



OPEN

Explaining the emergence of complex networks through log-normal fitness in a Euclidean node similarity space

Keith Malcolm Smith

Networks of disparate phenomena—be it the global ecology, human social institutions, within the human brain, or in micro-scale protein interactions—exhibit broadly consistent architectural features. To explain this, we propose a new theory where link probability is modelled by a log-normal node fitness (surface) factor and a latent Euclidean space-embedded node similarity (depth) factor. Building on recurring trends in the literature, the theory asserts that links arise due to individualistic as well as dyadic information and that important dyadic information making up the so-called depth factor is obscured by this essentially non-dyadic information making up the surface factor. Modelling based on this theory considerably outperforms popular power-law fitness and hyperbolic geometry explanations across 110 networks. Importantly, the degree distributions of the model resemble power-laws at small densities and log-normal distributions at larger densities, posing a reconciliatory solution to the long-standing debate on the nature and existence of scale-free networks. Validating this theory, a surface factor inversion approach on an economic world city network and an fMRI connectome results in considerably more geometrically aligned nearest neighbour networks, as is hypothesised to be the case for the depth factor. This establishes new foundations from which to understand, analyse, deconstruct and interpret network phenomena.

Theories and models of the emergence of complex networks allow us to gather insights into their potential generative mechanisms^{1,2}. The seminal prototype of network models is the Erdős-Rényi (ER) random graph where all links have equal probability, p , of appearing in the graph. A realisation of this random graph is generated by assigning uniformly random values to all node pairs and substantiating the existence of those links whose values lie above the probability threshold, p ³. For a large enough number of nodes, each distinct graph topology (i.e. graph isomorphism class) has roughly equal probability of appearing from this model⁴. Yet, the topological characteristics of real-world networks substantially and consistently deviate from ER random graphs⁵, telling us that real-world networks occupy a relatively small and highly uncommon set of graph isomorphism classes. Subsequently, a proliferation of network models have been developed in attempts to understand or reproduce common real-world network characteristics.

We can broadly classify network models either as being generative or non-generative. Non-generative models such as configuration models^{5,6}, stochastic block models⁷, and complex hierarchy models⁸ attempt to target or emulate real-world network properties, focused on practical issues such as providing null models of specific network properties. Generative models, on the other hand, seek to derive complex network-like topologies from proposed generative mechanisms, the aim of which is to provide plausible physical explanations for the non-arbitrary topological features found in real-world networks from first principles. A popular branch of generative modelling derives from the theory of preferential attachment, where new nodes entering the network have greater probability of linking to nodes with greater numbers of existing links. Such mechanisms have been shown to generate scale-free degree distributions, which have been observed also in some real-world networks². It has also been shown that scale-free networks can instead develop from power-law node ‘intrinsic fitness’, where each node has a probability of forming connections according to a power-law distribution⁹.

There is public disagreement among network scientists about how common scale-free degree distributions really are in networks¹⁰. Recent work analysing what kinds of distributions best fit degree distributions from a corpus of hundreds of real-world networks suggested that power-law degree distributions accounted for less

¹Usher Institute, University of Edinburgh, Edinburgh, UK. ²Health Data Research UK, London, UK. email: k.smith@ed.ac.uk

than 5% of the corpus, while fitting log-normal distributions achieved equivalent or better results for 88%¹¹. This quickly generated counter-arguments from scale-free network proponents¹⁰. Foremost of which was a work stating that a broader classification of what constituted a scale-free network was required, namely that power-laws need only be present in the right-tail of the degree distribution, rather than the whole distribution (denoted as pure power-laws), for the network to be classified as scale-free¹². Indeed, it has been known for some time that pure power-law degree distributions are necessarily only found in sparse networks¹³.

One part of the current work demonstrates that the log-normal distribution may be the key to reconciling these viewpoints. First of all, we argue that distributions of abilities or tendencies, such as those proposed in the idea of intrinsic fitness, tend to be log-normal rather than power-law¹⁴. Secondly, the right tails of log-normal distributions approximate power-laws¹⁵, satisfying the previously mentioned more relaxed definition of scale-free¹². Thirdly, using modelling we seek to establish if log-normal fitness creates power-law degree distributions at sparse densities and log-normal degree distributions in more dense networks.

Another branch of generative models considers nodes existing in a latent space and connections occurring where those nodes are close together in the space. The idea that nodes which are similar to each other are more likely to form connections, otherwise described as homophily, is intuitively sensible. By extension, this has led to the theory that some latent space of node similarities underlies the development of network structure¹⁶. A prototype of this approach can be seen as the random geometric graph, where nodes are random samples of an n -dimensional Euclidean space and where links form between the closest samples¹⁷. This model has some relevant properties to real world networks such as a high modularity and clustering, but does not display the degree heterogeneity implicated by hub nodes typical of complex networks. Further to this, Serrano et al. proposed an elegant hyperbolic geometric model where nodes randomly sampled on the unit circle were attached geometrically with constraints for the expected degree distribution of the network^{18,19}. Utilising this model, it was then proposed that a trade-off of popularity and similarity was an alternative explanation of network evolution²⁰. Although this combination of ‘popularity’ and ‘similarity’ is an attractive proposition, and one that will be echoed in the theory of this paper, these works do not provide an explanation for how the degree distributions of complex networks themselves arise.

The literature suggests two major themes in explaining the emergence of complex networks: (1) heavy-tailed node fitness—an individual aspect describing general potentials of nodes for interactivity and (2) homophily—a pair-wise aspect describing the suitability of pairs of nodes for making links. These here are combined in a new theory, called surface-depth theory, which proposes to model link probability using factors of log-normal fitness (the surface factor) and node similarity embedded in a high dimensional Euclidean space (the depth factor). ‘Surface’ and ‘depth’ here are terms chosen to reflect superficial and meaningful information, respectively, in a dyadic (i.e. pairwise) sense. One of the overarching goals of network science is to capture dyadic phenomena, whereas this theory, building on previous literature, emphasises that the links gathered from real-world data depend not only on dyadic phenomena but also on individual properties of nodes (or node fitness). In this way, we emphasise that these individual properties—while certainly creating interesting and important structure—are not intrinsically dyadic and yet would seem to tend to dominate the network and obfuscate much of the interesting dyadic information. We rigorously test our theory against prevailing theories of power-law distributions and hyperbolic geometry across over 100 real world networks, showing that our theory significantly and consistently achieves much greater accuracy in emulating real world network topologies. We then describe an application of this theory for recovering the depth factor of weighted complex networks and validate this on pertinent economic and brain networks.

Theory

In the following we combine a number of key existing ideas in the network science literature with novel insights to produce a coherent and simple theory of how complex networks develop their characteristic topologies. To aid the reader, an illustration of the different parts of the theory and how they are used to generate a network model is provided in Fig. 1.

Surface factor. Let $\mathcal{V} = \{1, \dots, n\}$ be a set of nodes representative of individual components of a network. Then, suppose that these components have individual tendencies to make links to the other components. Consider in social networks that the tendencies of people to make new friends is the result of a number of psychological variables—such as extroversion and charisma—which are general attributes held by individuals. In economics, more open and wealthy countries are more likely to make stronger international ties and have the capacity to maintain more ties. For an example in biology, recent computational experiments indicate plausibility that gene-expression (which influences the concentration of proteins within cells) may aid in the formation of protein–protein interaction networks²¹. In each case, the collection of tendencies to make links of each node will form some kind of distribution. Whether and what generality of distribution type is possible across such disparate phenomena is a necessary consideration for a universal approach to generative modelling of networks.

