ARTICLE

Sizing complex networks

Gorka Zamora-López [1,2]* & Romain Brasselet[3]

Among the many features of natural and man-made complex networks the small-world phenomenon is a relevant and popular one. But, how small is a small-world network and how does it compare to others? Despite its importance, a reliable and comparable quantification of the average pathlength of networks has remained an open challenge over the years. Here, we uncover the upper (ultra-long (UL)) and the lower (ultra-short (US)) limits for the pathlength and efficiency of networks. These results allow us to frame their length under a natural reference and to provide a synoptic representation, without the need to rely on the choice for a null-model (e.g., random graphs or ring lattices). Application to empirical examples of three categories (neural, social and transportation) shows that, while most real networks display a pathlength comparable to that of random graphs, when contrasted against the boundaries, only the cortical connectomes prove to be ultra-short.

[1] Department of Information and Communication Technologies, Pompeu Fabra University, Barcelona, Spain. [2] Center for Brain and Cognition, Pompeu Fabra University, Barcelona, Spain. [3] Scuola Internazionale Superiore di Studi Avanzati (SISSA), Trieste, Italy. *email: gorka@Zamora-Lopez.xyz

The small-world phenomenon has fascinated popular culture and science for decades. Discovered in the realm of social sciences during the 1960s, it arises from the observation that any two persons in the world are connected through a short chain of social ties[1]. Since then many real networks have been found to be small-world as well[2–4], from natural to manmade systems. But, how small is a small-world network and how does it compare to others? In the physical world we evaluate and compare the size of objects by contrasting them to a common reference, usually a standard metric system defined and agreed by the community. In the case of complex networks the difference is that every network constitutes its own metric space. Thus, the question of whether a network is smaller or larger than another implies the comparison of two different spaces with each other, rather than the more familiar situation in which two objects are contrasted within the space they live.

The quantification of the small-world property relies on the computation of the average pathlength—the average graph distance between all pairs of nodes. As it happens for any graph metric, its outcome depends on the number of nodes and links of the network under study. Consider two empirical networks. G_1 is a small social network, e.g., a local sports club of $N_1 = 100$ members. A link between two members implies they are friends. G_2 is an online social network with ten million users ($N_2 = 10^7$) where two profiles are connected if users follow each other. If we found that the average pathlength l_1 of G_1 is smaller than the length l_2 of G_2, we would not be able to conclude that the internal topology of the local sports-club is shorter, or more efficient, than the structure of the large online network. The observation that $l_1 < l_2$ could be a trivial consequence of the fact that $N_1 \ll N_2$. In order to fully interpret this result we need to disentangle the contribution of the network's internal architecture to the pathlength from the incidental influence contributed by the number of nodes and links.

In practice, the typical strategy seen in the literature to deal with this problem consists in comparing empirical networks to well-known graph models, e.g., random graphs, scale-free networks and regular lattices[2,5–8]. It shall be emphasised that these architectures represent different null-models, generated upon particular hypotheses and thus useful to answer a variety of questions we may have about the data. However, their role as absolute references is restricted since they are not representative of the limits that network metrics, such as pathlength and efficiency, can take[9,10].

Here, we establish a reference framework under which the length of networks can be reliably interpreted and compared. To do so, we have first identified the upper and the lower boundaries for the pathlength which, in turn, depend on the specific number of nodes and links. We have extended these results to the global efficiency[11], an analogous metric that is measurable also when the networks are sparse and disconnected. We have found that these limits are given by families of characteristic configurations which we term ultra-short (US) and ultra-long (UL) networks. With these boundaries at hand, we can now assess the average pathlength of a network—of a given size and density—by evaluating how much it deviates from the smallest and the largest pathlength it could possibly take. This framework allows us to evaluate both empirical networks and graph models together under the same reference. We show that typical null-models (random, scale-free and ring networks) undergo a transition as their density increases: initially, when they are sparse, their pathlength is well above the lower boundary, but then they all become US at sufficient density. The rate of convergence, however, differs across models. Finally, we compare a sample set of well-known empirical networks (neural, social and transportation). We find that, while most of these systems display a pathlength comparable to that of random

graphs, only cortical connectomes reveal to be US when contrasted to the boundaries.

Results

In this paper we will refer to small-world networks in the classical sense, where the typical distance between nodes is much smaller than the size of the network. While our focus is on the properties of the pathlength, the clustering coefficient of the graphs at the boundaries is shown in Supplementary Note 6. In order to avoid the ambiguous meaning of the term "size" in the literature, in the following we will use "size" only to refer to the number of nodes N in a network and we will correspondingly employ the adjectives small and large. We will refer to the average pathlength l of a network as its "length" and use corresponding adjectives short and long. We will denote the properties of directed graphs adding a tilde to the symbols. For example, if L is the number of undirected edges in a graph, $\tilde{L}$ will be the number of directed arcs in a directed graph (digraph).

Ultra-short and ultra-long networks. The main result of this paper is the identification of the lower and upper limits for the average pathlength and global efficiency for (di)graphs of arbitrary number of nodes N and links L. We identify these boundaries as families of (di)graph configurations which we name US and UL networks, summarised in Fig. 1; see Methods and Supplementary Notes 1–3 for details. These families arise from a few simple building blocks, Fig. 1. The sparsest connected graphs that can be constructed are named trees, i.e., graphs without cycles. All trees of size N contains $L = N - 1$ edges. Among trees, star and path graphs are the configurations with the shortest and the longest pathlength, respectively. In a star graph, any two nodes can reach each other jumping through the central hub while in a path graph, the whole network needs to be traversed to travel from one end to the other. When the links are directed, however, directed rings are the sparsest connected digraphs, which consist of $\tilde{L} = N$ arcs, all pointing in the same orientation. Finally, a complete graph is the network in which all nodes are connected to each other, thus containing $L_o = \frac{1}{2}N(N - 1)$ edges or $\tilde{L}_o = N(N - 1)$-directed arcs. The average pathlength of a complete graph is $l_o = 1$, the shortest of all networks.

US and UL graphs of arbitrary L can be achieved by adding edges to star and path graphs, respectively, Figs. 1a–c. In the case of digraphs, both US and UL configurations are obtained by adding arcs to a directed ring, Figs. 1d–h. The precise order of link addition differs from case to case. Two findings deserve a special mention. (1) US and UL graphs can be generated adding edges one-by-one to the initial configurations, Fig. 1a, b, but construction of extremal digraphs is often non-Markovian. That is, an US or an UL digraph with $\tilde{L} + 1$ arcs cannot always be achieved by adding one arc to the extremal digraph with $\tilde{L}$ arcs. For example, Fig. 1d shows that the digraphs with shortest pathlength initially transition from a directed ring onto a star graph following unique configurations we named flower digraphs. (2) For a given size and number of links, all configurations with diameter $\text{diam}(G) = 2$ have exactly the same pathlength[9,10,12] and they are US, regardless of how their links are internally arranged. Their pathlength is $l = 2 - \rho$, where ρ is the link density. See the US graph theorem (Theorems 1 and 4) in Supplementary Notes 1 and 2.

