PLOS COMPLEX SYSTEMS

RESEARCH ARTICLE

Statistical complexity of heterogeneous geometric networks

Keith Malcolm Smith[1]*, **Jason P. Smith**[2]

1 Department of Computer and Information Sciences, University of Strathclyde, Glasgow, United Kingdom, **2** Department of Mathematics, Nottingham Trent University, Nottingham, United Kingdom

* keith.m.smith@strath.ac.uk

Abstract

Degree heterogeneity and latent geometry, also referred to as popularity and similarity, are key explanatory components underlying the structure of real-world networks. The relationship between these components and the statistical complexity of networks is not well understood. We introduce a parsimonious normalised measure of statistical complexity for networks. The measure is trivially 0 in regular graphs and we prove that this measure tends to 0 in Erdös-Rényi random graphs in the thermodynamic limit. We go on to demonstrate that greater complexity arises from the combination of heterogeneous and geometric components to the network structure than either on their own. Further, the levels of complexity achieved are similar to those found in many real-world networks. However, we also find that real-world networks establish connections in a way which increases complexity and which our null models fail to explain. We study this using ten link growth mechanisms and find that only one mechanism successfully and consistently replicates this phenomenon– probabilities proportional to the exponential of the number of common neighbours between two nodes. Common neighbours is a mechanism which implicitly accounts for degree heterogeneity and latent geometry. This explains how a simple mechanism facilitates the growth of statistical complexity in real-world networks.

OPEN ACCESS

Citation: Smith KM, Smith JP (2025) Statistical complexity of heterogeneous geometric networks. PLOS Complex Syst 2(1): e0000026. https://doi.org/10.1371/journal.pcsy.0000026

Editor: Réka Albert, Pennsylvania State University, UNITED STATES OF AMERICA

Received: February 29, 2024

Accepted: October 25, 2024

Published: January 3, 2025

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcsy.0000026

Data Availability Statement: The data that support the findings of this study are publicly available from https://github.com/JasonPSmith/hierarchical_complexity.

Funding: The author(s) received no specific funding for this work.

Author summary

A statistically complex system is one which is neither regular nor random, but contains diversity in components and structure. This departs from algorithmic complexity which describes how difficult it is to explain information, but which is maximal for uniformly random information. We provide a definition of statistical complexity for networks and propose a normalised measure which satisfies that definition. We go on to explore the relationship between statistical complexity and the two major components thought to underlie network structure– the popularity of nodes in making connections (degree heterogeneity) and the geometric similarity of nodes. We find that the statistical complexity of real-world networks agrees with a model which combines both components. We then notice a positive relationship between the density of links in real-world networks and their statistical complexity, which is not present in our modelling. We find that we can

Competing interests: The authors have declared that no competing interests exist.

replicate this relationship, however, by growing the network using link probability which is based on a pair of nodes' number of common neighbours. We conclude that statistical complexity is a natural by-product of uncomplicated network mechanisms, returning to the old adage that complexity arises from simplicity.

Introduction

Complexity is a word used often in its common meaning within various scientific disciplines to describe the size and multiplicity of facets and scales within a given real-world system. In such cases, it is often used without reference to a specific measurement, or measurements are focused on counting the numbers of such facets and scales which are apparent in that system. In computer science, complexity has two specific definitions. Firstly, computational complexity describes the shortest amount of processing time (as a function of input size) it takes to derive the desired output from an algorithm [1]. Secondly, Kolmogorov complexity is a measure of the complexity of information based on the size of the smallest piece of code required to derive that information as an output [2]. There is no systematic way of finding such a shortest piece of code and proving that it is the shortest piece of code to compute a generic piece of information [3], however Kolmogorov complexity is related to measures of entropy, based on the predictability of information. For a given size of information, *n*, it is understood that randomly generated pieces of information would require the largest amount of code to be deterministically reproduced, while completely regular information, e.g. aaa. . .a written *n* times, would take the least amount of code to reproduce– "Write 'a' *n* times". In this way there is a significant interest in framing complexity in terms of information entropy, since entropy similarly dictates a scale between regular and random structures.

Yet, while randomly generated information may be difficult to deterministically reproduce, it is not structurally complex in a statistical sense. Indeed, the statistical properties of randomly generated information are defined a-priori and are evidently simple. This led the field of dynamical systems to lay out a different conceptual framework of complexity. In this view, a measure of complexity should go to zero for regular and random structures in the thermodynamic limit (as number of components goes to infinity), while being higher for systems presenting non-trivial and diverse correlations [4, 5]. One particularly important point of developing measures of statistical complexity is that using a scale between regular and random with complexity somewhere in the middle, a common approach from an information theoretic angle, does not allow for a useful measure of complexity itself [4, 5]. Instead, we need a scale between the simple (regularity and randomness both having uniform generational principles) and the complex, allowing us to directly measure the extent of complexity in any given system.

When it comes to studying complexity in networks, we are concerned with the complexity of interactions– essentially, how diverse the connectivity patterns in the network are. While others borrow from the algorithmic view of complexity [6, 7], here we are concerned with the statistical complexity of networks. A notable early work on statistical complexity of networks introduced a measure called the network diversity score and provided a comprehensive overview of other complexity and entropy measures of networks and their limitations [8]. Another work considered statistical complexity in networks from an information theoretic angle, multiplying Jensen Shannon divergence of a network with network entropy [9]. Neither of these works, however, provides a treatment of statistical complexity as previously described, and the measurements have limitations partly owing to lack of normalisation to network size and/or density. In this study, we establish a normalisation of Hierarchical Complexity (NHC) as a network statistical complexity measure. In contrast to the network diversity score, NHC is a parsimonious calculation of a single feature of a network, rather than the product of four different measures [10]. In contrast to the measure in [9] it does not require a reference graph, displays strong independence to network density and vanishes in the thermodynamic limit for Erdös-Rényi random graphs.

The hierarchy referred to here is the degree hierarchy of the network. A hierarchically complex system is one for which diversity of connectivity patterns are found across hierarchical levels (either individual degrees or ranges of degrees called tiers). Its introduction was motivated by the need to measure the complicated hierarchical networks of brain function and structure where it was expected that diverse functionality would be reflected in diverse connectivity patterns. HC has so far seen limited application in fairly small ($n < 100$) macro-scale human brain networks [10–14] and a corpus of real-world networks of varied origin ($n < 5000$) [15].

A major issue with the generalisability of the HC measure is that it is not normalised by number of nodes or number of edges. Two similarly derived networks with different numbers of nodes or edges can be expected to have different values of HC. This paper addresses this issue by introducing a normalised HC (NHC) measure. We show mathematically that this measure is bounded above by 2 and satisfies the statistical complexity definition of being asymptotically zero for Erdös-Rényi random graphs (an appropriate equivalent to randomness in dynamical systems). We then go on to explore results of this normalisation on different types of random graphs with the two most evident structural properties relevant to real-world networks, hierarchy and geometry. Hierarchy here relates to the distribution of node fitness [16] or popularity [17]. Geometry relates to the latent space of similarities between nodes [18, 19]. Combining hierarchy and geometry successfully captures many of the properties of real-world networks [20], but whether these properties are enough to explain the statistical complexity of networks is not known. After this we go on to explore NHC in real-world networks. Here, an unexpected relationship between NHC and density is noted. Finally, we explore explanations for this relationship by applying different kinds of link growth mechanisms based on degrees and overlap of node neighbourhoods. Combined, the results reveal non-trivial hierarchical complexity in real-world networks and we open the way for more reliable and robust applications of hierarchical complexity across network domains.

Theory

Key themes within this work are encapsulated within the image in Fig 1 which illustrates hierarchical complexity arising from a combination of geometric and hierarchical structure. Henceforth, unless specified otherwise, let G be a graph with n nodes, m edges and density $d = 2m/n(n - 1)$.

Hierarchical complexity

To compute hierarchical complexity, we first define the *Neighbourhood Degree Sequence* (NDS) of a node i of degree k as

$$s_i = \{s_{i1}, s_{i2}, \cdots, s_{ik}\} \tag{1}$$

where the s_{ij}'s are the degree of the nodes to which i is connected such that $s_{i1} \leq s_{i2}, \ldots, \leq s_{ik}$.