Work on understanding the emergence of power-laws in the tails of degree distributions has gravitated towards power-laws themselves as the distribution of such tendencies, referred to as ‘scale-free node fitness’⁹. Power-laws tend to crop up in relationships between variables such as in allometry or in dimensions of cities²², although caution is widely advised in postulating such relationships from observation²³. In most cases, however, empirical evidence suggests singular variables consist of a large bell shaped concentration of values with a heavy right tail and are well suited to modelling with the log-normal distribution¹⁴. This, in turn, suggests that such variables come from the product of more than one independent random variable, since the product of independent positive random variables tends to the log-normal distribution (via the central limit theorem in the log-scale). Note, a log-normal distribution is typically defined as the distribution resulting from a normally

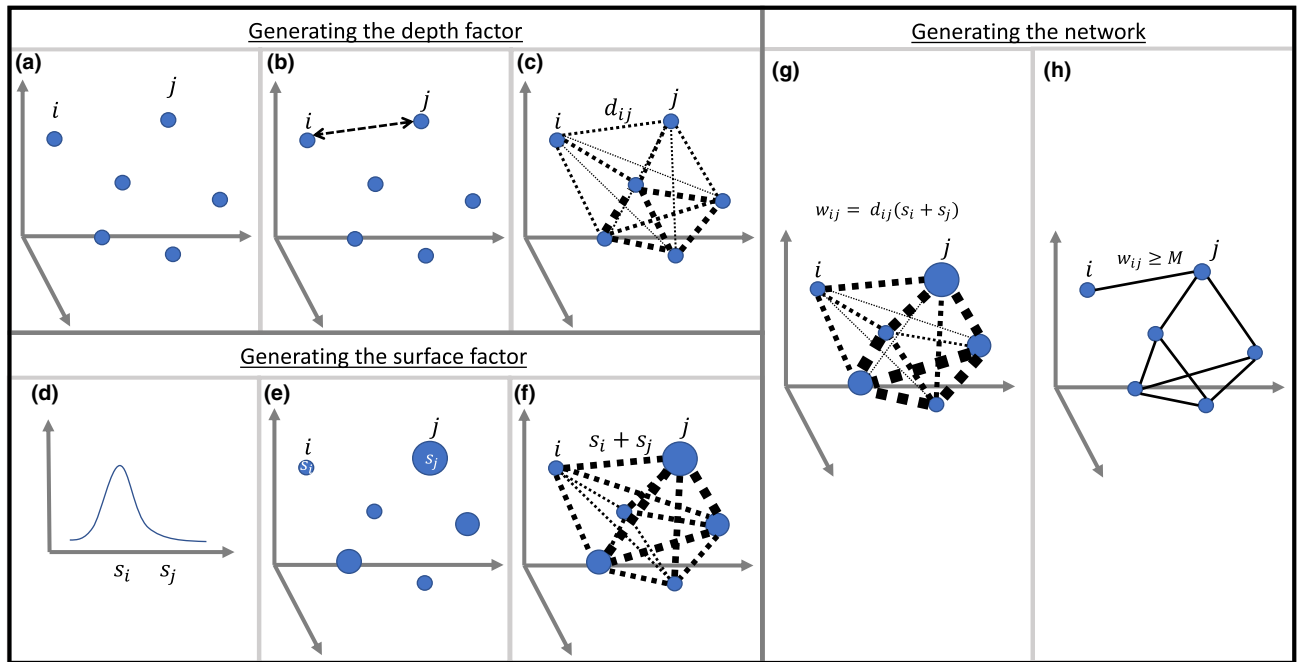


Figure 1. Generating a surface-depth model. (a) Nodes are randomly given coordinates in a q -dimensional Euclidean space. (b) Distances between each pair of nodes are measured. (c) distances are inverted to provide the depth factor of the model (weight of links indicates magnitude of d_{ij}). (d) Nodes are randomly given samples, s_i , of a log-normal distribution $LN(\mu, \sigma)$. (e) These s_i are taken as the fitness of the nodes (size of nodes indicate magnitude of each s_i). (f) The surface factor of a node pair is then taken as the sum of these values (weight of links indicates magnitude of $s_i + s_j$). (g) The link probability is then proportional to the product of depth and surface factors. (h) Links are established for the largest m values where m is the number of links in the network to be modelled (where the m th largest value is M).

distributed variable as the argument of the exponential function, $s = \exp(x)$ where $x \sim N(\mu, \sigma)$. Then, we propose to model the tendency of components to make links as a variable distributed log-normally, $s \sim \text{LogN}(\mu, \sigma)$. This is particularly promising given that recent evidence suggests most observed degree distributions of complex networks appear better approximated by log-normal distributions than power-laws¹¹.

Moreover, it is known that the tail of the log-normal distribution resembles a power-law¹⁵, i.e. a straight line on a log-log plot. The log of a log-normally distributed variable, x , is normally distributed, $y = \ln(x) \sim N(\mu, \sigma)$, while the log of the probability density function of this normal distribution is a quadratic in $y - \mu$,

$$\ln(\text{pdf}(y)) = \ln\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}\right) \tag{1}$$

$$= -\ln(\sigma\sqrt{2\pi}) - \frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2. \tag{2}$$

Then the rate of change of this is linear in $y - \mu$ and as the distribution moves further from the mean the fractional change in increase from one point to the next (i.e. $(y_{i+1} - y_i)/y_{i+1}$) decreases and the plot tends to a straight line.

Now, we relate to the variable s as the surface factor of the network, since it does not really help to describe why any two nodes are connected together beyond that either or both have strong or weak tendencies to make connections. We could consider whether such tendencies are additive or multiplicative for pairs of nodes, i.e. is the combined tendency of s_i and s_j ($s_i + s_j$) or $s_i s_j$? This is not of immediate importance since the product of two log-normally distributed variables is log-normal, while the addition of two log-normally distributed variables, x and y , with the same parameters μ and σ is approximated by the log-normal distribution $x + y \approx z \sim LN(\hat{\mu}, \hat{\sigma})$, where

$$\hat{\sigma}^2 = \ln((e^{\sigma^2} + 1)/2) \tag{3}$$

and

$$\hat{\mu} = \mu + \ln(2) + (\sigma^2 - \hat{\sigma}^2)/2, \tag{4}$$

as described in²⁴. However, we are concerned primarily with the effect this factor has on the degrees of the network rather than on individual links. In this case, the sum turns out to be more tractable. Consider,

$$u_i = \sum_{j \neq i} (s_i + s_j) \quad (5)$$

$$= (n-1)s_i + \left(\sum_{i=1}^n s_j \right) - s_i \quad (6)$$

$$= (n-2)s_i + \sum_{i=1}^n s_j \quad (7)$$

$$= As_i + B, \quad (8)$$

where $A = n - 2$ and $B = \sum_{i=1}^n s_j$. This is precisely linear in s_i , noting that A and B are exactly the same for all i . On the other hand,

$$v_i = \sum_{j \neq i} s_i s_j \quad (9)$$

$$= s_i \sum_{j \neq i} s_j, \quad (10)$$

and so there is no such exact linear relationship with s_i . We could only say that it is approximately Bs_i for large enough n and small enough s_i . Since the sum is more practical for our purposes, we shall here stick with $s_i + s_j$ as the surface factor for the existence probability of link (i, j) .

Note, for the log-normal distribution, we can arbitrarily fix μ and allow the shape parameter σ to vary to produce the different shapes of the distribution, thus essentially, the surface factor has a single parameter, σ .

Depth factor. Below this surface, we follow the homophily principle by assuming that there are similarities between components which make it more likely for connections to occur between them. In this way, we incorporate the idea of latent spaces encoding similarities between nodes¹⁶. Thus, we suppose that components are distinguishable by some number, q , of independent latent variables, x_1, x_2, \dots, x_q . Then, the similarity of nodes i and j across these variables can be described by some inverse distance function (to be consistent with the surface factor 'closer' nodes should attain larger values)

$$d_{ij} = f(x_1(i), x_1(j), x_2(i), x_2(j), \dots, x_q(i), x_q(j)). \quad (11)$$

A very obvious and important consideration of such latent variables is simply the geometry within which the components are set. If two components are proximal to one another, it stands to reason they are more likely to share a link than to share links with components which are further away, disregarding other variables. It is important to point out that latent variables could also be categorical. For instance, in a social network, people who belong to the same club, A say, are more likely to be linked than to others in another club, B .

The geometry of the latent space is an important consideration. Serrano et al.¹⁸ developed a latent space model in hyperbolic geometry. Nodes were placed on the unit disc (equivalent to the latent space of the model), parameterised by the angle to some arbitrary axis, while the degree distribution of the network was used to parameterise the radius of the node on the disc. While an elegant model, choosing the unit circle as the latent space is problematic as it restricts the dimensionality of the space.

