Spatial network surrogates for disentangling complex system structure from spatial embedding of nodes

Marc Wiedermann,[1,2,*] Jonathan F. Donges,[1,3] Jürgen Kurths,[1,2,4,5] and Reik V. Donner[1]

[1] *Potsdam Institute for Climate Impact Research — P.O. Box 60 12 03, 14412 Potsdam, Germany, EU*
[2] *Department of Physics, Humboldt University — Newtonstr. 15, 12489 Berlin, Germany, EU*
[3] *Stockholm Resilience Centre, Stockholm University — Kräftriket 2B, 114 19 Stockholm, Sweden, EU*
[4] *Institute for Complex Systems and Mathematical Biology, University of Aberdeen — Aberdeen AB24 3FX, UK, EU*
[5] *Department of Control Theory, Nizhny Novgorod State University — Gagarin Avenue 23, 606950 Nizhny Novgorod, Russia*

(Dated: October 1, 2015)

Networks with nodes embedded in a metric space have gained increasing interest in recent years. The effects of spatial embedding on the networks' structural characteristics, however, are rarely taken into account when studying their macroscopic properties. Here, we propose a hierarchy of null models to generate random surrogates from a given spatially embedded network that can preserve global and local statistics associated with the nodes' embedding in a metric space. Comparing the original network's and the resulting surrogates' global characteristics allows to quantify to what extent these characteristics are already predetermined by the spatial embedding of the nodes and links. We apply our framework to various real-world spatial networks and show that the proposed models capture macroscopic properties of the networks under study much better than standard random network models that do not account for the nodes' spatial embedding. Depending on the actual performance of the proposed null models, the networks are categorized into different classes. Since many real-world complex networks are in fact spatial networks, the proposed approach is relevant for disentangling underlying complex system structure from spatial embedding of nodes in many fields, ranging from social systems over infrastructure and neurophysiology to climatology.

PACS numbers: 89.75.-k, 89.75.Kd, 89.75.Hc

I. INTRODUCTION

Many, if not most complex systems that exhibit a network structure are spatially embedded in some metric space [1]. Examples of large interest include on the one hand structural networks, such as social [2], transportation and distribution [3–5], communication [6], or electricity networks [7] with links representing connections between the entities represented by the networks' nodes. On the other hand, functional networks with links indicating functional, mostly statistical, interdependencies between individual nodes have been studied in the context of functional brain [8, 9] or climate networks [10, 11].

A variety of network measures ranging from individual node properties such as degree and shortest-path betweenness to global characteristics such as clustering coefficients and average path length are commonly utilized to quantify the structural properties of a system under study [12, 13]. Many studies aim to classify the investigated networks into different categories, such as small-world networks [14] and subclasses thereof [15, 16] by means of the aforementioned topological characteristics.

In fact, many of the complex systems commonly studied by means of network theoretical methods are in fact spatial networks with nodes and links embedded in some metric space [1], e.g., the Earth's surface for infrastructure or climate networks [17]. Most studies, however, do not take into account the possible influence of a network's spatial structure on its resulting micro- or macroscopic characteristics. Thus it often remains unquantified whether a certain categorization of a network, such as a small-world property, is to some extent already explicable as emerging from the network's spatial embedding alone. Specifically, established random network models that may be used to assess whether a network follows a certain rule of construction solely preserve topological characteristics, such as the link density in *Erdős-Rényi random graphs* [18] or the degree sequence in the *configuration model* [19].

To classify possible types of spatially embedded networks, several models have been proposed that generate random surrogates from a given spatially embedded node sequence by, e.g., randomly distributing links according to the spatial distance between nodes [20, 21], setting a prescribed linking probability between nodes to address boundary effects in climate networks [22] or optimizing the length-dependent costs related to the construction of a link in power grids [23]. These models, however, were primarily designed to assess and reproduce construction principles behind certain types of complex networks and their underlying mechanism are usually tailored to a specific application.

In order to explicitly study the general influence of a network's spatial embedding on its resulting macroscopic characteristics we propose here a set of random network models to create surrogates that preserve certain geographical and topological features of these given

* marcwie@pik-potsdam.de

networks. The surrogates are constructed by iteratively rewiring the original network while preserving a set of its geographical features. In particular, one model, which will be called *GeoModel I* hereafter, preserves, in addition to the degree sequence, the global link-length distribution. A second model referred to as *GeoModel II* additionally preserves for each node the length distribution of the links connected to it and, hence, imposes an even stronger spatial constraint on the rewiring process. The resulting surrogate networks allow for evaluating to what extent observed macroscopic properties of a given network are explicable by geometric constraints inflicted on the network's structure.

We apply our method to a number of real-world complex networks: the US airline network, the US interstate network, the Internet [24], the Scandinavian power grid [25], a world trade network [26], and the road network of a German city (obtained from `http://www.openstreetmap.org`). Additionally, we study the application of our models to a random geometric graph [27] and an Erdős-Rényi network [18]. For comparison, we construct iteratively rewired surrogate networks that only preserve topological characteristics of the given networks, namely the mean degree on the one hand and the degree distribution on the other hand.

Our study reveals that the macroscopic properties of a certain set of networks are only reproduced by applying either of the two geometrically constrained models proposed in this work, while the consideration of topological features alone is not sufficient. Generally, preserving the global link length distribution and, hence, applying GeoModel I already reproduces well the average path length of a given network. In order to additionally reproduce the global clustering coefficient, the per-node link length distributions also need to be taken into account and, hence, the application of GeoModel II becomes necessary.

The remainder of this paper is organized as follows. Section II introduces the algorithms behind the random network models proposed in this work as well as the network characteristics that are used to evaluate their performances. Section III gives an overview on the network data that is investigated and Sec. IV showcases the results of the study. Section V presents our conclusions and an outlook on future directions of research.