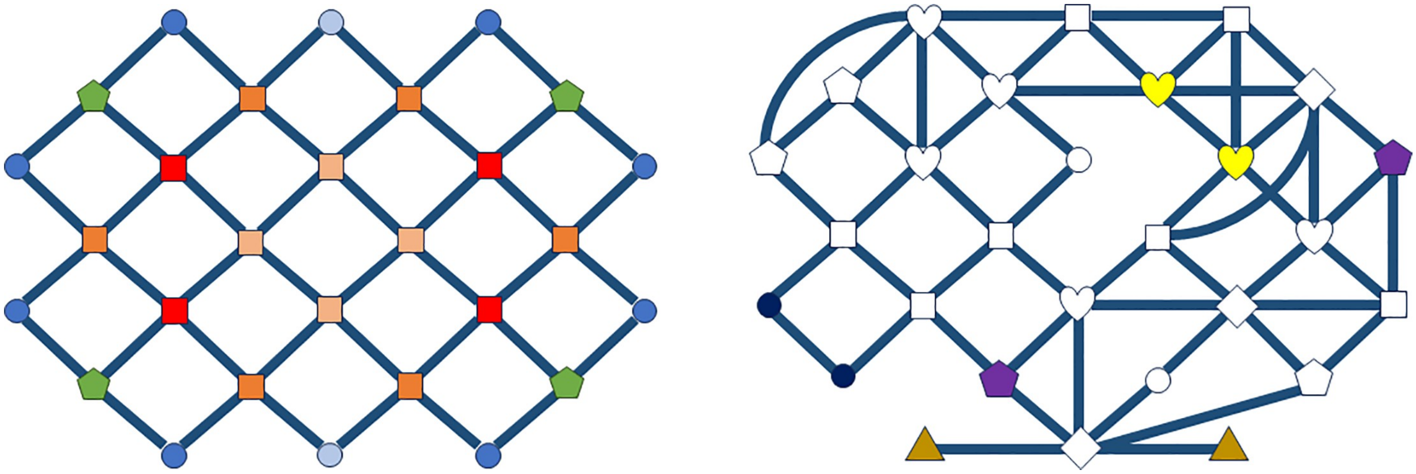


Fig 1. On the left we see a geometric graph with a regular structure. Node shapes indicate distinct degrees while colours indicate distinct, repeating neighbourhood degree sequences. On the right, nodes are randomly assigned different numbers of connections. These connections are made to the closest nodes, maintaining a geometric nature, but now we see the diversity of structure this opens up. Again, different shapes indicate distinct degrees, but now there are many unique neighbourhood degree sequences which remain colourless. This diversity reflects a higher hierarchical complexity.

https://doi.org/10.1371/journal.pcsy.0000026.g001

Then, for all ℓ nodes of degree k, we can stack their NDSs into an $\ell \times k$ matrix:

$$\mathbf{S}_k(G) = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1k} \\ s_{21} & s_{22} & \cdots & s_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ s_{l1} & s_{l2} & \cdots & s_{lk} \end{bmatrix} \tag{2}$$

The original definition of hierarchical complexity for degree k takes the variance over the columns of this matrix and then averages across columns:

$$R_k(G) = \frac{\sum_{j=1}^{k} \sigma_j^2}{k} \tag{3}$$

where σ_j^2 is the variance of the jth column of $\mathbf{S}_k(G)$. The global measure is then the average of this over degrees:

$$R(G) = \frac{1}{|\mathcal{D}_2|} \sum_{k \in \mathcal{D}_2} R_k(G). \tag{4}$$

where $\mathcal{D}_2$ is the set of degrees in the graph taken by at least two nodes (since variance is only meaningful over at least two elements).

Normalised hierarchical complexity

It is clear that the above definition of hierarchical complexity depends on network size. Networks of larger size have greater potential for larger degrees, which will influence the variances within Eq (3). Indeed, the maximum variance of numbers in $[1, n - 1]$ can occur with the sample $\{1, n - 1\}$ (or any equal number of 1s and $n - 1$s) which has variance $((n - 2)/2)^2$. Furthermore, computations have demonstrated that hierarchical complexity also correlates with number of edges.

Lack of normalisation to number of nodes and/or density is not unusual in network science. Indeed, the latter is essentially common place– consider two of the most widely considered network metrics, the global clustering coefficient and global efficiency which are both maximised in complete graphs. It is then of note that we can propose the following as a normalised measure of hierarchical complexity to both n and d.

Definition 1. *Let G be a network, we define the k-th normalised hierarchical complexity as*:

$$\hat{R}_k(G) = \begin{cases} \dfrac{\sum_{j=1}^{k} \sigma_j}{(1-d)m}, & \textit{if } m \neq 0 \text{ and } \mathrm{d} \neq 1 \\ 0, & \textit{if } m = 0 \textit{ or } d = 1 \end{cases} \tag{5}$$

where d is the density of G, m is the number of edges in G, and σ_j is the standard deviation of the jth column of $\mathbf{S}_k(G)$, that is, the matrix where each row is the ordered degree sequence of a node of degree k. Thus we propose the normalised global measure as

$$\hat{R}(G) = \frac{1}{|\mathcal{D}_2|} \sum_{k \in \mathcal{D}_2} \hat{R}_k(G). \tag{6}$$

To justify this normalisation, we note that the normalised hierarchical complexity is bounded.

Theorem 1. *The normalised hierarchical complexity is bounded above by 2, that is,* $\hat{R}(G) \leq 2$ *for every graph G.*

Proof. Consider $\hat{R}_k$. Let d_j be the difference of the max and min elements of the j-th column of $\mathbf{S}_k(G)$, so $\sigma_j \leq \frac{d_j}{2}$. Let $\alpha = \sum_{j=1}^{k} d_j$.

Now, let x_j be the maximum degree of the j'th column of $\mathbf{S}_k(G)$. We claim that $2m \geq \sum_{j=1}^{k} x_j$. If $x_1, \ldots, x_k$ are degrees of distinct vertices, this follows immediately, but it could occur that x_i and x_j correspond to the same vertex. Note that $x_1 \leq x_2 \leq \ldots \leq x_k$, since the rows are ordered in increasing order. Also note that if $x_i = x_j$, with $i > j$, and x_j is in row r, then the i-th entry in row r must also equal x_i, since it cannot be smaller than something to it's left x_j, and cannot be bigger than the max of it's column x_i. And since the degrees in each row correspond to distinct vertices (as they are given by the neighourhood), if $x_i = x_j$ then there are at least two vertices of degree x_i. And generalising this, if x_i appears as the degree of the same vertex ℓ times, then we can find ℓ distinct vertices of the same degree, so $2m \geq \sum_{j=1}^{k} x_j$. Also note that $x_j \geq d_j$, for all j, therefore

$$2m \geq \sum_{j=1}^{k} x_j \geq \sum_{j=1}^{k} d_j = \alpha.$$

Considering the minimal elements of each column and applying a similar argument to above we can also deduce that $2m' \geq \alpha$, where $m' = \frac{n(n-1)}{2} - m$ is the number of non-edges. Therefore, $\min(m, m') \geq \frac{\alpha}{2}$. Moreover, $\max(m, m') \geq \frac{n(n-1)}{4}$, since either at least half the edges are there or half are not.

So we can bound $(1 - d)m$ below by:

$$\begin{aligned} (1-d)m &= \frac{2mm'}{n(n-1)} = \frac{2\max(m, m')\min(m, m')}{n(n-1)} \\ &\geq \frac{\left(\frac{n(n-1)}{4}\right)\alpha)}{n(n-1)} = \frac{\alpha}{4} \end{aligned}$$

Therefore,

$$\hat{R}_k(G) = \frac{\sum_{j=1}^{k} \sigma_j}{(1-d)m} \leq \frac{\frac{\alpha}{2}}{(1-d)m} \leq \frac{\frac{\alpha}{2}}{\frac{\alpha}{4}} = 2$$

So $\hat{R}_k(G)$ is bounded above by 2 for all k, and $\hat{R}(G)$ is the average of values bounded above by 2, hence $\hat{R}(G) \leq 2$.□

There are several things to note about the formula (5). Firstly, instead of variance across neighbourhood degree sequences, we here opt for standard deviation. The distribution of the standard deviation over multiple samples will in general be more symmetric than that of the variance which will be right-skewed and standard deviation is generally a more appropriate measure when normalising.

Secondly, the division by $(1 - d)m$ is the term which acts to normalise the measure. This term was borrowed from the normalisation of degree variance [21]. There, it was shown to normalise degree variance and bound it below 1 for all graphs.

Thirdly, instead of taking the mean over the standard deviations this normalisation takes the sum. In actuality we can consider this as a multiplication of the mean by the degree of the neighbourhood degree sequences k (cancelling out the k on the denominator of the average). This effectively takes account of the sampling error of taking the mean over the σ_j's. The sampling error of the mean over k samples is:

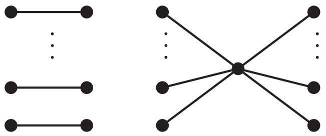
$$\frac{\sigma}{\sqrt{k}}, \tag{7}$$

where σ here is the standard deviation of the k element-wise standard deviations, so that the accuracy of the mean depends on the degree k.

Note that in the normalisation presented we do not multiply by $\sqrt{k}$ to standardise these measurements, but by k itself. This is because it is also linked with the division by m.

We do not believe the bound of 2 given in Theorem 1 is tight. Through considered construction of a disconnected graph which exploits variances of 1 degree nodes, we find the largest value for $\hat{R}$ tends to $\frac{1}{3}$, which occurs with the following graph family:

Example 1. *Consider the graph*:



where each column of nodes consists of $\frac{n-1}{4}$ *nodes.*

Note that $R_1 = \frac{n-3}{4}$, *no other degree contributes to R*, $m = \frac{3(n-1)}{4}$ *and* $d = \frac{3}{2n}$. *So*

$$\hat{R} = \hat{R}_1 = \frac{\frac{n-3}{4}}{\left(1 - \frac{3}{2n}\right)\frac{3(n-1)}{4}} \to \frac{1}{3}, \qquad \text{as } n \to \infty$$

We believe that the above family of graphs gives the largest value for $\hat{R}$, but we leave this as a conjecture:

Conjecture 1. *For any graph G we have* $\hat{R}(G) < \frac{1}{3}$.

While we in no way claim the above family of graphs is statistically or otherwise complex, the intended application of the measure is for connected graphs with many different degrees making a contribution. We can ensure this edge case goes to zero by using the corrective term for multiply-ordered degrees as described in [15], but for most intended purposes this is not necessary.

