The average size of independent sets of graphs

Eric O. D. Andriantiana[1] (ID) · Valisoa Razanajatovo Misanantenaina[2] ·
Stephan Wagner[2]

Abstract
We study the average size of independent (vertex) sets of a graph. This invariant can be
regarded as the logarithmic derivative of the independence polynomial evaluated at 1.
We are specifically concerned with extremal questions. The maximum and minimum
for general graphs are attained by the empty and complete graph respectively, while
for trees we prove that the path minimises the average size of independent sets and the
star maximises it. Although removing a vertex does not always decrease the average
size of independent sets, we prove that there always exists a vertex for which this is
the case.

Keywords Independent sets · Average size · Trees · Extremal problems

Mathematics Subject Classification 05C35 · 05C05 · 05C07

1 Introduction

The number of independent sets is a graph invariant that has been studied extensively.
It has been dubbed the *Fibonacci number* of a graph [10] due to the fact that the
number of independent sets in a path is always a Fibonacci number, and is now known

✉ Eric O. D. Andriantiana
    E.Andriantiana@ru.ac.za

    Valisoa Razanajatovo Misanantenaina
    valisoa@sun.ac.za

    Stephan Wagner
    swagner@sun.ac.za

1   Department of Mathematics (Pure and Applied), Rhodes University, PO Box 94,
    Grahamstown 6140, South Africa

2   Department of Mathematical Sciences, Stellenbosch University, Private Bag X1, Matieland 7602,
    South Africa

as *Merrifield–Simmons index* in mathematical chemistry in honour of the work of chemists Merrifield and Simmons [8]. Moreover, its connection to the *hard core model* in statistical physics is well established (see [1] for a general reference).

Extremal problems (concerned with finding maximum or minimum values) regarding the number of independent sets have been of particular interest, especially in the aforementioned context of mathematical chemistry. Graphs with various restrictions have been studied, as well as graph classes such as trees, unicyclic or bicyclic graphs; see [14] for a recent survey.

In this paper, we study similar questions for the average size of independent sets. This invariant, while interesting in its own right, comes up in various contexts: in [3], an asymptotic lower bound is given for triangle-free graphs, which can be used to obtain bounds on Ramsey numbers. In [2], the same authors established an upper bound as a tool to prove that the disjoint union of complete bipartite graphs $K_{d,d}$ maximises the number of independent sets of a d-regular graph.

An invariant of a similar nature is the average order of a subtree, as introduced to the literature by Jamison [5,6]. In particular, he proved that the average order of subtrees of an n-vertex tree is at least $(n + 2)/3$, with equality for the path, which parallels our Theorem 4.4. There has been a fair amount of recent activity around this invariant [4,9,13,15] and its generalisations [12].

This paper is structured as follows: in the next section, we collect some basic results that will be needed for our analysis. In Sect. 3, we consider the behaviour of the average size of the independent sets under removal of vertices or edges. It turns out that the average size of independent sets is not monotonic under these operations, as we will show by explicit counterexamples. This is in contrast to the number of independent sets, but also to the aforementioned average subtree order [6]. However, we prove that it is always possible to find a vertex whose removal decreases the average size of independent sets. We also show that—not unexpectedly—the empty and complete graph attain the maximum and minimum respectively among graphs of a given order. Finally, we focus on trees in Sect. 4, where it is shown that the path and the star are extremal. While the proof for the star is fairly short and straightforward, the situation for the path is much more complex. The paper concludes with a brief discussion of a generalised invariant.

2 Preliminaries

Let G be a graph, and let $i(G, k)$ be the number of independent sets of size k. The independence polynomial of G is defined by

$$I(G, x) = \sum_k i(G, k) x^k,$$

see [7] for a survey on the independence polynomial and its properties. The total number of independent subsets of G is simply the value of the independence polynomial at 1:

$$I(G, 1) = \sum_k i(G, k),$$

and the first derivative, evaluated at $x = 1$, is

$$I'(G, 1) = \sum_k k i(G, k).$$

Hence the average size of independent vertex subsets in G is the logarithmic derivative

$$\text{avi}(G) = \frac{I'(G, 1)}{I(G, 1)}.$$

For ease of notation, we will write $I(G)$ instead of $I(G, 1)$, as well as $T(G)$ instead of $I'(G, 1)$ (total size of all independent sets).

Example 2.1 Let us compute the average size of independent sets of the n-vertex edgeless graph E_n and the n-vertex star S_n. We have

$$I(E_n, x) = (1 + x)^n, \qquad I(S_n, x) = (1 + x)^{n-1} + x,$$

which give

$$I(E_n) = 2^n \qquad \text{and} \qquad I(S_n) = 2^{n-1} + 1,$$
$$T(E_n) = n2^{n-1} \qquad \text{and} \qquad T(S_n) = (n - 1)2^{n-2} + 1,$$

and hence

$$\text{avi}(E_n) = \frac{n}{2}, \qquad \text{avi}(S_n) = \frac{n-1}{2} + \frac{3-n}{2^n + 2}. \tag{1}$$

The following standard recursion, which is obtained by distinguishing independent sets containing a vertex v and those not containing it, is very useful in calculating the independence polynomial of graphs.

Proposition 2.2 *Let v be a vertex of G and $N[v] = \{u : uv \in E(G)\} \cup \{v\}$ its closed neighbourhood. We have*

$$I(G, x) = I(G - v, x) + x I(G - N[v], x).$$

As an immediate consequence, we obtain recursions for the aforementioned invariants $I(G)$ and $T(G)$.