For our modelling, we need a description of the properties of the latent variables, x_i . We know that geometry is a key consideration of networks, and thus we have up to three variables which can be approximated using a random geometric graph where coordinates are chosen uniformly at random over the interval $[0, 1]$. For simplicity we shall prescribe all variables as independent and identically distributed (i.i.d.), thus we shall simply model similarities between nodes as distances of a random geometric graph in q dimensions. Of course, it is likely that different variables will have different distributive properties in reality, but, as we shall demonstrate, this simple assumption actually works quite well in practice for modelling a diverse range of complex networks. Taking into account that smaller distances should indicate greater probability of attachment, we have, for each link, a depth factor of

$$d_{ij} = \exp \left(- \sqrt{\sum_{k=1}^q (x_{ik} - x_{jk})^2} \right) \quad (12)$$

for each $x_i \sim U(0, 1)$ and independent.

One important detail of i.i.d. latent variables is that the limit of the distribution of their sum as $q \rightarrow \infty$ is a normal distribution, by the central limit theorem. This extends to Euclidean distances between samples: take two randomly sampled points in q -dimensional space, $\mathbf{x} = \{x_1, x_2, \dots, x_q\}$ and $\mathbf{y} = \{y_1, y_2, \dots, y_q\}$ with each $x_i, y_i \sim U(0, 1)$. Then let

$$z_i = (x_i - y_i)^2, \quad (13)$$

so that each z_i is also i.i.d and, by the central limit theorem, $\sum_{i=1}^q z_i$ has a normal distribution in the limit as $q \rightarrow \infty$. From the delta method²⁵, this holds also for functions of the distribution such as the square root— $\sqrt{\sum_{i=1}^q z_i}$ —which is just the Euclidean distance between \mathbf{x} and \mathbf{y} and this further extends to Eq. (12). This property will be of use later in attempts to invert the surface factor of observed networks.

Combining factors. From the above, the probability of a connection being established between nodes i and j of a network is proportional to both the similarity of the nodes (depth factor) and the combined fitness of the nodes (surface factor), giving

$$p_{ij} \sim d_{ij}(s_i + s_j). \quad (14)$$

Assuming that these are the only considerations of the probability of existence of a link, we can take the weights of links in our network as

$$w_{ij} = d_{ij}(s_i + s_j) \quad (15)$$

up to linearity. For a complex binary network with m links, we can then, for example, take the m largest weights as extant, use a nearest neighbours connectivity approach²⁶, or use a combination of the two to specify the exact number of links while ensuring there are no isolated nodes. The only parameters of this model are the number of dimensions of the depth factor, q , and the shape parameter for the log-normal distribution of the surface factor, σ and, for a network, G , with n nodes and m links, we can describe its surface-depth model as $G_{s-d}(q, \sigma)$. Note, we intentionally avoid normalising weights to provide exact formula for p_{ij} , because we wish to model networks using the same number of nodes and links to avoid the confounding effects of network size and density on network metrics.

Estimating the surface factor in a weighted network. Given the above theory, it would be of high interest to uncover the depth factor of real networks as this would help to determine and analyse the similarity structure of nodes beyond the somewhat confounding tendencies for attachment. However, recovering the depth factor of sparse binary networks poses a very challenging problem, as it would seem intractable to determine which links are stronger to a given node than any other from the binary links. What we can do, however is to apply our methods to weighted networks by assuming that the weights of the network are approximately linearly proportional to the underlying link probabilities of the network. This is motivated by the fact that, for example, thresholded functional brain networks display the consistent topological characteristics of binary real world networks²⁷.

We saw that distances in Euclidean space have a distribution tending to normal as $q \rightarrow \infty$, and thus approximate the normal distribution for large q . Importantly, the normal distribution is a symmetric distribution with 0 skewness. On the other hand, degree distributions of real world networks and those coming from our model are right-skewed (at least for densities $d < 0.5$, relevant to most real-world networks). We must presume then, that if our model holds, the majority of this skewness is attributed to the surface factor of the network, while the distribution of depth factor weights has minimal skew. Therefore, we propose here an optimisation algorithm to determine an estimate of the log-normal surface factor of a network by minimising the skewness of network weights after inverting estimated surface factors determined by an array of log-normal distributions. In this case, the argument of the minimisation is the shape parameter σ of the log-normal distribution. Supplementary material Section III demonstrates (1) that distances between random samples in a q -Euclidean geometric space have highly symmetric distributions even for fairly small q , and (2) simulation experiments showing correlations between the real and estimated depth factor weights are inversely related to skewness. Note, without knowledge of the degree distribution of the hypothetical depth factor, we are left with the practical assumption that the ranks of the n random samples of the log-normal distribution align with the ranks of the weighted degrees of the given weighted network.

Materials and methods

Here, we detail the data used in our studies; the details of our modelling approach for real-world networks, alongside the tests and comparisons conducted; and the details of the surface factor optimisation algorithm. For methodological details of more basic exploratory experiments on the model, see Section III of the “Supplementary material”.

Real-world network data. Two datasets of networks were used for the modelling experiments. The first consisted of 25 networks taken from the Network Repository (NR) across different domains²⁸. This consisted of eight social networks—karate club, hi-tech firm, dolphins, wikivote, Hamsterster, Enron email, Dublin contact, and Uni email; six biological networks—mouse brain, macaque cortex, c elegans metabolism, mouse, plant, and yeast proteins; three ecological networks—Everglades, Mangwet and Florida; three infrastructure networks—US airports, euroroads and power grid; and three economic networks—global city network (binarised at 20% density), US transactions 1979 commodities and industries. Many of these were classic benchmark networks.

The second network dataset was the corpus used in²⁹ from the Colorado Index of Complex Networks (ICON). Of this dataset, we looked at the 184 static networks and, for the sake of computational time, chose to look only at those between 20 and 500 nodes in size. Further, we discarded bipartite networks as these have 0 clustering

and thus obviously need a different depth factor consideration than the random geometric graph which has a large clustering coefficient. This provided a final count of 85 networks.

For the surface inversion examples, we used two well-established weighted networks. The first is the world city network, available from the Globalisation and World Cities research network^{30,31}, constructed using relationships of producer service firms at the forefront of economic influence within each city. Here, each link weight is the sum over service firms of the product of the size of the service firm's offices in the two locations, normalised by the value of the maximum possible linkage in the network. In this way it relates how similar the economies of the cities are while having bias towards strength of the economy in the city. Full details are available in³⁰.

The second was the fairly sparse (link density of 0.0917) weighted group average fMRI network available freely from the brain connectivity toolbox³², the foremost resource for brain network analysis algorithms. This fMRI network was derived from a group of 27 healthy individuals. Grey matter was parcellated into 638 regions and the Blood Oxygen-Level Dependent (BOLD) time series was derived for each region. From these, Pearson's correlations of the time-series between pairs of regions were computed and normalised using the Fisher transform. The average values across the 27 individuals were then taken. For full details, see³³.

Modelling real-world networks. For a given network, we found optimal parameters of the surface depth-model based on the Root Mean Squared Error (RMSE) of topological network metrics. We compared our model against two popular existing theories of power-law fitness and hyperbolic geometry. These could be easily incorporated into our analysis by a switching of factors (switching log-normal for power-law in the surface factor and switching Euclidean geometry for spherical geometry in the depth factor). The details are described below.