When studying large networks it is common to find that these are sparse and fragmented into many components. While the pathlength in these cases is infinite, these networks can still be characterised by their efficiency, which remains a finite quantity allowing to zoom into the sparse regime. We remind that the global efficiency E of a network is defined as the average of the
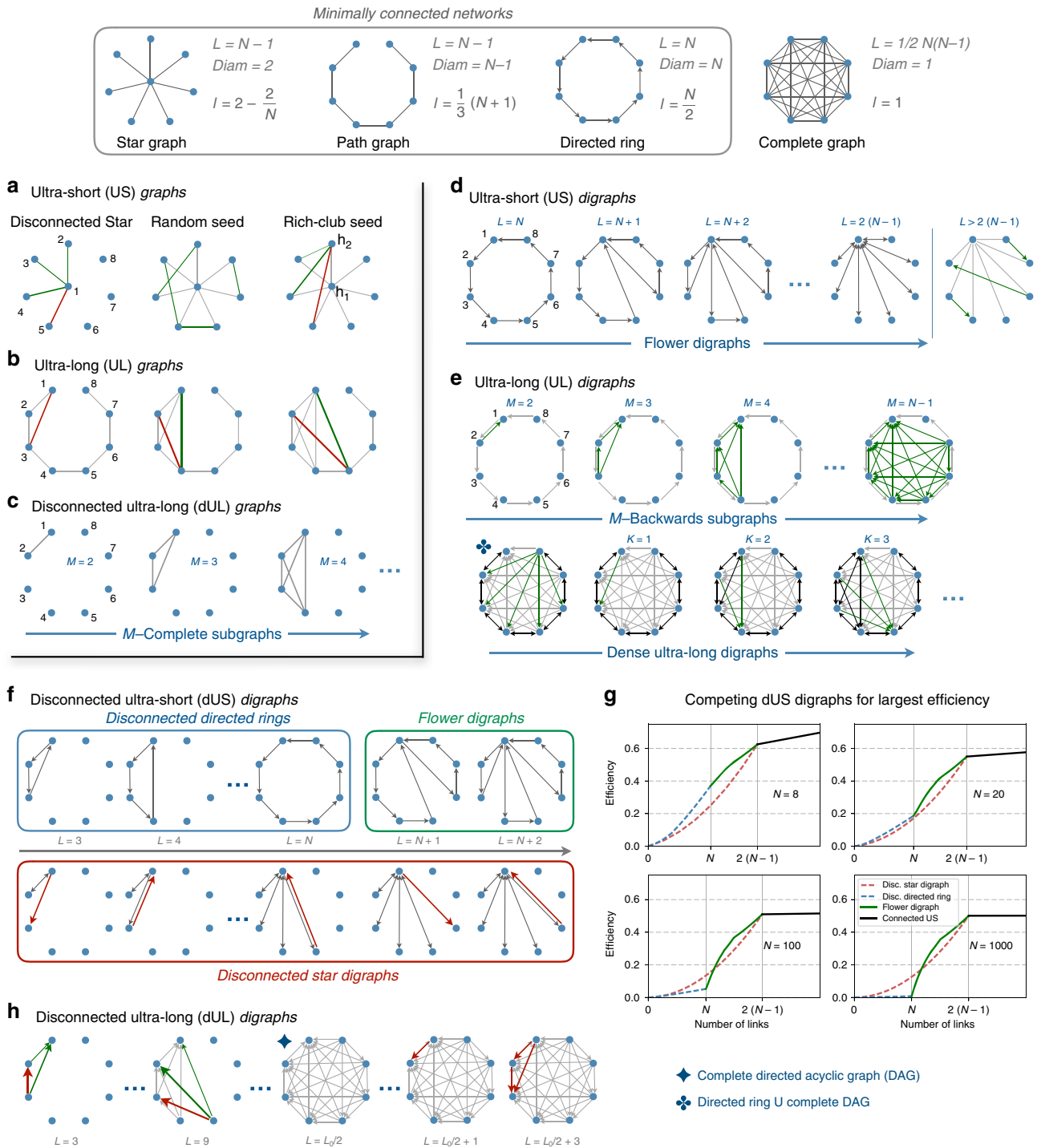
Fig. 1 Construction of ultra-short and ultra-long (di)graphs. Star graphs, path graphs, directed rings and complete graphs serve as the starting references to construct (di)graphs of arbitrary density with extremal average pathlength. **a–c** Procedures to build ultra-short and ultra-long graphs, both connected and disconnected. Edge colour denotes the order of edge addition. Red edges are the last added and green links the ones in the previous steps. **d–h** Generation of directed graphs with extremal pathlength or efficiency. These cases are often non-Markovian and reveal novel structures. **d** In the sparse regime ultra-short digraphs are characterised by a collection of directed cycles converging at a single hub. We name these flower digraphs. Every arc added leads to a flower digraph with an additional "petal". **e** Although several configurations may lead to digraphs with longest pathlength, we introduce here an algorithmic approximation to the upper bound: M-backwards subgraphs. **f** Up to three different digraph configurations compete for largest efficiency. **g** The winner depends on network density. Finally, **h** digraphs with smallest efficiency are achieved by constructing the densest directed acyclic graphs possible, to minimise the contribution of cycles to the path structure of the network

inverses of the pairwise distances. Thus the contribution of disconnected pairs (with infinite distance) is null. We could identify sparse configurations [$L < N − 1$ or $\tilde{L} < 2(N − 1)$] with the largest efficiency, whose efficiency transitions from zero

(for an empty network) to that of a star graph. In the case of graphs, Fig. 1a, there is a unique optimal configuration but for digraphs we found that up to three different structures compete for the largest efficiency when $\tilde{L} < 2(N − 1)$, Fig. 1f, g. See Fig. 1c,

h for the graphs and digraphs with smallest possible efficiency, which equals the link density, $E = \rho$.

The length of common network models. In the following we illustrate how knowledge of the US and UL boundaries frame the space of lengths that networks take, both empirical and models. We start by investigating the null-models which over the years have dominated the discussions on the topic of small-world networks: random graphs (Erdős–Rényi type), scale-free networks[13] and ring lattices. We consider undirected and directed versions with $N = 1000$ nodes and study the whole range of densities; from empty ($\rho = 0$) to complete ($\rho = 1$). The results are shown in Fig. 2. Find comparative results for $N = 100$ and $N = 10,000$ in Supplementary Note 4. Shaded areas mark the values of pathlength and global efficiency that no network can achieve. Solid lines represent the ranges in which the models are connected and dashed lines correspond to the efficiency when the networks are disconnected. The location of the original building-blocks (star graphs, path graphs, directed rings and complete graphs) are also shown over the maps for reference.

The pathlength of random, scale-free and ring networks decays with density, as expected, with the three eventually converging onto the lower boundary and becoming US, Fig. 2a, b. But, the decay rates differ across models. Scale-free networks are always shorter than random graphs in the sparser regime, where the length of both models is well above the lower boundary. However, the two models converge simultaneously onto the US limit at $\rho \approx 0.08$. On the other hand, the ring lattices decay much slower and only becomes US at $\rho \approx 0.5$.

Figure 2c, d reproduces the same results in terms of efficiency. An advantage of efficiency is that it always takes a finite value, from zero to one, regardless of whether a network is connected or not. Zooming into the sparser regime, we observe that the efficiency of both random (E_r) and scale-free (E_{SF}) graphs undergoes a transition, shifting from the UL to the US boundary,

Fig. 2d. They are nearly identical except for a narrow regime in between $\rho \in (4 \times 10^{-4}, 2 \times 10^{-3})$. Here, E_{SF} grows earlier than E_r, reaching a peak difference of $E_{SF} \approx 5 \times E_r$ at $\rho = 10^{-3}$. The reason for this is that SF graphs percolate earlier than random graphs[14]. Indeed, the onset of a giant component in random graphs of size $N = 1000$ happens at $\rho \approx 10^{-3}$.

The results for the directed versions of the random and scale-free networks, Fig. 2e, f, are very similar. The main difference is that when $\tilde{L} = N$ both the upper and the lower boundaries are born from the same point corresponding to a directed ring, Fig. 2e.