Proposition 2.3 *Let v be a vertex of G and $N[v]$ its closed neighbourhood. We have*

$$I(G) = I(G - v) + I(G - N[v]), \tag{2}$$
$$T(G) = T(G - v) + T(G - N[v]) + I(G - N[v]).$$

Proof The first equation (2) is obtained from Proposition 2.2 by putting $x = 1$, the second by differentiating first and plugging in $x = 1$ afterwards.   □

Thus, we get the following identities for the average size of independent sets:

Proposition 2.4 *Let v be a vertex of G and $N[v]$ its closed neighbourhood. We have*

$$\mathrm{avi}(G) = \frac{\mathrm{T}(G-v) + \mathrm{T}(G-N[v]) + \mathrm{I}(G-N[v])}{\mathrm{I}(G-v) + \mathrm{I}(G-N[v])}$$
$$= \frac{\mathrm{avi}(G-v)\,\mathrm{I}(G-v) + (\mathrm{avi}(G-N[v])+1)\,\mathrm{I}(G-N[v])}{\mathrm{I}(G-v) + \mathrm{I}(G-N[v])}.$$

We conclude this section with the following lemma, which will be useful later.

Lemma 2.5 *If $G_1, G_2, \ldots, G_k$ are the disjoint components of a graph G, then*

$$\mathrm{avi}(G) = \sum_k \mathrm{avi}(G_k).$$

Proof It is well known that

$$\mathrm{I}(G, x) = \prod_k \mathrm{I}(G_k, x).$$

Now take the logarithm on both sides, differentiate and plug in $x = 1$ to obtain the desired result.   □

3 Vertex or edge removal

Many graph invariants satisfy a monotonicity property with respect to vertex or edge removal. This means that removing a vertex (or alternately, removing an edge) either always decreases or always increases the value of the invariant. The total number of independent sets is a simple example for such monotonicity properties: clearly, we have

$$\mathrm{I}(G-v) < \mathrm{I}(G) \quad \text{and} \quad \mathrm{I}(G-e) > \mathrm{I}(G)$$

for every vertex v and every edge e of a graph G. As we will see in this section, the average size of independent sets is not a monotonic invariant. However, we will show that there always exists a vertex in the graph whose removal decreases avi. Then, we characterize extremal graphs among all n-vertex graphs.

Let us first show that the average size of independent sets is not monotonic with respect to vertex removal. If v and u are respectively a leaf and the centre of the 4-vertex star S_4, then we have (see (1))

$$\mathrm{avi}(S_4 - v) = \mathrm{avi}(S_3) = 1 < \frac{13}{9} = \mathrm{avi}(S_4),$$

Fig. 1 A tree T
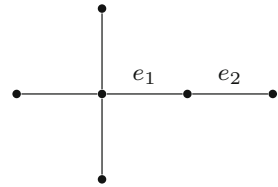
but

$$\text{avi}(S_4 - u) = \text{avi}(E_3) = \frac{3}{2} > \frac{13}{9} = \text{avi}(S_4).$$

The average size of independent sets is also not monotonic with respect to edge removal: considering the tree in Fig. 1, we have $I(T, x) = 2x^4 + 7x^3 + 10x^2 + 6x + 1$ and

$$\text{avi}(T - e_1) = \text{avi}(S_4) + \text{avi}(S_2) = \frac{13}{9} + \frac{2}{3} = \frac{19}{9} < \frac{55}{26} = \text{avi}(T),$$

but

$$\text{avi}(T - e_2) = \text{avi}(S_5) + \text{avi}(S_1) = \frac{33}{17} + \frac{1}{2} = \frac{83}{34} > \frac{55}{26} = \text{avi}(T).$$

While the inequality $\text{avi}(G - v) < \text{avi}(G)$ does not always hold (as the example of the star shows), we can show that for every graph G, there exists a vertex v with this property. To this end, Theorem 3.2 will be needed.

Definition 3.1 Let X be a nonempty finite set, and $\mathcal{P}(X)$ its powerset. For a nonempty subset $\mathcal{A}$ of $\mathcal{P}(X)$, we define

$$\text{av}(\mathcal{A}) = \frac{1}{|\mathcal{A}|} \sum_{A \in \mathcal{A}} |A|.$$

Theorem 3.2 *Let $\mathcal{B}$ be a subset of the powerset $\mathcal{P}(X)$ of a nonempty set X, such that the cardinalities of the elements of $\mathcal{B}$ are not all the same. Then there exists $x_0 \in X$ such that $\mathcal{B} \cap \mathcal{P}(X - \{x_0\})$ is not empty and*

$$\text{av}(\mathcal{B}) > \text{av}(\mathcal{B} \cap \mathcal{P}(X - \{x_0\})).$$

Proof It is convenient to use the abbreviations

$$n_k(\mathcal{A}) = |\{A \in \mathcal{A} : |A| = k\}| \quad \text{and} \quad S(\mathcal{A}) = \sum_{A \in \mathcal{A}} |A| = \sum_{k \geqslant 0} k \cdot n_k(\mathcal{A}).$$

We will prove that

$$\text{av}(\mathcal{B}) > \frac{\sum_{x \in X} S(\mathcal{B} \cap \mathcal{P}(X - \{x\}))}{\sum_{x \in X} |\mathcal{B} \cap \mathcal{P}(X - \{x\})|}. \tag{3}$$

Note here that the denominator is not 0: if $\mathcal{B} \cap \mathcal{P}(X - \{x\})$ were empty for all x, then $\mathcal{B}$ could only contain the set X and nothing else, contradicting our assumption that the cardinalities of the elements of $\mathcal{B}$ are not all the same. The claim of the theorem follows immediately, since there must be an $x_0 \in X$ such that

$$\frac{\sum_{x \in X} S(\mathcal{B} \cap \mathcal{P}(X - \{x\}))}{\sum_{x \in X} |\mathcal{B} \cap \mathcal{P}(X - \{x\})|} \geqslant \frac{S(\mathcal{B} \cap \mathcal{P}(X - \{x_0\}))}{|\mathcal{B} \cap \mathcal{P}(X - \{x_0\})|}.$$