Expected complexity values

Next we show that $\hat{R}$ satisfies the conditions of a statistical complexity measure of being 0 in the thermodynamic limit for the Erdös-Rényi random graph. It is known (see [10]) that if a graph is regular, that is, every node has the same degree, then the hierarchical complexity is 0–since we have n nodes of degree k all with neighbourhood degree sequences $\{k, k, \ldots, k\}$.

We can also show that for the other end of the entropic spectrum, Erdös-Rényi random graphs, that the hierarchical complexity tends to 0, as n tends to infinity. To do so we need a formula for the standard deviation of the i-th largest sample from this distribution, this is known as the i-th order statistic, see [22] for background on order statistics. We can use known results on order statistics to derive the following:

Theorem 2. *Fix* $d \in [0, 1]$. *The normalised hierarchical complexity of an Erdös-Rényi graph* $\mathrm{ER}(n, d)$ *tends to* 0 *as n tends to infinity, that is,*

$$\lim_{n\to\infty} \hat{R}(\mathrm{ER}(n, d)) = 0.$$

Proof. For brevity let $\hat{R}_{n,d} \coloneqq \hat{R}(\mathrm{ER}(n, d))$. Note that if $d = 0$ or $d = 1$, then $\mathrm{ER}(n, d)$ is regular, thus $\hat{R}_{n,d} = 0$ for all n, and the result holds. So assume $d \in (0, 1)$. We begin by giving an approximation of $\hat{R}_{n,d}$.

First note that the node degrees of an Erdös-Rényi graph are sampled from the binomial distribution $B(n - 1, d)$. Whilst this sampling is not strictly independent, the dependence is very weak, and the correlation tends to zero as n tends to infinity [23], and thus can be disregarded in our approximation and asymptotic analysis.

In [24] an approximate formula for the standard deviation of the i-th order statistic of k samples for a continuous distribution with PDF $\phi(x)$ and CDF $\Phi(X)$ is given by:

$$\sigma_i \approx \frac{1}{\phi\left(\Phi^{-1}\left(\frac{i}{k+1}\right)\right)} \sqrt{\frac{i(k-i+1)}{(k+1)^2(k+2)}}. \tag{8}$$

By [25, Equation (1.1)] this approximation also holds for discrete distributions.

So to get the k'th normalised hierarchical complexity of $\mathrm{ER}(n, d)$, we let X be the binomial distribution $B(n - 1, d)$, sum (8) across $i = 1, \ldots, k$ and divide by our normalisation, which gives:

$$\begin{aligned}\hat{R}_k(\mathrm{ER}(n, d)) &\approx \frac{\sum_{i=1}^{k} \frac{1}{\phi\left(\Phi^{-1}\left(\frac{i}{k+1}\right)\right)} \sqrt{\frac{i(k-i+1)}{(k+1)^2(k+2)}}}{(1-d)m} \\ &= \frac{2\sum_{i=1}^{k} \frac{\sqrt{i(k-i+1)}}{\phi\left(\Phi^{-1}\left(\frac{i}{k+1}\right)\right)}}{d(1-d)n(n-1)(k+1)\sqrt{k+2}}.\end{aligned}$$

The second equality is because the expected number of edges is $m = \frac{n(n-1)d}{2}$.

We can approximate the expected smallest and largest degree in $\mathrm{ER}(n, d)$, using [Equation 4.5.1] [22] which gives an approximation for the i'th order statistic as $\Phi^{-1}\left(\frac{i}{n+1}\right)$, for sufficiently large n. This gives our lower and upper summands a and b, and the global formula is given by averaging $\hat{R}_k$ between a and b.

So for large n the global hierarchical complexity can be approximated by

$$\hat{R}(\mathrm{ER}(n,d)) \approx \frac{2\sum_{k=a}^{b} \frac{\sum_{i=1}^{k} \frac{\sqrt{i(k-i+1)}}{\phi\left(\Phi^{-1}\left(\frac{i}{k+1}\right)\right)}}{(k+1)\sqrt{k+2}}}{(b-a)d(1-d)(n-1)n}, \tag{9}$$

where ϕ and Φ are the PMF and CDF, respectively, of the binomial distribution $B(n-1, d)$, and $a = \lfloor \Phi^{-1}\left(\frac{1}{n}\right) \rfloor$ and $b = \lceil \Phi^{-1}\left(\frac{n-1}{n}\right) \rceil$. Note that this approximation is not particularly close, particularly for small n, but it is sufficient to consider the limit of the complexity as n grows.

The binomial distribution can be approximated by the normal distribution, for which the quantile function Φ^{-1} is known in terms of the inverse error function. Combining this with an approximation of the inverse error function [26, Equation 13] we get:

$$\begin{aligned} \Phi^{-1}(x) &\approx (n-1)d + \sqrt{2(n-1)d(1-d)}\ \mathrm{erf}^{-1}(2x-1) \\ &\approx (n-1)d + \sqrt{-\ln\left(4x(1-x)\right)2(n-1)d(1-d)} \end{aligned} \tag{10}$$

Combining Eq (10) with the De Moivre-Laplace approximation of the binomial PMF we get:

$$\begin{aligned} \phi(\Phi^{-1}(x)) &\approx \frac{\exp\left(-\frac{(\Phi^{-1}(x)-(n-1)d)^2}{2(n-1)d(1-d)}\right)}{\sqrt{2\pi(n-1)d(1-d)}} \\ &\approx \frac{\exp\left(-\frac{((n-1)d+\sqrt{-\ln\left(4x(1-x)\right)2(n-1)d(1-d)}-(n-1)d)^2}{2(n-1)d(1-d)}\right)}{\sqrt{2\pi(n-1)d(1-d)}} \\ &= \frac{\exp(\ln(4x(1-x)))}{\sqrt{2\pi(n-1)d(1-d)}} = \frac{4x(1-x)}{\sqrt{2\pi(n-1)d(1-d)}} \end{aligned} \tag{11}$$

We can then use Eq (11) to show that asymptotically $\hat{R}_{n,d}$ is bounded above by zero:

$$\begin{aligned} \hat{R}_{n,d} &\approx \frac{2}{(b-a)d(1-d)n(n-1)} \sum_{k=a}^{b} \frac{\sum_{i=1}^{k} \frac{\sqrt{i(k-i+1)}}{\phi\left(\Phi^{-1}\left(\frac{i}{k+1}\right)\right)}}{(k+1)\sqrt{k+2}} \\ &\approx \sum_{k=a}^{b} \frac{2\sum_{i=1}^{k} \frac{\sqrt{2\pi(n-1)d(1-d)}\sqrt{i(k-i+1)}}{4\frac{i}{k+1}\left(1-\frac{i}{k+1}\right)}}{(b-a)d(1-d)n(n-1)(k+1)\sqrt{k+2}} \\ &= \sum_{k=a}^{b} \frac{\sqrt{\pi}\sum_{i=1}^{k} \frac{k+1}{\sqrt{i(k-i+1)}}}{n(b-a)\sqrt{2d(1-d)(n-1)(k+2)}} \\ &= \frac{\sqrt{\pi}\sum_{k=a}^{b}\sum_{i=1}^{k} \frac{k+1}{\sqrt{i(k-i+1)(k+2)}}}{n(b-a)\sqrt{2d(1-d)(n-1)}} \\ &\leq \frac{\sqrt{\pi}\sum_{k=a}^{b}\sum_{i=1}^{k} 1}{n(b-a)\sqrt{2d(1-d)(n-1)}} \\ &\leq \frac{\sqrt{\pi}(b-a)b}{n(b-a)\sqrt{2d(1-d)(n-1)}} \underset{n\to\infty}{\to} 0 \end{aligned} \tag{12}$$

The final step follows since d is fixed and b is always smaller than n, thus the denominator grows at least $\sqrt{n-1}$ faster than the numerator.□

We conjecture that a similar result holds for random geometric graph (RGG) (see Section for the definition of RGG's) with a fixed average degree of b (note in this case, $d \to 0$ as $n \to \infty$). In particular, when we randomly position n nodes on the unitary Euclidean plane and connect two nodes whenever they are within radius $r = \sqrt{b/(n\pi)}$ of each other. This radius r is selected to ensure we obtain a graph with the required average degree. In this version of an RGG, the degree distribution is also the binomial distribution (see [27]), so we can apply a similar technique as used in the proof of Theorem 2. However, RGGs have non-trivial degree correlations violating statistical assumptions used in the Erdös-Rényi case [27]. Nevertheless, we conjecture that a different bound can be obtained that does tend to 0, we leave this as an open problem.

Conjecture 2. *Let* $RGG(n, r)$ *be the random geometric graph in the unitary plane with* n *nodes and radius* $r = \sqrt{\frac{b}{n\pi}}$, *for a fixed* $b \in \mathbb{N}$. *We have*

$$\lim_{n\to\infty} \hat{R}(RGG(n, r)) = 0.$$

We've seen that the formula for hierarchical complexity is closely linked to order statistics and the quantile function of probability distributions. Very few closed formulas exist for quantile functions, and they are known to be difficult to analyse. Due to this we are otherwise limited in our analytical treatment of this measure. The remainder of the paper shows the validity of this normalisation in application and we use it to derive novel insights from models and real data. Particularly, we pursue the hypothesis that statistical complexity arises naturally through a combination of hierarchical and geometric components to network connectivity.