Length and global efficiency of empirical networks. We now elucidate how the knowledge of the boundaries allows us to interpret the length of real networks faithfully. First, we will illustrate how different references may give rise to subjective interpretations. Then, we will show how the boundaries help framing both empirical and model networks into a unified representation.

Given two networks G_1 and G_2 with pathlength $l_1 < l_2$ (the absolute values), we could affirm that G_1 is shorter than G_2. But, if G_2 is bigger ($N_1 < N_2$) then the fact that $l_1 < l_2$ does not necessarily imply that the topology of G_1 is more efficient than the topology of G_2. An approach commonly followed in the literature to face this problem consists in comparing networks G_1 and G_2 with well-known network models, or null-models. For example, random graphs (Erdős–Rényi) and ring lattices are typically employed as the references to characterise the "small-worldness" of complex networks. In some cases, the relative pathlength $l' = l/l_r$, is used which considers the length l_r of random graphs as the lower boundary[5]. This measure takes $l' = 1$ when the length of the network matches that of random graphs. In other cases, a two-point normalisation using ring lattices as the upper boundary[6,7] was proposed, $l' = (l - l_r)/(l_{latt} - l_r)$. Here, $l' = 0$ if the length of the real network equals that of random
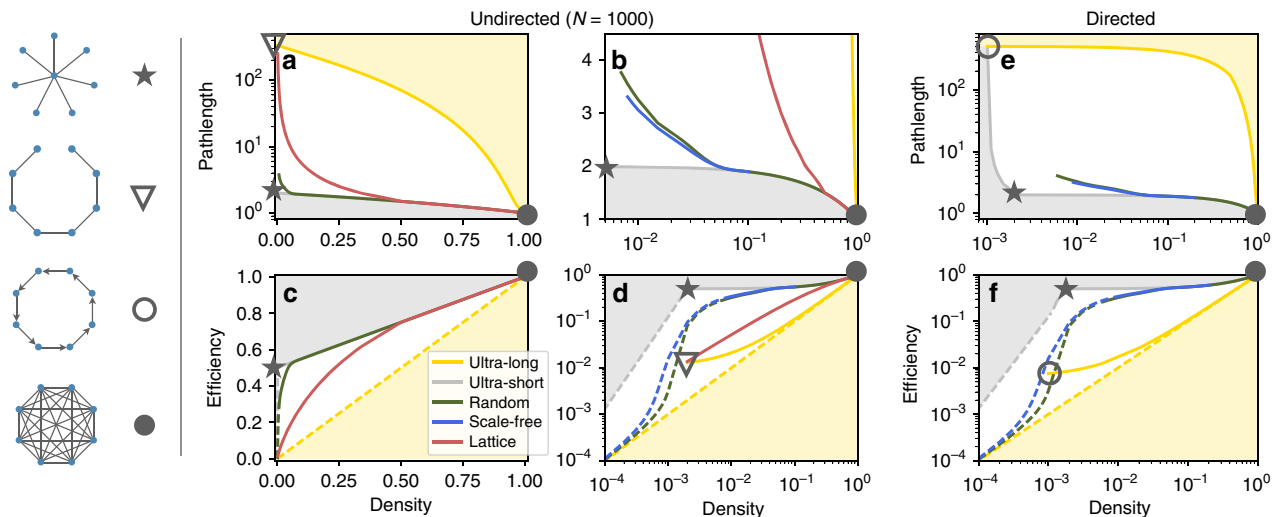

Fig. 2 Pathlength and efficiency of characteristic network models. Ring lattices (red), random (Erdős–Rényi type, green) and scale-free[13] graphs (blue) of $N = 1000$ nodes, plotted together with the corresponding upper and lower boundaries ultra-long (yellow) and ultra short (grey). Shaded areas mark values of pathlength and efficiency that no (di)graph of the same size can achieve. **a** The average pathlength of the three models decay towards the ultra-short limit at sufficiently large density. **b** Same as panel (**a**) with density re-scaled for visualisation. **c** Global efficiency of the network models. The lower boundary (ultra-long) is represented by two lines: a dashed line representing the efficiency of disconnected graphs $E_{dUL} \approx \rho$ and a solid line for connected graphs. **d** Same as panel (**c**) with density re-scaled for visualisation. The efficiency of random and scale-free graphs undergoes a transition from ultra-long to ultra-short centered at their percolation thresholds. **e** Pathlength of random and scale-free digraphs. In this case, the two boundaries emerge from the same point corresponding to a directed ring (open circle). **f** Efficiency of random and scale-free digraphs. Curves for random and scale-free networks are averages over 1000 realisations. Dashed lines represent ranges of density for which the models are disconnected and solid lines represent (di)graphs which are connected. Results for sizes $N = 100$ and $10,000$ are shown in Supplementary Note 4

graphs (the lower boundary) and $l' = 1$ if it matches the length of ring lattices (the upper boundary). Considering the US and UL boundaries, one could also propose the following re-scaling

$$l' = \frac{l}{l_{US}}. \tag{1}$$

$$l' = \frac{l - l_{US}}{l_{UL} - l_{US}}. \tag{2}$$

At this point, it shall be stressed that the use of relative metrics may lead to biased interpretations, depending on the question(s) we are asking about the data. Relative pathlengths such as $l' = l/l_r$ and $l' = l/l_{latt}$ are conceptually different from the ones in Eqs. (1) and (2) because l_r and l_{latt} are expectation values associated to given null-models while l_{US} and l_{UL} are actual limits. Thus, they are meant to answer different questions. On the one hand, null-models are constructed upon particular sets of constraints and following generative assumptions on the rules governing how nodes link with each other. The role of expectation values is thus to test whether those hypotheses explain the pathlength observed in the real networks. On the other hand, limits correspond to the extremal values (di)graphs could take and are independent of generative assumptions. The reader should keep in mind that here, our intention is neither to identify the key topological factors that "explain" the pathlength observed in real networks nor to model the real networks themselves. Our goal is to address the question "how short or how long is a given network?" This allows us to position network models in the space of reachable pathlength and to compare these models with potentially different numbers of nodes and links.

For practical illustration, we study a set of empirical datasets from three different domains: neural and cortical connectomes, social networks and transportation systems, see Table 1. These examples represent a small but diverse set of real networks with sizes ranging from $N = 34$ to 4941 and densities ranging from $\rho \approx 10^{-4}$ to 0.330. First, we show that the ranking of these networks with respect to pathlength is very much altered depending on the relative reference chosen, Fig. 3. According to the absolute pathlength, Fig. 3a, cortical and neural connectomes tend to be shorter than social and transportation networks. The network of prison inmates is directed and weakly connected,

therefore it has an infinite pathlength. See Supplementary Note 5 for the results presented in terms of global efficiency. We could now ask whether these observations are a trivial consequence of the different sizes of those networks. For that, we consider the relative pathlength $l' = l / N$, Fig. 3b. In this case, the ranking changes considerably. The short length of the cortico-cortical connectomes seems to be partly explained by their small size ($N < 100$).

