MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **54** (2005) 341-350

ISSN 0340 - 6253

On the Variance 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 February 2, 2005)

Abstract

We present results on the variance of 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 connectivity index or the 2nd Zagreb index are examples for indices of this kind. For random graphs on n vertices, we show that $Var(\mathcal{I})$ increases linearly in the number of vertices while $Var(1/\sqrt{n}\mathcal{I})$ remains bounded. Experiments with chemical structures and random graphs confirm our results. With a bounded variance, a better separation of size-dependent and size-independent properties is obtained. The results are important for the processing of descriptor data with neural nets.

1 Introduction

This paper complements a previous paper [1] in which we examined the covariance of topological indices that have the form

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

for a function $f : \mathbb{N}^2 \to \mathbb{R}$. In this paper we consider the variance of $\mathcal{I}_{\mathbf{X}}$ in a random graph model with a fixed number of vertices n. We show that the variance of $\mathcal{I}_{\mathbf{X}}$ increases at least linearly in the number of vertices n while it is bounded for $\frac{1}{\sqrt{n}}\mathcal{I}_{\mathbf{X}}$. Experiments with structures from the NCI 127k database [2] and random graphs in section 4 confirm our results.

A uniform variance of descriptors throughout a data set is important if the data are clustered by neural nets such as self-organizing maps or learning vector quantization [3, 4, 5]. Neural nets are trained by adjusting weight vectors according to the input data presented, trying to minimize an error function that measures the distance between weight vectors and data vectors. If descriptor data with a variance increasing in the number of atoms n is clustered this way, large molecules have a higher impact on the formation of the weight vectors during training than small molecules. More control on the clustering can be exercised if the variance of the descriptors, as a function of n, remains constant or in $\Theta(1)$ (Landau symbols are explained below) and if the number of atoms is encoded as a separate descriptor. Thus, a better separation of size-dependent and size-independent properties is obtained. As the results in [6] indicate, the factor $1/\sqrt{n}$ accounting for a uniform variance might further reduce the mutual dependence for already uncorrelated descriptors.

2 Preliminaries

We use the same random graph model as described in [1], sections 2 and 5, for a fixed number of vertices n: The space of random graphs $\mathscr{G}(n, p_n)$ consists of graphs on a fixed set of vertices $\{1, \ldots, n\}$ whose edges are chosen independently with probability

$$p_n = \alpha \frac{n-2}{\binom{n}{2}} \tag{2.1}$$

so that the expected number of edges is $\boldsymbol{E}|E| = \alpha(n-2)$. The parameter $\alpha > 0$ describes the branching of the graphs. The special choice of p_n results in properties that proved useful in [1].

We consider the topological index

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

where $G = (\{1, \ldots, n\}, E)$ is a random graph in $\mathscr{G}(n, p_n)$,

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

is the indicator function for event $\{\{u,v\} \in E\}$ and

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

are the *vertex pair properties*.

It is convenient to use the following *Landau symbols* [7], which are defined as:

$$O(f) = \{h : \mathbb{N} \to \mathbb{N} \mid |h(n)| \leq cf(n) \text{ for all but finitely many } n \text{ and a constant } c > 0\}$$

$$\Omega(f) = \{h : \mathbb{N} \to \mathbb{N} \mid |h(n)| \geq cf(n) \text{ for all but finitely many } n \text{ and a constant } c > 0\}$$

$$\Theta(f) = O(f) \cap \Omega(f)$$

As usual, we also use these symbols in equations to denote an element of the respective set. Thus, $\Theta(1)$ denotes a *set* of functions h as well as a *single* function h with $c_1 < |h(n)| < c_2$ for constants $c_2, c_2 > 0$ and all but finitely many n. Observe that, by definition, all functions in $\Theta(1)$ are bounded, but not vice versa.

3 Theoretical Results

For any reasonable topological index $\mathcal{I}_\mathbf{X}$ the variance of X_{12} is positive for adjacent vertices $1, 2$. The technical requirement in the following theorem is therefore no restriction. Theorem 1 also holds if $\mathcal{I}_\mathbf{X}$ is transformed according to [1], theorem 7, which results in $\mathcal{I}_\mathbf{X}$ being uncorrelated to $|E|$.

Theorem 1.

If $Var(X_{12}|1_{12} = 1) > 0$ *then*

1. $Var(\mathcal{I}_\mathbf{X}) \in \Theta(n)$
2. $Var\left(\frac{1}{\sqrt{n}}\mathcal{I}_\mathbf{X}\right) \in \Theta(1)$

in random graph space $\mathcal{G}(n, p_n)$ *with* p_n *defined by* (2.1).