Now let us prove (3). In the sum

$$\sum_{x \in X} S(\mathcal{B} \cap \mathcal{P}(X - \{x\})) = \sum_{x \in X} \sum_{k \geqslant 0} k \cdot n_k(\mathcal{B} \cap \mathcal{P}(X - \{x\})),$$

the size of each $B \in \mathcal{B}$ contributes $|X| - |B|$ times. Hence

$$\sum_{x \in X} S(\mathcal{B} \cap \mathcal{P}(X - \{x\})) = \sum_{k \geqslant 0} (|X| - k) k \cdot n_k(\mathcal{B}) = |X| S(\mathcal{B}) - \sum_{k \geqslant 0} k^2 n_k(\mathcal{B})$$

$$= |X| \operatorname{av}(\mathcal{B}) |\mathcal{B}| - \sum_{k \geqslant 0} k^2 n_k(\mathcal{B}).$$

Similarly,

$$\sum_{x \in X} |\mathcal{B} \cap \mathcal{P}(X - \{x\})| = \sum_{x \in X} \sum_{k \geqslant 0} n_k(\mathcal{B} \cap \mathcal{P}(X - \{x\}))$$

$$= \sum_{x \in X} \sum_{k \geqslant 0} (|X| - k) n_k(\mathcal{B}) = |X| |\mathcal{B}| - S(\mathcal{B}).$$

By the Cauchy–Schwarz inequality, we have

$$S(\mathcal{B})^2 = \left(\sum_{k \geqslant 0} k \cdot n_k(\mathcal{B}) \right)^2 \leqslant \sum_{k \geqslant 0} n_k(\mathcal{B}) \cdot \sum_{k \geqslant 0} k^2 n_k(\mathcal{B}) = |\mathcal{B}| \sum_{k \geqslant 0} k^2 n_k(\mathcal{B}),$$

and equality holds if and only if there is only one k such that $n_k(\mathcal{B}) \neq 0$, which means all the elements of $\mathcal{B}$ have the same size. Since this is ruled out by our assumptions, we actually have

$$\operatorname{av}(\mathcal{B}) S(\mathcal{B}) < \sum_{k \geqslant 0} k^2 n_k(\mathcal{B}).$$

Therefore we get

$$\frac{\sum_{x \in X} S(\mathcal{B} \cap \mathcal{P}(X - \{x\}))}{\sum_{x \in X} |\mathcal{B} \cap \mathcal{P}(X - \{x\})|} = \frac{|X| \mathrm{av}(\mathcal{B}) |\mathcal{B}| - \sum_{k \geqslant 0} k^2 n_k(\mathcal{B})}{|X||\mathcal{B}| - S(\mathcal{B})}$$

$$< \frac{|X| \mathrm{av}(\mathcal{B}) |\mathcal{B}| - \mathrm{av}(\mathcal{B}) S(\mathcal{B})}{|X||\mathcal{B}| - S(\mathcal{B})} = \mathrm{av}(\mathcal{B}),$$

which concludes the proof of (3) and thus the theorem. □

Corollary 3.3 *If G is a nonempty graph, then there exists a vertex v in G such that*

$$\mathrm{avi}(G - v) < \mathrm{avi}(G).$$

Proof Apply Theorem 3.2, with $\mathcal{B}$ being the set of independent vertex subsets of G. □

We have seen that there is always a vertex in a graph whose removal decreases the average size of independent sets avi. However, the dual statement for edge removal does not hold, namely there is not always an edge whose removal increases avi (nor is there always an edge whose removal decreases avi). As counterexamples, we can consider the stars S_6 and S_4: for any edge e in S_6 (S_4, respectively) we have

$$\mathrm{avi}(S_6) = \frac{27}{11} > \frac{83}{34} = \mathrm{avi}(S_6 - e),$$

$$\mathrm{avi}(S_4) = \frac{13}{9} < \frac{3}{2} = \mathrm{avi}(S_4 - e).$$

So every edge removal in S_6 decreases avi, while every edge removal in S_4 increases avi. Despite this, the edgeless graphs and the complete graphs are extremal:

Theorem 3.4 *For every n-vertex graph G that is not the edgeless graph E_n or the complete graph K_n, $n/(n+1) = \mathrm{avi}(K_n) < \mathrm{avi}(G) < \mathrm{avi}(E_n) = n/2$.*

Proof The first inequality is straightforward from the fact that the only independent sets of K_n are n independent sets of size 1 and the empty set: all other graphs with n vertices have these independent sets and some larger ones. We prove the second inequality by induction. For $n = 1$, there is no possible graph different from E_n, so there is nothing to prove. Now, assume that the inequality is true for all $n \leqslant k$, for some $k \geqslant 1$. Let G be a $(k+1)$-vertex graph that is not edgeless. Let $v \in V(G)$ be a vertex such that $\deg(v) \geqslant 1$. We have

$$\mathrm{avi}(G) = \frac{\mathrm{T}(G - v) + \mathrm{T}(G - N[v]) + \mathrm{I}(G - N[v])}{\mathrm{I}(G - v) + \mathrm{I}(G - N[v])}.$$

Using the induction hypothesis, we obtain

$$
\begin{aligned}
\operatorname{avi}(G) &\leqslant \frac{(k/2)\,\mathrm{I}(G-v) + ((k-1)/2+1)(\mathrm{I}(G-N[v]))}{\mathrm{I}(G-v) + \mathrm{I}(G-N[v])} \\
&= \frac{k}{2} + \frac{\mathrm{I}(G-N[v])/2}{\mathrm{I}(G-v)+\mathrm{I}(G-N[v])} < \frac{k}{2} + \frac{1}{2} = \frac{k+1}{2} = \operatorname{avi}(E_{k+1}).
\end{aligned}
$$

4 Trees

In this section, we are concerned with extremal trees regarding the average size of independent sets. Let us first consider the problem of maximizing the average size of independent sets among all n-vertex trees.

Theorem 4.1 *For every n-vertex tree T, $\operatorname{avi}(S_n) \geqslant \operatorname{avi}(T)$.*

Proof In the cases where $n = 1, 2, 3$, we must have $T = S_n$, and thus the claim holds.