II. METHODS

A. Preliminaries

Consider a network $G = (V, E)$ with given sets of nodes (V) and links (E). Each node is labeled with a natural number $i = 1, 2, \ldots, N$, with N being the total number of nodes in the network. The network is represented by its adjacency matrix $\mathbf{A}$ with entries $A_{ij} = 1$, if $\{i, j\} \in E$ and $A_{ij} = 0$ otherwise. Thus, we study here the case of undirected and unweighted networks without self-loops and multiple links between nodes. Additionally, each node is assigned a position $\mathbf{x}_i$ in some metric space of dimension d. In the applications presented in this work, nodes are either embedded on the surface of a sphere, i.e., the Earth's surface, or in a Cartesian coordinate system. In the first case, the position of a node is determined by its latitudinal and longitudinal coordinates λ_i and ϕ_i and, hence, $\mathbf{x}_i = (\lambda_i, \phi_i)$. In the second case, a node's position is given by its Cartesian coordinates x_i and y_i with respect to some arbitrarily chosen origin, $\mathbf{x} = (x_i, y_i)$. The $N \times N$ distance matrix $\mathbf{D}$ then gives the distance between all nodes in the network. For the case of a spherical coordinate system, its entries d_{ij} are computed as the great circle distances between nodes,

$$d_{ij} = R \arccos(\sin\lambda_i \sin\lambda_j + \cos\lambda_i \cos\lambda_j \cos\Delta_{ij}) \quad (1)$$

with $\Delta_{ij} = \phi_j - \phi_i$. R denotes the radius of the sphere, which is rescaled to unit length in all applications and, hence, we set $R = 1$. For the case of a Cartesian coordinate system, the entries of $\mathbf{D}$ are given by the Euclidean distance between two nodes,

$$d_{ij} = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}. \quad (2)$$

From this the local cumulative distribution function $P_i(l)$ of link lengths l of node i follows directly as

$$P_i(l) = \frac{1}{k_i} \sum_j A_{ij} \Theta(l - d_{ij}). \quad (3)$$

Here, $k_i = \sum_j A_{ij}$ is the degree, e.g., the number of neighbors, of a node i and $\Theta(\cdot)$ is the Heaviside function. The global cumulative link length distribution $P(l)$ in the network is then given as

$$P(l) = \frac{1}{2M} \sum_i k_i P_i(l), \quad (4)$$

with $M = |E|$ denoting the total number of links in the network.

B. Complex network characteristics

To characterize the macroscopic structure of the networks under study as well as their corresponding random surrogates, we rely on two commonly used global network measures, the global clustering coefficient $\mathcal{C}$ and the average path length Ł [12]. An evaluation of both measures is commonly used to classify a network under study as, e.g., a small-world network, which is defined to display a high clustering coefficient while at the same time showing a low average path length [14].

Global clustering coefficient. The global clustering coefficient $\mathcal{C}$ gives the probability to find connected triples, i.e., closed triangles formed by links in the network adjacent to a randomly selected node [14]. It is

defined as the arithmetic mean taken over all local clustering coefficients,

$$\mathcal{C} = \frac{1}{N}\sum_i \mathcal{C}_i \tag{5}$$

with

$$\mathcal{C}_i = \frac{1}{k_i(k_i-1)}\sum_{j\neq k} A_{ij}A_{jk}A_{ki}. \tag{6}$$

Note that $\mathcal{C}_i$ is only defined if $k_i > 1$. Otherwise we set $\mathcal{C}_i = 0$.

Average path length. The average path length $\mathcal{L}$ gives the average number of edges along shortest paths between two randomly chosen nodes. Given that $\mathcal{L}_{ij}$ denotes the number of such steps between two nodes i and j the average path length follows as

$$\mathcal{L} = \frac{1}{N(N-1)}\sum_{i\neq j} \mathcal{L}_{ij}. \tag{7}$$

In the case when there exists no path between i and j we set $\mathcal{L}_{ij} = N-1$.

Hamming distance. Consider two undirected networks $G = (V, E)$ and $G' = (V, E')$ with a common set of nodes and the same number of links $M = |E| = |E'|$ represented by adjacency matrices $\mathbf{A}$ and $\mathbf{A}'$. The Hamming distance $\mathcal{H}$ then provides a measure of the dissimilarity between the two sets of links E and E' [28, 29],

$$\mathcal{H} = \frac{1}{4M}\sum_{i,j} |A'_{ij} - A_{ij}| \in [0, 1]. \tag{8}$$

A Hamming distance of $\mathcal{H} = 0$ implies that the two networks G and G' have identical sets of links ($E = E'$), while $\mathcal{H} = 1$ indicates that the two sets of links have entirely dissimilar entries ($E \cap E' = \emptyset$). In the scope of this work, $\mathcal{H}$ is utilized to assess the dissimilarity between a network under study and the surrogate networks that are created from it. Hence, we aim to maximize the Hamming distance between a network and its surrogates while at the same time evaluating the degree of similarity between the global clustering coefficient and average path length of the original and the random networks.

C. Random network models

We generate random network surrogates from a given real-world network by applying four different algorithms. Two of them, random link switching and random rewiring do not take into account any spatial embedding of the network's nodes, whereas this consideration is an explicit part of the two novel models, GeoModel I and GeoModel II.

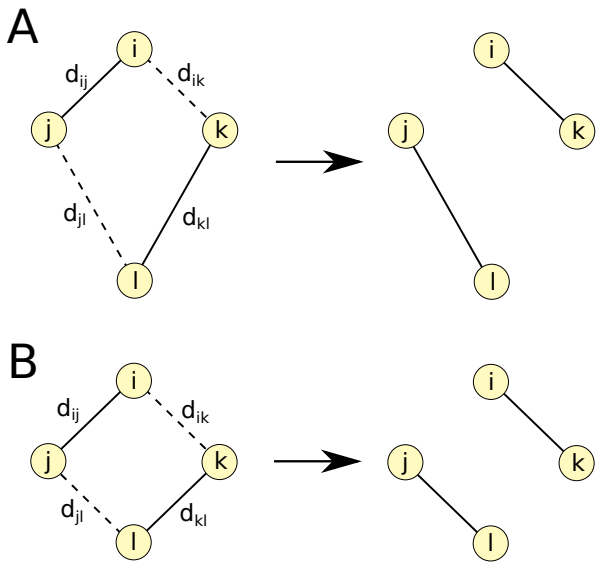

FIG. 1. (Color online) Sketch of the rewiring process that generates randomized surrogates of a given original network by applying GeoModel I (A) and GeoModel II (B). Nodes i, j, k and l are drawn at random, such that i is linked to j and k is linked to l (solid lines) but no links are present between i and k, and j and l (dashed lines). According to the chosen network model the distances $d_{\bullet\bullet}$ between the nodes are evaluated. If the nodes form an approximate kite (A) or diamond (B) with the connections between them present as depicted, previous links are replaced by links connecting i with k and j with l.

The general structure of the algorithm is described as follows. Starting from a copy $\mathbf{A}'$ of the original network's adjacency matrix $\mathbf{A}$:

(i) Draw four distinct nodes i, j, k, l uniformly at random from V.

(ii) Depending on the applied random network model under study, check whether a certain condition $\mathbf{C}$ is TRUE. If $\mathbf{C}$ is FALSE return to step (i).