When considering random graphs as the reference, Fig. 3c, we find that all the networks take relative values $l' \approx 1$; with the neural networks, the Zachary karate club and the airports network being the closest to random graphs. With these results at hand we may tend to believe that, e.g., the airport transportation network is shorter than the dolphins social network. This interpretation is only part of the picture because l_r is an expectation value. Here, l' provides a ranking of the networks whose pathlength is better explained by the generative assumptions of random graphs. If the US limit is taken as reference, a different scenario is found, Fig. 3d. In this case, the cortical networks (cat, macaque and human) lie marginally above $l' = 1$, meaning that they are practically US. The dolphins and the facebook circles are twice as long as the lower boundary and the transportation networks deviate even further. The comparison to random graphs was not possible for three transportation networks (London transportation, Chicago transportation and the U.S. power grid) because their densities lie below the percolation threshold for Erdős–Rényi graphs and thus no connected graphs could be constructed with the same N and L. Calculated in terms of global efficiency, this comparison shows that these three networks are less efficient than random graphs and lie notably far from the optimal, see Supplementary Note 5. Notice that transportation networks, in reality, are embedded in space and are nearly planar. The planarity helps sparse networks to be connected[15], while they would easily be disconnected if this constraint were ignored. However, here we are discarding any constraints besides N and L whether they concern internal features (e.g., the degree distribution) or external (e.g., spatial embedding), and focus solely on the theoretical limits the pathlength and efficiency can take.

Figure 3e shows the two-point relative pathlengths when both random graphs and ring-lattices are taken as references. With these results the Zachary Karate Club and the airports network would appear to be shorter than the brain connectomes, the collaboration network of jazz musicians and the dolphin's network. Since l_r and l_{latt} are expectation values, the results should be seen as which of the two null-models, random graphs or ring-lattices, better explains the length of each real network. Finally, considering the US and the UL boundaries as references, Fig. 3f, it becomes clear that all the networks are closer to the US boundary than to the UL.

So far, we have evidenced that the use of relative metrics to rank networks according to their length may lead to biased interpretations because the outcome depends on whether the reference is a null-model or a limit. Therefore, a more informative solution than relying on relative metrics is to display all the relevant results, both for real networks and for null-models, into a unified representation. Figure 4 shows the global efficiency of the thirteen empirical networks (+), together with their corresponding US (grey bars) and UL (gold bars) boundaries, and the expectation values for equivalent random graphs (blue lines) and ring lattices (red lines). This representation elucidates the results reported in Fig. 3c–f. First, the space of accessible efficiencies (delimited in between the lower and the upper boundaries) differs from case to case, depending on the size and density of each network. Second, the position that both real networks and

Table 1 Characteristics of the sample real networks investigated: For illustrative purposes we analyse twelve networks from three different domains: neural, social and transportation. Networks marked with and asterisk (*) are directed

Class	Network	N	L	Density
Neural	Cat[34,35]	53	826*	0.300
	Macaque[36,37]	85	2356*	**0.330**
	Human[38]	66	590	0.275
	C. elegans[39]	274	2956*	0.040
Social	Jazz[40]	279	2742	0.141
	Zachary[41]	**34**	78	0.139
	Dolphins[42]	62	159	0.084
	FB circles	4039	88,234	0.011
	Prison[43]	67	182*	0.041
Transport	London[44]	317	370	0.007
	Chicago	376	402	0.006
	Airports[45]	3618	14,142	0.002
	Power grid[2]	**4941**	6594	**0.0006**

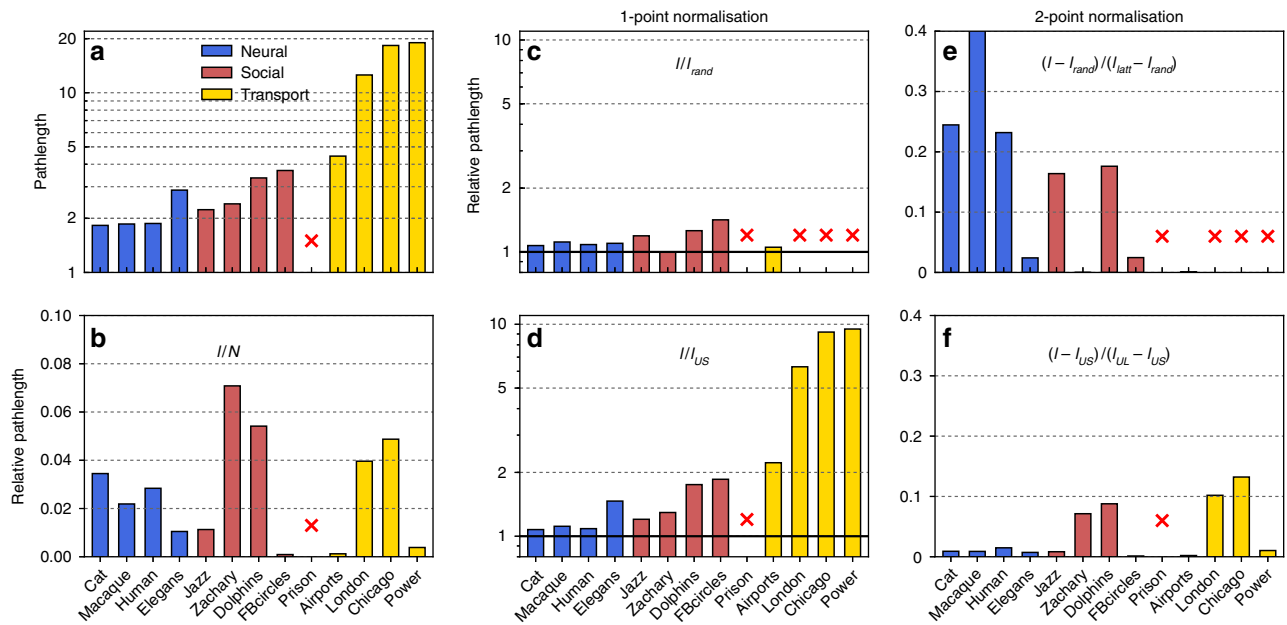Values in bold represent largest and smallest size or density

Fig. 3 Comparison of absolute and relative pathlength for selected neural, social and transportation networks. **a** Absolute average pathlength of the empirical networks, **b**–**f** different relative pathlength definitions. **b** Relative to network size N, **c** relative to equivalent random graphs, **d** relative to the pathlength to the ultra-short boundary. **e**, **f** Two-point normalisations considering random graphs and ring lattices as benchmarks (**e**), and relative to the ultra-short and ultra-long boundaries (**f**). Red crosses indicated cases for which all random graphs generated as benchmark were disconnected and had thus an infinite pathlength. The Prison social network is weakly connected and can thus only be studied by characterising efficiency, see Supplementary Note 5
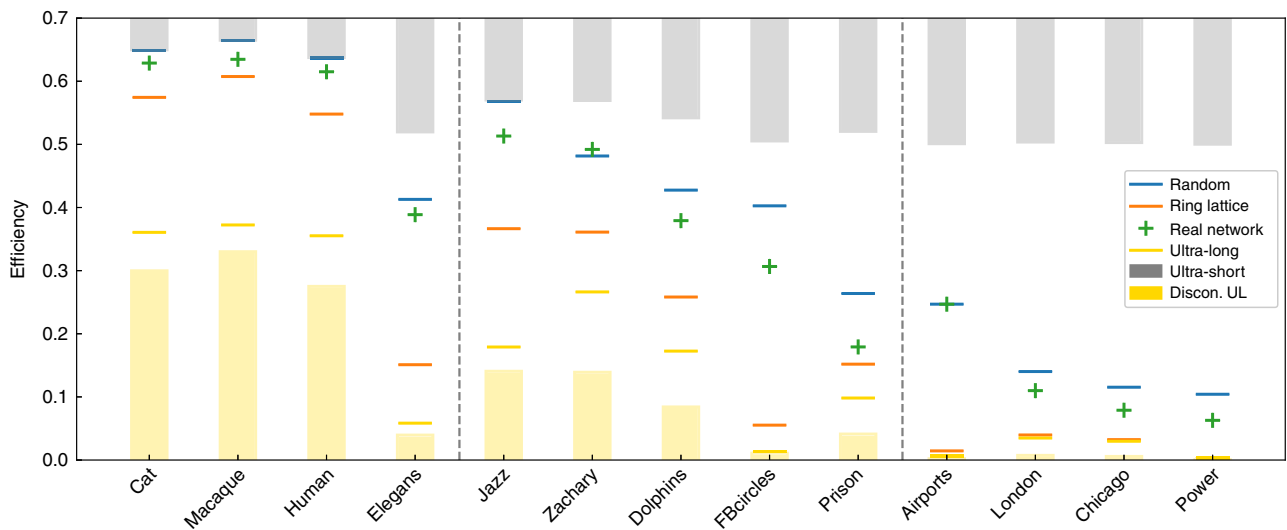