Materials and methods

Network models

Our understanding of this normalisation is aided by its application to different network models and studying the behaviour of our normalisation of HC as we change the size and density of the network. In the following, a graph refers to a mathematical object of a set of nodes with adjoining edges. A network refers to a graph representation of the relationships or connections between components of a real-world complex system.

Firstly, we use **Erdös-Rényi (E-R) Random Graphs**. E-R random graphs are generated using random uniform edge probabilities in [0, 1] [28]. They give an indication of the behaviour of an 'average' graph of a given size and density. That being said, they do not give any indication of the behaviour of an 'average' network as it lacks many of the basic characteristics common to networks such as a relatively high clustering coefficient and degree heteregoneity.

Random Geometric Graphs (RGG) are generated from randomly sampled co-ordinates in the unit cube (i.e. 3D) [29]. These samples then represent nodes and the inverse distances between node pairs are the weights of the edges between them. For a desired network density, we select the m largest weights as our graph edges. RGGs have properties of high clustering desirable for networks, however they also lack the characteristic degree heterogeneity of networks.

Surface-Depth (S-D) models provide geometric graphs with heterogeneous degree distributions which show distinct similarities to many real-world networks [20]. We shall refer to these models throughout as **Random Hierarchical Geometric Graphs (RHGG)** since they combine a geometric component to connections with a hierarchical component, both of which we also want to study in isolation. These are generated using two parameters, the σ of a log-normal distribution and the number of dimensions, q, of a random geometric graph. The weights of the edges are then defined as

$$w_{ij} = d_{ij}(s_i + s_j), \tag{13}$$

where d_{ij} is the inverse distance between nodes i and j in a random geometric graph with q dimensions and each s_i is a random sample from a log-normal distribution $LN(\mu, \sigma)$. Again, for a desired network density, we select the m largest weights as our graph edges. In this study, we perform a basic exploration with $q = 3$, $\mu = 0$, and σ fixed at 0.2 as this produces graphs with a suitable heterogeneity. We then also explore the effect of varying σ, and so heterogeneity of geometric graphs, on hierarchical complexity.

Configuration models of these RHGGs allow us to probe the extent to which hierarchical complexity of the RHGGs can be attributed merely to the hierarchical structure of the network. We refer to these as **Random Heterogeneous Graphs (RHGs)** to emphasise the relationship with our other models. Briefly, configuration models work by fixing the degree distribution of a network but otherwise randomising the connections. Each node is provided with a number of stubs equal to its degree. These stubs are then randomly paired between nodes to establish edges [30].

Link growth mechanisms

We used ten different link growth mechanisms to observe the effect on NHC of increasing density in different ways on real-world networks.

For all node pairs without edges, $(i, j) \notin \mathcal{E}$ we considered:

Random growth. Edge probabilities are uniform.

Popularity growth. Similar in fashion to preferential attachment for addition of new nodes to a network, the probability of a new edge is proportional to the sum of the degrees of the nodes:

$$p_{ij} \sim k_i + k_j. \tag{14}$$

Similarity growth. For g_i, g_j the neighbourhoods of nodes i and j, the probability of edge (i, j) occurring is proportional to the Jaccard index of their neighbourhoods:

$$p_{ij} \sim J(g_i, g_j) = \frac{|g_i \cap g_j|}{|g_i \cup g_j|}. \tag{15}$$

Popularity × similarity growth. The probability of a new edge is proportional to the intersection of their neighbourhoods, essentially removing the size normalisation of the Jaccard index:

$$p_{ij} \sim |g_i \cap g_j| \tag{16}$$

so that the probability of connection is dependent on the size of the neighbourhoods and the overlap of the neighbourhoods. Note, popularity and similarity growth is what is commonly known as the common neighbours algorithm, as it just counts the number of shared neighbours two nodes have to predict whether they will become connected in the future [31, 32]. While simple at face value, we can clearly see from the above that this works well to account for both popularity and similarity components of a network.

For each of the three latter approaches (popularity, similarity, and popularity × similarity) we took three different approaches to deciding on links: random probabilistic selection, random exponentiated probabilistic selection, and deterministic rank-based selection.

Probabilistic. To get the probabilities, p_{ij}, we divide each individual measurement (e.g., $k_i + k_j$ for hierarchical attachement) by the sum over all available measurements for example,

$$p_{ij} = (k_i + k_j)/\sum_{(i,j)\notin\mathcal{E}}(k_i + k_j). \tag{17}$$

Edges are then randomly selected based on these probability spaces $\{p_{ij}\}_{(i,j)\notin\mathcal{E}}$.

Exponentiated probabilistic. This approach is taken to better differentiate between the strong and weak potential links in the probability space and so make it more likely for stronger potential links to be selected. Here, we simply take the exponentials of the probabilities before normalisation: $p_{ij} = \exp(k_i + k_j)/\sum_{(i,j)\notin\mathcal{E}}\exp(k_i + k_j)$.

Deterministic. Here, we simply take the top x probabilities as new links. This is typically how link prediction would be done.

These three approaches span from the more randomised, to the more rigid, with exponentiated probabilistic growth taking the middle ground.

Data

We obtained data for twenty large networks from two databases– the SNAP database [33] and the Network Repository [34]. While hierarchical complexity has been applied to several different types of networks (including social networks, protein networks and infrastructure networks) with mixed results compared to configuration models [15], it has yet to be applied to larger sized networks. Further, we have so far been unable to directly compare hierarchical complexity of networks of different sizes due to the lack of a normalisation.

Network were chosen to cover a wide range of sizes (1912–36692) and types (protein interaction networks, social networks, infrastructure networks, collaboration networks), and also to include groups of certain types of networks to explore relationships within and between network types. The number of nodes, edges and the network density for each network are shown in Table 1.

Protein-protein interaction networks generated from co-expression correlations were taken from the Network Repository, which in turn derived these graphs from data from wormnet [35]. These were obtained for Homo Sapiens (HS), Derio Rario (DR)– zebrafish, Drosophilia Melanogaster (DM)– fruit fly, and caenorabditis elegans (CE)– a nematode. All of these are exceptionally well studied, model species for which the data is most extensive and reliable.

Also from the Network repository we took two infrastructure networks, one being the widely studied network of the Western States power grid of the US (power grid) [36] and the other being a network of international flights between airports where nodes are airports and edges are established where there are flights between those airports (open flights) [37]. All other networks were obtained from the SNAP repository.

We studied five collaboration networks within Physics disciplines, constructed from arXiv data. For these, edges are established between co-authors of papers. Topics are self-selected by authors during arXiv manuscript uploads. These topics are astrophysics (collab AstroPh), condensed matter physics (collab CondMat), general relativity (collab GrQc), high energy physics (collab HepPh), and high energy physics theory (collab HepTh).

We studied six social networks constructed from the twitch platform for six different languages– English (twitch ENGBE), French (twitch FR), German (twitch DE), Portuguese

Table 1. Statistics for twenty real world networks.

Network	*n*	*m*	*d*
email Enron	36692	367662	0.0005
Facebook	22470	171002	0.0007
collab CondMat	22167	186936	0.0008
collab AstroPh	16000	396160	0.0031
protein CE	15229	245952	0.0021
collab HepPh	12008	237010	0.0033
collab HepTh	9877	51971	0.0011
twitch DE	9498	153138	0.0034
lastFM Asia	7624	27806	0.0010
twitch ENGBE	7126	35324	0.0014
twitch FR	6549	112666	0.0053
collab GrQc	5242	28980	0.0021
power grid	4941	6594	0.0005
twitch ES	4648	59382	0.0055
protein HS	4413	108818	0.0112
twitch RU	4385	37304	0.0039
protein DM	4040	76717	0.0094
protein DR	3289	84940	0.0157
open flights	2939	30501	0.0071
twitch PTBR	1912	31299	0.0171

https://doi.org/10.1371/journal.pcsy.0000026.t001

(twitch PTBR), Russian (twitch RU), and Spanish (twitch ES) [38]. Nodes are the users of twitch and edges are friendships between them. Facebook page-page is another social network where nodes represent official Facebook pages while edges are mutual likes between sites, collected through the Facebook Graph API in November 2017 and restricted to four categories of pages- politicians, governmental organisations, television shows and companies [38]. The LastFM social network is the network where nodes are users from Asian countries and edges are mutual follower relationships between them [39]. We also analysed a large email network between Enron employees (email Enron), originally made public by the Federal Energy Regulatory Commission during its investigation. Here, nodes are email addresses and edges are established wherever there are emails sent between addresses.

Allen brain model for use in large network experiment. For studying the effect of normalisation in very large graphs, we used the mouse V1 model from the Allen Institute for Brain Science [40]. This model contains approximately 230,000 neurons, and can be considered as a network where each neuron is a node and there is an edge between two nodes if they are connected by a synapse. We can sample geometric cylinders from this model using the associated code provided at [41], this allows us to construct networks that should be very similar structurally, but vary in size from anything up to 230,000 nodes. Due to computational limitations we only compute complexity values on up to 95,000 nodes.

Data for replication in link growth mechanism experiment. For studying the link growth mechanisms, we used a replication dataset from the ICON corpus consisting of 139 networks mostly describing biological, social and technological phenomena [15, 42]. These ranged from $n = 50$ to $n = 3155$ with a mean of 341 ± 462. Densities ranged from $d = 0.0011$ to $d = 0.3884$ with a mean of 0.0578 ± 0.0717.