(iii a) (applies to random rewiring) If $\mathbf{C}$ is TRUE, break the link connecting i with j and establish a link connecting k with l. Hence, $A'_{ij} = A'_{ji} = 1 \to 0$ and $A'_{kl} = A'_{lk} = 0 \to 1$.

(iii b) (applies to all other random network models) If $\mathbf{C}$ is TRUE, break the links connecting i with j and k with l and establish links connecting i with k and j with l. Hence, $A'_{ij} = A'_{ji} = 1 \to 0$, $A'_{kl} = A'_{lk} = 1 \to 0$, $A'_{ik} = A'_{ki} = 0 \to 1$, and $A'_{jl} = A'_{lj} = 0 \to 1$.

(iv) As long as a certain number of rewirings r is not reached return to (i) with the modified adjacency matrix $\mathbf{A}'$.

The resulting modified copy $\mathbf{A}'$ of the original network's adjacency matrix $\mathbf{A}$ is then returned for further evaluation. In the following, we introduce the explicit form of the conditions $\mathbf{C}$ for a rewiring process to take place.

1. Random Rewiring

Random rewiring, the simplest case, takes place if a link exists between the randomly drawn nodes i and j, but no link exists between k and l. Hence,

$$\mathbf{C} = A'_{ij} \wedge \neg A'_{kl}. \tag{9}$$

The definition of $\mathbf{C}$ is then plugged into step (ii) and (iii a) in the above algorithm and depending on its value a new set of four nodes is drawn or a rewiring process takes place. Random rewiring solely preserves the average degree $K = N^{-1}\sum_i k_i$ in the network and, hence, after sufficiently many rewiring steps, the resulting surrogate network converges to an Erdős-Rényi random graph [18].

2. Random Link Switching

In addition to the mean degree K, random link switching preserves the local degree of each node in the network as well, but still neglects any aspect of a network's spatial embedding [30]. Hence, for the four nodes drawn in step (i) of the construction algorithm, we need to ensure that i is linked with j and k is linked with l, but i and k as well as j and l are not yet connected (Fig. 1). Hence, the condition $\mathbf{C}$ reads

$$\mathbf{C} = \mathbf{C}_1 = A'_{ij} \wedge A'_{kl} \wedge \neg A'_{ik} \wedge \neg A'_{jl}. \tag{10}$$

As the degree of each node is preserved, the resulting surrogate networks relate to the results one would obtain from the configuration model [19]. However, for the present case the surrogate networks display no self-loops or multiple links between nodes.

3. GeoModel I

In addition to the above criterion $\mathbf{C}_1$, GeoModel I aims to also preserve the global link length distribution $P(l)$. Hence, the potentially newly established links must be of the same length as those that are removed from the network. This means that the four randomly drawn nodes i, j, k and l must form a kite with exactly one link present on each of the two sides of the same length (Fig. 1A). Since the nodes are usually embedded in a continuous domain this equality can only be fulfilled up to a certain accuracy. We hence demand that the newly established links have approximately the same length as the existing ones with some tolerance ϵ. In other words, the difference in lengths between the present and potentially established links should not exceed a certain fraction ϵ of the existing links' lengths. Thus, in addition to $\mathbf{C}_1$ the following condition must be fulfilled,

$$\mathbf{C}_2 = \Theta(\epsilon d_{ij} - |d_{ij} - d_{ik}|) \wedge \Theta(\epsilon d_{kl} - |d_{kl} - d_{jl}|) \tag{11}$$

$$\mathbf{C} = \mathbf{C}_1 \wedge \mathbf{C}_2. \tag{12}$$

Thus, ϵ measures the maximum allowed relative deviation in length between the existent and potentially newly established links. GeoModel I preserves the degree distribution in the same way as random link switching, but in addition approximately preserves the global link length distribution $P(l)$. The ensemble Ω_{GMI} of possible surrogates constructed by GeoModel I therefore forms a subset of the ensemble Ω_{rls} of all those surrogates possibly constructed from random link switching, $\Omega_{GMI} \subseteq \Omega_{rls}$. Generally, it is to be expected that with an increasing ϵ the Hamming distance $\mathcal{H}$ between the original networks and its surrogates increases. However, an increase in ϵ also induces larger deviations between the link length distributions of the original and surrogate networks. Hence, we aim to estimate the maximum meaningful value of ϵ by using a Kolmogorov-Smirnoff (KS) test [31], demanding that for an ensemble of n surrogate networks the resulting link length distributions are statistically indistinguishable from that of the original network in 95% of all cases under a confidence level of $\alpha = 95\%$ (see Appendix for details).

4. GeoModel II

In order to not only preserve the global but also the local, (per-node) link length distributions $P_i(l)$, we demand that the two links to be removed and the two links to be established all have approximately the same length, and hence, the nodes i, j, k and l form a diamond. That way, none of the lengths of links emerging from either of the four nodes is changed under each rewiring step. As above, in most situations this criterion can only be fulfilled approximately and we utilize, for convenience, the same parameter ϵ to extend the conditions $\mathbf{C}_1$ and $\mathbf{C}_2$ by

$$\mathbf{C}_3 = \Theta(\epsilon \max(d_{ik}, d_{jl}) - |d_{ik} - d_{jl}|) \tag{13}$$

$$\mathbf{C} = \mathbf{C}_1 \wedge \mathbf{C}_2 \wedge \mathbf{C}_3. \tag{14}$$

Thus, the difference in length of the newly established links (and therefore also the difference in lengths of the existing links) must not be larger than a certain fraction ϵ of their respective maximum length. For our studies we decided to depict the maximum of d_{ik} and d_{jl} as the scale of the tolerance window. However, other choices, such as the minimum value or the arithmetic mean of the two, might also be possible and would result in different optimal values of the tolerance parameter ϵ. A detailed investigation on the effect of the actual definition of the link length that is chosen as a reference remains as a subject of future research.

Again, the ensemble Ω_{GMII} of all possible surrogates constructed from GeoModel II forms a subset of all possible surrogates constructed from GeoModel I and random link switching since it only imposes a further condition in addition to the already employed ones, $\Omega_{GMII} \subseteq \Omega_{GMI} \subseteq \Omega_{rls}$.