Five topological network metrics were chosen on which to base the optimisation of the model to a real world network. For a network G with node set $\mathcal{V} = \{1, 2, \dots, n\}$ and link set $\mathcal{E} = \{(i, j) : i, j \in \mathcal{V}\}$, $|\mathcal{E}| = 2m$ these were

1. The clustering coefficient, C . This measures fraction of node triples, $\{i, j, k\} \in \mathcal{V}$, with all links present, $\{(i, j), (j, k), (k, i)\} \in \mathcal{E}$, in the network. This indicates how likely neighbouring nodes are to share neighbours.
2. Global efficiency³⁴, E . This measures the average of one over the shortest paths in the network

$$E = \frac{1}{n(n-1)} \sum_{i \neq j \in G} \frac{1}{d(i, j)}, \quad (16)$$

where $d(i, j)$ is the length of the shortest path between nodes i and j . This indicates how quickly on average one can get from one node to any other in the network. Normalised degree variance³⁵, V . This is a normalised measure of the variance of the degree distribution of the network,

$$V = \frac{\sum_{i=1}^n (k_i - \langle k \rangle)^2}{nm(1 - P)}, \quad (17)$$

where k_i is the degree of node i , $\langle k \rangle$ is the average degree of the network and P is the network density. This indicates the inequality of the degree distribution.

3. Modularity based on the Louvain algorithm³⁶, Q . This measures how strongly the network can be partitioned into groups of high connectivity, and with comparatively less connectivity between groups. The Louvain algorithm describes an optimisation of the partition of the network to maximise the modularity

$$Q = \frac{\delta(c_i - c_j)}{2m} \sum_{i \neq j \in G} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \quad (18)$$

where A_{ij} is the ij th entry of the adjacency matrix of the network, c_i is the community of node i (randomly initialised) and δ is the Kronecker delta function being 1 when $c_i = c_j$ and 0 otherwise. The modularity of the network is then taken as the optimised Q .

4. Assortativity³⁷, r , of network degrees. This is just a Pearson's correlation of the degrees between connected nodes and can be written

$$r = \frac{\sum_{t=1}^{2m} (k_{t1} - \langle k \rangle_{\mathcal{E}})(k_{t2} - \langle k \rangle_{\mathcal{E}})}{\sum_{t=1}^{2m} (k_{t1} - \langle k \rangle_{\mathcal{E}})^2}, \quad (19)$$

where $2m$ is the number of links in \mathcal{E} (each link is counted twice), k_{t1} and k_{t2} are the degrees of the first and second nodes in the t th link, and $\langle k \rangle_{\mathcal{E}}$ is the average degree of nodes turning up in all links in \mathcal{E} (so that node i 's degree, k_i , is counted precisely k_i times). This indicates how similar the degrees of connected nodes are across the network.

Each metric was chosen on the basis that (1) it covered a distinctly formulated topological aspect, and (2) its value was appropriately normalised with maximum possible magnitude of 1 so that the minimisation was not evidently biased to any particular index. This kind of minimisation has been previously used in e.g.^{38,39}. We assumed that for a node to exist in a sparse binary network, it would be required to be connected within it—consider that isolated nodes could exist in a system without the knowledge of the network constructor. Thus models (with the same number of nodes as their corresponding real-world networks) were ensured to have all nodes with at least degree 1 by including the nearest neighbours for each node. The rest of the links were then selected simply from the links with highest weights across all model weights until the number of links matched the real network.

After network metrics were computed for each generated model, the RMSE over all metrics between the real-world network and its model

$$RMSE = \sqrt{\frac{1}{T} \sum_{i=1}^T (M_i - \hat{M}_i)^2} \quad (20)$$

was computed, where, each M_i is the value of one of the five metrics defined above (arbitrarily) for the real-world network and \hat{M}_i is the corresponding value of that metric for the surface-depth model. In our case, then $T = 5$ —being the five metrics C , E , V , Q , and r . The RMSE was used for optimising the model by searching for the model parameters which produced the minimum RMSE. This optimisation was implemented using the following algorithm:

Algorithm 1 Modelling a network

- 1: Compute metrics C , E , V , Q and r of network G
 - 2: **for** $q \in \{1, 2, \dots, 10\}$ **do**
 - 3: Compute 20 realisations, $G_{s-d}(q, \sigma)$, of model with the same size and density as G with σ ranging from 0.05 up to 1 in steps of 0.05
 - 4: Compute C , E , V , Q and r of each of these models and take the mean over realisations for each
 - 5: Compute the RMSE between metrics of G and metric means of $G_{s-d}(q, \sigma)$
 - 6: Take σ' as the σ parameter of minimum RMSE model
 - 7: Compute 20 realisations of each surface-depth model with σ within 0.05 of σ' in steps of 0.01
 - 8: Take the model with the minimum RMSE value from this step as the minimum for the model with q dimensions
 - 9: The minimum across q of the minimum RMSEs across σ is then taken as the model of best fit to G
-

Importantly, it is not expected that the discretisation of the surface factor parameter causes any problems here. It is reasonable to assume in this instance that there are no local minima that would confound the optimization because of the discretisation, since the distributions of the surface-factors are smooth, the right-skew of the distributions are monotonic functions (increasing with log-normal and decreasing with power-law) of the parameters, and the distributions themselves have only global maxima and minima. Note also, we took a maximum of $q = 10$ arbitrarily to save on time as we assume the topological properties of the model are asymptotic with q , as demonstrated in the Supplementary Material Section I.A. Figure E in Section II of the “Supplementary material” plots the index values of 10 networks and their models alongside results obtained for models utilising surface and depth factors separately, illustrating how the model adapts to each network.

We compared this model against competing theories of power-law fitness⁹ and hyperbolic geometry (alongside higher dimensional spherical surface geometries)¹⁸. The same algorithm was used for power-law fitness and spherical surface geometry by substituting the log-normal parameter, $\sigma \in [0, 1]$, for a power-law parameter, $\gamma \in [2, 3]$ (the interval within which most scale-free networks are found to follow), and by substituting q -dimensional Euclidean geometry for q -dimensional spherical surface geometry, respectively.

For power-law fitness, the link weights were computed as:

$$d_{ij}(s_i + s_j) \quad (21)$$

with s_i sampled randomly from a power-law distribution with parameter γ . Again, γ was first checked in steps of 0.05 in the interval $[2, 3]$ in the first stage of the Algorithm 1 and then steps of 0.01 in the second stage.

For spherical surface geometry, random samples of a q -dimensional spherical surface were generated where coordinates for a single sample were obtained from normalising q normally distributed samples and distances between two samples, $x = [x_1, x_2, \dots, x_q]$ and $y = [y_1, y_2, \dots, y_q]$, computed per the formula

$$d(x, y) = a \cos \left(\sum_{i=1}^q x_i y_i \right). \quad (22)$$

Then the negative of the exponential was taken, following Eq. (12), and dimensions of spherical geometry were directly substituted for dimensions of Euclidean geometry in Algorithm 1.

Once the best performing parameters for each model were obtained, the RMSE of these best-performing models were compared to assess which model's topology was closest to the real-world network. We also calculated the Spearman correlation coefficient and its p -value between each network's best-fit surface factor parameter and depth factor parameter to test the assumption that these parameters should be independent. Next, degree distributions of the log-normal and power-law models were compared against those of the real-world networks by computing the effect sizes (as the normalised z -statistic, $z/\sqrt{n^2/2n} = z/\sqrt{n/2}$) and p -values (the null hypothesis, that the distributions were not different, was rejected in the case that $p \leq 0.05$) for the Kolmogorov–Smirnov (KS) two-sample test. This allowed us to assess whether log-normal surface factors could explain the degree distributions of real world networks and how this compared to the popular power-law theory.

Surface factor optimisation. To test the validity of the model in weighted networks, we assessed to what extent an attempted surface inversion of the weights (i.e. dividing the weights in (15) by $(s_i + s_j)$ to recover d_{ij}) outputted weights with stronger geometric qualities and similarity relationships between the nodes.