Results

Normalisation results on random models

Fig 2 shows the results of the normalisation applied to our chosen random graph models. For each random model (Erdös-Rényi, RGG, RHG, and RHGG) we generated 100 realisations with $n \sim U[50, 10000]$ and $d \sim U[0, 1]$ and plot results against both n and d.

Notably, Erdös-Rényi random graphs have the lowest complexity of the models studied. Above this, the complexity of RGGs and RHGs is similar across all n and d. The greatest complexity is clearly observed in the RHGGs, particularly at lower densities, indicating that greater complexity arises naturally through the combination of hierarchical and geometric structure.

While Erdös-Rényi random graphs, RGGs and RHGs have similar levels of complexity across density (highlighting the normalisation features of our measure across density), there is a different behaviour noted in RHGGs. Higher complexity at low densities compared to high densities in this instance can be considered a structural feature present in the graphs. We can understand very high densities as regarding the connectivity of the least important connections (specifically, the complement of the graphs) which in this instance are those between nodes with low degrees which are geometrically distant. It is reasonable to expect complexity here to be as low as for RHGs and RGGs.

As per the theoretical results for E-R random graphs, the experiments across the different models shows an inverse relationship for $\hat{R}$ with n but little to no dependency on d. The reason for hierarchical complexity decreasing with increasing n may be a true relationship of the complexity of these models as n increases, rather than a normalisation issue. Indeed, we can expect that larger sample sizes of neighbourhood degree sequences given by larger random graphs would result in more homogeneous ordered sequences as they better approximate the global degree distributions.

Interestingly, for the RHGGs– which more accurately model real-world network structure– there is very little if any decrease with increasing n for $n > 1000$. This indicates that

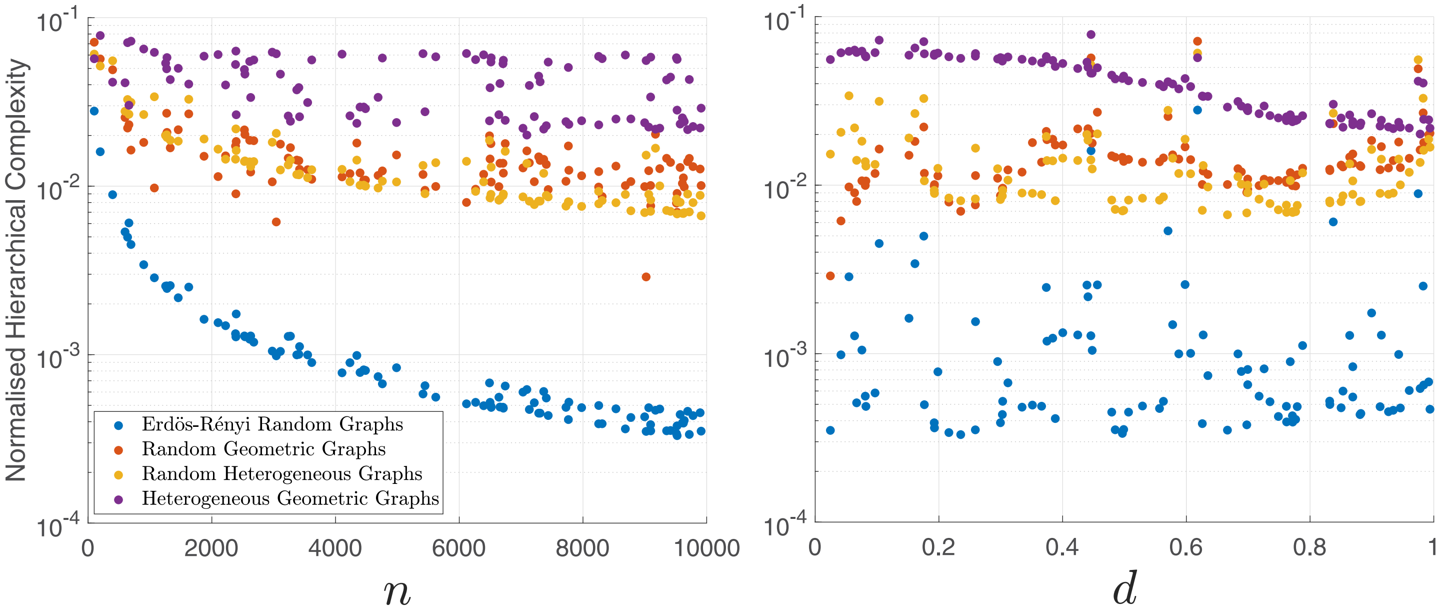


Fig 2. Results show measurements for random realisations ($n \sim U[50, 10000]$, $d \sim U[0, 1]$) of different random graphs as denoted in the legend, against size n and density d.

https://doi.org/10.1371/journal.pcsy.0000026.g002

comparisons of $\hat{R}$ in large networks are reliable, but caution must still be taken when making comparisons in smaller networks.

In the Supplementary Information section SI1 in S1 File, we explore the suitability of two relevant previously proposed network measures for assessing statistical complexity of networks using similar analyses as above. We find that neither meet the requirements expected of being 0 for regular and ER random graphs. We also find that they are not useful for distinguishing between different types of models which should be distinguishable in terms of statistical complexity.

Normalisation results on increasing number of nodes

To extend our observations of the normalisation with respect to network size we studied the behaviour of the normalisation in very large graphs. In this case we use the mouse V1 model from the Allen Institute for Brain Science, see [40], and Erdös-Rényi random graphs, for which we know the normalised complexity value tends to 0 by Theorem 2.

We would expect networks of the same structure to have the same, or at least similar, NHC. However, what do we mean by the "same structure"? If two networks have a different number of nodes they inherently have a different structure. In fact, the size of the network is related to the complexity, because as the size of random graphs increase their uniformity increases, this is due to the inverse relationship between variance and the sample mean. As such, we would expect that the NHC will decrease slightly as the number of nodes increases, but two sufficiently large networks of similar structure, but different size, would have similar complexity values.

The effect of this is well demonstrated in Fig 3. We see that the Erdös-Rényi random graph tends to 0 with increasing n, as expected. At the same time, for the Allen Brain for small n the complexity is higher, but for larger n, roughly $n \geq 2000$, the complexity is very close between samples, appearing to tend to a non-zero limit in n. This demonstrates a behaviour expected of statistical complexity in dynamical systems, that randomness (aswell as regularity) vanishes to 0 complexity in the limit of n, while non-zero complexity is maintained in diverse structure [5].

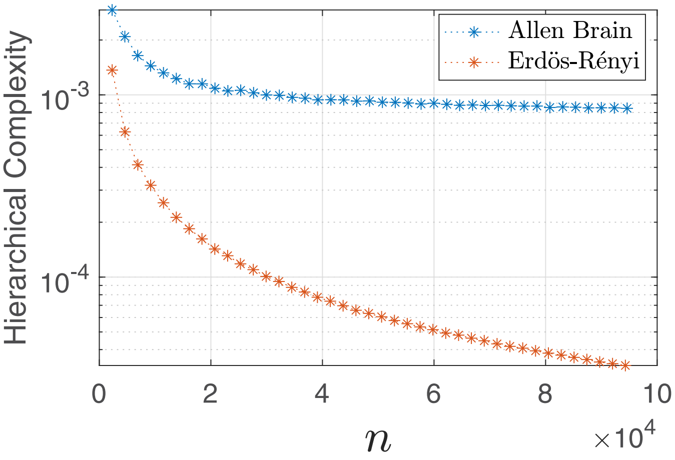


Fig 3. The normalised hierarchical complexity of cylinders of increasing sizes of the Allen Brain V1 mouse model vs ER graphs of the same size and density.

https://doi.org/10.1371/journal.pcsy.0000026.g003

Effect of degree heterogeneity on hierarchical complexity

Here we study the change in complexity as we increase heterogeneity among the RHGGs. We generated RHGGs of size $n = 1000$ and $d \sim U[0, 1]$. The heterogeneity was determined with 100 realisations of the model for each $\sigma_h = 0.01, 0.02, \ldots, 2$. Results are shown in Fig 4.

The results show that NHC has a consistent range of values across most degrees for heterogeneous geometric graphs, tapering off quickly at either end towards 1 and n. It is beneficial that the measure does not have a positive or negative relationship with degree so that the measure does not overly emphasise any particular range of degrees. Further, the fact the measure quickly tapers off towards very small degrees protects the measure from being influenced by the high levels of uncertainty of sampling at these small sample sizes.

NHC is generally strongest in models with σ_h between 0.2 and 0.4 and decreases towards low and high heterogeneity. When σ_h is high the network becomes dominated by the hierarchical relationships, which should make the network more ordered. On the other hand, with low σ_h the model gets closer to a random geometric graph which has low values of NHC as shown in Fig 2. This is consistent with our expectations of NHC being a statistical measure of complexity in networks.

In section SI2 in S1 File of the Supplementary Information, we apply NHC to the non-uniform Popularity Similarity Optimisation model [43] to see how it fairs on another model which directly utilises degree heterogeneity and latent geometry. We find similar patterns as for RHGGs, that the model achieves highest NHC in a middle ground of degree heterogeneity.

