Normalised degree variance: a network heterogeneity index unbiased to size and density

Keith Smith¹,* & Javier Escudero² Member, IEEE

Abstract—Finding graph indices which are unbiased to network size and density is of high importance both within a given field and across fields for enhancing comparability of modern network science studies. The degree variance is an important metric for characterising network heterogeneity. Here, we provide an analytically valid normalisation of degree variance to replace previous normalisations which are either invalid or not applicable to all networks. It is shown that this normalisation provides equal values for graphs and their complements; it is maximal in the star graph (and its complement); and its expected value is constant with respect to density for random networks of the same size. We strengthen these results with model observations in weighted random networks, random geometric networks and resting-state brain networks, showing that the proposed normalisation is unbiased to both network size and density. The closed form expression proposed also benefits from high computational efficiency and straightforward mathematical analysis. In an application of a subnetwork comparability problem of nationwide and within state US airport networks, the nationwide US airport network is shown to be much more heterogeneous than most within-state networks, illustrating the importance of the increased reliability of this true normalisation.

1 Introduction

The most fundamental parameters of a network are its size (number of nodes) and density (ratio of number of edges to total possible). The richness of topologies elicited by these simple building blocks has fascinated mathematicians for centuries and science, in general, for decades since the explosion of interest in understanding real-world complex networks. Comparing networks of different size and/or density is difficult because of the dependency of many network features on these fundamental parameters [1], yet the ability to compare between different sizes and densities is necessary to gain an unbiased understanding of the topologies in a given field as well as across fields. Indeed, the fact that differently sized networks explain similar phenomena in a given area is commonplace. For example, social networks depend on the number of participants in the study [2]; infrastructural networks within cities will vary in size according to the size of the city [3]; and networks constructed from sensors, such as in electrophysiology, depend on the number of sensors used [4].

The degree distribution of a network is a key property for assessing its topology [5]. One important characteristic of the degree distribution is the degree variance [6], \( \text{var}(k) \), which can aid in understanding the relationship of degree hierarchies and hub dominance [7]. However, variance depends on the size of samples used, which varies accordingly for fixed density and varying size—larger size needing more edges to account for the same density. Similarly, the degree distribution is dependent on the number of edges in the network since the maximum possible degree is fixed by network size. Thus, we need to work towards effective and efficient control of the scaling of network size and density in the measurement of degree variance. This is of particular necessity to support increasingly critical interdisciplinary areas using large multi-dimensional datasets.

Here, we propose a mathematically rigorous and computationally efficient normalisation of degree variance, \( \bar{v} \), with a closed form expression. Particularly, we will prove that \( \bar{v} \in [0,1] \) for all graphs. Not only this, we will show that graph complements achieve equivalent values of \( \bar{v} \); that \( \bar{v} \) only satisfies unity asymptotically for the star graph (and its complement) as \( n \to \infty \); and that \( \bar{v} \) is independent of network density for random graphs [18]. Furthermore, our normalisation is also well defined for graphs with isolated nodes and thus, we argue, of broader scope than the other relevant normalised measure of heterogeneity [16]. Demonstrations are provided that our normalisation is the least variable normalisation with respect to network size and density for a number of network types and that it is also computationally efficient. We will then show an application of the utility of the new normalisation where we compare within-state airport networks of flight times in the USA [23].

2 Background

A simple network is defined by a set of nodes, \( V = \{1, \ldots, n\} \), connected together by a set of edges, \( E = \{(i,j) : i,j \in V\} \). Network size is then \( |V| = n \). The convention is that \( |E| = 2m \) including each of the \( m \) edges twice ((i,j) & (j,i)). The largest possible size of \( E \) is obtained in the complete graph with \( 2m = n(n-1) \). Thus, network density \( d = 2m/n(n-1) \). The degree of a node, \( k_i \), is simply the number of edges adjacent to it, and we denote the set of degrees of a graph as \( k = \{k_1, \ldots, k_n\} \).

Degree variance is seen as a measure of graph heterogeneity which is essentially conceptually equivalent to

---

¹Centre for Medical Informatics, Usher Institute, University of Edinburgh, 9 Edinburgh BiSQ, EH16 4UX, UK
²School of Engineering, Institute for Digital Communications, University of Edinburgh, West Mains Rd, Edinburgh, EH9 3FB, UK
*e-mail: k.smith@ed.ac.uk
graph irregularity. Regular graphs have been of interest to mathematicians for many years, at least since the pioneering work of Petersen in 1891 [8]. However, it was not until the seminal paper of Collatz and Sinogowitz in 1957 [9] that indices to characterise the irregularity of graphs became a topic of interest. In [9] it was proposed to study the difference between the largest eigenvalue and the average degree, \( \epsilon = \lambda_1(G) - \frac{1}{2} \sum_i k_i / n \), observing that this measure was only zero for regular graphs. It was conjectured that star graphs attained the highest value of \( \epsilon \), but this was disproved by Cvetković & Rowlinson [10].