To do this, we first required a method to best approximate the log-normal distribution which could hypothetically be the distribution of the surface factor. In the “Theory” section, we noted that random Euclidean distances in a hypercube tend to a normal distribution as the number of dimensions, q , tends to infinity. Section III demonstrates that, indeed, even for fairly small q , the distribution of distances looks normal and certainly has negligible skewness. Therefore, we proposed to approximate the hypothetical surface factor of a real world weighted network by finding the parameter, σ , which minimised the skewness after its inversion from the network weights. Then, for a weighted network with adjacency matrix \mathbf{W} of size n with entries W_{ij} , the shape parameter of a log-normal surface factor was estimated, up to two decimal places, by the following algorithm:

Algorithm 2 Estimating the surface factor

- 1: **for** $\sigma \in \{0.01, 0.02, \dots, 1\}$ **do**
 - 2: Compute 1000 realisations, $\{s_k\}_{k=1}^{1000}$, where each s_k is a vector of n samples from log-normal distribution $LN(0.5, \sigma)$
 - 3: For each s_k , order the samples according to the ranks of the weighted degrees of \mathbf{W} (e.g. largest sample goes in position i where node i has largest weighted degree)
 - 4: For newly arranged s_k , compute the surface factor matrix, \mathbf{S} , whose entries $S_{ij} = s_k(i) + s_k(j)$
 - 5: Compute the depth factor estimation matrix, \mathbf{D} , with entries $D_{ij} = W_{ij}/S_{ij}$
 - 6: Compute the skewness of the non-diagonal entries of \mathbf{D}
 - 7: For each σ , average the skewness over the 1000 realisations
 - 8: The value of σ which achieves minimum average skewness is taken as the optimised estimate of the hypothetical surface factor of \mathbf{W}
-

From this, the estimated depth factor matrix \mathbf{D} of the real-world weighted network was obtained as that with the minimum skewness of its entries. To assess the plausibility of \mathbf{D} as a depth factor, we compared the 5-Nearest Neighbour (5NN) graphs of \mathbf{W} and \mathbf{D} . Considering that the weighted degrees may be seen as a simpler approximation of any underlying surface factor distribution, without the need to assume log-normality, we also compared our approach with the network of weights obtained by simply dividing weights, W_{ij} , by the average of the weighted degrees of the pair of adjacent nodes (i.e. a ‘weighted degree inversion’), obtaining the matrix \mathbf{H} with entries

$$H_{ij} = \frac{2W_{ij}}{\sum_{k=1}^n W_{ik} + \sum_{k=1}^n W_{jk}}. \quad (23)$$

The resulting 5NN graphs of \mathbf{W} , \mathbf{D} and \mathbf{K} were assessed in terms of the associations of the nodes. For the world city network, we assessed the proximity of the nearest neighbours on the globe and performed community detection using Louvain’s modularity algorithm³⁴ to assess to what extent communities were composed of proximal groups of cities. For full details see Supplementary Section V. For the fMRI network, we used the provided geometric information of the nodes to assess proximity of nearest neighbours. We also employed community detection and assessed (1) the normalised mutual information between modules in the 5NN networks and the 5NN of the geometric graph of the brain, (2) to what extent communities (or modules) were symmetric across the brain (i.e. in what percentage of cases was a right hemisphere region in the same community as a left hemisphere region), and (3) the average longest distance found within communities. For full details see Supplementary Section IV.

Experiments

Section I.A of the “Supplementary material” provides some initial explorations of the topology of the model covering topological differences between surface-depth models and random geometric graphs and the behaviour of degree distribution with increasing network density. Importantly, we found that surface-depth models have general characteristics associated with real-world networks, such as high clustering coefficient and modularity, high degree heterogeneity, and disassortativity. Furthermore, Section I.B goes on to show that degree distributions of surface-depth models with $n = 1000$ and $q = 4$ exhibited power-laws at densities of 1–4% and log-normal distributions at densities of 4–40% (specifically, null hypotheses of two-sample KS tests with power-law and log-normal degree distributions could consistently not be rejected at the 5% level in these cases).

To validate Algorithm 1, 1000 surface-depth models were generated with randomly selected parameters and fed into the algorithm. The error of the estimated parameters produced from the algorithm were then assessed. The interquartile range (i.e. 50% of the distribution) of the error estimated number of dimensions of the depth factor, q , was from $[0,1]$ dimension of the true parameter, while the interquartile range of the estimated shape parameter of the surface factor, σ , was $[-0.02, 0.02]$. For both it was seen that there were positive correlations between the error and the magnitude of the parameters, indicating that the larger the parameters produced by the algorithm were, the larger their error from the true parameters are likely to be. Full details and results can be found in Section I.C of the “Supplementary material”.

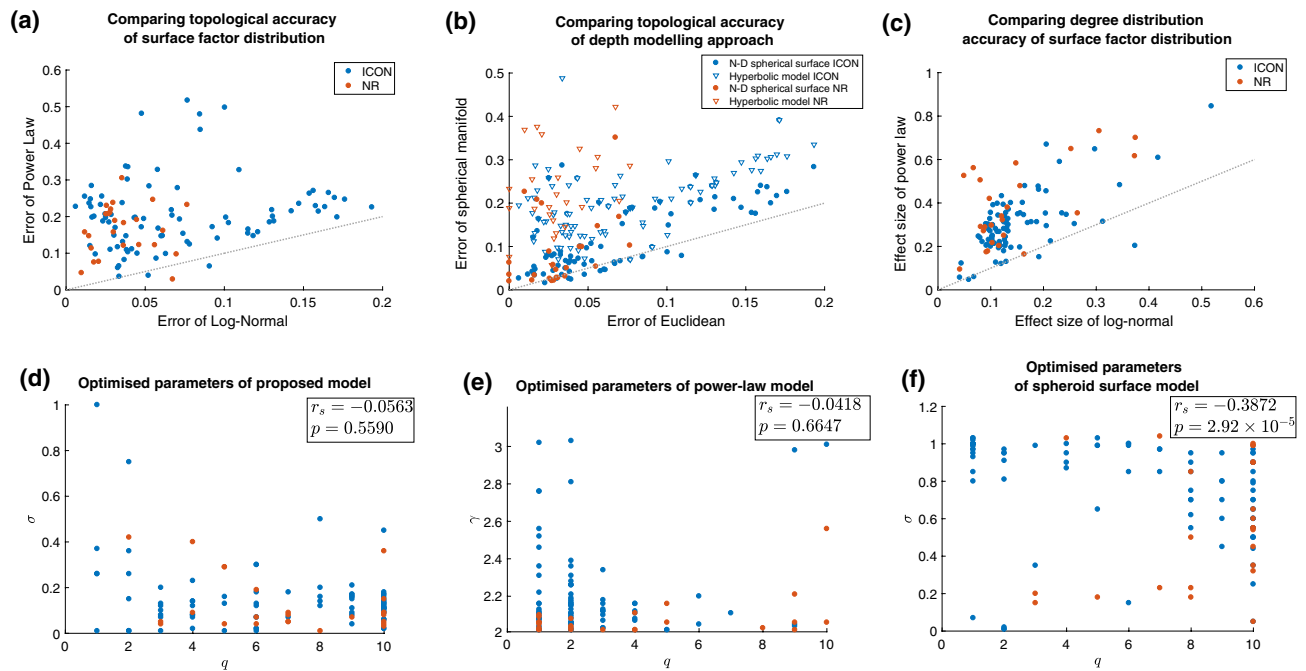


Figure 2. Plots (a) and (b) show root mean squared errors of the proposed model against power-law attachment spherical surface geometry (including the hyperbolic model), respectively. (c) Effect sizes of degree distributions between model and network (log-normal versus power-law attachment). Dotted lines show the line of parity while ICON is dataset from the Index of Complex Networks and NR is the dataset from the network repository. Plots (d–f) show the surface model parameter plotted against the depth model parameter for the proposed theory, power-law attachment theory and spherical surface geometry theory, respectively. Spearman's correlation coefficients, r_s , and their p -values between the parameters indicate how correlated the model parameters are across the 110 networks.

We shall continue with the most pertinent results regarding the modelling of real world networks. We modelled 110 real world binary networks collected from two difference sources. The most accurate surface-depth model was then chosen by optimising for the two model parameters, σ and q , following Algorithm 1. Note, in each case, the number of nodes and links in the resulting model were kept the same as in the original network. We then did the exact same approach with parameter substitutions for (1) power-law fitness instead of log-normal fitness in the surface factor, and, separately, (2) spherical surface geometry for node similarity instead of Euclidean space in the depth factor.