Name	N	M	K	ρ	$\mathcal{C}$	$\mathcal{L}$	ϵ_I	ϵ_{II}	Grid Type
US airline	190	837	8.86	0.0466	0.679	2.176	0.04	0.07	Spherical
Internet	13.372	28.253	4.23	0.0003	0.423	3.630	0.04	0.01	Spherical
US interstate	935	1.315	2.82	0.0030	0.107	20.207	0.17	0.24	Spherical
Scandinavian power grid	236	318	2.71	0.0115	0.084	9.156	0.16	0.27	Cartesian
World trade	186	7.043	76.14	0.4094	0.815	1.594	0.02	0.04	Spherical
Urban roads (Eschwege)	855	1.174	2.75	0.0032	0.050	18.313	0.15	0.22	Spherical
Random geometric graph	2.000	5.493	5.50	0.0027	0.588	30.428	0.11	0.13	Cartesian
Erdős-Rényi graph	2.000	5.493	5.50	0.0027	0.003	4.643	0.01	0.01	Cartesian

TABLE I. Overview of all networks investigated in this study including their number of nodes N and links M, average degree K, link density ρ, global clustering coefficient $\mathcal{C}$, and average path length $\mathcal{L}$. ϵ_I and ϵ_{II} denote the relative tolerances that are chosen for generating random network surrogates from GeoModel I and GeoModel II, respectively.

III. DATA

We consider different real-world networks to illustrate the performance of our algorithms and test to what extent macroscopic characteristics are recaptured by random network surrogates that take into account spatial constraints on the distribution of links in the network. We first investigate three different previously studied infrastructure networks [24]: the US airline network with nodes displaying airports and links indicating flights scheduled between them, the US interstate network with links representing highways and nodes serving as country borders, termination points and intersections between highways, and the Internet with nodes corresponding to autonomous systems around the globe where links stand for data connections between them. Contrasting the case of the interstate network, we also study an infrastructure network of smaller spatial scale by retrieving the urban road network of a German small-town (Eschwege) from `www.openstreetmap.org` (accessed 2012-01-30). Here, nodes again represent intersections and links are roads connecting them. Moreover, we apply our framework to the Scandinavian power grid, where links represent high voltage transmission lines and nodes are transformation stations or power plants [25]. These types of networks have been intensively studied in the framework of complex network theory and the understanding of their global properties has been reported as crucial since these strongly determine their local behavior, e.g., the robustness to failures of single nodes [32–34]. Finally, we study the world trade network of 2009 with nodes representing the center of a country and links indicating trade between them [26] as a representative of a non-physical, yet spatially embedded transaction network.

For comparison with these real-world networks, we also study synthetic networks with known properties, which serve as a benchmark for our analysis. Particularly, we consider a random geometric graph with nodes put randomly on a plain unit square [35, 36]. All nodes with a spatial distance of less than 0.03 are connected to yield a manageable density of links. We expect that this network's macroscopic properties are only explainable by considering random network models that take into account the spatial embedding of the nodes. For the sake of comparison, we construct one realization of an Erdős-Rényi random graph with the same number and position of nodes and the same number of links randomly put between them as in the random geometric graph. As links are put without any relation to spatial distances, the simplest network model, i.e., random rewiring, should already capture this network's macroscopic features.

A summary of all networks included in this study together with each network's number of nodes N and links M as well as further network parameters is presented in Tab. I.

IV. RESULTS

We now apply the four random network models introduced above to the different real-world and synthetic networks under study. In a first step we illustrate how to estimate a proper tolerance parameter ϵ for GeoModel I and GeoModel II. Specifically, we illustrate the procedure for the example of the US interstate network and the application of GeoModel I. We then discuss in detail the results of all four network models applied to the US interstate and the airline network and show to what extent macroscopic network characteristics are reproduced by the different network models. Finally, we present a comprehensive intercomparison between all networks investigated in this study by applying the different models to each real world network. We evaluate, which macroscopic features of a network can be reproduced by which model and sort the real-world networks into different classes, i.e., those for which spatial embedding plays a minor role when estimating macroscopic properties and those where the spatial structure explicitly needs to be taken into account.

For all cases discussed from now on we construct an ensemble of $n = 100$ surrogate networks for each network under study and iteratively rewire each random model for $r = 20M$ steps.

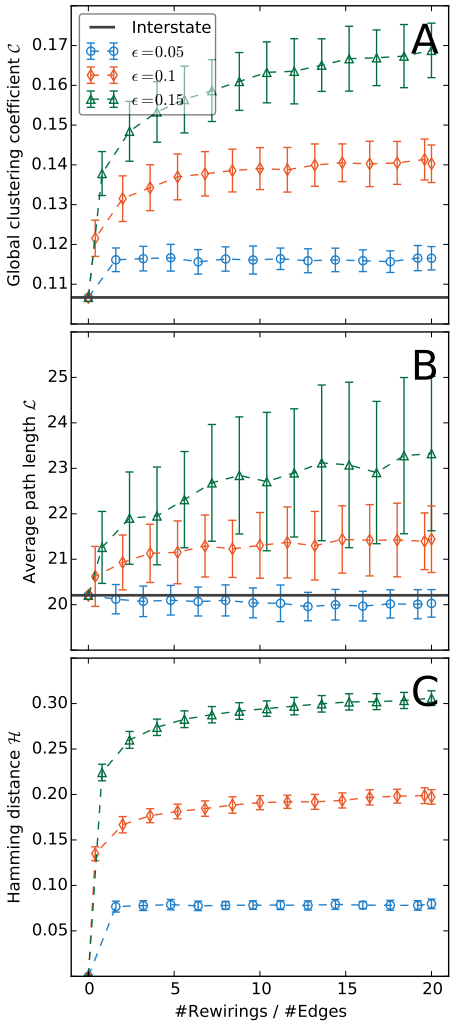

FIG. 2. (Color online) Evolution of global clustering coefficient $\mathcal{C}$ (A) and average path length Ł (B) with the number of rewirings for an ensemble of $n = 100$ surrogates generated from the interstate network by applying GeoModel I and using different tolerances ϵ_I. Solid lines indicate the respective value of $\mathcal{C}$ and Ł of the interstate network itself. (C) The Hamming distance $\mathcal{H}$ between the surrogate networks and the original network. Scatter symbols denote the mean value and error bars indicate one standard deviation of each measure.