Assume the inequality holds for all $n \leqslant k$, for some $k \geqslant 3$. Now suppose that $T \neq S_n$ is a tree with $n = k+1$ vertices. Let $v \in V(T)$ be a leaf of T and u its neighbour. Then $T - v$ is still a tree and by Proposition 2.3

$$
\mathrm{I}(T-v) - \mathrm{I}(T-v-u) = \mathrm{I}(T-v-N[u]) > 1.
$$

Moreover, the star minimises the number of independent sets among all n-vertex trees (see [10]), i.e., $\mathrm{I}(T-v) \leqslant \mathrm{I}(S_{n-1}) = 2^{n-2} + 1$. Thus

$$
\frac{\mathrm{I}(T-N[v])}{\mathrm{I}(T-v)} = 1 - \frac{\mathrm{I}(T-v) - \mathrm{I}(T-N[v])}{\mathrm{I}(T-v)} < 1 - \frac{1}{\mathrm{I}(T-v)} \leqslant 1 - \frac{1}{2^{n-2}+1}.
\tag{4}
$$

Using the recursion of Proposition 2.4 and Theorem 3.4, we obtain

$$
\begin{aligned}
\operatorname{avi}(T) &= \frac{\operatorname{avi}(T-v)\,\mathrm{I}(T-v) + (\operatorname{avi}(T-N[v])+1)\,\mathrm{I}(T-N[v])}{\mathrm{I}(T-v)+\mathrm{I}(T-N[v])} \\
&\leqslant \frac{\operatorname{avi}(S_{n-1})\,\mathrm{I}(T-v)/\mathrm{I}(T-N[v]) + (n-2)/2 + 1}{\mathrm{I}(T-v)/\mathrm{I}(T-N[v]) + 1}.
\end{aligned}
$$

Since $\operatorname{avi}(S_{n-1}) < (n-2)/2 + 1$, $\frac{\operatorname{avi}(S_{n-1})x + (n-2)/2 + 1}{x+1}$ is decreasing as a function of x for $x \geqslant 0$. Combined with the inequality in (4), this shows that

$$
\operatorname{avi}(T) < \frac{\operatorname{avi}(S_{n-1})(2^{n-2}+1)/2^{n-2} + (n-2)/2 + 1}{(2^{n-2}+1)/2^{n-2} + 1} = \operatorname{avi}(S_n).
$$

It is not unexpected that the star maximises avi, since it has the greatest independence number among trees. It is also the tree with the greatest number of independent sets (a well-known fact first established in [10]). The path, on the other hand, has the smallest number of independent sets among all trees of a given size. One would therefore expect that the average size of independent sets also attains its minimum for the path, which

is indeed the case. Proving this fact requires more effort. Let us first find an explicit formula for the average size of independent sets of a path. We denote the golden ratio by $\phi = (\sqrt{5}+1)/2$ and obtain the following result.

Lemma 4.2 *The average size of independent sets of the n-vertex path P_n is*

$$\operatorname{avi}(P_n) = \frac{5-\sqrt{5}}{10} n + \frac{3-\sqrt{5}}{5} - \frac{n+2}{\sqrt{5}((-\phi^2)^{n+2}-1)}. \tag{5}$$

In particular,

(a) $\lim_{n\to\infty} \operatorname{avi}(P_n) - \frac{5-\sqrt{5}}{10} n = \frac{3-\sqrt{5}}{5}$,

(b) $\operatorname{avi}(P_n) \geqslant \frac{5-\sqrt{5}}{10} n + \frac{1}{\sqrt{5}} - \frac{1}{3}$, *with equality only for $n = 2$. For all positive integers $n \neq 2$, we even have $\operatorname{avi}(P_n) \geqslant \frac{5-\sqrt{5}}{10} n + \frac{2}{\sqrt{5}} - \frac{3}{4}$, with equality only for $n = 4$.*

Proof It is well known that the number of independent sets of P_n is the Fibonacci number $F_{n+2} = \frac{1}{\sqrt{5}}(\phi^{n+2} - (-\phi)^{-n-2})$ (see [10]). The total number of vertices $T(P_n)$ in all independent sets of P_n is determined by the recursion

$$T(P_n) = T(P_{n-1}) + T(P_{n-2}) + I(P_{n-2})$$

that follows from Proposition 2.3, and the initial values $T(P_1) = 1$ and $T(P_2) = 2$. The solution to this recursion is easily found to be

$$T(P_n) = \left(\frac{1+\sqrt{5}}{10} n + \frac{2\sqrt{5}}{25}\right)\phi^n + \left(\frac{1-\sqrt{5}}{10} n - \frac{2\sqrt{5}}{25}\right)(-\phi)^{-n}.$$

The formula (5) for the quotient $\operatorname{avi}(P_n) = T(P_n)/I(P_n)$ follows immediately, as does the limit in (a).

Now we show that the absolute value of the error term is decreasing: for $n \geqslant 2$, we have

$$\left| \frac{n+2}{\sqrt{5}\,((-\phi^2)^{n+2}-1)} \cdot \frac{\sqrt{5}((-\phi^2)^{n+1}-1)}{n+1} \right| \leqslant \left(1 + \frac{1}{n+1}\right) \cdot \frac{\phi^{2(n+1)}+1}{\phi^{2(n+2)}-1}$$

$$= \phi^{-2}\left(1 + \frac{1}{n+1}\right) \cdot \frac{\phi^{-2(n+1)}+1}{1-\phi^{-2(n+2)}} \leqslant \phi^{-2} \cdot \frac{4}{3} \cdot \frac{\phi^{-6}+1}{1-\phi^{-8}} = \frac{4(\sqrt{5}-1)}{9} < 1.$$

Therefore, the difference

$$\left| \operatorname{avi}(P_n) - \frac{5-\sqrt{5}}{10} n - \frac{3-\sqrt{5}}{5} \right|$$

is decreasing in n. Moreover, note that the sign of $\frac{n+2}{\sqrt{5}((-\phi^2)^{n+2}-1)}$ alternates, so that $\operatorname{avi}(P_n)$ is alternatingly greater and less than $\frac{5-\sqrt{5}}{10} n + \frac{3-\sqrt{5}}{5}$. It follows that the