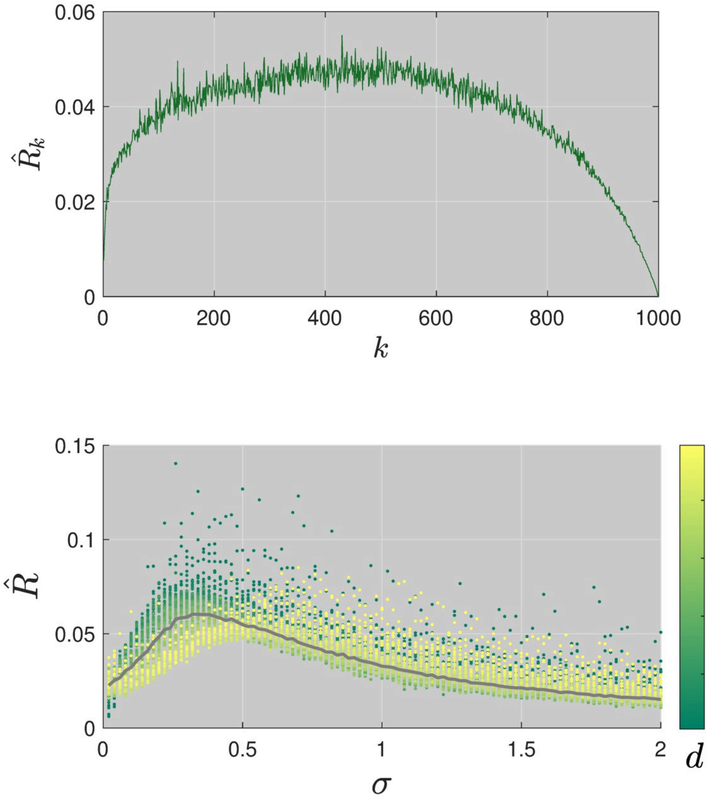


Fig 4. Top, average normalised hierarchical complexity per degree of heterogeneous geometric graphs with with $n = 1000$, $d \sim U[0, 1]$, and $\sigma_h = 0.01, 0.02, \ldots, 2$. One hundred realisations are created for each σ. Bottom, global normalised hierarchical complexity plotted against heterogeneity of these networks.

https://doi.org/10.1371/journal.pcsy.0000026.g004

We also explore the effect of varying the clustering in the network and the number of communities through additional parameters of the model. We find that decreasing the clustering by randomising more connections decreases NHC, and that introduction of strong community blocks in the geometry also decreases NHC. These are broadly consistent with our expectations of a statistical complexity metric.

Hierarchical complexity in large real-world networks

We compared unnormalised and NHC in the twenty large networks detailed in Table 1. The results are shown in Table 2. The values of the non-normalised HC range from 4 in the power grid network to 87858 in the German twitch network, a difference by a factor of nearly 22000. This range highlights the lack of utility of an unnormalised and unbounded measure for comparisons of different networks. On the other hand, the values of NHC range from 0.0022 in the power grid network up to 0.0830 in the Portugese twitch network, a difference by a factor of just under 38.

We can see that the normalised measure allows us to compare between these networks more clearly. For example, consider the twitch DE, twitch FR and twitch RU networks, which have the largest non-normalised HC values. These three networks are all constructed in the same way (friendships in twitch) and have similar densities (0.0034,0.0053, and 0.0039, respectively), and yet the twitch DE non-normalised HC value is 6 times bigger than the twitch RU non-normalised HC, suggesting that the twitch DE network is significantly more complex than the twitch RU network. However, when normalised these three networks all have very similar NHC values, suggesting they have very similar levels of complexity, as one would expect.

Table 2. Rankings of hierarchical complexity and normalised hierarchical complexity for twenty real world networks.

R	Network	$\hat{R}$	Network
87858	twitch DE	0.0830	twitch PTBR
35996	twitch FR	0.0684	twitch RU
14328	twitch RU	0.0612	twitch DE
11099	email Enron	0.0586	twitch FR
7625	twitch ES	0.0526	open flights
7167	twitch PTBR	0.0454	twitch ES
4216	twitch ENGBE	0.0360	protein DR
3275	facebook	0.0334	protein HS
2207	collab HepPh	0.0297	email Enron
1283	protein HS	0.0279	protein DM
1015	protein DR	0.0272	twitch ENGBE
873	collab AstroPh	0.0258	collab HepPh
755	protein CE	0.0182	facebook
752	protein DM	0.0154	collab GrQc
582	open flights	0.0147	lastFM Asia
293	LastFM Asia	0.0124	protein CE
261	collab CondMat	0.0105	collab AstroPh
68	collab GrQc	0.0056	collab HepTh
39	collab HepTh	0.0054	collab CondMat
4	power grid	0.0022	power grid

https://doi.org/10.1371/journal.pcsy.0000026.t002

Next consider the other three twitch networks: ES, PTBR and ENGBE. We can see that twitch ES and twitch ENGBE have very similar unnormalised HC values, but when normalised the twitch PTBR is twice as complex as twitch ES. This is not surprising considering they have markedly different densities, and we will see in the next section that even after normalisation the network density is correlated to the NHC value in real world networks. We also see that the twitch ENGBE has much lower complexity than the other twitch networks, which is likely explained by the lower density of the network.

So we can see that our normalisation reduces the scale of the difference between complexities of networks, and allow us to better compare networks of different sizes. However, we also see that the density of the network still correlates with the NHC value.

Growth of hierarchical complexity in real world networks

We found that hierarchical complexity was positively correlated with network density in the twenty networks from Table 1 ($\rho = 0.7203$, $p = 0.0005$), Fig 5, left. We confirmed this observation with a network dataset of 139 smaller networks ($n \in [50, 3155]$, $\rho = 0.5325$, $p = 1.5 \times 10^{-11}$) [15, 42], Fig 5, right. To discount the potential confounding effect of network size, n, on these correlations, we implemented linear regression on network density with network size as a predictor and found that the correlations of the residuals of the regression with NHC were still significant in both cases– $\rho = 0.6526$, $p = 0.0023$ for the 20 large networks and $\rho = 0.2294$, $p = 0.0066$ for the 139 small-to-medium sized networks.

At the same time, we have shown that NHC shows strong normalisation with respect to density for many types of graph. The relationship between density and NHC in real world networks is therefore unlikely due to a lack of normalisation, but is a true relationship requiring a mechanistic explanation. To try to explain this relationship we applied ten link growth algorithms, as described in section III.B, to real-world networks to artificially increase their density and see if any would consistently lead to the targeted increase in NHC.

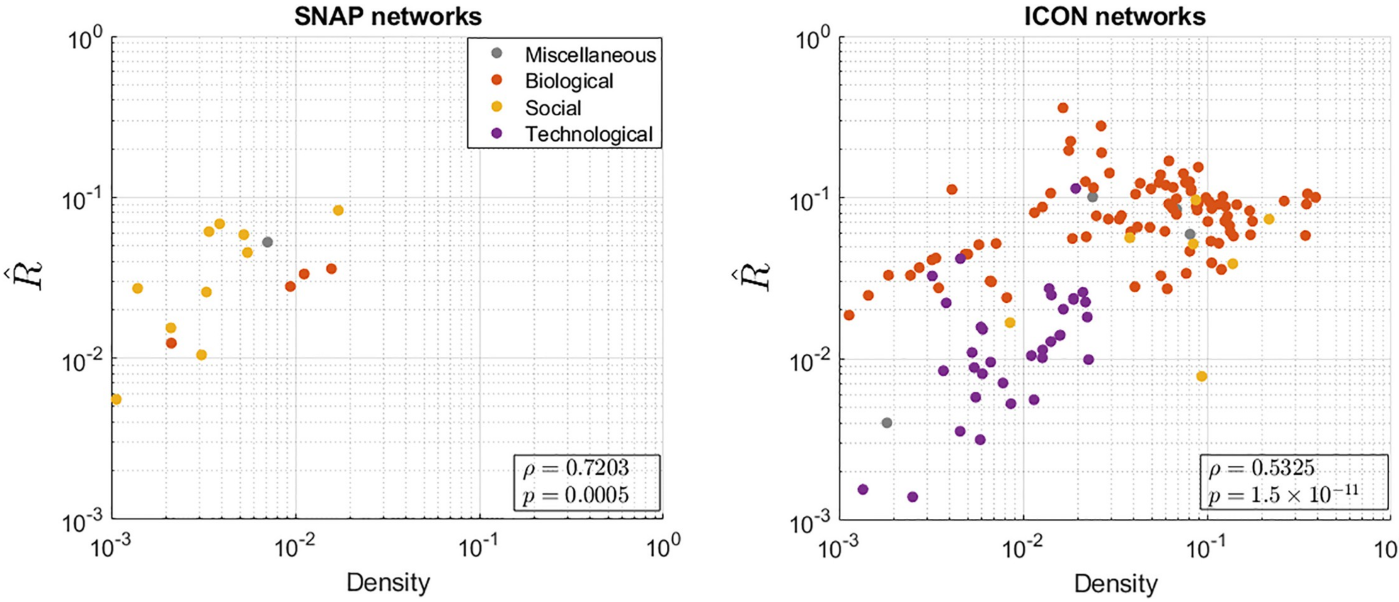


Fig 5. Scatterplots visualising the positive association between density and hierarchical complexity in real-world networks. Spearman's correlation coefficient and associated p-value shown inset. Bottom row shows average results of the values of NHC as we increase density of networks according to the link growth mechanisms as described in the legend.