With the booming interest in applications of graph theory to real-world networks, the degree variance, \( v(G) = \text{var}(k) \), was proposed as a measure of graph heterogeneity in 1981 [6]. Later, Bell [11] compared \( v(G) \) and \( \epsilon \) and found incompatibilities between them, noting that the measures had different relative values for certain pairs of graphs. Using the constructs of quasi-complete and quasi-star graphs introduced by Ahlsewade & Katona [12], Bell then proved that \( \epsilon \) was in fact maximal only for quasi-complete graphs, whereas \( v(G) \) was maximal either for a quasi-complete or a quasi-star graph for any given network size and density (although not necessarily uniquely) [11].

A quasi-star graph is a graph consisting of \( p \) dominant nodes and at most one node of degree \( q > p+1 \) connected to other non-dominant nodes, here referred to as a developing node, for a graph with \( p(n-1)+q \) edges (Fig 1, left). If \( q = 1 \) then the remaining edge connects two non-dominant nodes so that they both have degree \( p+1 \). A quasi-complete graph is a graph consisting of a complete subgraph of order \( q \) and at least \( n - q - 1 \) isolated nodes, with the remaining node connected to \( p \) nodes of the complete subgraph for a graph with \( q(q-1) + p \) edges, see Fig 1, right. We regard here as perfect quasi-complete and quasi-star graphs as such graphs without any remainder, \( q \). These are special in that the perfect quasi-complete graph of order \( p \) is the complement of the perfect quasi-star graph of order \( n - p \).

In a related topic, Abrego et al. [13] proved which of either the quasi-star or quasi-complete graphs with \( n \) nodes and \( m \) edges obtained the maximum sum of squares of degrees. It was then shown by Smith & Escudero [7] that maximising the sum of squares of degrees for fixed \( n \) and \( m \) was equivalent to maximising the degree variance by the equation

\[
v(G) = \frac{||k||^2}{n-1} - \frac{(2m)^2}{n(n-1)}. \tag{1}
\]

where \( ||.|| \) is the \( L_2 \)-norm, and thus Abrego et al.’s result was also seen to hold for \( v(G) \).

In the work by Snijders [6], a normalisation of the degree variance was attempted through division by the value achieved for the quasi-star graph with same size and density because it was thought that the quasi-star graph was always maximal. For the quasi-star graph with same size and density as a graph \( G, G_{qs} \), this can be written

\[
J = \frac{v(G)}{v(G_{qs})}. \tag{2}
\]

But the previously noted results render such a normalisation invalid. Apart from this, one needs to construct the quasi-star graph to compute it, rendering a general mathematical analysis of this property infeasible and causing computational efficiency problems for very large graphs.

Normalisation of the degree variance has also been proposed through division by the average degree, \( \langle k \rangle \) of the network [14],

\[
s^2(G) = \frac{v(g)}{\langle k \rangle}. \tag{3}
\]

However the average degree depends on network density and thus network size.

The other notable class of measure of graph irregularity are those based on absolute differences of degrees. This was originated by Albertson [15] with the measure

\[
irr(G) = 1/2 \sum_{i,j} |k_i - k_j|. \tag{4}
\]

Estrada [16] then provided a normalised measure based instead on inverse square roots of degrees:

\[
\rho(G) = \frac{\sum_{i,j} \left(k_i^{-1/2} - k_j^{-1/2}\right)^2}{n - 2\sqrt{n-1}}, \tag{5}
\]

by relating it to the Randić index and using its known upper bounds [17]. Estrada argued that this measure was maximised only by star graphs and not quasi-complete graphs. However, in actuality the measure is not well defined for any graph with isolated nodes, such as quasi-complete graphs, since this leads to division by zero in the terms of \( k_i^{-1/2} \) for isolated \( i \). On top of this, as we shall see, this measure is biased to network density where graphs with larger densities can be expected to obtain lower values than graphs with small densities.

Thus, a normalised heterogeneity index which is applicable to all graph types and is invariant to network size and density is of great interest to help resolve these outstanding issues.

3 Normalised Degree Variance

We will begin directly with stating and proving the proposed normalisation of degree variance.

**Proposition 1.** For any graph \( G \) with density \( d = \frac{2m}{n(n-1)} \), the identity

\[
\bar{v}(G) = \frac{n-1}{nm(1-d)} v(G) \tag{6}
\]

is bounded in the interval \( [0, 1] \).