Table 1 Values of $c_1, c_2, \ldots, c_5$ for independent sets

n	1	2	3	4	5
c_n	$\frac{1}{2\sqrt{5}} \approx 0.2236$	$\frac{1}{\sqrt{5}} - \frac{1}{3} \approx 0.1139$	$\frac{3}{2\sqrt{5}} - \frac{1}{2} \approx 0.1708$	$\frac{2}{\sqrt{5}} - \frac{3}{4} \approx 0.1444$	$\frac{\sqrt{5}}{2} - \frac{25}{26} \approx 0.1565$

minimum of the difference $\mathrm{avi}(P_n) - \frac{5-\sqrt{5}}{10}n$ is attained for $n = 2$. Among all $n \neq 2$, the minimum occurs when $n = 4$. The values of $\mathrm{avi}(P_n)$ are easily calculated in both cases, and the two inequalities in (b) follow. $\qquad\square$

For ease of notation, we set

$$a = \frac{5 - \sqrt{5}}{10} \approx 0.27639320 \qquad (6)$$

and $c_n = \mathrm{avi}(P_n) - an$. Table 1 gives values of c_n for small n.

Before we prove our main result, we require one more lemma:

Lemma 4.3 *For every tree T and every vertex v of T, we have*

$$\frac{1}{2} \leqslant \frac{\mathrm{I}(T - v)}{\mathrm{I}(T)} < 1.$$

Proof Note first that $\mathrm{I}(T) = \mathrm{I}(T - v) + \mathrm{I}(T - N[v])$. Since $T - N[v]$ is a subgraph of $T - v$, we have $\mathrm{I}(T - N[v]) \leqslant \mathrm{I}(T - v)$, hence $2\,\mathrm{I}(T - v) \geqslant \mathrm{I}(T)$, which proves the first inequality. The second inequality simply follows from the fact that $T - v$ is an induced proper subgraph of T. $\qquad\square$

Theorem 4.4 *For every tree T of order n that is not a path, we have the inequality* $\mathrm{avi}(T) \geqslant an + b$, *where $b = (79\sqrt{5} - 165)/70 \approx 0.16641957$ and a is as in* (6). *Consequently, the path minimises the value of* $\mathrm{avi}(T)$ *among all trees of order n.*

Proof We prove the inequality by induction on n. For $n \leqslant 3$, there is nothing to prove since the only trees with three or fewer vertices are paths. Thus assume now that $n \geqslant 4$, and consider a vertex v of the tree T whose degree is at least 3 (which must exist if T is not a path). Denote the neighbours of v by $v_1, v_2, \ldots, v_k$ and the components of $T - v$ by $T_1, T_2, \ldots, T_k$ (in such a way that v_j is contained in T_j). By Proposition 2.4, we have

$$\begin{aligned}
\mathrm{avi}(T) &= \frac{\mathrm{T}(T)}{\mathrm{I}(T)} = \frac{\mathrm{T}(T - v) + (\mathrm{I}(T - N[v]) + \mathrm{T}(T - N[v]))}{\mathrm{I}(T)} \\
&= \frac{\mathrm{I}(T - v)}{\mathrm{I}(T)} \cdot \frac{\mathrm{T}(T - v)}{\mathrm{I}(T - v)} + \frac{\mathrm{I}(T - N[v])}{\mathrm{I}(T)} \cdot \left(1 + \frac{\mathrm{T}(T - N[v])}{\mathrm{I}(T - N[v])}\right) \\
&= \frac{\mathrm{I}(T - v)}{\mathrm{I}(T)}\,\mathrm{avi}(T - v) + \frac{\mathrm{I}(T) - \mathrm{I}(T - v)}{\mathrm{I}(T)}\,(1 + \mathrm{avi}(T - N[v])) \\
&= \frac{\mathrm{I}(T - v)}{\mathrm{I}(T)} \sum_{j=1}^{k} \mathrm{avi}(T_j) + \left(1 - \frac{\mathrm{I}(T - v)}{\mathrm{I}(T)}\right)\left(1 + \sum_{j=1}^{k} \mathrm{avi}(T_j - v_j)\right).
\end{aligned}$$

Assume first that $k \geqslant 5$, and let $T' = T - T_k$ be the tree obtained by removing T_k from T. Repeating the calculation, we also have

$$\text{avi}(T') = \frac{\text{I}(T' - v)}{\text{I}(T')} \sum_{j=1}^{k-1} \text{avi}(T_j) + \left(1 - \frac{\text{I}(T' - v)}{\text{I}(T')}\right)\left(1 + \sum_{j=1}^{k-1} \text{avi}(T_j - v_j)\right).$$

For simplicity, let us introduce the notations $\rho = \frac{\text{I}(T-v)}{\text{I}(T)}$ and $\rho' = \frac{\text{I}(T'-v)}{\text{I}(T')}$. Note that

$$\rho = \frac{\text{I}(T - v)}{\text{I}(T)} = \frac{\prod_{j=1}^{k} \text{I}(T_j)}{\prod_{j=1}^{k} \text{I}(T_j) + \prod_{j=1}^{k} \text{I}(T_j - v_j)} = \left(1 + \prod_{j=1}^{k} \frac{\text{I}(T_j - v_j)}{\text{I}(T_j)}\right)^{-1} \quad (7)$$

and likewise

$$\rho' = \left(1 + \prod_{j=1}^{k-1} \frac{\text{I}(T_j - v_j)}{\text{I}(T_j)}\right)^{-1},$$

so that Lemma 4.3 implies $\rho > \rho'$. Now we write

$$\text{avi}(T) = \rho \sum_{j=1}^{k} \text{avi}(T_j) + (1 - \rho)\left(1 + \sum_{j=1}^{k} \text{avi}(T_j - v_j)\right)$$