https://doi.org/10.1371/journal.pcsy.0000026.g005

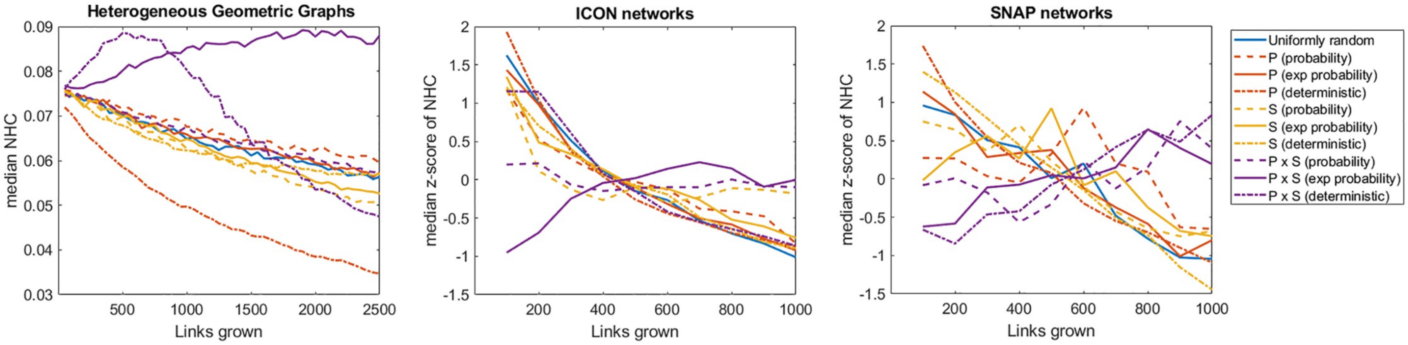


Fig 6. Average results of the values of NHC as we increase density of networks according to the link growth mechanisms as described in the legend. P stands for Popularity, S for Similarity and P × S for Popularity × Similarity. For the Heterogeneous geometric graphs we take the median over 100 iterations. For the ICON and SNAP networks we take the median of the z-scores of trajectories to avoid bias of networks with high NHC.

https://doi.org/10.1371/journal.pcsy.0000026.g006

Firstly, we implemented link growth on 100 HRGGs of size $n = 250$, density $d = 0.05$ and density $\sigma = 0.2$. Each link growth algorithm in Section III.B was applied for 50 iterations with 50 new links established each iteration. For the real-world networks, we applied the link growth algorithms to all of the 139 ICON networks and 17 of the 20 SNAP networks (omitting the 3 largest due to computational limitations). For these we used 10 iterations with 100 new links established each iteration. Results are shown in Fig 6.

There are two clear observations to be made from these results. Firstly, the only kind of link growth which provided increased NHC were those based on the common neighbours (here referred to as Popularity × Similarity) algorithm. This shows a clear favoured strategy for link growth in complex networks.

All of the random, Popularity only and Similarity only algorithms worked to decrease NHC and in similar amounts as the baseline uniformly random algorithm. There was no particular observable consistent difference between the Popularity and Similarity only mechanisms. However, the deterministic mechanisms typically did worse most of the time. Infact, in the HRGGs the Popularity only deterministic mechanism did much worse than the uniformly random approach. Anytime the algorithm showed a greater decrease than the uniformly random mechanism we can assume this is caused by a more ordered structure being enforced on the network, since the only NHC smaller than ER random graphs are regular graphs and highly organised graphs where all nodes of the same degree have the same or similar neighbourhood degree sequences [15].

Secondly, the only algorithm which consistently increased the NHC across the three datasets was the exponentiated probabilistic Popularity × Similarity algorithm. This algorithm takes the middle ground between the more random standard probabilistic algorithm and the deterministic link ranking algorithm. The probabilistic mechanism increased for the SNAP networks over 10 iterations, but decreased slightly in the ICON networks and moreso in the RHGGS with a similar trajectory as for uniformly random growth. The deterministic mechanism intially increased alot for the RHGG model before going into a steep decline, while it increased in the SNAP networks but decreased in line with the uniformly random mechanism in the ICON networks.

Discussion

Our modelling demonstrated greater statistical complexity arising through the combination of hierarchical and geometric components. While the space of random geometric graphs is regular, the nodes being placed randomly in that space opens up pockets of higher and lower connectivity, as quantified through degree correlations [27]. This may be understood as a classic instance where randomness and regularity interact to generate some degree of complexity. Similarly, heterogeneous random graphs contain an ordered structure in terms of the hierarchy of node degrees, combined with randomness of connections established through the configuration model procedure. By giving a randomly allocated log-normal node fitness to the nodes randomly placed in Euclidean space we see an amplification in terms of complexity. This is particularly interesting since heterogeneous geometric networks closely model many aspects of real-world networks [20]. Future work will explore whether access to heterogeneous connectivity patterns in networks facilitates the formation of more heterogeneous functionality and therefore assess the advantages conferred by the combination of hierarchy and geometry which real-world networks appear to incorporate almost universally, as found in e.g. [11] (although there are a subset of networks which buck this trend [32]).

In the results of NHC among the 20 large real-world networks some patterns appear. The general trends among the three groups of networks we have (6 twitch networks, 4 protein networks, 5 physics collaboration networks) indicate that the twitch networks tend to have highest complexity, while protein networks have fairly high complexity and physics collaboration networks have low complexity. The power grid having the lowest complexity can be expected as it is a network with high geometrical constraints and we might expect some specific universal design principles in its construction. On the other hand, online social networks of twitch are largely free from geometrical constraints (although will still have a latent similarity space, but this can be of arbitrarily large dimension) and may reflect the diversity of social relationships. However, sample sizes would need to be increased to provide stronger evidence for any such generalisations.

The results from the link growth mechanism experiments showed that explanations for the positive relationship between density and NHC were not given by link growth mechanisms for popularity or similarity of nodes separately, but again the combination of the two. It indicates that growth of statistical complexity requires a trade-off of randomness and determinism. Too much randomness and no structure is developed. Too little randomness and the structure becomes too rigid. It also highlights how real-world networks may naturally grow to develop higher statistical complexity simply through nodes more likely to (but not with certainty) form links with nodes with a lot of common neighbours. It also highlights some amount of futility in trying to ever perfectly predict links– there is randomness in connectivity, and in fact networks may well benefit from that randomness in breaking rigidity of patterns and becoming more diverse.

The evidence that random geometric graphs have non-zero NHC in the thermodynamic limit gives an interesting insight into the geometrical nature of NHC. It tentatively points towards a definitive notion of statistical complexity of networks. Essentially, if we take seriously the established notion of networks as embedded in a latent geometrical space [19, 44], then we can start to conceptualise measures of statistical complexity of networks such as NHC as attempting to measure the irregularity of the distribution of points over that space.

It is worth noting that in all of the NHC values we computed, we rarely find anything above 0.1. This means, we expect 0.1 is a very high value of NHC for a network. Here, we recall that a normalisation does not require values to be in any particular range, just that the values are comparable for the same phenomena, for example model with all other parameters fixed, with changes to the targeted normalisation parameter. Further, while we provide an upper bound of 2 to demonstrate there is no possible case of the measure exploding to infinity, we do not believe this is a tight upper bound. From our experience, it is likely a tight upper bound is even below 1.

In the future, we intend to thoroughly explore the application of NHC for analysis of brain networks, across scales, across different types of networks, and eventually, across species. We will also explore applications to protein-protein interaction networks utilising the vast datasets available through the STRING database in order to begin to answer questions regarding the relationship between NHC of protein-protein interactions and evolutionary parameters. Applications to social networks hold obvious appeal given the consistently high complexity we noticed in the twitch social networks. There is also serious scope for extension and improvement of NHC. For example, we could consider generalising the measure to considering neighbourhoods of neighbourhoods, up to an arbitrary depth. Generally speaking, our understanding of statistical complexity of networks in different fields is limited by a lack of parsimonious measurements for its quantification. The tools offered here can therefore help researchers to begin to answer fundamental questions regarding the existence and extent of statistical complexity of real-world networks, especially in light of the larger and higher quality datasets becoming evermore available.

Conclusion

We proposed and demonstrated the utility of a normalisation for hierarchical complexity–NHC. We proved that this measure is bounded above by 2 and tends to zero for Erdös-Rényi random graphs with increasing size. This is analogous to a defining characteristic required of a statistical complexity measure in dynamical systems. We then demonstrated that, while random graph models containing degree heterogeneity and geometry individually had lower complexity, the combined components of degree heterogeneity and geometry is enough to create NHC of a similar level to real-world networks. However, we then found that real-world networks displayed an association between NHC and density that could not be explained solely by our models. Instead, we could manage to explain this consistently with a common neighbours link growth algorithm with exponentiated probabilities. Particularly, this was more consistent than a deterministic weight ranking algorithm and the more random non-exponentiated probability algorithm. All other algorithms tried failed to increase complexity. We therefore posit that real-world networks have a preference for growth which increases complexity of the interacting system. We provide a parsimonious measure for statistical complexity of networks which is ready to be applied to answering questions and gathering new insights into the degrees of complexity in various fields such as neuroscience, protein biology and social networks.

Supporting information

S1 File. This file contains additional analyses of other metrics, demonstrating how they are not adequate measurements of statistical complexity, and another model, the uPSO model which shows consistent behaviour as our other models with respect to statistical complexity. (PDF)

Acknowledgments

The authors would like to thank Dr April S. Kleppe for her thoughtful feedback on our work.

Author Contributions

Conceptualization: Keith Malcolm Smith, Jason P. Smith.

Formal analysis: Keith Malcolm Smith, Jason P. Smith.

