hat{f},\infty}\right| < \varepsilon$ for all $n > n_\varepsilon$ and

$$\lim_{\beta\to\infty}\left|\boldsymbol{E}\left(\delta^{(1,2)}_{\hat{f},N+1}\right) - \delta^{(1,2)}_{\hat{f},\infty}\right| < \lim_{\varepsilon\to 0}\lim_{\beta\to\infty} P(N \le n_\varepsilon)\, O(1) + \varepsilon P(N > n_\varepsilon) = 0$$

Thus, the right hand side of (6.6) converges. □

7 Zero-Order Indices

As a second application, we now consider topological indices of the form

$$\mathcal{I}_{\mathbf{X}} = \sum_{v=1}^{N} X_v$$

with $X_v = f(\deg(v))$ for a function f with $|f(x)| \le b^x$ for a constant $b > 0$. Examples for indices of this form are the first Zagreb group index or the zero-order Kier & Hall index.

It follows similarly to the results of (6.1), (6.2) that $\mathcal{I}_{\tilde{\mathbf{X}}}$ with $\tilde{X}_v = X_v + c$ is uncorrelated to $\mathcal{I}_{\mathbf{1}} = N$ iff

$$c = -\frac{Cov(\mathcal{I}_{\mathbf{X}}, \mathcal{I}_{\mathbf{1}})}{Var(\mathcal{I}_{\mathbf{1}})} = -\boldsymbol{E}(X_1) \tag{7.1}$$

This holds for arbitrary random variables N. If we write $\delta_{f,n} = \boldsymbol{E}(X_1 \mid N = n)$, we have $c = -\boldsymbol{E}(\delta_{f,N})$ and it follows as in theorem 7 that $\lim_{\beta\to\infty} c$ exists and is finite.

Theorem 8.
For the zero-order index $\mathcal{I}_{\tilde{\mathbf{X}}}$ with $\tilde{X}_v = X_v + c$ and c defined according to (7.1) *holds:*

1. *$\mathcal{I}_{\tilde{\mathbf{X}}}$ is uncorrelated to $\mathcal{I}_{\mathbf{1}}$*
2. *$\lim_{\beta\to\infty} c$ exists and is finite*

8 Discussion

Theorem 7 shows that a topological index $\mathcal{I}_{\mathbf{X}}$ of the form (2.2) can be transformed to an index $\mathcal{I}_{\tilde{\mathbf{X}}}$ that is uncorrelated to $\mathcal{I}_{\mathbf{1}} = \boldsymbol{E}|E|$ within model $\mathscr{G}(n, p_n)$. A similar assertion holds for zero-order indices. This transformation should not introduce a non-linear dependence as illustrated in figure 2. This is formally stated by (6.6) and the fact that c does not depend on β for $\beta \to \infty$, or equivalently, for $\boldsymbol{E}(N) \to \infty$. Thus, the dependence among $\mathcal{I}_{\mathbf{X}}$ and $\boldsymbol{E}|E|$ is much reduced. However, the number of bonds may still influence the variance of $\mathcal{I}_{\tilde{\mathbf{X}}}$.

The values ϵ_n introduced in (6.4) are needed to prove assertion (2) of theorem 7. If, in a practical application, $\delta^{(1,1)}_{f,n}$ converges rapidly or if the size of the molecules in the data set does not vary much it may not be necessary to transform X_{uv} to $\hat{X}_{uv}$ since (6.3) will not change much either. In this case, the same is true for c.

To show the necessity of this transformation, it still remains to show when $\mathcal{I}_{\mathbf{X}}$ and $\boldsymbol{E}|E|$ are correlated within this setting, as it was shown for independent vertex properties with non-zero expectation [4, 5, 6]. Further, a computer experiment with chemical data is needed to see the difference this transformation makes to the correlations among topological indices of the kind we consider here.

References

[1] I. Motoc, A. Balaban, O. Mekenyan, and D. Bonchev. Topological indices: Inter-relations and composition. *MATCH Commun. Math. Comput. Chem.*, 13:369–404, 1982.

[2] S. Basak, V. Magnuson, G. Niemi, R. Regal, and G. Veith. Topological indices: Their nature, mutual relatedness, and applications. *Math. Model.*, 8:300–305, 1987.

[3] B. Bollobás. *Modern Graph Theory*. Springer, 1998.

[4] B. Hollas. Correlation properties of the autocorrelation descriptor for molecules. *MATCH Commun. Math. Comput. Chem.*, 45:27–33, 2002.

[5] B. Hollas. An analysis of the autocorrelation descriptor for molecules. *J. Math. Chem.*, 33(2):91–101, 2003.

[6] B. Hollas. Correlations in distance-based descriptors. *MATCH Commun. Math. Comput. Chem.*, 47:79–86, 2003.

[7] R. Todeschini and V. Consonni. *Handbook of Molecular Descriptors*. Wiley, 2000.