Proof. First, we have to determine

$$\boldsymbol{E}\left(\mathcal{I}_\mathbf{X}^2\right) = \sum_{u<v} \sum_{u'<v'} \boldsymbol{E}(X_{uv} X_{u'v'} 1_{uv} 1_{u'v'})$$

To do so, we use the sets

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

for which we proved in [1], lemma 2

$$|S_0| = \binom{n}{2}\binom{n-2}{2} \tag{3.1}$$

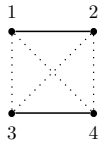
$$|S_1| = 6\binom{n}{3} \tag{3.2}$$

$$|S_2| = \binom{n}{2} \tag{3.3}$$

Thus, we have

$$\begin{aligned}\boldsymbol{E}\big(\mathcal{I}_{\mathbf{X}}^2\big) &= |S_0|\boldsymbol{E}(X_{12}X_{34}1_{12}1_{34})\\ &= |S_1|\boldsymbol{E}(X_{12}X_{13}1_{12}1_{13})\\ &= |S_2|\boldsymbol{E}\big(X_{12}^2 1_{12}\big)\end{aligned} \tag{3.4}$$

Observe that X_{12}, X_{34} are not independent since they have random variables $1_{13}, 1_{14}, 1_{23}, 1_{34}$ in common. However, X_{12}, X_{34} are independent if we condition for fixed values of $1_{13}, 1_{14}, 1_{23}, 1_{34}$ as follows.



Let $I = \{(1,3),(1,4),(2,3),(2,4)\}$ and $A = \{(i_{uv})_{(u,v)\in I} \mid i_{uv} \in \{0,1\}\}$. Then

$$\begin{aligned}\boldsymbol{E}(X_{12}X_{34}1_{12}1_{34}) &= \sum_{(i_{uv})\in A} \boldsymbol{E}\big(X_{12}X_{34}1_{12}1_{34}1_{\{(1_{uv})=(i_{uv})\}}\big)\\ &= \sum_{(i_{uv})\in A} \big[\boldsymbol{E}(X_{12} \mid 1_{12}=1 \wedge (1_{uv})=(i_{uv}))\,\boldsymbol{E}(X_{34} \mid 1_{34}=1 \wedge (1_{uv})=(i_{uv}))\\ &\qquad \cdot p_n^2 P((1_{uv})=(i_{uv}))\big]\end{aligned} \tag{3.5}$$

Let $q_n = 1 - p_n$. We consider the following cases:

1. For $i_{uv} = 0$ for all $(u,v) \in I$ we have

$$\begin{aligned}a_1 &:= \boldsymbol{E}(X_{12} \mid 1_{12}=1 \wedge (1_{uv})=(i_{uv}))\\ &= \boldsymbol{E}\left(X_{12}1_{12}\prod_{(u,v)\in I}(1-1_{uv})\right)\frac{1}{p_n q_n^4}\\ &= \big(\boldsymbol{E}(X_{12}|1_{12}=1)\,p_n + 4\boldsymbol{E}(X_{12}|1_{12}1_{13}=1)\,p_n^2 + O(p_n^3)\big)\frac{1}{p_n q_n^4}\\ &= \Big(\delta_{f,n}^{(1,1)} + 4p_n\delta_{f,n}^{(1,2)}\Big)\frac{1}{q_n^4} + O(p_n^2)\end{aligned}$$

where $\delta_{f,n}^{(1,1)} = \boldsymbol{E}(X_{12}|1_{12}=1)$ and $\delta_{f,n}^{(1,2)} = \boldsymbol{E}(X_{12}|1_{12}1_{13}=1)$ as in [1], (3.1).

2. For $i_{13} = 1$ and $i_{uv} = 0$ for $(u, v) \neq (1, 3)$ we have

$$\begin{aligned} a_2 &:= \boldsymbol{E}(X_{12} \mid 1_{12} = 1 \wedge (1_{uv}) = (i_{uv})) \\ &= \boldsymbol{E}\left(X_{12} 1_{12} 1_{13} \prod_{(u,v)\neq(1,3)} (1 - 1_{uv}) \right) \frac{1}{p_n^2 q_n^3} \\ &= \delta_{f,n}^{(1,2)} \frac{1}{q_n^3} + O(p_n) \end{aligned}$$

The cases $i_{14} = 1$, $i_{uv} = 0$ for $(u, v) \neq (1, 4)$ etc. yield the same result for symmetry reasons.