Fig. 4 Comparison of global efficiency for selected neural, social and transportation networks. The global efficiency of the thirteen empirical networks ($+$) is shown together with their ultra-short and ultra-long boundaries. The span of the boundaries very much differs from case to case because of the different sizes and densities of the networks studied. For the denser networks (e.g., cortical connectomes) the efficiency of random graphs (blue lines) lie almost on top of the largest possible efficiency (ultra-short boundary). On the contrary, for the sparser networks (e.g., the transportation systems) the efficiency of random graphs very much divert from ultra-short

null-models take within this space is revealed. In the case of the three brain connectomes (cat, macaque and human) their equivalent random graphs match the US boundary. Thus comparing these networks to random graphs is the same as comparing them to the lower limit. However, in sparser networks this is no longer the case. For example, the efficiency of the neural network of the *Caenorhabditis elegans* is close to that of random graphs, but both values depart from the US boundary. In this case, the network is far from ring lattices (red lines) and from the

UL boundary. The opposite is found for transportation networks: their global efficiency and the efficiency of corresponding random graphs lie closer to the UL boundary than to the US.

In summary, the synoptic representation in Fig. 4 tells us that: (1) The pathlength (efficiency) of empirical networks is usually comparable to the length of random graphs; meaning that the random connectivity hypothesis partly explains the empirical values. But (2) the position networks take with respect to the limits very much differs from case to case, exposing how close

each network—real and models—lie from the optimal efficiency or from the worst-case scenario. In future practical studies, Fig. 4 can be complemented with the results from other null-models, each accounting for specific sets of constraints, e.g., conserving the degree distribution or considering the limitations imposed by the spatial embedding. The expectation value calculated for each null-model will add one datapoint per network and thus allow to visualise the contribution of every set of constraints. Here, we restricted to random graphs and ring lattices for the illustrative purposes.

Discussion

Among the many descriptors to characterise complex networks, the average pathlength is a very important one. It lies at the heart of the small-world phenomenon and also plays a crucial role in network dynamics, as short pathlength facilitate global synchrony[4,16] or the diffusion of information and diseases[17,18]. Unfortunately, the pathlength is also difficult to treat mathematically and most analytic results so far are restricted to statistical approximations on scale-free and random graphs, at the thermodynamic limit[19,20]. Here, we have taken a considerable step forward by identifying and formally calculating the limits for the average pathlength and for the global efficiency in networks of any size and density. We provide results for both directed and undirected networks, whether they are sparse (disconnected) or dense (connected), thus delivering solutions that are useful for the whole range of real networks and (di)graph models studied in practice.

We have found that these boundaries are given by specific architectures which we generically named US and UL networks. The optimal configurations are not always unique and may vary according to size or density. US and UL networks are thus characterised by a collection of models as summarised in Fig. 1.

We have studied empirical networks from three scientific domains—neural, social and transportation. The comparison evidences that cortical connectomes are the shortest of the three classes. In fact, they are practically as short as they could possibly be and any alteration of their structure, e.g., a selective rewiring of their links, would only lead to negligible decrease of the pathlength. Over the last decade it has been discovered that brain and neural connectomes are organised into modular architectures with the cross-modular paths centralised through a rich-club[21–25]. Recently, it has been shown that this type of organisation supports complex network dynamics as compared to the capabilities of other hierarchical architectures[26,27]. Now, we find that cortical connectomes are also quasi-optimal in terms of pathlength. On the other extreme, transportation networks are more than five times longer than the corresponding lower limit. This contrast between cortical and transportation networks is rather intriguing since both are spatially embedded. While the aim of neural networks might be the rapid and efficient access to information within the network, transportation networks are planar and developed to service vast areas surrounding a city. Thus they are often characterised by long chains spreading out radially from a rather compact centre.

From a practical point of view, our theoretical findings solve the problem of assessing and comparing how short or how long a network is. Evaluating the length of a network with a single number—whether absolute or relative—has strong limitations and often involves arbitrary choices. The usual approach consists in defining relative metrics that normalise empirical observations l with the pathlength from a model, e.g., random graphs, ring lattices or rewired networks which conserve the node degrees. The choice of the null-model depends on the particular question we may be asking about the data. We have stressed that relative

metrics as $l' = l/l_r$ and $l' = l/l_{US}$ serve different purposes because l_r is an expectation value (for a given generative assumption and conditional on a set of constraints) while l_{US} is a limit. Thus, l_{US} is free from assumptions on how empirical networks may have been generated. Relative metrics based on null-models are better suited to answer questions like "why does an empirical network take the observed value l?" or "how surprising is that a network takes pathlength l?" rather than to resolve how short or how long a network is. The US and the UL limits allow us to anchor the values of pathlength and efficiency into well-defined regions of the space of accessible outcomes, which depend on N and L alone. A more telling approach is then to display all the results (both the empirical observations and the expectation values arising from different null-models, e.g., generated by the recent nPSO model[28]) into a unified representation together with the corresponding limits. Figure 4 offers a synoptic way to assess the position every network (real or model) takes in the space of global efficiencies, and thereby discloses the relations needed to interpret the results[29].

In addition to the limits, one may want to obtain the full distribution of pathlength and efficiencies for a given N and L. Such a distribution would be instrumental in answering statistical questions such as "what is the most likely pathlength for networks of size N and L links?". On the one hand, for fixed N and L, the distribution will exist between the boundaries we have reported here. On the other hand, the description of the full distribution still depends on the choice of a generating mechanism, e.g., whether we consider non-isomorphic (non-labelled) networks alone, or isomorphic (labelled) ones. Regarding the likelihood of the US and the UL (di)graph families, for now we can only affirm that the UL limit is a very unlikely outcome in the case of undirected graphs (Fig. 1b) because the configurations are unique: the rewiring of a single link will reduce their pathlength. On the contrary, the US boundary is highly degenerate since any configuration with diameter diam = 2 is US.

The aim of the present paper was the study of the limits for average pathlength and efficiency. We have thus referred to small-world networks in its classical sense, paying attention only to the graph distance. Although the US and UL (di)graphs were not optimised for clustering, US graphs possess a rather small clustering (comparable to random graphs). On the contrary, the clustering of UL graphs grows very rapidly reaching values close to $C = 1$, since most of their links are contained within a complete subgraph. See Supplementary Note 6.

The implications of our results transcend the purely structural study of networks. Given that the underlying patterns of connectivity very much influence dynamic phenomena, e.g., synchrony and diffusion, it remains an open question to investigate the emergent collective behaviour supported by the families of US and UL (di)graphs, and to assess their use as benchmarks for the study of dynamics in real systems.

The limits reported here are the solutions to the most basic optimisation problem for the average pathlength and global efficiency, restricted to the minimal set of constraints (number of nodes and links) inherent to all networks. While the analytic boundaries are well-defined and can be calculated, the configurations are usually degenerate. For example, in Fig. 1a we illustrated two different strategies to construct US graphs: (1) by randomly seeding links to a star graph or (2) by deliberately constructing a rich-club. While the pathlength of these constructs is the same, their dynamic properties may significantly differ.