The Root Mean Squared Error (RMSE) in topology of the models for each network—calculated through five distinct and widely used normalised topological metrics, C , E , V , Q and r —is scatter plotted against RMSE using (1) a power-law surface factor and (2) spherical surface depth factor in Fig. 2a,b, respectively. The proposed model clearly outperformed models of theories of both power-law attachment and hyperbolic geometry, with a median RMSE of just 0.0449 compared with 0.1932 and 0.2012 for power-law attachment and hyperbolic geometry, respectively. It also clearly outperformed general q -dimensional spherical surface geometry with a median RMSE of 0.0813. In fact, RMSE was smaller in the proposed model than hyperbolic geometry in 99.09% of networks, power-law fitness in 97.27% of networks and general spherical surface geometry in 80% of networks studied. Furthermore, the average sizes of RMSE were a remarkable 293.4%, 287.5% and 170.4% the size of the proposed model for hyperbolic geometry, power-law fitness and general spherical surface geometry models, respectively.

Next, for each real-world network we compared the degree distributions of the best-fit model with the real-world networks using KS two-sample tests. This was done fifty times for each network and median results recorded. Of the 110 networks studied, 68.2% had no significant median p -value, while 81.8% had no noticeable effect size (≤ 0.2), with all but one of the remainder (17.27%) having only small effect sizes ($\in [0.2, 0.5]$). Again, these compared very favourably against the power-law fitness model, see Fig. 2c. Indeed, the average effect size of the power-law model was 225.7% that of the average log-normal model.

We then tested to see whether any correlation or anti-correlation was established between the optimised parameters, q and σ , of the model. The existence of any significant correlation would indicate that the parameters were not independent and thus would negate the claims of the theory that independent surface and depth factors existed to make up link probability. Scatter plots of σ against q for all networks are shown for the proposed model, the power-law attachment model and the general spherical surface model in Fig. 2d–f, respectively. Spearman's correlation coefficient, r_s , was used to assess levels of correlation between q and σ . There was no correlation found between σ and q of the proposed theory's model ($r_s = -0.0563$, $p = 0.5590$), validating the independence assumption of surface and depth factors of complex networks. On the other hand, a significant anti-correlation was found between σ and q when spherical surface geometry was used ($r_s = -0.3872$, $p = 2.92 \times 10^{-5}$),

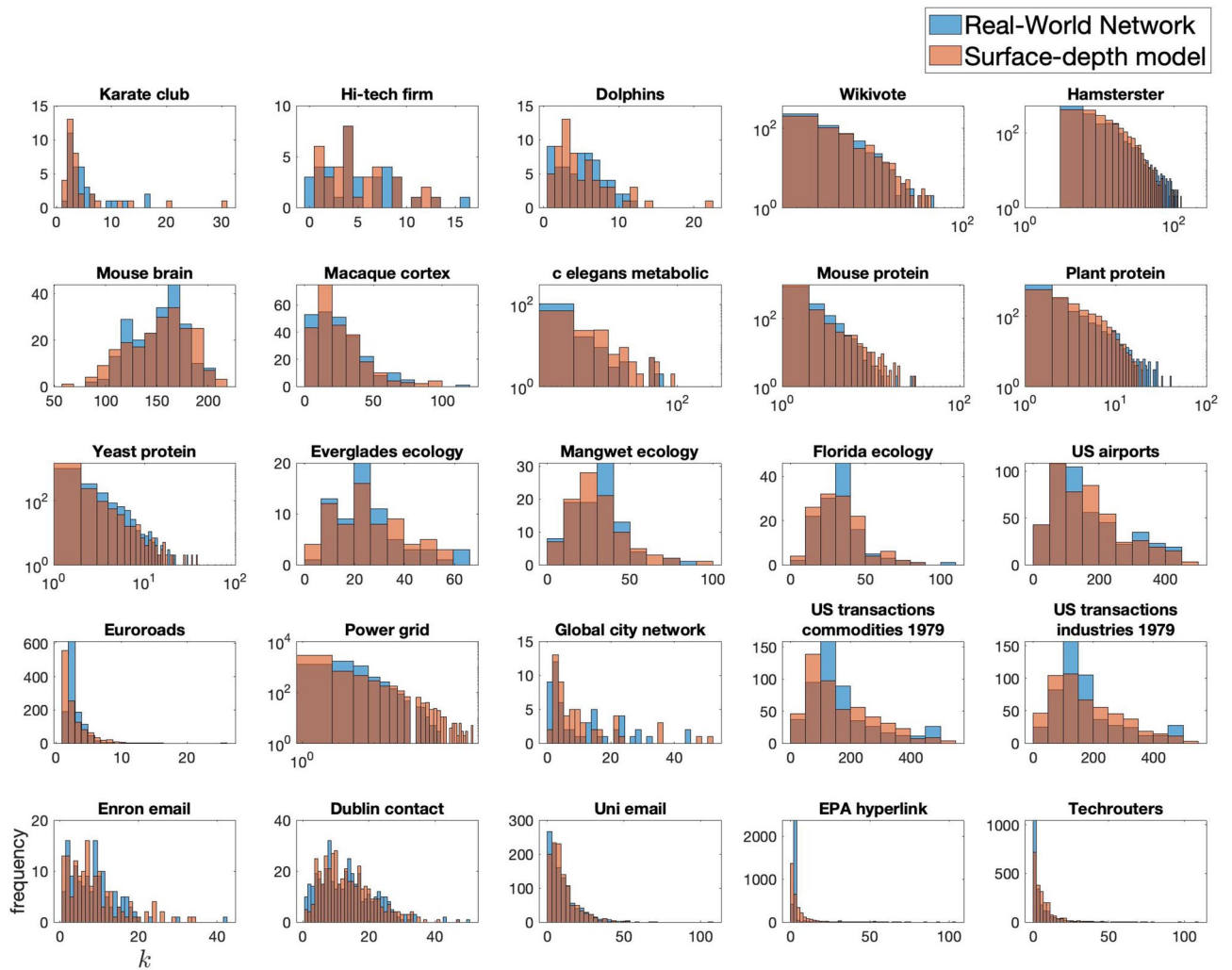


Figure 3. Comparison of the degree distributions between real-world networks and their respective closest fit surface-depth model. These are log–log plots where there is a clear scaling distribution. Axes as in bottom left plot—degree, k , against frequency.

indicating that this model, and the hyperbolic geometry model of which it is a generalisation, was not as appropriate a theoretical foundation for network topology emergence.

Figure 3 shows comparisons of the degree distributions of the network repository networks and their best-fit surface-depth models. The similarity between distributions across all networks of various size, density and domain is striking. From all of these results, the surface-depth model appears as a good candidate for a unifying theory of attachment in complex network topologies, achieving scale-free like distributions in networks at sparse densities and log-normal like distributions in networks of larger densities, as can be seen in real-world networks in¹¹ for example.

Interestingly, there was a particular class of networks that proved to have large errors for all models even though their degree distributions were on the whole largely indistinguishable from those of the proposed model. These were food web networks. Looking more closely, it appeared there was an exceptional difference in the clustering coefficients in this case. Median differences for each index across food web networks were as follows: 0.2753 for C , 0.0206 for E , 0.0593 for V , 0.0185 for r , and 0.0449 for Q . The very low relative clustering in food web networks makes sense since we can expect that it is uncommon for predators of the same prey to hunt one another as well. This suggests that better modelling of the depth factor may help to better capture the information here.