**Proof.** From (1), normalisation for degree variance can be achieved given a known upper bound of \( ||k||^2 \) for a graph. Fortunately, this exists from a result by de Caen [24],

\[
||k||^2 \leq m \left( \frac{2m}{n-1} + n - 2 \right). \tag{7}
\]
Substituting this into (1), we have

\[
v(G) \leq \frac{m}{n(n-1)} \left( \frac{2m}{n-1} + n - 2 \right) - \frac{(2m)^2}{n(n-1)}
\]

\[
\iff n - \frac{1}{m} \leq \frac{2m}{n(n-1)} + n - 2 - \frac{4m}{n(n-1)}
\]

\[
\iff n - \frac{1}{nm} \leq \frac{2m}{n(n-1)} + 1 - \frac{2}{n} - \frac{4m}{n^2(n-1)}
\]

\[
\iff 1 - \frac{2}{n} + \frac{4m}{n^2(n-1)} \leq \frac{2m}{n(n-1)} - \frac{2mn}{n^2(n-1)}
\]

\[
= 1 - \frac{2m}{n(n-1)}
\]

\[
\iff \bar{v}(G) \leq 1,
\]

where * comes from the fact that \(2m \leq n(n-1)\).

We now have a valid normalisation of degree variance, but its properties with respect to quasi-star and quasi-complete graphs remains to be seen. Further, we must check its behaviour with respect to network size and density. The following results prove that \(\bar{v}(G)\) for a star graph of size \(n\) tends to 1 as \(n \to \infty\), and that perfect quasi-star graphs are a decreasing function of density, thus the star graph is always maximal amongst them. These results validate degree variance as a relevant measure of heterogeneity and, as far as we are aware, the only normalised measure of heterogeneity that is valid for all simple graphs, including those with isolated nodes.

**Corollary 1.** Let \(G^*(n, p)\) be the perfect quasi-star graph with \(n\) nodes and \(p\) dominant nodes, then the following statements hold:

1) \(\bar{v}(G^*(n, 1)) \to 1\) as \(n \to \infty\)

2) \(\bar{v}(G^*(n, p))\) is a monotonically decreasing function with respect to \(p\) for \(n \geq 1, p \geq 0 \in \mathbb{R}\)

**Proof.**

1) For a perfect quasi-star graph, \(m\) is made up of \(p\) dominant nodes of degree \(n-1\) and \(n-p\) nodes of degree \(p\). The general degree sequence of a perfect quasi-star graph is thus

\[
\{p, \ldots, p, n-1, \ldots, n-1\}
\]

and

\[
2m = (n-p)p + p(n-1) = p(2n-p-1),
\]

so that

\[
1 - d = 1 - \frac{p(2n-p-1)}{n(n-1)} = \frac{n(n-1) - p(2n-p-1)}{n(n-1)} = \frac{n^2 + p^2 - 2pn - (n-p)}{n(n-1)} = \frac{(n-p)(n-p-1)}{n(n-1)}.
\]

Then

\[
\bar{v}(G^*(n, p)) = \frac{1}{1 - \frac{n(n-1)}{n} - \frac{(2m)^2}{n^2m}} = \frac{(n-p)p^2 + p(n-1)^2 - (2m)^2}{nm}.
\]

\[
= \frac{n(n-1)}{n} - \frac{2(p^2(n-p)p + p(n-1)^2 - (2m)^2)}{n^2m}.
\]

and for a the star graph with \(p = 1\), this simplifies to

\[
\bar{v}(G^*(n, 1)) = \frac{n^2(n-1) - (2n-2)^2}{(n-2)(n-1)n} = \frac{n^3 - 5n^2 + 8n - 4}{n^3 - 3n^2 + 2n} \to 1 \text{ as } n \to \infty,
\]

as required.

2) It is well known that a function is monotonically decreasing if and only if its derivative is always less than or equal to zero. Thus, for \(f(p) = \bar{v}(G^*(n, p))\), we prove that \(f'(p) \leq 0 \forall n \geq p \geq 0\).
First we factorise the numerator of (22). Since \( p = n \) is the complete graph with 0 degree variance, it must be that \( p = n \) is a root of this polynomial. Then, by long polynomial division we find that \( p = n - 1 \) is a double root of this polynomial and the numerator factorises to give

\[
\bar{v}(G^a(n, p)) = \frac{2n(n-1)(n-p)(n-p-1)^2}{(n-p)(n-p-1)(2n-p-1)n^2},
\]

\[
= \frac{2(n-1)(n-p-1)}{(2n-p-1)n}
\]

(25)

We now differentiate with respect to \( p \) using the quotient rule to get

\[
\frac{\partial}{\partial p} \bar{v}(G^a(n, p)) = \frac{-2(n-1)(2n-p-1)n + n(2(n-1)(n-p-1))}{(2n-p-1)^2n^2}
\]

\[
= \frac{-2(n-1)(2n-p-1)n - n(n-1)n + n + 1}{(2n-p-1)^2n}
\]

\[
= \frac{-2(n-1)}{(2n-p-1)^2} \leq 0 \forall n, p \geq 0 \in \mathbb{R}
\]

(26)

(27)

(28)

(29)

as required.

While the degree variance gives equivalent values for graphs and their complements, it is clear that the heterogeneity index does not, since the value of a star graph is 1 while the value of its complement, the quasi-complete graph of order \( n - 1 \), is undefined. We will now show that the proposed normalisation of degree variance also provides equivalent values for graphs and their complements.