$$= \rho\,\text{avi}(T_k) + (1 - \rho)\,\text{avi}(T_k - v_k)$$

$$+ \rho \sum_{j=1}^{k-1} \text{avi}(T_j) + (1 - \rho)\left(1 + \sum_{j=1}^{k-1} \text{avi}(T_j - v_j)\right)$$

$$= \rho\,\text{avi}(T_k) + (1 - \rho)\,\text{avi}(T_k - v_k)$$

$$+ \frac{1 - \rho}{1 - \rho'}\left(\rho' \sum_{j=1}^{k-1} \text{avi}(T_j) + (1 - \rho')\left(1 + \sum_{j=1}^{k-1} \text{avi}(T_j - v_j)\right)\right)$$

$$+ \frac{\rho - \rho'}{1 - \rho'} \sum_{j=1}^{k-1} \text{avi}(T_j).$$

By Lemma 4.2 and the induction hypothesis, we have $\text{avi}(T_j) \geqslant a|T_j| + \frac{1}{\sqrt{5}} - \frac{1}{3}$ for all j. It follows that

$$\sum_{j=1}^{k-1} \text{avi}(T_j) \geqslant \sum_{j=1}^{k-1}\left(a|T_j| + \frac{1}{\sqrt{5}} - \frac{1}{3}\right) = a(|T'| - 1) + (k - 1)\left(\frac{1}{\sqrt{5}} - \frac{1}{3}\right)$$

$$\geqslant a|T'| + 4\left(\frac{1}{\sqrt{5}} - \frac{1}{3}\right) - a > a|T'| + b.$$

Moreover, the induction hypothesis gives us $\mathrm{avi}(T') \geqslant a|T'| + b$. Finally,

- If $|T_k| \geqslant 4$, then by the induction hypothesis, Lemmas 4.2 and 4.3 , we have

$$
\rho \, \mathrm{avi}(T_k) + (1 - \rho) \, \mathrm{avi}(T_k - v_k)
$$
$$
\geqslant \rho \left(a|T_k| + \frac{2}{\sqrt{5}} - \frac{3}{4} \right) + (1 - \rho) \left(a(|T_k| - 1) + \frac{2}{\sqrt{5}} - \frac{3}{4} \right)
$$
$$
= a|T_k| + \frac{2}{\sqrt{5}} - \frac{3}{4} - (1 - \rho)a \geqslant a|T_k| + \frac{2}{\sqrt{5}} - \frac{3}{4} - \frac{a}{2} > a|T_k|.
$$

- If $|T_k| = 3$, then $\rho \, \mathrm{avi}(T_k) + (1-\rho) \, \mathrm{avi}(T_k - v_k) \geqslant \rho + (1-\rho) \cdot \frac{2}{3} = \frac{2+\rho}{3} \geqslant \frac{5}{6} > 3a$ (by Lemma 4.3).
- If $|T_k| = 2$, then $\rho \, \mathrm{avi}(T_k) + (1 - \rho) \, \mathrm{avi}(T_k - v_k) = \rho \cdot \frac{2}{3} + (1 - \rho) \cdot \frac{1}{2} = \frac{3+\rho}{6} \geqslant \frac{7}{12} > 2a$ (by Lemma 4.3).
- If $|T_k| = 1$, then $\rho \, \mathrm{avi}(T_k) + (1 - \rho) \, \mathrm{avi}(T_k - v_k) = \rho \cdot \frac{1}{2} + (1 - \rho) \cdot 0 = \frac{\rho}{2}$, and since $\frac{\mathrm{I}(T_k - v_k)}{\mathrm{I}(T_k)} = \frac{1}{2}$ in this case, we have $\rho \geqslant \frac{2}{3}$ by (7). Thus

$$
\rho \, \mathrm{avi}(T_k) + (1 - \rho) \, \mathrm{avi}(T_k - v_k) \geqslant \frac{1}{3} > a.
$$

In conclusion, $\rho \, \mathrm{avi}(T_k) + (1 - \rho) \, \mathrm{avi}(T_k - v_k) > a|T_k|$. Combining all inequalities, we obtain

$$
\mathrm{avi}(T) > a|T_k| + \frac{1 - \rho}{1 - \rho'} \left(a|T'| + b \right) + \frac{\rho - \rho'}{1 - \rho'} \left(a|T'| + b \right)
$$
$$
= a(|T'| + |T_k|) + b = a|T| + b.
$$

This completes the case that $k \geqslant 5$, so we are left with the cases $k = 3$ and $k = 4$. We return to the representation

$$
\mathrm{avi}(T) = \rho \sum_{j=1}^{k} \mathrm{avi}(T_j) + (1 - \rho) \left(1 + \sum_{j=1}^{k} \mathrm{avi}(T_j - v_j) \right). \tag{8}
$$

Now we distinguish different cases depending on how many of the branches T_j have one, two or three vertices respectively. If T_j has three vertices, we also distinguish whether v_j is the centre vertex or a leaf of T_j. This gives us a total of 35 cases for $k = 3$ and 70 cases for $k = 4$, corresponding to the solutions of

$$
x_1 + x_2 + x_3 + x_4 + x_5 = k.
$$

Here, x_1 and x_2 stand for the number of T_j's with one and two vertices respectively, x_3 and x_4 for the number of T_j's with three vertices and v_j the centre (x_3) or a leaf (x_4) respectively, and x_5 is the number of T_j's with four or more vertices. In each of the cases, we use the following explicit values and estimates:

$$
\operatorname{avi}(T_j)
\begin{cases}
= \frac{1}{2} & |T_j| = 1, \\
= \frac{2}{3} & |T_j| = 2, \\
= 1 & |T_j| = 3, \\
\geqslant a|T_j| + \frac{2}{\sqrt{5}} - \frac{3}{4} & \text{otherwise,}
\end{cases}
$$