Depth factor recovery through estimated surface factor inversion. To probe further whether surface-depth factors could really be observed in real-world networks, we applied depth factor recovery and subsequent analysis of the recovered depth factor's geometric qualities on two important cases of weighted networks: an economic world city network and a group average fMRI functional brain network, as described in the “[Methods and materials](#)”. In both cases, we optimised the log-normal distributions of the surface factors following the network weight skewness minimisation Algorithm 2 in the methods, based on the fact that Euclidean

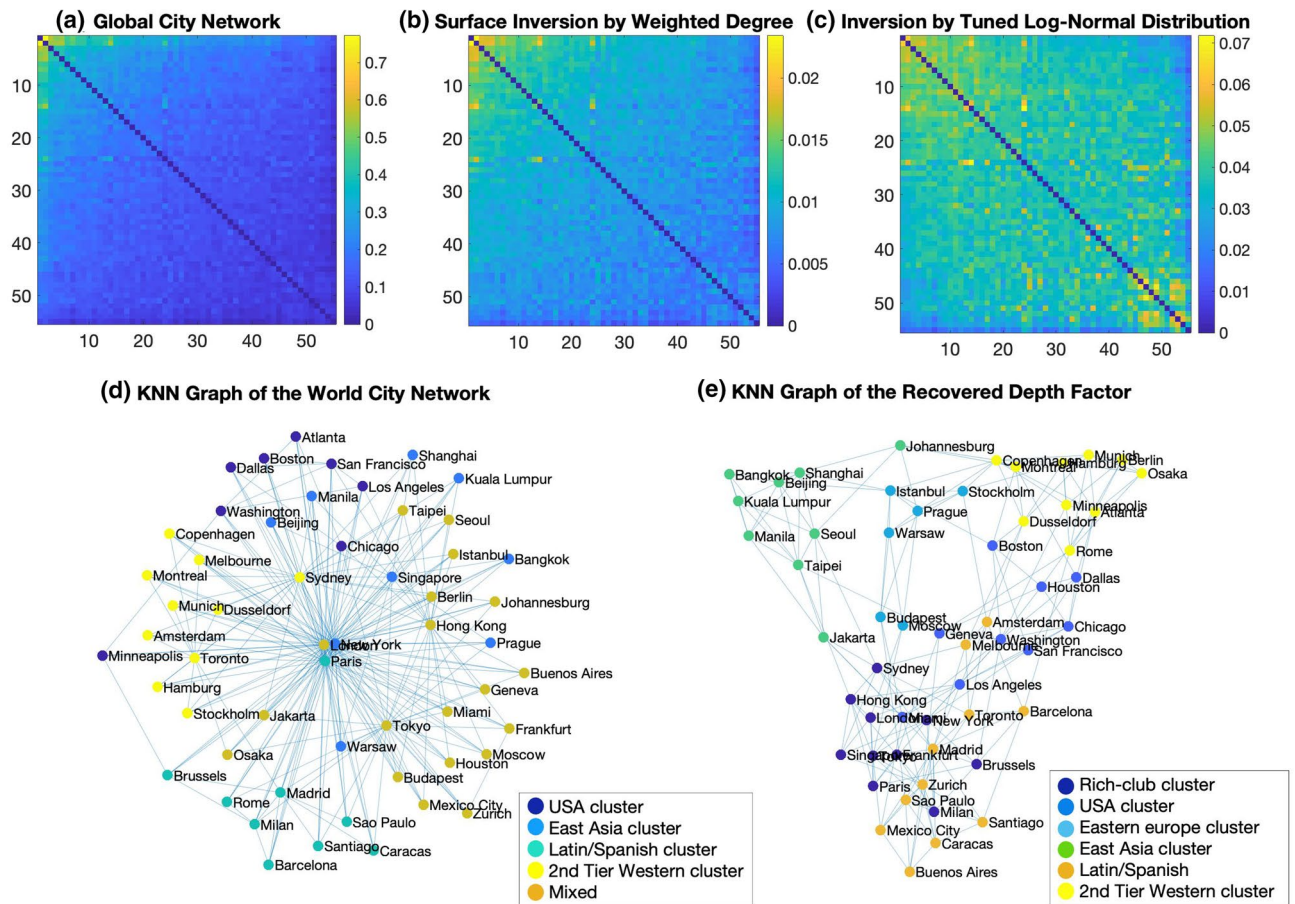


Figure 4. (a) Weighted adjacency matrices (ordered by weighted degree) of the global city network, (b) an estimated depth factor of the network using the weighted degree and (c) an estimated depth factor using a tuned log-normal distribution, respectively. Both axes here represent the rows and columns of the adjacency matrix. (d) Plot of the five-nearest neighbours graph of the world city network (left) and (e) its recovered depth factor (right) with detected modules shown in different colours. Modules in the depth factor are observably more distinguishable from the shape of the network, whereas relationships between the nodes in the original network are dominated by a few nodes.

distances in the q -dimensional hypercube tends towards the symmetric normal distribution as $q \rightarrow \infty$ by the central limit theorem, and on the observations in Supplementary material Section III.

For the global city network, the optimal log-normal distribution was found at $\sigma = 0.59$. K-Nearest Neighbour (KNN) graphs with $K = 5$ were then computed from the global city network and its estimated depth factor. We also compared this with just using the weighted degree distribution as an estimate of the surface factor. Figure 4a–c show the weighted adjacency matrices of the original network and the estimated depth factors from the weighted degree and tuned log-normal distribution surface inversion approaches, respectively.

Modules were computed using Louvain's modularity method³⁴. The 5NN graphs were then plotted using the same force-based algorithm where connected nodes are attracted and non-connected nodes repelled from one another⁴⁰, Fig. 4d,e. Remarkably, surface inversion of the hub-centric world city network produced a highly modular network with geometric qualities. On inspection, spaces within the network layout were notable by their global proximity and cultural ties. We analysed this statistically in the case of global proximity. Section V of the “Supplementary material” contains these details alongside tables of the five nearest neighbours of each city for each approach. Of these, 180 (65.45%) were found to be proximal on the globe (either being in the same continent or otherwise geographically close) for the tuned log-normal inversion compared to 50.55% for the degree-based inversion and just 37.82% for the original network. Furthermore, the five cities with greatest weighted degree (London, New York, Paris, Tokyo and Hong Kong) appeared in just 10.56% of the tuned log-normal inversion compared with 76.64% of the nearest neighbours in the original network and 46.18% in the degree-based inversion, with 9.27% being that expected by random chance. In addition, 52 of the 55 cities were found within the 5 nearest neighbours of all cities in the tuned log-normal inversion approach, whereas this number was just 15 for the original network and 38 for the degree-based inversion. All in all, the tuned log-normal inversion provided a remarkably more geometrically congruent network, with a clear elimination of rich-club-style⁴¹ bias in nearest neighbours. Some qualitative observations are also worth noting. Barcelona and Madrid were found to be in the same community as all Latin American cities, appealing to their cultural ties, whereas Latin American cities were

not even all found in the same community in the original network. Further, Eastern Europe and East Asia both had clearly distinct communities in the recovered depth factor but not so in the original network.

For the fMRI network, the optimal log-normal distribution was found at $\sigma = 0.27$. The availability of the 3D coordinates of the nodes representing brain regions allowed us to construct a geometric graph for comparison. The sparsity of the network posed a significant confounding factor in this instance as only those links which already existed could be chosen in the resulting 5NN graph. Nonetheless, we considered four measurements of the geometric appropriateness of the resulting depth factor: (1) the percentage of overlapping links with the 5NN graph of the geometric network, (2) the normalised mutual information between the modules of the network and the modules of the geometry (3) the proportion of symmetric nodes across brain hemispheres appearing in the same module, and (4) the average largest distance within modules. Details of these analyses are in the Supplementary material Section IV. In all cases the estimated depth factor outperformed the original network. The depth factor achieved consistently greater geometric overlap, normalised mutual information and module symmetry, and smaller average largest distance within modules. This again demonstrates the relationship of the estimated depth factor with underlying geometry of the considered networks.

The combined evidence from the world city and fMRI networks provides promising evidence of the real existence of surface and depth factors in complex networks, substantiating the real-world applicability of the proposed theory and opening up new avenues for discovery in weighted network analysis particularly.

Limitations and future work

The theory put forward is topologically accurate in modelling most of the complex networks studied here, yet we made no attempt to take into account dynamically changing networks and network evolution. That being said, it would seem that evolution and dynamics of networks could be incorporated in our theory by shifts occurring in surface and depth factors. For instance, a node may take on different values of its latent variables thus changing the nodes to which it is most similar which would result in a change to the links the node makes. Otherwise, the node may increase or decrease its fitness giving it a higher/lower tendency to make connections, again resulting in a dynamic change of the network. New nodes could be assumed to appear somewhere within the latent variable space but with an initially low tendency to make the connections. Such processes could be stochastically encoded.