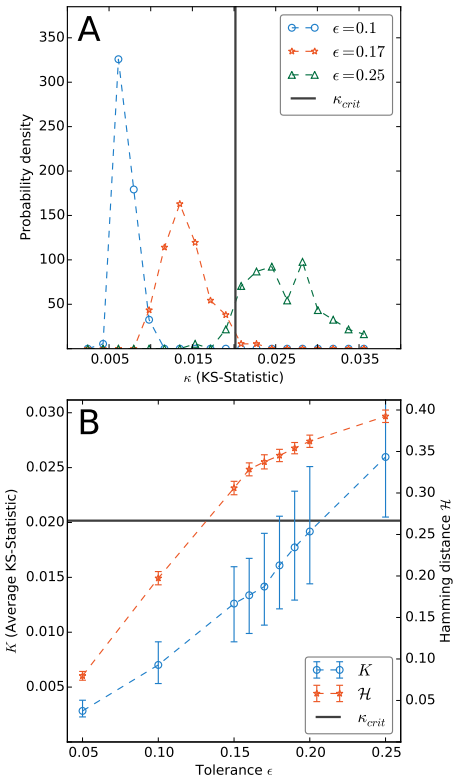

FIG. 3. (Color online) (A) Distribution of KS statistics κ measuring the maximum distance in the global cumulative link length distribution $P(l)$ between the interstate network and each of the $n = 100$ network surrogates obtained by applying GeoModel I under different tolerances ϵ and $20M$ rewirings. (B) The average KS statistics K and Hamming distance $\mathcal{H}$ after $20M$ successful rewirings depending on the choice of tolerance ϵ. Error bars denote one standard deviation for the Hamming distance and the 5th and 95th percentile of the distribution of KS statistics. The solid line indicates the critical value κ_{crit} below which the surrogates' and the original network's link length distribution are considered statistically indistinguishable under a confidence level of $\alpha = 0.95$.

A. Estimation of the tolerance parameter

The only free parameter in GeoModel I and GeoModel II is the tolerance parameter ϵ in Eqs. (11) and (13), that determines which link lengths are treated as being sufficiently similar. To illustrate the influence of ϵ on our results, we apply GeoModel I to the US interstate network and create $n = 100$ surrogate networks that display the same degree sequence and approximately the same global link length distribution $P(l)$ as the original network. Figure 2 shows the mean evolution of global clustering coefficient $\mathcal{C}$, average path length Ł and Hamming distance $\mathcal{H}$ for different choices of ϵ. As expected, we note that for the lowest choice of ϵ ($\epsilon = 0.05$) the surrogate networks' $\mathcal{C}$ and Ł are closest to the values for the original network (Fig. 2A,B). However, in that case, the Hamming distance displays low values around 0.075 meaning that 92.5% of links in the original network are also present in the surrogate networks (Fig. 2C). With increasing ϵ the values of $\mathcal{H}$ also increase and, hence, the surrogate networks become increasingly dissimilar from the original network. At the same time $\mathcal{C}$ and Ł also differ more from their target values (Fig. 2A,B).

As GeoModel I aims to approximately preserve the global link length distribution $P(l)$ we examine also the distribution of the KS statistics κ for the ensemble of surrogate networks at different tolerance parameters ϵ (Fig. 3A). For low values of ϵ, all cumulative link length distributions are statistically indistinguishable with 95%

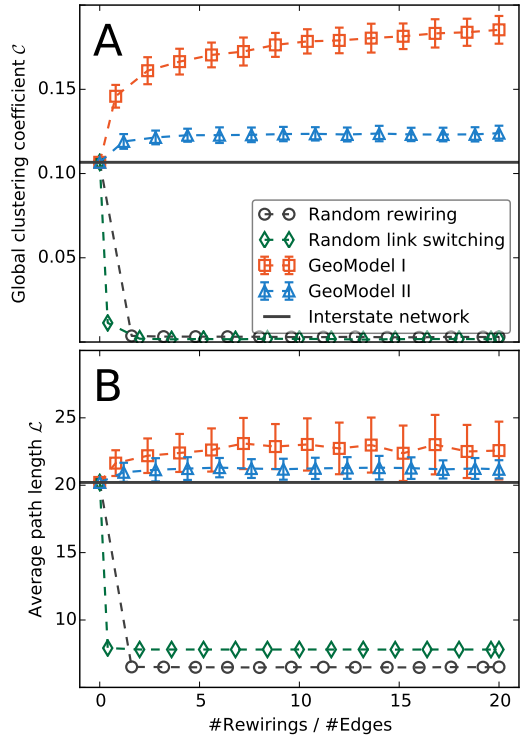

FIG. 4. (Color online) Evolution of (A) global clustering coefficient $\mathcal{C}$ and (B) average path length Ł with the number of rewirings averaged over ensembles of $n = 100$ surrogates generated from the interstate network by applying the different random network models (dashed lines). For GeoModel I and GeoModel II the tolerances are set to $\epsilon_I = 0.17$ and $\epsilon_{II} = 0.24$, respectively. Scatter symbols denote the mean value of each measure. Error bars indicate one standard deviation and are shown if their size exceeds that of the symbol. Solid lines indicate the value of $\mathcal{C}$ and Ł in the original network.

confidence, which results in values of κ being smaller than the critical value κ_{crit}. This value indicates the upper bound of the confidence interval (Fig. 3A) and is determined as the largest possible value that satisfies Eq. (A.2). However, as already discussed above, the Hamming distance $\mathcal{H}$ becomes very low for low ϵ and only a few links differ between the original and the surrogate networks (Fig. 3B). On the other hand, for large ϵ most link length distributions are dissimilar under the desired confidence level and, hence, the purpose of GeoModel I is not fulfilled. We find that for $\epsilon = 0.17$, 95% of all distributions are statistically equivalent with 95% confidence and, hence, GeoModel I achieves its highest possible Hamming distance (Fig. 3).

Following the same procedure, the optimal tolerance can be obtained for GeoModel II as well as for all other networks under study. It is important to note that the values of ϵ generally differ between the two random network models as for GeoModel II the additional criterion $\mathbf{C}_3$ must be fulfilled. We therefore denote ϵ_I the optimal tolerance for GeoModel I and ϵ_{II} the respective optimal tolerance for GeoModel II. A summary of all tolerances for each network and random network model is given in Tab. I. We note that the obtained values of ϵ differ between $\epsilon = \epsilon_{II} = 0.01$ for the Internet and $\epsilon = \epsilon_{II} = 0.27$ for the power grid. Moreover, in most cases we find that $\epsilon_I < \epsilon_{II}$.

The heterogeneity in the distribution of links in the original network seems to seems to play a crucial role for the resulting value of ϵ. Further studies on the interplay between the tolerance parameter and the shape of the cumulative distribution function of link lengths as well as the number of nodes N and links M should be addressed in future research. Further, we note that the length distributions of the networks under study are not necessarily symmetric. Thus, more advanced statistical tests such as Anderson-Darling or Shapiro-Wilk tests with potentially larger power than the Kolmogorov-Smirnoff test might improve the assessment of the statistical equivalence between the surrogates' and the original networks' distributions [37]. An assessment of these quantifiers is, however, beyond the scope of this work and remains as a subject of future studies.