A class of optimisation problems with practical and economic implications is that of network navigability, which is tightly related to the small-world phenomenon and the travelling salesman problem. In this case, the optimal solutions depend both on the network structure and the routing algorithm. Recent studies

showed that navigability on empirical networks is efficient due to a hidden (hyperbolic) space many real networks are embedded in ref. [30]. Thus optimal navigability with Greedy Routing is achieved when packages are forwarded from the destination to its neighbours at closest hyperbolic distance.[28,31] It remains an open question whether among the diversity of US (di)graphs, some configurations optimise navigability with Greedy Routing against others. Finally, if the networks are embedded in space, e.g., transportation networks, identifying optimal structures implies considering a balance between their efficiency and the resources needed to build the infrastructures, or the costs associated with navigating them[32], as well as the fact that in these cases the definition of shortest path and efficiency take euclidean distances into account[15,33].

In conclusion, network optimisation problems involve the maximisation of a variety of parameters. Our results are the solutions to the simplest case with a minimal set of constraints. They can serve as the starting point to inform more complex problems which include additional constraints beyond N and L.

Methods

Length boundaries of complex networks. We now describe the principles behind the generation of US and UL (di)graphs. The formal mathematical results, including all definitions, propositions and theorems we refer to in this section are found in Supplementary Notes 1 and 2, together with the complete set of equations to analytically calculate the US and the UL boundaries. Supplementary Note 3 includes systematic numerical searches for some UL (di)graph configurations.

Undirected graphs. Our first goal is to generate US graphs, that is, graphs of arbitrary number of nodes and edges with the shortest possible pathlength. This can be achieved by adding edges to a star graph. Indeed, any arbitrary order followed to add edges to a star graph will result in an US graph. Figure 1a illustrates two different examples. One consists of seeding edges at random while in the other links are orderly planted favouring the creation of new hubs; a procedure that would eventually lead to the formation of a rich-club. The reason why the order in which edges are added is irrelevant for the value the average pathlength takes, is that the diameter of the star graph is $\delta = 2$. Any further edge (i,j) added results in converting an entry $d_{ij} = 2$ in the distance matrix to $d_{ij} = 1$. As a consequence, at fixed density, all graphs with diameter $\delta = 2$ are US and have the same pathlength. See the US graph theorem (Theorem 1) for a formal statement, and refs. [9,10,12] for alternative proofs. The pathlength and efficiency of a US graph are given by

$$l_{US} = 2 - \frac{L}{L_o} = 2 - \rho \qquad (3)$$

$$E_{US} = \frac{1}{2}\left(1 + \frac{L}{L_o}\right) = \frac{1}{2}(1 + \rho), \qquad (4)$$

where $L_o = \frac{1}{2}N(N-1)$ and $\rho := L/L_o$ is the density of the network.

To generate connected graphs of arbitrary L with the longest possible pathlength, namely UL graphs, we consider the path graph as a starting point. Any link added to a path graph reduces its diameter, i.e., the distance between the nodes at the two ends. The key is thus to add new links, one-by-one, such that the diameter of the resulting network is minimally reduced at every step. This can be achieved by orderly accumulating all new edges at one end of the chain, Fig. 1b. The procedure creates complete subgraphs of size N_c as L grows, with $N_c = \lfloor \frac{1}{2}(3 + \sqrt{9 + 8(L-N)}) \rfloor$ where $\lfloor \cdot \rfloor$ stands for the floor function. The remainder of the network consists of a tail of size $N_t = N - N_c$. The complete subgraph contains $L_c = \frac{1}{2}N_c(N_c-1)$ edges and the tail $L_t = N_t$. If $L \neq L_c + L_t$, the remaining edges are placed connecting the first node of the tail to the complete subgraph. We find that the average pathlength of an UL graph can be approximated as

$$l_{UL} \approx 2 + \rho - 2\sqrt{\rho} + \frac{N}{3}(1 - 3\rho + 2\rho\sqrt{\rho}). \qquad (5)$$

The approximation improves as N increases, incurring a relative error smaller than 1% for $N > 122$. See Theorem 3 and Supplementary Eqs. (7)–(9) for the exact solutions, and also ref. [9].

So far, we have only considered connected networks. When $L < N - 1$ the shortest architecture (largest possible efficiency) consists of an incomplete star graph of size $N' = L + 1$. This leaves the remaining $N - N'$ nodes isolated, Fig. 1a and Definition 1. We refer to these networks as disconnected US (dUS) graphs (Theorem 2). Once $L \geq N$, the solutions for the most efficient and US graphs are identical (i.e., star graphs with added links).

The construction of disconnected graphs with smallest efficiency is a non-Markovian process. Smallest efficiency is achieved by never having a pair of nodes indirectly connected. This can be realised by forming complete subgraphs which are mutually disconnected (Definition 4 and Proposition 1). In the special cases when $L = \frac{1}{2}M(M-1)$ for $M = 2, 3, \ldots, N$, the network with smallest efficiency consists of a complete subgraph of size M, and $N - M$ isolated nodes, Fig. 1c. The distance between two nodes in the complete subgraph is $d_{ij} = 1$ while all other distances are infinite. Therefore, the efficiency in these cases is exactly $\rho = \frac{L}{L_o}$. The efficiency can also be equal to ρ in intermediate cases (Definition 5 and Proposition 2). We refer to these networks as disconnected UL (dUL) graphs. In summary, the efficiency of dUS and of dUL graphs are given by

$$E_{dUS} = \frac{1}{4L_o}\left[L^2 + 3L\right] : L < (N-1), \qquad (6)$$

$$E_{dUL} = L/L_o = \rho : 0 < L < L_o. \qquad (7)$$

Directed graphs. We will denote the properties of digraphs with a tilde, e.g., $\tilde{L}, \tilde{l}$ and $\tilde{E}$. Following standard notation, we will refer to directed links as arcs. The identification of US and UL digraphs is more intricate because the conditions for a digraph to be connected are more flexible, distinguishing between weakly and strongly connected. We have found three major differences with the results for graphs. (1) The minimally connected digraph is a directed ring (DR) instead of star or path graphs. Thus, DRs are the origin for both US and UL connected digraph families. (2) The construction of US and UL digraphs is often a non-Markovian process. (3) In certain regimes of density more than one configuration compete for the optimal pathlength or efficiency.

The US graph theorem guarantees that any graph with diameter $\delta = 2$ has the shortest possible pathlength regardless of its precise configuration. This result also applies to digraphs (Theorem 4) and thus any set of arcs added to a star graph will lead to an US digraph. The difference is that a star graph contains $\tilde{L} = 2(N-1)$ arcs. Hence, the result holds for $\tilde{L} \geq 2(N-1)$. However, in the range $N \leq \tilde{L} < 2(N-1)$ strongly connected digraphs exist, whose diameter is always larger than two. In this range the digraphs with the shortest pathlength consist of a set of directed cycles overlapping at a single hub, Fig. 1d. We name these networks flower digraphs (Definition 6). Notice that flower digraphs represent the natural transition between a DR and a star graph. The DR is the flower made of a single cycle of length N and a star graph is the flower digraph with $N - 1$ "petals" of length 2. Hence, in this regime US digraph generation is non-Markovian (Proposition 3).