Also there are evident limitations in the modelling of the depth factor, most clearly seen in the generally higher clustering coefficient of the model. To improve the model's accuracy, new methods would be required for more accurate depth factors and the fusion of different types of latent variables, including categorical variables and variables with different distributions, as well as weighting variables for their importance. Mechanisms which may account for lower clustering should be explored. The current assumptions don't allow for factors which mitigate the inherently strong homophily of Euclidean geometry, such as repulsion between nodes. The proposal that a depth factor of weight similarities can be extracted has clear implications in terms of geometric deep learning⁴². Along similar lines, a recent study considered using machine learning approaches on a hyperbolic network model⁴³. It seems that such methods can be fairly straightforwardly translated to the geometries of the proposed depth factor and we expect our study will open up interesting future research along these lines. Immediate applications of the theory include surface inversion to other weighted networks and the consideration of this theory to advance efforts in important network problems such as community detection and link prediction.

Data availability

Datasets used are readily available and as referenced in this article.

Code availability

Code used is freely available at DOI <https://doi.org/10.17605/OSF.IO/PMXU7>.

Received: 29 September 2020; Accepted: 29 December 2020

Published online: 21 January 2021

References

1. Watts, D. & Strogatz, S. Collective dynamics of small-world networks. *Nature* **393**, 440–442 (1998).
2. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509 LP–512 (1999).
3. Erdős, P. & Rényi, A. On random graphs. *Publicationes Mathematicae Debrecen* **6**, 290–297 (1959).
4. Bollobás, B. *Random Graphs*, ch. 8 of *Modern Graph Theory. Graduate Texts in Mathematics* (Springer, New York, 1998).
5. Newman, M. E. J. *Random Graphs as Models of Networks*, ch. 2 of *Handbook of Graphs and Networks: From the Genome to the Internet* (Wiley, New Jersey, 2006).
6. Maslov, S. & Sneppen, K. Specificity and stability in topology of protein networks. *Science* **296**, 910–913 (2002).
7. Holland, P., Laskey, K. & Leinhardt, S. Stochastic block models: First steps. *Social Netw.* **5**, 109–137 (1983).
8. Smith, K. & Escudero, J. The complex hierarchical topology of EEG functional connectivity. *J. Neurosci. Methods* **276**, 1–12 (2017).
9. Caldarelli, G., Capocci, A., De Los Rios, P. & Munoz, M. Scale-free networks from varying vertex intrinsic fitness. *Phys. Rev. Lett.* **89**, 258702 (2002).
10. Holme, P. Rare and everywhere: Perspectives on scale-free networks. *Nat. Commun.* **10**, 1016 (2019).
11. Broido, A. & Clauset, A. Scale-free networks are rare. *Nat. Commun.* **10**, 1017 (2019).
12. Voitalov, I., van der Hoorn, P., van der Hofstad, R. & Krioukov, D. Scale-free networks well done. *Phys. Rev. Res.* **1**, 033034 (2019).
13. Del Genio, C., Gross, T. & Bassler, K. All scale-free networks are sparse. *Phys. Rev. Lett.* **107**, 178701 (2011).
14. Limpert, E. & Stahel, W. The log-normal distribution. *Significance* **14**, 8–9 (2017).
15. Mitzenmacher, M. A brief history of generative models for power law and lognormal distributions. *Internet Math.* **1**, 1358 (2004).
16. Hoff, P., Raferty, A. & Handcock, M. Latent space approaches to social network analysis. *J. Am. Statistical Assoc.* **97**, 1090–1098 (2002).
17. Dall, J. & Christensen, M. Random geometric graphs. *Phys. Rev. E* **66**, 016121 (2002).

18. Serrano, A., Krioukov, D. & Boguñá, M. Self-similarity of complex networks and hidden metric spaces. *Phys. Rev. Lett.* **100**, 078701 (2008).
19. Allard, A., Serrano, M., García-Pérez, G. & Boguñá, M. The geometric nature of weights in real complex networks. *Nat. Commun.* **8**, 14103 (2017).
20. Papadopoulos, F., Kitsak, M., Serrano, M., Boguna, M. & Krioukov, D. Popularity versus similarity in growing networks. *Nature* **489**, 537–540 (2012).
21. Klein, B. *et al.* Resilience and evolvability of protein–protein interaction networks. *bioRxiv*. <https://doi.org/10.1101/2020.07.02.184325> (2020).
22. West, G. *Scale: The Universal Laws of Growth, Innovation, Sustainability, and the Pace of Life in Organisms, Cities, Economies, and Companies* (Penguin Press, New York, 2017).
23. Stumpf, M. & Porter, M. Critical truths about power laws. *Science* **335**, 665–666 (2012).
24. Marlow, N. A normal limit theorem for power sums of independent normal random variables. *Bell Syst. Tech. J.* **46**, 2081–2089 (1967).
25. Doob, J. The limiting distributions of certain statistics. *Ann. Math. Stat.* **6**, 160–169 (1935).
26. Eppstein, D., Paterson, M. & Yao, F. On nearest-neighbor graphs. *Discrete Comput. Geometry* **17**, 263–282 (1997).
27. Bullmore, E. & Sporns, O. Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nat. Neurosci. Rev.* **10**, 186–198 (2009).
28. Rossi, R. A. & Ahmed, N. K. The network data repository with interactive graph analytics and visualization. in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 4292–4293 (2015).
29. Ghasemian, A., Hosseinmardi, H. & Clauset, A. Evaluating overfit and underfit in models of network community structure. *IEEE Trans. Knowl. Data Eng.* **32**, 1722–1735 (2019).
30. Taylor, P. Specification of the world city network. *Geogr. Anal.* **33**, 181–194 (2001).
31. Taylor, P. & Walker, D. World city network: Data matrix construction and analysis. <http://www.lboro.ac.uk/gawc/datasets/da7.html>. Accessed 11 January 2021.
32. Rubinov, M. & Sporns, O. Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage* **52**, 1059–1069 (2010).
33. Crossley, N. *et al.* Cognitive relevance of the community structure of the human brain functional coactivation network. *PNAS* **110**, 11583–11588 (2013).
34. Latora, V. & Marchiori, M. Efficient behavior of small-world networks. *Phys. Rev. Lett.* **87**, 198701 (2001).
35. Smith, K. & Escudero, J. Normalised degree variance. *Appl. Netw. Sci.* **5**, 32 (2020).
36. Blondel, V., Guillaume, J., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **10**, P10008 (2008).
37. Newman, M. Assortative mixing in networks. *Phys. Rev. Lett.* **89**, 208701 (2002).
38. Betzel, R. F. *et al.* Generative models of the human connectome. *Neuroimage* **124**, 1054–1064 (2016).
39. Topirceanu, A., Udrescu, M. & Marculescu, R. Weighted betweenness preferential attachment: A new mechanism explaining social network formation and evolution. *Sci. Rep.* **8**, 10871 (2018).
40. Fruchterman, T. & Reingold, E. Graph drawing by force-directed placement. *Softw. Practice Exp.* **21**, 1129–1164 (1991).
41. Colizza, V., Flammini, A., Serrano, M. & Vespignani, A. Detecting rich-club ordering in complex networks. *Nat. Phys.* **2**, 110–115 (2006).
42. Bronstein, M., Bruna, J., LeCun, Y., Szlam, A. & Vandergheynst, P. Geometric deep learning: Going beyond Euclidean data. *IEEE Signal Process. Mag.* **34**, 18–42 (2017).
43. Muscoloni, A., Thomas, J., Ciucci, S., Bianconi, G. & Cannistraci, C. Machine learning meets complex networks via coalescent embedding in the hyperbolic space. *Nat. Commun.* **8**, 1615 (2017).

Acknowledgements

The author is thankful to Anton Pichler for useful discussions concerning economic networks. This work was supported by Health Data Research UK (MRC ref Mr/S004122/1), which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, National Institute for Health Research (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome.

Author contributions

K.S. is the sole author and did all of the presented work.

Competing interests

The author declares no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-81547-3>.

Correspondence and requests for materials should be addressed to K.M.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021