**Proposition 2.** For a graph \( G \) with complement \( \hat{G} \),

\[
\bar{v}(G) = \bar{v}(\hat{G}).
\]

(30)

**Proof.** For a graph \( G \) with edges and degrees \( k_1, \ldots, k_n \) so that \( \sum_{i=1}^n k_i = 2m \), its complement, \( \hat{G} \), has \( \hat{m} = (n(n - 1) - 2m)/2 \) edges and degrees \( \hat{k}_i = n - 1 - k_i \). Then

\[
\bar{v}(G) = \frac{n(n-1)\left(\sum_{i=1}^n k_i\right) - (n-1)(2m)^2}{nm(n(n-1) - 2m)},
\]

(31)

and

\[
\bar{v}(\hat{G}) = \frac{n(n-1)\left(\sum_{i=1}^n \hat{k}_i\right) - (n-1)(2\hat{m})^2}{n\hat{m}(n(n-1) - 2n)}
\]

\[
= \frac{n(n-1)\left(\sum_{i=1}^n (n - 1 - k_i)^2\right) - (n-1)(n(n-1) - 2m)^2}{2n\hat{m}(n(n-1)-2m)}
\]

\[
= \frac{n(n-1)(n(n-1)^2 + \sum k_i^2 - (n-1)\sum k_i)}{2n\hat{m}(n(n-1)-2m)}
\]

\[
= \frac{n(n-1)(n(n-1)^2 + \sum k_i^2 - (n-1)^2\sum k_i)}{2n\hat{m}(n(n-1)-2m)}
\]

\[
= \frac{n(n-1)(n(n-1)^3 + 4n(n-1)^2m - (n-1)(2m)^2)}{2n\hat{m}(n(n-1)-2m)}
\]

\[
= \frac{n(n-1)\sum k_i^2 - (n-1)(2m)^2}{nm(n(n-1) - 2m)} = \bar{v}(G),
\]

(32)

(33)

(34)

(35)

(36)

as required.

\[\square\]

**3.1 Normalised degree variance of random networks**

The well-known statistical properties of the degree distributions of random networks allows us to calculate \( \bar{v} \) for random graph ensembles \( G(n, q) \) where \( q \) is the probability of the existence of edges.

**Corollary 2.** For the random graph ensemble \( G(n, q) \)

\[
\bar{v}(G(n, q)) = \frac{2(n-1)}{n^2} \to 0 \text{ as } n \to \infty \quad \forall q
\]

(37)

**Proof.** Recall that \( G(n, q) \) has a binomial degree distribution with variance \((n-1)q(1-q)\) and also that \( q = 2m/n(n-1) \) [18]. Then

\[
\bar{v}(G(n, p)) = \frac{n-1}{nm(1-q)}(n-1)q(1-q)
\]

\[
= \frac{(n-1)^2}{nm(n-1)} \quad \text{as } n \to \infty,
\]

(38)

(39)

(40)

as required.

\[\square\]

Importantly, it is clear from this result that \( \bar{v}(G) \) is unbiased to network density for random graphs since \( \bar{v}(G(n, p)) \) is independent of \( p \), i.e. attains the same value for all densities. Furthermore, it decays fairly slowly towards 0 at a rate of \( 1/n \).

**3.2 Network size-independent minimum of normalised degree variance for graphs with a fixed proportion of vertices of a given degree**

Here we consider lower bounds for the normalisation’s invariance to network size. Suppose for a graph, \( G \), we guarantee \( x \) nodes of degree \( a \), having degrees \( k = \{a, a, \ldots, a, k_{x+1}, \ldots, k_n\} \). Then

\[
\bar{v}(G) = \frac{1}{1-\hat{d}} \left( \frac{\sum_{i=1}^n k_i^2}{nm} - \frac{4m}{n^2} \right)
\]

\[
= \frac{1}{1-\hat{d}} \left( \frac{nx^2 + \sum_{i=1}^{x+1} k_i^2}{nm} - \frac{4m}{n^2} \right)
\]

\[
\geq \frac{1}{1-\hat{d}} \left( \frac{xa^2 + (n-x)\left(\frac{2m-xa}{n-x}\right)^2}{nm} - \frac{4m}{n^2} \right)
\]

\[
= \frac{1}{1-\hat{d}} \left( \frac{x(a^2n^2 - 4amn + 4m^2)}{n^2(n-x)m} \right)
\]

\[
= \frac{1}{1-\hat{d}} \left( \frac{x(a^2n^2 - 4amn + 4m^2)}{n^2(n-x)m} \right)
\]

\[
= \frac{1}{1-\hat{d}} \left( \frac{x(2n^2 - 4amn + 4m^2)}{n^2(n-x)m} \right),
\]

where * comes from the fact that \( \sum_{i=1}^n k_i = 2m \) and that the minimum value of the sum of squares come from all elements being equal \( \implies k_{x+i} = (2m-x)/(n-x) \forall i \).