Investigation: Keith Malcolm Smith, Jason P. Smith.

Methodology: Keith Malcolm Smith, Jason P. Smith.

Visualization: Keith Malcolm Smith.

Writing – original draft: Keith Malcolm Smith, Jason P. Smith.

Writing – review & editing: Keith Malcolm Smith, Jason P. Smith.

References

1. Hartmanis J, Hopcroft JE. An overview of the theory of computational complexity. Journal of the ACM (JACM). 1971; 18(3):444–475. https://doi.org/10.1145/321650.321661
2. Kolmogorov AN. On tables of random numbers. Theoretical Computer Science. 1998; 207(2):387–395. https://doi.org/10.1016/S0304-3975(98)00075-9
3. Chaitin GJ, Arslanov A, Calude C. Program-size complexity computes the halting problem. Department of Computer Science, The University of Auckland, New Zealand; 1995.
4. Huberman BA, Hogg T. Complexity and Adaptation. Physica D: Nonlinear Phenomena. 1986; 22 (1):376–384. https://doi.org/10.1016/0167-2789(86)90308-1
5. Feldman DP, Crutchfield JP. Measures of statistical complexity: Why? Physics Letters A. 1998; 238 (4):244–252.
6. Morzy M, Kajdanowicz T, Kazienko P. On Measuring the Complexity of Networks: Kolmogorov Complexity versus Entropy. Complexity. 2017; p. 3250301.
7. Zenil H, Kiani HA, Tegnér J. A Review of Graph and Network Complexity from an Algorithmic Information Perspective. Entropy. 2018; 20(8):551. https://doi.org/10.3390/e20080551 PMID: 33265640
8. Emmert-Streib F, Dehmer M. Exploring Statistical and Population Aspects of Network Complexity. PLOS ONE. 2012; 7(5):1–17. https://doi.org/10.1371/journal.pone.0034523 PMID: 22590495
9. Wiedermann M, Donges JF, Kurths J, Donner RV. Mapping and discrimination of networks in the complexity-entropy plane. Phys Rev E. 2017; 96:042304. https://doi.org/10.1103/PhysRevE.96.042304 PMID: 29347608
10. Smith K, Escudero J. The complex hierarchical topology of EEG functional connectivity. Journal of Neuroscience Methods. 2017; 276:1–12. https://doi.org/10.1016/j.jneumeth.2016.11.003 PMID: 27856276
11. Smith KM, Bastin ME, Cox SR, Valdés Hernández MC, Wiseman S, Escudero J, et al. Hierarchical complexity of the adult human structural connectome. Neuroimage. 2019; 191:205–215. https://doi.org/10.1016/j.neuroimage.2019.02.028 PMID: 30772400
12. Blesa M, Galdi P, Cox SR, Sullivan G, Stoye DQ, Lamb GJ, et al. Hierarchical Complexity of the Macro-Scale Neonatal Brain. Cerebral Cortex. 2021; 31(4):2071–2084. https://doi.org/10.1093/cercor/bhaa345 PMID: 33280008
13. Valdés Hernández MC, Smith KM, Bastin ME, Amft EN, Ralston SH, Wardlaw JM, et al. Brain network reorganisation and spatial lesion distribution in systemic lupus erythematosus. Lupus. 2021; 30(2):285–298. https://doi.org/10.1177/0961203320979045 PMID: 33307988
14. Smith KM, Starr JM, Escudero J, Ibãnez A, Parra MA. Abnormal functional hierarchies of EEG networks in familial and sporadic prodromal Alzheimer's disease during visual short-term memory binding. Frontiers in Neuroimaging. 2022; 1:883968. https://doi.org/10.3389/fnimg.2022.883968 PMID: 37555153
15. Smith KM. On neighbourhood degree sequences of complex networks. Scientific Reports. 2019; 9:8340. https://doi.org/10.1038/s41598-019-44907-8 PMID: 31171806
16. Caldarelli G, Capocci A, De Los Rios P, Munoz MA. Scale-free networks from varying vertex intrinsic fitness. Physical Review Letters. 2002; 89:258702. https://doi.org/10.1103/PhysRevLett.89.258702 PMID: 12484927
17. Papadopoulos F, Kitsak M, Serrano M, Boguna M, Krioukov D. Popularity versus similarity in growing networks. Nature. 2012; 489:537–540. https://doi.org/10.1038/nature11459 PMID: 22972194
18. Hoff PD, Raferty AE, Handcock MS. Latent space approaches to social network analysis. Journal of the American Statistical Association. 2002; 97:1090–1098. https://doi.org/10.1198/016214502388618906
19. Smith AL, Asta DM, Calder CA. The Geometry of Continuous Latent Space Models for Network Data. Statistical Science. 2019; 34(3):428–453. https://doi.org/10.1214/19-sts702 PMID: 33235407
20. Smith KM. Explaining the emergence of complex networks through log-normal fitness in a Euclidean node similarity space. Scientific Reports. 2021; 11:1976. https://doi.org/10.1038/s41598-021-81547-3 PMID: 33479422

21. Smith KM, Escudero J. Normalised degree variance. Applied Network Science. 2020; 5:32. https://doi.org/10.1007/s41109-020-00273-3 PMID: 32626822

22. David HA, Nagaraja HN. Order statistics. John Wiley & Sons; 2004.

23. Alarfaj B, Taylor C, Bogachev L. The joint node degree distribution in the Erdös-Rényi network; 2023.

24. Baglivo JA. Mathematica laboratories for mathematical statistics: Emphasizing simulation and computer intensive methods. SIAM; 2005.

25. Nagaraja H. Order statistics from discrete distributions. Statistics. 1992; 23(3):189–216. https://doi.org/10.1080/02331889208802365

26. Strecok A. On the calculation of the inverse of the error function. Mathematics of Computation. 1968; 22 (101):144–158. https://doi.org/10.1090/S0025-5718-1968-0223070-2

27. A Antonioni MT. Degree Correlations in Random Geometric Graphs. Physical Review E. 2012; 86:037101. https://doi.org/10.1103/PhysRevE.86.037101 PMID: 23031054

28. Erdös P, Rényi A. On random graphs. Pubilcationes Mathematicae Debrecen. 1959; 6:290–297.

29. Dall J, Christensen M. Random geometric graphs. Physical Review E. 2002; 66:016121. https://doi.org/10.1103/PhysRevE.66.016121 PMID: 12241440

30. Maslov S, Sneppen K. Specificity and stability in topology of protein networks. Science. 2002; 296 (5569):910–913. https://doi.org/10.1126/science.1065103 PMID: 11988575

31. Newman MEJ. Clustering and preferential attachment in growing networks. Physical Review E. 2001; 64:025102(R). https://doi.org/10.1103/PhysRevE.64.025102 PMID: 11497639

32. Cannistraci C, Alanis-Lobato G, Ravasi T. From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks. Scientific Reports. 2013; 3:1613. https://doi.org/10.1038/srep01613 PMID: 23563395

33. Leskovec J, Krevl A. SNAP Datasets: Stanford Large Network Dataset Collection; 2014. http://snap.stanford.edu/data.

34. Rossi RA, Ahmed NK. The Network Data Repository with Interactive Graph Analytics and Visualization. In: AAAI; 2015. Available from: https://networkrepository.com.

35. Cho A, Shin J, Hwang S, Kim C, Shim H, Kim H, et al. WormNet v3: a network-assisted hypothesis-generating server for Caenorhabditis elegans. Nucleic acids research. 2014; 42(W1):W76–W82. https://doi.org/10.1093/nar/gku367 PMID: 24813450

36. Watts DJ, Strogatz SH. Collective dynamics of small-world networks. nature. 1998; 393(6684):440–442. https://doi.org/10.1038/30918 PMID: 9623998

37. Opsahl T. Why anchorage is not (that) important: Binary ties and sample selection; 2011. Available from: https://toreopsahl.com/2011/08/12/why-anchorage-is-not-that-important-binary-ties-and-sample-selection.

38. Rozemberczki B, Allen C, Sarkar R. Multi-scale Attributed Node Embedding; 2019.

39. Rozemberczki B, Sarkar R. Characteristic Functions on Graphs: Birds of a Feather, from Statistical Descriptors to Parametric Models. In: Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM'20). ACM; 2020. p. 1325–1334.

40. Billeh YN, Cai B, Gratiy SL, Dai K, Iyer R, Gouwens NW, et al. Systematic integration of structural and functional data into multi-scale models of mouse primary visual cortex. Neuron. 2020; 106(3):388–403. https://doi.org/10.1016/j.neuron.2020.01.040 PMID: 32142648

41. Models of the Mouse Primary Visual Cortex;. https://portal.brain-map.org/explore/models/mv1-all-layers.

42. Ghasemian A, Hosseinmardi H, Clauset A. Evaluating overfit and underfit in models of network community structure. IEEE Transactions on Knowledge and Data Engineering. 2019;.

43. Muscoloni A, Cannistraci CV. A nonuniform popularity-similarity optimization (nPSO) model to efficiently generate realistic complex networks with communities. New Journal of Physics. 2018; 20:052002. https://doi.org/10.1088/1367-2630/aac06f

44. Krioukov D, Papadopoulos F, Kitsak M, Vahdat A, Bogñá M. Hyperbolic geometry of complex networks. Physical Review E. 2010; 82:036106. https://doi.org/10.1103/PhysRevE.82.036106 PMID: 21230138