B. Interstate network

With the two tolerances $\epsilon_I = 0.17$ and $\epsilon_{II} = 0.24$ estimated for applying GeoModel I and GeoModel II to the interstate network, we now investigate the evolution of $\mathcal{C}$ and Ł with an increasing number of rewiring steps for the four different random network models (Fig. 4). Generally, we note that random rewiring and random link switching converge towards a state where there is hardly any further fluctuation in the evolution of $\mathcal{C}$ and Ł after less than $2M$ steps of rewiring (Fig. 4). Similarly, GeoModel II converges after $5M$ steps. Only for GeoModel I, we note small fluctuations in the average evolution of Ł (Fig. 4B) and a slow saturation in the average evolution of $\mathcal{C}$ (Fig. 4A) up to the maximum value of $r = 20M$ steps of rewiring.

We note that surrogate networks obtained from random rewiring and random link switching do not capture well the macroscopic characteristics of the interstate network indicated by large deviations of $\mathcal{C}$ and Ł from their original values (Fig. 4A,B). In fact, with respect to the global clustering coefficient $\mathcal{C}$, the two models perform equally badly (Fig. 4A). For the average path length Ł, the additional constraint of a preserved degree sequence when applying random link switching yields a slight improvement over the process of random rewiring as in average the surrogate networks' values of Ł are closer to that of the original network.

Additionally taking into account spatial constraints on the lengths of links in the random networks and, hence, applying GeoModel I and GeoModel II yields macroscopic characteristics of the surrogates that are much closer to those of the original network (Fig. 4). Specif-

ically, GeoModel I already estimates a value of Ł very close to that of the original interstate network (Fig. 4B), while the estimated value of $\mathcal{C}$ still diverges strongly from its target (Fig. 4A). The additional constraint of a preserved local link length distribution $P_i(l)$ overcomes this issue and GeoModel II provides surrogate networks that, in addition to Ł, also approximate $\mathcal{C}$ in good agreement with the original network. However, slight differences in the two quantities estimated by GeoModel II in comparison with the original network's characteristics are still present. Additionally constraining the algorithm to also preserve a network's degree-degree correlation [30] might further improve the agreement between the surrogates and the original network. An investigation of such higher order effects remains as a subject of future research. We also note that GeoModel I and GeoModel II tend to overestimate the values of Ł and $\mathcal{C}$ at least for the particular case of the interstate network. This effect might be related to optimization principles, such as the minimization of intersection crossings for road networks, underlying the original network that are not accounted for by the surrogate networks' construction mechanism. Future studies should address in more detail, for what types of networks GeoModel I and GeoModel II over- or underestimate the respective target values of, e.g., global clustering coefficient and average path length.

C. Airline network

We now apply the same procedure as discussed before to the airline network and compute the evolution of global clustering coefficient $\mathcal{C}$ and average path length Ł with an increasing number of rewiring steps for the four different random network models (Fig. 5). We note a fast convergence towards a state with no more fluctuations in the average evolution of $\mathcal{C}$ and Ł for all four random network models. As for the interstate network, we find that random rewiring does not produce surrogate networks, which capture well the macroscopic characteristics of the airline network. However, in contrast to the former case, random link switching already reproduces very well both macroscopic quantities $\mathcal{C}$ and Ł. GeoModel I and GeoModel II slightly improve these results, but for the present case of the airline network a prescribed degree sequence already produces surrogate networks with properties close to those of the original network. Thus, for the airline network the spatial embedding of the nodes and the resulting characteristic distribution of link lengths is of less importance for its macroscopic properties as compared to the US interstate network.

D. Intercomparison between different spatial networks

As in the previous sections, we now compute for each of the networks under study (Tab. I) the evolution of global

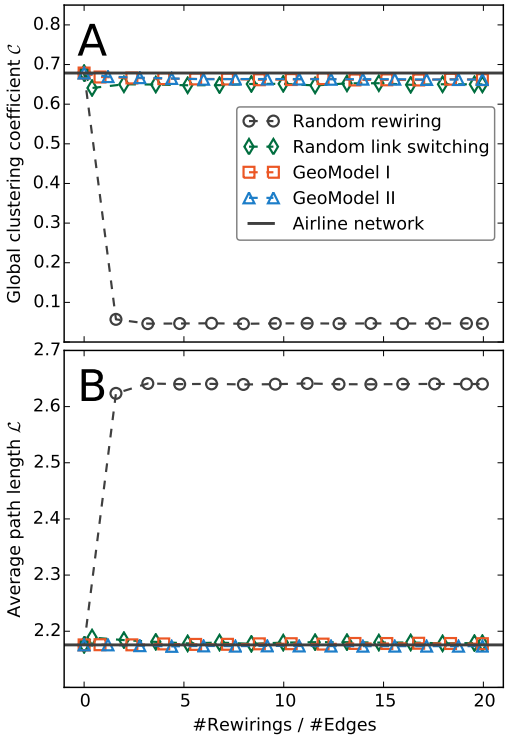

FIG. 5. (Color online) Same as Fig. 4 but for the airline network and tolerances of $\epsilon_I = 0.04$ and $\epsilon_{II} = 0.06$. Error bars are not shown as they do not exceed the size of the symbols.

clustering coefficient $\mathcal{C}$ and average path length Ł by evaluating ensembles of $n = 100$ realizations of each network model and using $r = 20M$ rewiring steps. To give a comprehensive summary, we compute for each network and network model the average relative deviation $\Delta\mathcal{C}$ and ΔŁ from the respective original network's values,

$$\Delta\mathcal{C} = \frac{\langle \mathcal{C}_{sur} \rangle - \mathcal{C}_{orig}}{\mathcal{C}_{orig}} \tag{15}$$

$$\Delta\text{Ł} = \frac{\langle \text{Ł}_{sur} \rangle - \text{Ł}_{orig}}{\text{Ł}_{orig}}. \tag{16}$$

Figure 6 summarizes the results for all spatial networks under study. As expected, the Erdős-Rényi network's topological features are already well reproduced by random rewiring, while all other network surrogates display large deviations from its original values (Fig. 6A). In all cases, the global clustering coefficients $\mathcal{C}$ of the surrogate networks are lower than those of the respective original networks (indicated by negative values of $\Delta\mathcal{C}$ in Fig. 6A), which is in accordance with the fact that networks generated from random rewiring are expected to display a clustering coefficient close to their link density [38]. Remarkably, the average path length of the world trade network is also already well reproduced by random rewiring (resulting in ΔŁ close to zero in Fig. 6A), which is likely