Now, we fix density so that \( \hat{d} = 2m/n(n-1) \), i.e. \( 2m = dn(n-1) \) and we get

\[
\frac{2x(an - dn(n-1))^2}{d(1-\hat{d})n^2(n-1)(n-x)}.
\]

(41)
Now, \[ \Box \] 0

\[ a(n - 1) = d = \frac{2m}{n(n - 1)} \]

\[ a = \frac{2m}{n} \]

which is the average degree of the graph, which makes sense since regular graphs have degree variance 0 where each node has degree \( \frac{2m}{n} \). Taking \( a \) to be small, say 1, the leading terms in \[ \Box \] are

\[ \frac{2xdn^4}{(1 - d)n^5} = \frac{x}{n} \frac{2d}{1 - d}, \]

(42)

Noting that \( x/n \) is the proportion of nodes which have degree 1, we thus have a constant minimum value of degree variance for such graphs with known density \( \forall n \).

We can also consider \( a = n - 1 \), i.e. graphs with a known proportion of dominant vertices. Then, \[ \Box \] becomes

\[ \frac{2x(n(n - 1))^2}{d(n - x)} \approx \frac{x}{n} \frac{2(1 - d)}{d}, \]

(43)

considering the leading terms. Notably, this is precisely 42 if we take \( d = 1 - d \).

Therefore, if the proportions of a given degree of a network is known, then a lower bound of \( V \) for such networks can be established. This could help, for example, in questions of hub dominance in real world networks. One can fix the number of dominant nodes in the network and know the range of \( V \) possible for such networks with greater accuracy. We shall now go on to apply our normalisation to network models and real world data to show its relevance to problems of network comparability and computational efficiency.

4 EXPERIMENTS

4.1 Validity and stability with respect to network size of normalisation approaches

Counter examples for the previously proposed normalisations are not difficult to come by. To demonstrate this we shall consider values for quasi-star graphs, quasi-complete graphs, weighted Erdős-Rényi random graphs \[ \Box \] and weighted random geometric graphs \[ \Box \], and EEG networks of size \( n = 16, 32, 64 \) and 128.

The EEG networks are derived from a 129 node EEG eyes open dataset. This is available online under an Open Database License via the Neurophysiological Biomarker Toolbox tutorial \[ \Box \]. It consists of data for 16 volunteers. We have previously used the data for which full processing details can be found in \[ \Box \]. Weighted connectivity adjacency matrices were computed using the phase-lag index (PLI) \[ \Box \]. To get a network of size \( n \) from these EEG networks, \( n \) electrodes were chosen at random 100 times and results averaged.

For each \( n \), the perfect quasi-star graphs and perfect quasi-complete graphs for each \( d \), and integer percentage binarisation thresholds of random graphs, random geometric graphs and EEG networks are computed and we take the proposed normalisation, \( \bar{v} \), quasi-star normalisation, \( J \), average degree normalisation, \( \sigma^2 \), as well as the heterogeneity index, \( \rho \), of these networks. Fig. 2 shows the obtained values plotted against density.

The proposed normalisation works as expected with all values in [0, 1]. The maximum value is achieved by the largest (\( n = 128 \)) star graph (perfect quasi-star with \( p = 1 \)) and its complement perfect quasi-complete graph (\( p = n - 1 \)). The normalisation by quasi-star graphs, Fig. 2, centre, shows clear violations of the [0, 1] normalisation by the quasi-complete graphs— as is expected in cases where the quasi-complete graph has larger degree variance than the quasi-star graph— as well as low densities of thresholded weighted graph models. The normalisation by average degree increases proportionally with \( n \) thus is critically flawed as a normalisation. We shall refrain from using it further and focus on the other indices.

All of the graphs show a remarkable stability with respect to network size of small networks, suggesting that this normalisation is very suitable for comparing networks of different sizes and similar density. Furthermore, for non-extremal values of the EEG networks, there is a marked stability with respect to density also, supporting the claim that this normalisation is unbiased to network density. To quantify these statistically, we compute the coefficient of variance (i.e. the ratio of the standard deviation to the mean) of the normalisations with respect to network size for each density, averaging over densities, and the coefficient of variance of the normalisations with respect to density, averaging over network sizes, respectively.

The resulting average coefficients of variation are reported in Table 1. Smaller values show less variability, with bold indicating best performance for each network. The proposed normalisation has the least variability with respect to network size for quasi-complete graphs, random geometric graphs and EEG networks. For quasi-star graphs, normalisation by quasi-star achieves a variability of 0, redundantly. For random graphs, \( \bar{v} \) is second only to \( \sigma^2 \). The result for \( \sigma^2 \) here is quite anomalous, but it is probably achieved actually because it is such a poor normalisation of network size for general networks— for our normalisation, for instance, random graph values are inversely proportional to \( n \).