3. For all other cases $\boldsymbol{E}(X_{12} \mid 1_{12} = 1 \wedge (1_{uv}) = (i_{uv})) = O(1)$.

Thus, by (3.5)

$$\boldsymbol{E}(X_{12} X_{34} 1_{12} 1_{34}) = a_1^2 p_n^2 q_n^4 + 4 a_2^2 p_n^3 q_n^3 + O(p_n^4)$$

The remaining terms in (3.4) are easily handled. By the Cauchy-Schwartz inequality,

$$|\boldsymbol{E}(X_{12} X_{13} 1_{12} 1_{13})| = |\boldsymbol{E}(X_{12} X_{13} \mid 1_{12} 1_{13} = 1)| p_n^2 \leq \delta_{f^2,n}^{(1,2)} p_n^2 = O(p_n^2)$$

Finally,

$$\boldsymbol{E}(X_{12}^2 1_{12}) = \delta_{f^2,n}^{(1,1)} p_n$$

In [1], lemma 1 we proved $\boldsymbol{E}(\mathcal{I}_{\mathbf{X}}) = \delta_{f,n}^{(1,1)} \boldsymbol{E}|E|$. Together with (3.1)-(3.3) and (3.4), we get

$$\begin{aligned} Var(\mathcal{I}_{\mathbf{X}}) = \boldsymbol{E}|E| \Bigg[&\binom{n-2}{2} \left(a_1^2 p_n q_n^4 + 4 a_2^2 p_n^2 q_n^3 \right) + O(p_n) + \delta_{f^2,n}^{(1,1)} \\ &- \binom{n}{2} p_n \left(\delta_{f,n}^{(1,1)} \right)^2 + O(p_n) \Bigg] \end{aligned}$$

A Taylor series expansion for $1/n$ gives

$$\begin{aligned} Var(\mathcal{I}_{\mathbf{X}}) = \boldsymbol{E}|E| \Bigg[&4\alpha \left(\delta_{f,n}^{(1,1)} \right)^2 (2\alpha - 1) \\ &+ 8\alpha^2 \delta_{f,n}^{(1,2)} \left(2\delta_{f,n}^{(1,1)} + \delta_{f,n}^{(1,2)} \right) + \delta_{f^2,n}^{(1,1)} + O(1/n) \Bigg] \\ = \boldsymbol{E}|E| \cdot 8\alpha^2 \Bigg[&\left(\delta_{f,n}^{(1,1)} + \delta_{f,n}^{(1,2)} \right)^2 + Var(X_{12} \mid 1_{12} = 1) + O(1/n) \Bigg] \end{aligned}$$

since $\delta_{f^2,n}^{(1,1)} = \left(\delta_{f,n}^{(1,1)} \right)^2 + Var(X_{12} \mid 1_{12} = 1)$. By [1], theorem 3, $\lim_{n\to\infty} \delta_{f,n}^{(i,j)}$ and $\lim_{n\to\infty} \delta_{f^2,n}^{(i,j)}$ exist for all i, j. Thus, the terms in the square brackets are in $\Theta(1)$ and the claim follows with $\boldsymbol{E}|E| = \Theta(n)$. □

Remark. A similar proof shows that this result also holds for zero-order indices $\sum_{v=1}^{n} g(\deg(v))$.

4 Experimental Results

To validate our results, we used both chemical graphs as well as random graphs to compute values for the Zagreb indices M_1, M_2, the modified Zagreb index M_2' [8], the Platt number F, the connectivity index χ, and the zero-order Kier & Hall-index ${}^0\chi$. These indices are defined as follows [8, 9, 10]:

$$M_1 = \sum_{v=1}^{n} \deg(v)^2$$

$$M_2 = \sum_{\{u,v\}\in E} \deg(u)\deg(v)$$

$$M_2' = \sum_{\{u,v\}\in E} \frac{1}{\deg(u)\deg(v)}$$

$$F = \sum_{\{u,v\}\in E} \deg(u) + \deg(v) - 2$$

$$\chi = \sum_{\{u,v\}\in E} \frac{1}{\sqrt{\deg(u)\deg(v)}}$$

$${}^0\chi = \sum_{v=1}^{n} \frac{1}{\sqrt{\deg(v)}}$$

Chemical Graphs

Structures were taken from the freely available NCI 127k database [2], which contains connection tables for about 127,000 molecules. Some of these structures were not connected and were therefore discarded, leaving 126,674 molecular graphs for the computer experiment. The mean vertex degree is 1.88, some vertices however have degrees greater than 4.