$$
\operatorname{avi}(T_j - v_j)
\begin{cases}
= 0 & |T_j| = 1, \\
= \frac{1}{2} & |T_j| = 2, \\
= \frac{2}{3} & |T_j| = 3 \text{ and } v_j \text{ is a leaf of } T_j, \\
= 1 & |T_j| = 3 \text{ and } v_j \text{ is the centre of } T_j, \\
\geqslant a(|T_j| - 1) + \frac{2}{\sqrt{5}} - \frac{3}{4} & \text{otherwise,}
\end{cases}
$$

$$
\frac{\mathrm{I}(T_j - v_j)}{\mathrm{I}(T_j)}
\begin{cases}
= \frac{1}{2} & |T_j| = 1, \\
= \frac{2}{3} & |T_j| = 2, \\
= \frac{3}{5} & |T_j| = 3 \text{ and } v_j \text{ is a leaf of } T_j, \\
= \frac{4}{5} & |T_j| = 3 \text{ and } v_j \text{ is the centre of } T_j, \\
\in \left[\frac{1}{2}, 1\right] & \text{otherwise.}
\end{cases}
$$

The bounds on avi in the case that T_j has four or more vertices follow from the induction hypothesis (if $T_j - v_j$ is disconnected, applied to all components), while the last inequality is simply Lemma 4.3.

We plug these bounds into (8) and also use the identity (7) again. Since the expression (8) is linear in ρ, its minimum is either attained for the largest or smallest possible value of ρ. This gives us a lower bound for $\operatorname{avi}(T)$ in each of the aforementioned 105 cases, which can all be checked easily with a computer. As an example, let us consider the case that gives us the worst bound: it is obtained for $x_1 = x_3 = x_4 = 0$, $x_2 = 1$ and $x_5 = 2$ (i.e., one branch with two vertices, two branches with four or more vertices). Let T_1 and T_2 both have four or more vertices, so that the third branch T_3 consists of only two vertices. We have

$$
\operatorname{avi}(T_1) \geqslant a|T_1| + \frac{2}{\sqrt{5}} - \frac{3}{4},
$$

$$
\operatorname{avi}(T_2) \geqslant a|T_2| + \frac{2}{\sqrt{5}} - \frac{3}{4},
$$

$$
\operatorname{avi}(T_3) = \frac{2}{3}
$$

and

$$
\operatorname{avi}(T_1 - v_1) \geqslant a|T_1| - a + \frac{2}{\sqrt{5}} - \frac{3}{4},
$$

$$
\operatorname{avi}(T_2 - v_2) \geqslant a|T_2| - a + \frac{2}{\sqrt{5}} - \frac{3}{4},
$$

$$
\operatorname{avi}(T_3 - v_3) = \frac{1}{2}
$$

as well as

$$\rho = \left(1 + \frac{2}{3} \frac{\mathrm{I}(T_1 - v_1)\mathrm{I}(T_2 - v_2)}{\mathrm{I}(T_1)\mathrm{I}(T_2)} \right)^{-1} \in \left[\frac{3}{5}, \frac{6}{7} \right]$$

by Lemma 4.3. Thus

$$\mathrm{avi}(T_1) + \mathrm{avi}(T_2) + \mathrm{avi}(T_3) \geqslant a(|T_1| + |T_2|) + 2\left(\frac{2}{\sqrt{5}} - \frac{3}{4} \right) + \frac{2}{3}$$
$$= a(|T| - 3) + \frac{4}{\sqrt{5}} - \frac{5}{6} = a|T| + \frac{11}{2\sqrt{5}} - \frac{7}{3}$$

and likewise

$$\mathrm{avi}(T_1 - v_1) + \mathrm{avi}(T_2 - v_2) + \mathrm{avi}(T_3 - v_3)$$
$$\geqslant a(|T_1| - 1 + |T_2| - 1) + 2\left(\frac{2}{\sqrt{5}} - \frac{3}{4} \right) + \frac{1}{2}$$
$$= a(|T| - 5) + \frac{4}{\sqrt{5}} - 1 = a|T| + \frac{13}{2\sqrt{5}} - \frac{7}{2}.$$

Plugging all these inequalities into (8), we obtain

$$\mathrm{avi}(T) \geqslant \rho\left(a|T| + \frac{11}{2\sqrt{5}} - \frac{7}{3} \right) + (1 - \rho)\left(1 + a|T| + \frac{13}{2\sqrt{5}} - \frac{7}{2} \right)$$
$$= a|T| + \frac{13}{2\sqrt{5}} - \frac{5}{2} + \rho\left(\frac{1}{6} - \frac{1}{\sqrt{5}} \right) \geqslant a|T| + \frac{13}{2\sqrt{5}} - \frac{5}{2} + \frac{6}{7}\left(\frac{1}{6} - \frac{1}{\sqrt{5}} \right)$$
$$= a|T| + b.$$

The other cases are treated in the same fashion and give lower bounds with larger constant terms. To complete the proof of the theorem, it only remains to prove an upper bound on $\mathrm{avi}(P_n)$. However, we already know from Lemma 4.2 that

$$\mathrm{avi}(P_n) = an + \frac{3 - \sqrt{5}}{5} - \frac{n + 2}{\sqrt{5}((-\phi^2)^{n+2} - 1)}$$
$$\leqslant an + \frac{3 - \sqrt{5}}{5} - \frac{7}{\sqrt{5}((-\phi^2)^7 - 1)} = an + \frac{\sqrt{5}}{2} - \frac{25}{26}$$

for $n > 3$, and $\frac{\sqrt{5}}{2} - \frac{25}{26} \approx 0.15649553 < b$. Therefore, $\mathrm{avi}(P_n) < an + b \leqslant \mathrm{avi}(T)$ for every tree T with n vertices other than P_n.   □