With respect to density, the proposed normalisation has the least variability for quasi-complete graphs, random graphs and random geometric graphs. Again, for quasi-star graphs, normalisation by quasi-star achieves a variability of 0, redundantly. For EEG networks, \( \bar{v} \) is second only to \( J \). On inspection of Fig. 2, top left and right, however it would appear that this is due to the extremal values. For \( \bar{v} \), values fall steeply towards at lowest and highest densities for all network sizes. For \( J \) however, some fall steeply towards 0 and others fall steeply upwards, thus perhaps cancelling each other out in the coefficient of variation.

Indeed, this steep fall towards zero is a notable feature in values of \( \bar{v} \) of EEG networks. This means that at low densities, the networks are far from star-like, more similar to quasi-complete graphs, whereas at highest densities, the networks are far from quasi-complete, showing properties more similar to quasi-star networks. This can be interpreted in light of the well known “rich-club” phenomenon of brain networks— nodes with lots of connections are connected.
Fig. 2. Normalisations of degree variance of quasi-star graphs (QS), quasi-complete graphs (QC), Erdős-Rényi random graphs (E-R), Random Geometric Graphs (RGG) and EEG PLI networks (EEG) of sizes 16 (dotted lines), 32 (dashed lines), 64 (dash-dotted lines) and 128 (solid lines). The proposed normalisation shows smooth curves which are markedly stable with respect to $n$ for all graphs. Clearly, $J$ and $\sigma^2$ violate standard normalisation principles and are influenced by network size. On the other hand $\rho$ cannot be computed for graphs with isolated nodes which are particularly prominent at lower densities of the weighted models and values are squeezed at high densities. particularly strongly together [25]. This means that at sparse densities the rich-club evolves as an almost complete subnetwork, keeping heterogeneity low. On the other hand, the dominance of connections to hub nodes means that at very high densities, hub nodes and highly connected nodes become saturated (i.e. share edges with all other nodes), leaving, for want of a better term, a ‘poor-cub’ of isolated and weakly connected nodes, similar to quasi-star graphs.

Further to this, in section B of the Appendix, we provide an analytical result which says that, if we know the proportion of nodes in the network which have a given degree, $k_i$, we can find a fixed lower bound for $V$ which does not depend on network size. The ability to attain such results from the closed form expression of $V$ is another of its key benefits.

4.2 Computational efficiency of the closed-form expression

As the networks grow large, the previously analysed variability becomes negligible for $\bar{v}$ and $J$, where it appears there is a possible asymptotic convergence to a set limiting curve for each network type in Fig. 2. However, what becomes more important as networks grow large is rather the computational cost. The computational efficiency of $\bar{v}$ compared to $J$ can be garnered by comparing processing times of degree variance normalisations for larger graphs. We also compare with $\rho$ for completeness, which uses the graph Laplacian in its computation [16]. We use sparse scale-free graphs [20] at 1% density to demonstrate this. We will look at graphs of size 5,000, 10,000, 50,000 and 100,000. Scale-free graphs of density 1% can be obtained by selecting 1% of nodes as the core subgraph and the average degree of additional nodes as 0.5%n. The computation time using Matlab algorithms on a single core of a 3.6 GHz Intel Core i7 Processor 4274 HE with 32 GB 2400 MHz DDR4 is computed 25 times and the average time is shown in Fig. 3. This clearly demonstrates the increased computational efficiency of $\bar{v}$ over $J$ as a normalisation of degree variance. Indeed, $\bar{v}$ was computed an order of 10 times faster at each network size and the trend appears to show that this difference grows with greater network sizes. For example, $\bar{v}$ was computed for 10,000 node networks with an average speed of 0.003s while $J$ was computed with an average speed of 0.038s. Scaling to 100,000 node networks the respective average speeds were 0.300s and 8.163s– a factor of over 27

| Network          | $\bar{v}$ | $J$  | $\sigma^2$ | $\rho$ |
|------------------|----------|------|------------|--------|
| Quasi-star       | 0.0249   | 0    | 0.7647     | 0.5831 |
| Quasi-complete   | 0.0812   | 0.1600 | 0.8512 | n/a    |
| Random           | 0.0812   | 0.1600 | 0.8512 | n/a    |
| Random Geometric | 0.2240   | 0.3432 | 0.6418 | 0.4660 |
| EEG              | 0.0540   | 0.1825 | 0.7724 | 0.5485 |

| Network          | $\bar{v}$ | $J$  | $\sigma^2$ | $\rho$ |
|------------------|----------|------|------------|--------|
| Quasi-star       | 0.4320   | 0    | 1.1026     | 1.4377 |
| Quasi-complete   | 0.4509   | 0.4786 | 0.4877 | n/a    |
| Random           | 0.4509   | 0.4786 | 0.4877 | n/a    |
| Random Geometric | 0.3756   | 0.5469 | 0.4606 | 0.6287 |
| EEG              | 0.2610   | 0.2386 | 0.6085 | 0.8387 |

TABLE 1

TABLE 2

Coefficient of variation of network indices with respect to network size, averaged over density.