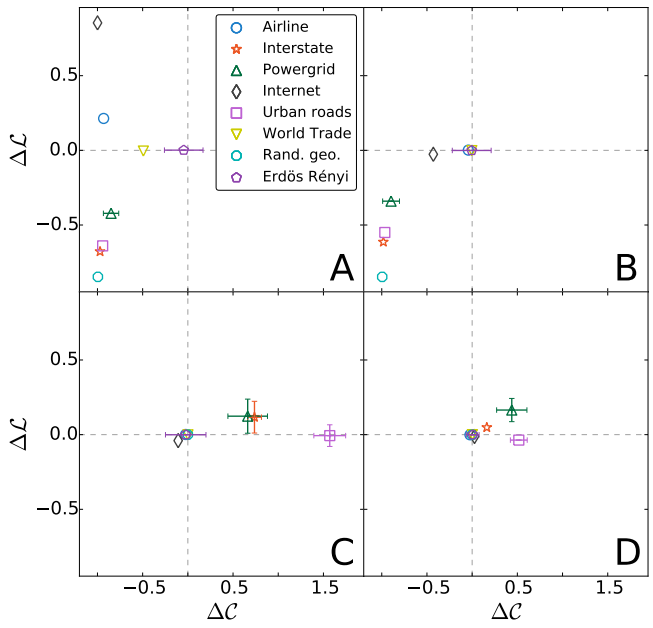

FIG. 6. (Color online) Average relative deviation of global clustering coefficient $\Delta\mathcal{C}$ and average path length $\Delta Ł$ from the respective original values computed over an ensemble of $n = 100$ surrogate networks after $20M$ successful rewirings by applying (A) random rewiring, (B) random link switching, (C) GeoModel I and (D) GeoModel II. The tolerances ϵ_I and ϵ_{II} used for each network and random network model are shown in Tab. I. Error bars denote the standard deviation in $\Delta\mathcal{C}$ and $\Delta Ł$ and are shown if their size exceeds that of the corresponding symbol.

due to its large link density of $\rho \approx 0.4$. We note that for the Internet and the airline network, the randomly rewired surrogates overestimate the average path length Ł, while for all remaining networks, this quantity is underestimated.

As discussed in Section IV C, the process of random link switching reproduces well the macroscopic properties of the airline network (Fig. 6B). The same observation also holds for the world trade and, as expected, for the Erdős-Rényi network. Additionally, the average path length of the Internet is already well captured by random link switching, too. Thus, the topological features of these networks are already well expressed in terms of their degree distribution and the spatial embedding of the nodes seems to have little influence on the average path length Ł and the global clustering coefficient $\mathcal{C}$. For the four other networks (power grid, urban roads, interstate, and random geometric graph), only slight improvements are visible when comparing the relative deviations $\Delta\mathcal{C}$ and $\Delta Ł$ obtained by applying random rewiring with those for random link switching (compare Fig. 6A and Fig. 6B).

Additionally taking the effects of the nodes' spatial embedding into account, we find that GeoModel I generates random surrogates of all networks under study for which the average path length Ł already becomes very close to its original value (Fig. 6C). However, while for the airline, Internet, world trade and Erdős-Rényi networks, the surrogates are also in good agreement with respect to deviations in the global clustering coefficient, we observe that GeoModel I still overestimates the values of $\mathcal{C}$ for the remaining networks. Thus, for the aforementioned networks, the global link length distribution $P(l)$ already determines the expected value of the average path length Ł, while the global clustering coefficient $\mathcal{C}$ is not yet explained sufficiently.

This mismatch is, however, to a large extent addressed by the usage of GeoModel II (Fig. 6D). We now find for all networks a lowering of the deviation in $\mathcal{C}$ as compared with the application of GeoModel I (compare Fig. 6C and Fig. 6D). This means that ultimately, in addition to the global link length distribution $P(l)$, the local link length distribution $P_i(l)$ predetermines in most cases and to a large extent the value of the global clustering coefficient $\mathcal{C}$.

In summary, we find a class of networks (including the Erdős-Rényi, airline, world trade and Internet network) for which random network models that do not take into account any spatial embedding of the nodes already generate surrogates with global clustering coefficients and average path lengths similar to those of the original networks. For a second class of networks (the power grid and random geometric graph as well as the interstate and urban road network) only taking the spatial structure of the original network explicitly into account in terms of GeoModel I and/or GeoModel II produces surrogates with global clustering coefficients and average path lengths comparable with those of the respective original networks. Remarkably, we find that GeoModel I serves to reproduce well the average path lengths of the aforementioned networks, while only the application of GeoModel II produces network surrogates for which the global clustering coefficient becomes also close to the respective original network's value.

We emphasize that the first class of networks, which includes the airline network, is generally non-planar. In contrast, those networks where spatial embedding is found to have a larger influence on macroscopic properties are almost or even fully planar. This hints to a direct relationship between the properties studied in this work and the planarity of networks for which a further investigation remains as a subject of future research.

V. CONCLUSION

We have introduced two novel models to generate random surrogates of a given spatial network that preserve either the global or the local distribution of link lengths between individual nodes and, hence, explicitly take into account the embedding of the network in some metric space. We have characterized the macroscopic properties of the resulting surrogates by means of the global

clustering coefficient and the average path length and compared these values to those of the original networks from which the surrogates were constructed. For reference, we have utilized iterative random rewiring and random link switching to produce random networks similar to Erdős-Rényi networks and the configuration model, respectively.

We have found that for a certain class of spatial networks random link switching already produces surrogates of comparable macroscopic structure as the original network. Thus, for these networks the spatial embedding of the nodes and links is not crucial for explaining their corresponding macroscopic properties. In contrast, we have identified another class of networks for which global clustering coefficients and average path lengths are only well reproduced when applying the newly introduced GeoModel I and/or GeoModel II that explicitly account for the spatial embedding of the nodes. Hence, for these networks information their geometric properties is needed to sufficiently explain their macroscopic structure. For the latter class of networks, we have found that their average path length can already be well reproduced by GeoModel I, while only using GeoModel II enables to also capture the global clustering coefficient to a large extent. Our findings align well with recent studies on the effect of the networks' spatial embedding on the small-world property of a system [39]. We confirmed that the two quantities that are commonly assessed when determining whether a network displays the small-world property are in many cases to a large extent already predetermined by the spatial distances between its nodes.