5 A more general setting

It is common in statistical physics to consider the hard-core distribution on the independent sets I of a graph G. That is, the study of a random independent set I with probability proportional to $\alpha^{|I|}$. In [2,3], the authors consider this model and prove bounds on the expected size of an independent set drawn from the hard-core model on G at fugacity α. When $\alpha = 1$, this expected size is precisely the invariant avi that we investigated in this paper. Recall that $\mathrm{i}(G, k)$ is the number of independent vertex subsets of size k in G. Now, choose a random independent set with probability proportional to α^k, where k is the size of the set and α is a positive number. We define the weighted total number of independent subsets of G, the weighted total size of independent subsets of G and the weighted average size of independent vertex subsets in G:

$$\mathrm{I}^\alpha(G) = \mathrm{I}(G, \alpha) = \sum_{k \geqslant 0} \mathrm{i}(G, k)\alpha^k,$$

$$\mathrm{T}^\alpha(G) = \mathrm{T}(G, \alpha) = \sum_{k \geqslant 0} k\,\mathrm{i}(G, k)\alpha^k,$$

$$\mathrm{avi}^\alpha(G) = \frac{\mathrm{T}(G, \alpha)}{\mathrm{I}(G, \alpha)}.$$

Example 5.1 For the n-vertex edgeless graph E_n and the star S_n, we have

$$\mathrm{I}^\alpha(E_n) = \sum_{k=0}^{n} \binom{n}{k} \alpha^k = (1 + \alpha)^n,$$

$$\mathrm{I}^\alpha(S_n) = \alpha + \sum_{k=0}^{n-1} \binom{n-1}{k} \alpha^k = \alpha + (1 + \alpha)^{n-1},$$

$$\mathrm{T}^\alpha(E_n) = \sum_{k=0}^{n} \binom{n}{k} k\alpha^k = \alpha n(1 + \alpha)^{n-1},$$

$$\mathrm{T}^\alpha(S_n) = \alpha + \alpha(n - 1)(1 + \alpha)^{n-2},$$

$$\mathrm{avi}^\alpha(E_n) = \frac{\alpha n}{1 + \alpha},$$

$$\mathrm{avi}^\alpha(S_n) = \frac{\alpha + \alpha(n - 1)(1 + \alpha)^{n-2}}{\alpha + (1 + \alpha)^{n-1}}.$$

All the results presented in this paper, except for the extremality of the path, generalise to this weighted average.

Theorem 5.2 *For any n-vertex graph G which is not the complete graph K_n and not the edgeless graph E_n we have* $\mathrm{avi}^\alpha(K_n) < \mathrm{avi}^\alpha(G) < \mathrm{avi}^\alpha(E_n)$.

Theorem 5.3 *For any n-vertex tree $T \neq S_n$, we have* $\mathrm{avi}^\alpha(T) < \mathrm{avi}^\alpha(S_n)$.

The proof that the path is extremal generalises to the case that $\alpha \in (0, 1]$, but not to all real values of α (in fact, the path is not extremal for large values of α).

Theorem 5.4 *For every $\alpha \in (0, 1]$ and every tree T of order n, we have the inequality* $\mathrm{avi}^\alpha(T) \geqslant \mathrm{avi}^\alpha(P_n)$.

To some extent, this also explains why proving extremality of the path is harder than proving extremality of the star. We refer to [11] for more details.

References

1. Brightwell, G.R., Winkler, P.: Hard constraints and the Bethe lattice: adventures at the interface of combinatorics and statistical physics. In: Tatsien, L.I. (ed.) Proceedings of the International Congress of Mathematicians, Vol. III (Beijing, 2002), pp. 605–624. Higher Education Press, Beijing (2002)
2. Davies, E., Jenssen, M., Perkins, W., Roberts, B.: Independent sets, matchings, and occupancy fractions. J. London Math. Soc. **96**(1), 47–66 (2017)
3. Davies, E., Jenssen, M., Perkins, W., Roberts, B.: On the average size of independent sets in triangle-free graphs. Proc. Amer. Math. Soc. **146**(1), 111–124 (2018)
4. Haslegrave, J.: Extremal results on average subtree density of series-reduced trees. J. Combin. Theory Ser. B **107**, 26–41 (2014)
5. Jamison, R.E.: On the average number of nodes in a subtree of a tree. J. Combin. Theory Ser. B **35**(3), 207–223 (1983)
6. Jamison, R.E.: Monotonicity of the mean order of subtrees. J. Combin. Theory Ser. B **37**(1), 70–78 (1984)
7. Levit, V.E., Mandrescu, E.: The independence polynomial of a graph—a survey. In: Bozapalidis, S., Kalampakas, A., Rahonis, G. (eds.) Proceedings of the 1st International Conference on Algebraic Informatics, pp. 233–254. Aristotle University of Thessaloniki, Thessaloniki (2005)
8. Merrifield, R.E., Simmons, H.E.: Topological Methods in Chemistry. Wiley, New York (1989)
9. Mol, L., Oellermann, O.: Maximizing the mean subtree order. J. Graph Theory. https://doi.org/10.1002/jgt.22434
10. Prodinger, H., Tichy, R.F.: Fibonacci numbers of graphs. Fibonacci Quart. **20**(1), 16–21 (1982)
11. Razanajatovo Misanantenaina, V.: Properties of Graph Polynomials and Related Parameters. PhD thesis, Stellenbosch University (2017)
12. Stephens, A.M., Oellermann, O.R.: The mean order of sub-k-trees of k-trees. J. Graph Theory **88**(1), 61–79 (2018)
13. Vince, A., Wang, H.: The average order of a subtree of a tree. J. Combin. Theory Ser. B **100**(2), 161–170 (2010)
14. Wagner, S.: Upper and lower bounds for Merrifield-Simmons index and Hosoya index. In: Gutman, I., Furtula, B., Das, K.C., Milovanović, E., Milovanović, I. (eds.) Bounds in Chemical Graph Theory—Basics. Mathematical Chemistry Monographs, vol. 19, pp. 155–187. University of Kragujevac, Kragujevac (2017)
15. Wagner, S., Wang, H.: On the local and global means of subtree orders. J. Graph Theory **81**(2), 154–166 (2016)