Coefficient of variation of network indices with respect to density, averaged over network size.
in difference. Furthermore, \( \bar{v} \) is has greater computational efficiency than \( \rho \) which showed average speeds of 0.006 for 10,000 nodes and 1.043 for 100,000 nodes, the latter being a factor of 3.5 in difference.

### 4.3 Illustration in a North American airport network

Let us demonstrate the proposed normalisation in the comparison of subnetworks in a real dataset. A North American airport network [23] consisting of airports in the USA and Canada was found in the Graph Algorithms in Matlab Code toolbox using command `load_gaimc_graph('airports')`. This consisted of 456 airports and the travel times of flights connecting them. We looked at the whole network and compared with subnetworks of the ten most populous US states (all nodes and connections within state)– California (CA), Texas (TX), Florida (FL), New York (NY), Pennsylvania (PA), Illinois (IL), Ohio (OH), Georgia (GA), North Carolina (NC) and Michigan (MI). The densities, network sizes and mean travel time are shown in Table 3 alongside several geo-economic indicators– landmass area, population density and gross domestic product (GDP). Importantly, note that the densities of these networks all lie within the more stable regions (parallel with the \( x \)-axis with respect to \( \bar{v} \) and \( J \) shown in Fig. 2).

Computing \( \bar{v} \), \( J \) and \( \rho \) for each binarised network, we get the values shown in Fig. 3. The values for the network as a whole are \( \bar{v} = 0.5012 \), \( J = 0.5687 \) and \( \rho = 0.1475 \). It is immediately clear that \( \rho \) is inappropriate in this situation. Five of the ten within state networks contain isolated nodes, rendering values of \( \rho \) indefinable. Also notable is that all values of \( J \) for networks of size less than or equal to 10 are higher than for the whole network. Using pearson’s correlation coefficient between network size and \( \bar{v} \) and \( J \) yields a value of -0.7963, whereas for \( v \) this value is much less at -0.5925. This agrees with the finding in Fig. 2 showing the greater reliability of \( \bar{v} \) in a real world example.

Using \( \bar{v} \) then, the heterogeneity of Pennsylvania, Illinois and Georgia are shown to be representative of the entire US airport network, suggesting a degree of self-similarity of the US airport with these states. North Carolina is the only state seen to have a considerably larger heterogeneity than the US airport network as a whole. Inspection of this network shows one isolated node and two star-like nodes of Asheville and Charlotte airports connected to all other nodes. The general trend though, shown in the remaining 6 out of 10 states, is a more even distribution of flights amongst the airports than is seen in the network of the entire US. Checking against some standard characteristics of states, degree spread is found to be only weakly correlated with population density (corr = -0.234) and land area (corr = -0.377). A moderate correlation is found, however, with GDP (corr = -0.518), indicating the tendency for more prosperous states to have more even distributions of connections between airports. The lack of strong correlations indicates that the information provided by \( \bar{v} \) is not redundant– differences in how airport networks are laid out across different states cannot be accounted for by generalisations of standard geographical and economic indicators.

On top of this, we calculated \( \bar{v} \) for all states with at least 5 airports. The following 14 states were thus excluded: Arkansas, Connecticut, Delaware, Indiana, Kentucky, Maryland, Massachusetts, Nevada, New Hampshire, New Jersey, New Mexico, Rhode Island, Utah and Vermont. For the rest, their mean \( x \) and \( y \) co-ordinates were computed and their relative heterogeneity, defined as

\[
\hat{v} = \frac{v(G) - \min(\bar{v}(G))}{\max(\bar{v}(G)) - \min(\bar{v}(G))},
\]

for \( G \in \mathcal{G} \), where \( \mathcal{G} \) is the set of all state airport networks, is shown in Fig. 3. The grey background co-ordinates, marking each individual airport, are coloured based on \( \hat{v} \) for the whole network.

The five states with largest heterogeneity are Oklahoma, Washington, Missouri, North Carolina and South Carolina. The five states with largest homogeneity are Kansas, West Virginia, North Dakota, Alaska and Nebraska. Notably, there is a much greater tendency for states to have lower heterogeneity than the entire network, indicated by
Fig. 4. Degree variance normalisations of the airport networks of the 10 most populous US states. The orange line indicates the heterogeneity for the whole USA and Canada airport network. Whereas $\bar{v}$ shows a lower degree of variability between states, $J$ appears biased to smaller networks and $\rho$ cannot be computed for states with isolated airports (i.e. airports not connected to any other airport in the state).

Fig. 5. Relative $\bar{v}$ of state airport networks plotted by geological co-ordinates. Black indicates the state with the smallest heterogeneity and white indicates the state with the largest heterogeneity. The grey circles in the background mark the co-ordinates for each airport.