Figures 1-12 show the variance as a function of the number of atoms n in the range $5 \leq n \leq 40$. In accordance with theorem 1,1., the variance of $M_1, M_2, M_2', F, \chi, {}^0\chi$ (figures 1-6) increases in n with considerable differences between large and small molecules. For example, $Var(M_2) = 10.3$ for $n = 5$ but $Var(M_2) = 1395.6$ for $n = 40$ which is a 135-fold increase in variance.

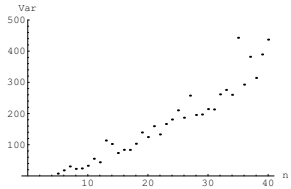


Figure 1: Variance of M_1

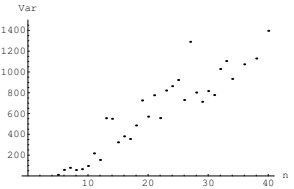


Figure 2: Variance of M_2

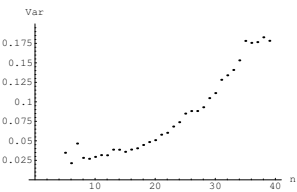


Figure 3: Variance of M_2'

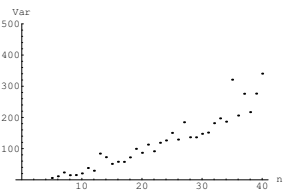


Figure 4: Variance of F

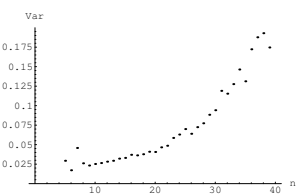


Figure 5: Variance of χ

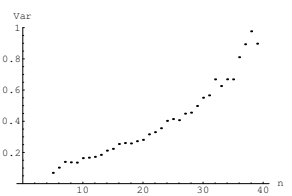


Figure 6: Variance of ${}^0\chi$

The corresponding results for $\frac{1}{\sqrt{N}}M_1, \frac{1}{\sqrt{N}}M_2, \frac{1}{\sqrt{N}}M_2', \frac{1}{\sqrt{N}}F, \frac{1}{\sqrt{N}}\chi, \frac{1}{\sqrt{N}}{}^0\chi$ are shown in figures 7-12. Some of these graphs still show an increase in variance with increasing n, but the overall increase is much reduced. The transformed zero-order Kier & Hall index $\frac{1}{\sqrt{N}}{}^0\chi$ shows only a slight increase towards the end while for the second Zagreb index we now have $Var\left(\frac{1}{\sqrt{N}}M_2\right) = 2.07$ for $n = 5$ and $Var\left(\frac{1}{\sqrt{N}}M_2\right) = 3.89$ for $n = 40$, hence a 16.9 fold increase. The larger part of this increase however occurs in the range $5 \leq n \leq 15$. This behavior is caused by the slow convergence of $Var\left(\frac{1}{\sqrt{N}}M_2\right)$ in n, as the following experiments with random graphs will show.