Construction of UL digraphs turns rather intricate and we will provide a partial solution here. Numerical exploration with small networks revealed that, in general, more than one configuration exist. See a summary of all the UL digraphs for networks of $N = 5$ in Supplementary Note 3. The process is divided into two regimes, with a transition happening at $\tilde{L}' = \frac{1}{2}N(N+1) - 1$, or $\tilde{\rho}' = \frac{1}{2} + \frac{1}{N}$. Given a DR with each node i pointing to node $i + 1$ (except the last is pointing to the first) any arc $i \to j$ added in the forward orientation ($i < j$) becomes a shortcut notably reducing the distance between several nodes. Arcs running in the opposite orientation ($i \to j$ with $i > j$) introduce cycles of length $j - i + 1$ which only reduce the distance between the nodes participating in the cycle. Thus the strategy is to add arcs to a DR such that each new arc causes the shortest cycle(s) possible. Despite the intricacy of the problem, a particular subclass of digraphs could be found which are guaranteed to be US. Given an integer M, the optimal configuration with $\tilde{L} = N + \frac{1}{2}M(M-1)$ arcs consist of the superposition of a DR and what we name an M-backwards subgraph or M-BS. An M-BS is formed by the first M nodes of the ring, with each node pointing to all its predecessors, Fig. 1e. Each M-BS contributes to reduce the pathlength of a DR by exactly $\Delta l_M = -\frac{1}{L_o}\frac{M(M-1)}{2}\left(N - \frac{M+4}{3}\right)$, see Lemma 3. After calculating the exact solution for these particular cases [Eqs. (S37)–(S39), and Proposition 4], we find that the pathlength l_{UL} of UL digraphs, of arbitrary $\tilde{L}$, can be approximated by

$$\tilde{l}_{UL} \approx 1 + \frac{3\tilde{\rho}}{2} - \left(\frac{\tilde{\rho}}{3} + 1\right)\sqrt{\frac{\tilde{\rho}}{2}} + N\left(\frac{1}{2} - \tilde{\rho} + \frac{\tilde{\rho}\sqrt{2\tilde{\rho}}}{3}\right). \qquad (8)$$

This approximation is valid when $\tilde{\rho} < \frac{1}{2} + \frac{1}{N}$.

In the particular case when $M = N$ (or $\tilde{\rho} = \frac{1}{2} + \frac{1}{N}$) the first node receives inputs from all other nodes and the last sends outputs to all the network. All the arcs of the original DR have become bidirectional except for the one pointing from the last to the first node, Fig. 1e. Its pathlength is $\tilde{l}_{UL} = \frac{N+4}{6}$, Proposition 5. From this point, any further arc added will create a reciprocal link. Then, the longest pathlength is maintained if the arcs of the M-backwards subgraphs are symmetrised in the same order they were created, (Proposition 6). In the specific cases where $\tilde{L} = \frac{N(N+1)}{2} - 1 + \frac{K(K-1)}{2}$, it is possible to completely bilateralise an M-BS with a K-forward subgraph of K-FS giving (Remark 12)

$$\tilde{l}_{UL} = \frac{4+N}{6} - \frac{K(K-1)}{2\tilde{L}_o}\left(N - \frac{2(K+1)}{3}\right). \qquad (9)$$

Finally, we focus on the efficiency of networks which may be disconnected. Regarding the US boundary up to three different network configurations compete for

the largest efficiency when $\tilde{L} < 2(N-1)$, Fig. 1f, g. One of the routes is non-Markovian. It consists in first creating DRs of growing size until $\tilde{L} = N$ which then naturally continues into flower digraphs. The second route is Markovian and corresponds to the directed version of the disconnected star procedure introduced for graphs. Both routes converge at $\tilde{L} = 2(N-1)$ where a star graph is formed. Figure 1g shows the competition of the three models for largest efficiency for different network sizes. At larger densities, when $\tilde{L} \geq 2(N-1)$, the US theorem applies.

To construct digraphs with minimal efficiency, we seed arcs to an initially empty network such that it contains as many weakly connected nodes as possible. We do so by adding M-backward subgraphs of increasing M to the empty graph, Fig. 1h. The distance matrix of such a digraph contains $\tilde{L}$ entries with $d_{ij} = 1$ and all remaining entries are infinite. Thus, its efficiency is $\tilde{E}_{dUL} = \tilde{L}/\tilde{L}_0 = \tilde{\rho}$, see Proposition 7. Arcs can be seeded following this procedure until $\tilde{L} = \tilde{L}_o/2$, corresponding to the largest M-BS, with $M = N$. At this point, the network consist of the densest possible directed acyclic graph. Any subsequent arc added will introduce at least one cycle. To conserve the lowest efficiency possible, new arcs need to cause cycles with a minimal impact over the path. This is achieved, again, by bilateralising the M-backwards subgraphs in the forward direction (Proposition 9). In these special cases, the efficiency of the digraphs equals their link density: $\tilde{E}_{dUL} = \tilde{\rho}$. Intermediate values of $\tilde{L}$ which do not meet these criteria, may display small departures from $\tilde{E}_{dUL} = \tilde{\rho}$, with the error decreasing as N grows.

Synthetic networks. Random graphs were generated following the random generator usually known as the $G(n, M)$ model, which guarantees all realisations have the same number of links. In our nomenclature $n \rightarrow N$ and $M \rightarrow L$ (or $M \rightarrow \tilde{L}$). Scale-free networks were generated following the method in ref. [13]. A power exponent of $\gamma = 3.0$ was used. The resulting SF digraphs would display correlated in- and out-degrees but not necessarily identical. The range of densities for scale-free networks was restricted to $\rho \in [0.0001, 0.1]$ because for $\rho > 0.1$ the power-law scaling of the degree distribution is lost due to saturation of the hubs. For each value of density an ensemble of 1000 realisations was generated. All synthetic networks were generated using the package `GAlib: a library for graph analysis in Python` (https://pypi.org/project/galib/).

Empirical datasets. The empirical networks employed are well known in the literature and have been often used as benchmarks, except for the local transportation of Chicago which we have assembled for the present manuscript. These datasets represent a heterogeneous sample of networks with a variety of sizes and densities, both directed and undirected, see Table 1 and Supplementary Note 7 for details. Those datasets are available online from different sources. We have constructed the local transportation network of Chicago for the present manuscript by combining the Chicago Transit Authority (CTA) and the METRA commuter rail systems based on the official transportation maps (http://www.transitchicago.com/). The network consists of 376 stations, of them 142 are serviced by the CTA system and 236 by the METRA railroad. We considered two station to be linked also if they were marked as accessible at a short walking distance, giving rise to a total 402 links and a density of $\rho = 0.006$. Since several stations in the network are named the same, an identifier to the line they belong was added.

Data availability

All empirical datasets employed in this work are well known networks available from different sources, except for the local transportation network of Chicago, which we have collected for the present paper. The data is provided as Supplementary Data, in file "TransportChicago.zip".

Code availability

Software to analytically calculate the ultra-short and ultra-long limits or to generate the network families discussed in the paper is available as a Python package (https://github.com/gorkazl/PathLims). The Examples folder contains scripts that allow to reproduce most of the results in Figs. 3 and 4. The package is released under the Apache License, v2.0. PathLims is compatible with the library for graph analysis GAlib (https://pypi.org/project/galib/) which was employed to generate synthetic networks and analyse all data, but it can be used separately.