The two otherwise proposed normalisations were shown to be invalid and another normalised heterogeneity index was shown to be ill-defined for graphs with isolated nodes. Beyond this, our normalisation had generally lower variability with respect to network size in quasi-complete graphs, random graphs, random geometric graphs, and EEG networks. Using sparse scale-free models, it was also shown to be more computationally efficient than other normalisation approaches. We then demonstrated the usefulness of the proposed normalisation in comparing differently sized subnetworks in a North American airport flight connections network, where we could reliably demonstrate variability in airport network node connections. Notably, it was found that within state airport networks had generally much lower heterogeneity than the entire US network, revealing that heterogeneity of the flight infrastructure across the US is generally much different and more imbalanced to more local flight infrastructures. The usefulness of the closed form expression for normalisation was also demonstrated in the production of two mathematical results, showing that normalised degree variance of quasi-star graphs decreases with respect to density and that a flexible lower bound for normalised degree variance was possible depending on proportions of given degrees within the network. All things considered, the proposed normalisation is put forward as the standard to be used in most scenarios for understanding heterogeneity in complex networks.

5 CONCLUSION

We introduced and mathematically justified a true normalisation for degree variance in networks, showing a large degree of invariance to network size and network density.

REFERENCES

[1] C.F. Dormann, J. Fründ, N. Blüthgen, B. Gruber, “Indices, Graphs and Null Models: Analyzing Bipartite Ecological Networks”, The Open Ecology Journal, 2: 7-24, 2009.
[2] F. Vega-Redondo, “Complex Social Networks”, Cambridge University Press, Cambridge, UK, 2007.
[3] M. Batty, “The Size, Scale, and Shape of Cities”, Science, 319(5864): 769-771, 2008.
[4] K. Smith, D. Abasolo, J. Escudero, “Accounting for the complex hierarchical topology of EEG phase-based functional connectivity in network binarisation”, PLOS ONE, 12(10): e0186164, 2017.
[5] M.E.J. Newman, “Networks”, Oxford University Press, Oxford, UK, 2010.
[6] T.A.B. Snijders, “The degree variance: an index of graph heterogeneity”, Social Networks, 3(3): 163-174, 1981.
[7] K. Smith, K., J. Escudero, “The complex hierarchical topology of EEG functional connectivity” Journal of Neuroscience Methods, 76: 1-12, 2017.
[8] J. Petersen, “Die theorie der regularen graphs”, Acta Math., 15: 193-220, 1891.
[9] L. Collatz, U. Sinogowitz, “Spektren endlicher Grafen”, Abh. Math. Sem. Uniu. Hamburg, 21: 63-77, 1957.
[10] D. Cvetković, P. Rowlinson, “On connected graphs with maximal index”, Publications de l’Institut Mathématique. Nouvelle Série, 44(58): 29-34, 1988.
[11] F.K. Bell, “A note on the irregularity of graphs”, Lin. Alg. Appl., 16: 45-64, 1992.
[12] R. Ahlswede, G.O.H. Katona, “Graphs with maximal number of adjacent pairs of edges”, Acta Math. Acad. Sci. Hungar., 32: 97-120, 1978.
[13] B.M. Ábrego, S. Fernández-Merchant, M.G. Neubauer, W. Watkins, “Sum of squares of degrees in a graph”, Journal of Inequalities in Pure and Applied Mathematics, 10(3): 64, 2009.
[14] M.G. Zimmermann, V.M. Eguiluz, M. San Miguel, “Coevolution of dynamical states and interactions in dynamic networks”, Physical Review E, 69: 065102, 2004.
[15] M.O. Alberston, “The Irregularity of a Graph”, Ars Combinatorica, 46: 219-225, 1997.
[16] E. Estrada, “Quantifying network heterogeneity” Physical Review E, 82: 066102, 2010.
[17] X. Li, Y. Shi, “A survey on the Randić index”, MATCH Commun. Math. Comput. Chem., 59: 127-156, 2008.
[18] P. Erdős, A. Rényi, “On Random Graphs”, Publicationes Mathematicae Debrecen, 6: 290-297, 1959.
[19] J. Dall, M. Christensen, “Random geometric graphs”, Phys. Rev. E, 66: 016121, 2002.
[20] A.L. Barábasi, R. Albert, “Emergence of scaling in random networks”, Science, 286(5439): 509-512, 1999.
[21] Neurophysiological Biomarker Toolbox, available from www.nbtwiki.net.
[22] C.J. Stam, G. Nolte, A. Daffertshofer, “Phase lag index: Assessment of functional connectivity from multi channel EEG and MEG with diminished bias from common sources”, Human Brain Mapping, 28(11): 1178-1193, 2007.
[23] The US airport network. Details available online at: https://www.mathworks.com/matlabcentral/mlc-downloads/downloads/submissions/24134/versions/1/previews/gaimc/.  
[24] D. de Caen, “An upper bound on the sum of squares of degrees in a graph”, Discrete Mathematics, 185(1-3): 245-248, 1998.
[25] M.P. van den Heuvel, O. Sporns, “Rich-club organization of the human connectome”, J. Neurosci., 31(44): 15775-15786, 2011.