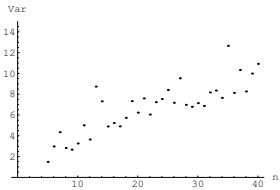


Figure 7: Variance of $\frac{1}{\sqrt{N}}M_1$

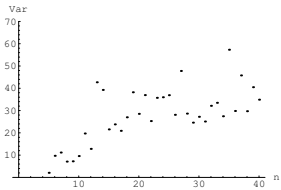


Figure 8: Variance of $\frac{1}{\sqrt{N}}M_2$

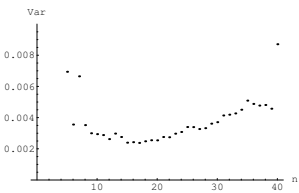


Figure 9: Variance of $\frac{1}{\sqrt{N}}M_2'$

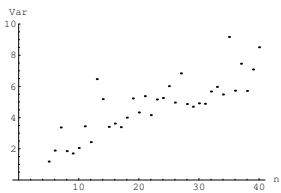


Figure 10: Variance of $\frac{1}{\sqrt{N}}F$

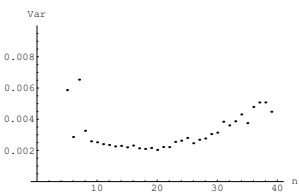


Figure 11: Variance of $\frac{1}{\sqrt{N}}\chi$

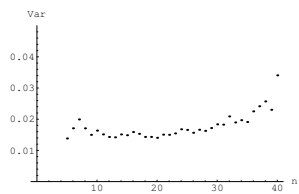


Figure 12: Variance of $\frac{1}{\sqrt{N}}{}^0\chi$

Random Graphs

With random graphs, we can generate a much larger number of graphs for each n than present in the NCI 127k database. For $10 \leq n \leq 100$ vertices and $\alpha = 1$ we equiprobably generated 100,000 labeled graphs each, using a high quality random number generator. The variance of the untransformed indices shows a steep linear increase in n and is not shown. The transformed indices $\frac{1}{\sqrt{N}}M_1$, $\frac{1}{\sqrt{N}}M_2$, $\frac{1}{\sqrt{N}}M_2'$, $\frac{1}{\sqrt{N}}F$, $\frac{1}{\sqrt{N}}\chi$, $\frac{1}{\sqrt{N}}{}^0\chi$ in figures 13-18 have variances that seemingly converge in n, thus supporting theorem 1.

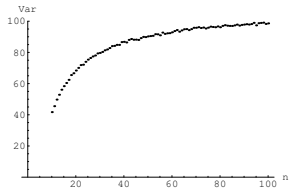


Figure 13: Variance of $\frac{1}{\sqrt{N}}M_1$

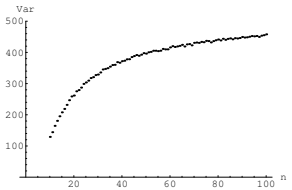


Figure 14: Variance of $\frac{1}{\sqrt{N}}M_2$

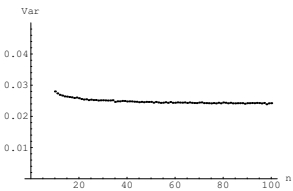


Figure 15: Variance of $\frac{1}{\sqrt{N}}M_2'$

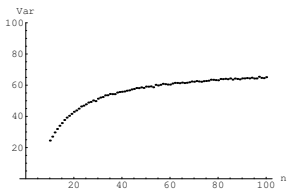


Figure 16: Variance of $\frac{1}{\sqrt{N}}F$

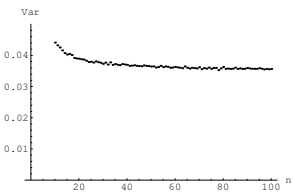


Figure 17: Variance of $\frac{1}{\sqrt{N}}\chi$

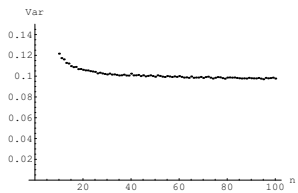


Figure 18: Variance of $\frac{1}{\sqrt{N}}{}^0\chi$

5 Conclusion

It is common practice to scale descriptor values to unity variance prior to further processing. However, this procedure does not even out the significantly different variance of descriptor values for differently sized molecules. This is most evident in the 135-fold increase in the variance in the data calculated for the 2nd Zagreb index M_2 (figure 3). In contrast, the variance of the indices divided by $\sqrt{N}$ remains in an interval that does not contain zero (i.e. the variance is in $\Theta(1)$), which is supported by experimental data with chemical graphs and, more clearly, random graphs.

References

[1] B. Hollas. The covariance of topological indices that depend on the degree of a vertex. *MATCH Commun. Math. Comput. Chem.*, 54(1):177–187, 2005.

[2] National Cancer Institute. Connection tables for 127,000 structures. ftp://helix.nih.gov/ncidata/2D/nciopen.mol.Z.

[3] J. Gasteiger and J. Zupan. *Neural Networks for Chemists*. Wiley-VCH, 1999.

[4] T. Kohonen. The self-organizing map. *Proc. IEEE*, 78:1464–1480, 1990.

[5] L. Fausett. *Fundamentals of Neural Networks*. Prentice Hall, 1994.

[6] B. Hollas. An asymptotically independent topological index on random trees. *J. Math. Chem.*, submitted.

[7] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1994.

[8] D. Vukičević and N. Trinajstić. Modified Zagreb M_2 index - comparison with the Randić connectivity index for benzenoid systems. *Croatica Chemica Acta*, 76(2):183–187, 2003.

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

[10] S. Nikolić, G. Kovačević, A. Miličević, and N. Trinajstić. The Zagreb indices 30 years after. *Croatica Chemica Acta*, 76(2):113–124, 2003.