References

1. Milgram, S. The small-world problem. *Psychol. Today* **2**, 60–67 (1967).
2. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).
3. Newman, M. E. J. The structure and function of complex networks. *SIAM Rev.* **45**(2), 167–256 (2003).
4. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. & Hwang, D.-U. Complex networks: structure and dynamics. *Phys. Rep.* **424**, 175–308 (2006).
5. Humphries, M. D., Gurney, K. & Prescott, T. J. Network 'small-world-ness': a quantitative method for determining canonical network equivalence. *PLoS ONE* **3**, e0002051 (2008).
6. Zamora-López, G., Zhou, C. S. & Kurths, J. Graph analysis of cortical networks reveals complex anatomical communication substrate. *Chaos* **19**, 015117 (2009).
7. Muldoon, S. F., Bridgford, E. W. & Basset, D. S. Small-world propensity and weighted brain networks. *Sci. Rep.* **6**, 22057 (2016).
8. Basset, D. S. & Bullmore, E. T. Small-world brain networks revisited. *Neuroscientist* **23**, 499–516 (2017).
9. Barmpoutis, D. & Murray, R.M. Extremal properties of complex networks. *arXiv* 1104.5532 (2011).
10. Gulyás, L., Horváth, G., Cséri, T. & Kampis, G. An estimation of the shortest and largest average path length in graphs of given density. *arXiv* 1101.2549 (2011).
11. Latora, V. & Marchiori, M. Efficient behaviour of small-world networks. *Phys. Rev. Lett.* **87**, 198701 (2001).
12. Barmpoutis, D. & Murray, R.M. Networks with th smallest average distance and the largest average clustering. *arXiv* 1007.4031 (2010).
13. Goh, K.-I., Kahng, B. & Kim, D. Universal behaviour of load distribution in scale-free networks. *Phys. Rev. Lett.* **87**, 27 (2001).
14. Cohen, R., Erez, K., benAvraham, D. & Havlin, S. Resilience of the internet to random breakdowns. *Phys. Rev. Lett.* **85**, 4626–4628 (2000).
15. Barthélemy, M. Spatial networks. *Phys. Reps.* **499**, 1–101 (2011).
16. Arenas, A., Díaz-Guilera, A., Kurths, J., Moreno, Y. & Zhou, C. Synchronization in complex networks. *Phys. Rep.* **469**, 93–153 (2008).
17. Pei, S. & Makse, H. A. Spreading dynamics in complex networks. *J. Stat. Mechs.* **12**, P12002 (2013).
18. Pastor-Satorras, R., Castellano, C., vanMieghem, P. & Vespignani, A. Epidemic processes in complex networks. *Rev. Mod. Phys.* **87**, 925–979 (2015).
19. Chung, F. & Lu, L. The average distance in random graphs with given expected degrees. *Internet Math.* **1**, 91–113 (2004).
20. Fronczak, A., Fronczak, P. & Holyst, J. Average path length in random networks. *Phys Rev. E* **70**, 056110 (2004).
21. Zamora-López, G. *Linking Structure And Function Of Complex Cortical Networks.* Ph.D. thesis (University of Potsdam, Potsdam, 2009).
22. Zamora-López, G., Zhou, C. S. & Kurths, J. Cortical hubs form a module for multisensory integration on top of the hierarchy of cortical networks. *Front. Neuroinform.* **4**, 1 (2010).
23. vandenHeuvel, M. P. & Sporns, O. Rich-club organization of the human connectome. *J. Neurosci.* **31**, 15775–15786 (2011).
24. Zamora-López, G., Zhou, C. S. & Kurths, J. Exploring brain function from anatomical connectivity. *Front. Neurosci.* **5**, 83 (2011).
25. Sporns, O. Network attributes for segregation and integration in the human brain. *Curr. Opi. Neurobiol.* **23**, 162–171 (2013).
26. Senden, M., Deco, G., deReus, M. A., Goebel, R. & vandenHeuvel, M. P. Rich club organization supports a diverse set of functional network configurations. *NeuroImage* **96**, 174–178 (2014).
27. Zamora-López, G., Chen, Y., Deco, G., Kringelbach, M. L. & Zhou, C. S. Functional complexity emerging from anatomical constraints in the brain: the significance of network modularity and rich-clubs. *Sci. Rep.* **6**, 38424 (2016).
28. Muscolini, A. & Cannistraci, C. V. A nonuniform popularity-similarity optimization (nPSO) model to efficiently generate realistic complex networks with communities. *New J. Phys.* **20**, 052002 (2018).
29. Nishikawa, T., Sun, J. & Motter, A. E. Sensitive dependence of optimal network dynamics on network structure. *Phys. Rev. X* **7**, 041044 (2017).
30. Boguñá, M., Krioukov, D. & Claffy, K. C. Navigability of complex networks. *Nat. Phys.* **5**, 74–80 (2009).
31. Muscolini, A., Thomas, J. M., Ciucci, S., Bianconi, G. & Cannistraci, C. V. Machine learning meets complex networks via coalescent embedding in the hyperbolic space. *Nat. Commun.* **8**, 1615 (2017).
32. Morer, I., Cardillo, A., Díaz-Guilera, A., Prignano, L. & Lozano, S. Comparing spatial networks: A 'one size fits all' efficiency-driven approach. *arXiv* 1807.000565 (2018).
33. Kartun-Giles, A.P., Barthélemy, M. & Dettmann, C.P. The shape of shortest paths in random spatial networks. *arXiv* 1906.04314 (2019).
34. Scannell, J. W. & Young, M. P. The connectional organization of neural systems in the cat cerebral cortex. *Curr. Biol.* **3**, 191–200 (1993).
35. Scannell, J. W., Blakemore, C. & Young, M. P. Analysis of connectivity in the cat cerebral cortex. *J. Neurosci.* **15**, 1463–1483 (1995).
36. Kötter, R. Online retrieval, processing, and visualization of primate connectivity data from the cocomac database. *Neuroinformatics* **2**, 127–144 (2004).
37. Kaiser, M. & Hilgetag, C. C. Nonoptimal component placement, but short processing paths, due to long-distance projections in neural systems. *PLoS Comput. Biol.* **2**, e95 (2006).
38. Hagmann, P. et al. Mapping the structural core of human cerebral cortex. *PLoS Biol.* **6**, e159 (2008).

39. Varshney, L. A., Chen, B. L., Paniagua, E., Hall, D. H. & Chklovskii, D. B. Structural properties of the caenorhabditis elegans neuronal network. *PLoS Comput. Biol.* **7**, e1001066 (2011).
40. Glaiser, P. & Danon, L. Community structure in jazz. *Adv. Complex Syst.* **6**, 565 (2003).
41. Zachary, W. W. An information flow model for conflict and fission in small groups. *J. Anthropol. Res.* **33**, 452–473 (1977).
42. Lusseau, D. et al. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. Can geographic isolation explain this unique trait? *Behav. Ecol. Sociobiol.* **54**, 396–405 (2003).
43. MacRae, D. Direct factor analysis of sociometric data. *Sociometry* **23**, 360–371 (1960).
44. DeDomenico, M., Solé-Ribalta, A., Gómez, S. & Arenas, A. Navigability of interconnected networks under random failures. *Proc. Natl Acad. Sci.* **111**, 8351–8356 (2013).
45. Guimerà, R., Mossa, S., Turtschi, A. & Amaral, L. The worldwide air transportation network: anomalous centrality, community structure and cities' global roles. *Proc. Natl Acad. Sci.* **102**, 7794–7799 (2005).

Acknowledgements

Author contributions
G.Z.-L. and R.B. contributed equally to the design of the study, to the theoretical developments and the mathematical proofs. G.Z.-L. performed the numerical simulations. Both authors wrote the paper.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is avaliable for this paper at this paper at https://doi.org/10.1038/s42005-019-0239-0.

Correspondence and requests for materials should be addressed to G.Z.-Lóp.

Reprints and permission information is available at http://www.nature.com/reprints

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