In summary, the surrogate network models introduced in this work provide an important step in assessing whether and to which extent global characteristics of a complex network are already predetermined by statistics associated with the spatial embedding of its nodes and links. For future work it would be of great importance to study in more detail which classes of networks are explicitly affected by the nodes' spatial embedding and which are already sufficiently quantified by some structural quantities such as the degree distribution. We observed that the optimal tolerance parameter ϵ (the only parameter of the models we introduced here) varies strongly depending on the specific networks under study. An assessment of the interplay between the networks' known topological properties and the estimated values of ϵ could help to directly estimate an optimal tolerance circumventing the need for the iterative process and the evaluation of KS statistics that is applied in the present work. Additionally, it is of interest to extend the models presented in this work to also conserve degree-degree correlations [30] and to be also applicable to weighted networks, such as airline networks, where the weight of each link scales with the number of passengers on the corresponding connection. For this purpose, our models could be combined with existing models for non-spatially embedded weighted networks [40] that follow a similar strategy of constrained rewiring of a given network as the models presented in this work.

ACKNOWLEDGMENTS

MW and RVD have been supported by the German Federal Ministry for Education and Research (BMBF) via the Young Investigators Group CoSy-CC^2 (grant no. 01LN1306A). JFD thanks the Stordalen Foundation and BMBF (project GLUES) for financial support. JK acknowledges the IRTG 1740 funded by DFG and FAPESP. MT Gastner is acknowledged for providing his data on the airline, interstate, and Internet network. P Menck thankfully provided his data on the Scandinavian power grid. We thank S Willner on behalf of the entire zeean team for providing the data on the world trade network. All computations have been performed using the Python package `pyunicorn` [41] that is available at `https://github.com/pik-copan/pyunicorn`.

[1] M. Barthélemy, Phys. Rep. **499**, 1 (2011).
[2] A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli, Phys. Rev. E **74**, 036116 (2006).
[3] J. R. Banavar, A. Maritan, and A. Rinaldo, Nature **399**, 130 (1999).
[4] S. H. Y. Chan, R. V. Donner, and S. Lämmer, Eur. Phys. J. B **84**, 563 (2011).
[5] A. Jarvis, S. Jarvis, and N. Hewitt, Earth Syst. Dynam. Discuss. **6**, 133 (2015).
[6] S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin, Nature **464**, 1025 (2010).
[7] L. Buzna, L. Issacharoff, and D. Helbing, International Journal of Critical Infrastructures **5**, 72 (2009).
[8] S. Achard, R. Salvador, B. Whitcher, J. Suckling, and E. Bullmore, J. Neurosci. **26**, 63 (2006).
[9] C. Zhou, L. Zemanov, G. Zamora, C. C. Hilgetag, and J. Kurths, Phys. Rev. Lett. **97**, 238103 (2006).
[10] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths, Eur. Phys. J. Spec. Top. **174**, 157 (2009).
[11] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths, Europhys. Lett. **87**, 48007 (2009).
[12] M. Newman, SIAM Rev. **45**, 167 (2003).
[13] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. Hwang, Phys. Rep. **424**, 175 (2006).
[14] D. J. Watts and S. H. Strogatz, Nature **393**, 440 (1998).
[15] L. a. N. Amaral, A. Scala, M. Barthélemy, and H. E. Stanley, PNAS **97**, 11149 (2000).
[16] A.-L. Barabasi, Science **325**, 412 (2009).
[17] A. A. Tsonis, K. L. Swanson, and P. J. Roebber, Bull. Amer. Meteor. Soc. **87**, 585 (2006).
[18] P. Erdős and A. Rényi, Publ. Math. Inst. Hung. Acad. Sci **5**, 17 (1960).
[19] M. Molloy and B. Reed, Random Struct. Alg. **6**, 161

(1995).
[20] L. Barnett, E. Di Paolo, and S. Bullock, Phys. Rev. E **76**, 056115 (2007).
[21] J. Heitzig, J. F. Donges, Y. Zou, N. Marwan, and J. Kurths, Eur. Phys. J. B **85**, 1 (2012).
[22] A. Rheinwalt, N. Marwan, J. Kurths, P. Werner, and F.-W. Gerstengarbe, EPL **100**, 28002 (2012).
[23] P. Schultz, J. Heitzig, and J. Kurths, Eur. Phys. J. Spec. Top. **223**, 2593 (2014).
[24] M. T. Gastner and M. E. J. Newman, Eur. Phys. J. B **49**, 247 (2006).
[25] P. J. Menck, J. Heitzig, J. Kurths, and H. J. Schellnhuber, Nat. Commun. **5** (2014).
[26] M. Lenzen, K. Kanemoto, D. Moran, and A. Geschke, Environ. Sci. Technol. **46**, 8374 (2012).
[27] M. Penrose, *Random Geometric Graphs. Vol. 5.* (Oxford University Press, Oxford, 2003).
[28] A. Radebach, R. V. Donner, J. Runge, J. F. Donges, and J. Kurths, Phys. Rev. E **88**, 052807 (2013).
[29] R. Hamming, AT&T Tech. J. **29**, 147 (1950).
[30] G. Zamora-Lpez, V. Zlati, C. Zhou, H. tefani, and J. Kurths, Phys. Rev. E **77**, 016106 (2008).
[31] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, Vol. 2 (1996).
[32] R. Albert, I. Albert, and G. L. Nakarado, Phys. Rev. E **69**, 025103 (2004).
[33] P. Crucitti, V. Latora, and M. Marchiori, Physica A **338**, 92 (2004).
[34] R. Kinney, P. Crucitti, R. Albert, and V. Latora, Eur. Phys. J. B **46**, 101 (2005).
[35] J. F. Donges, J. Heitzig, R. V. Donner, and J. Kurths, Phys. Rev. E **85**, 046105 (2012).
[36] C. Herrmann, M. Barthlemy, and P. Provero, Phys. Rev. E **68**, 026128 (2003).
[37] N. M. Razali and Y. B. Wah, Journal of Statistical Modeling and Analytics **2**, 21 (2011).
[38] R. Albert and A.-L. Barabsi, Rev. Mod. Phys. **74**, 47 (2002).
[39] S. Bialonski, M.-T. Horstmann, and K. Lehnertz, Chaos: An Interdisciplinary Journal of Nonlinear Science **20**, 013134 (2010).
[40] G. Ansmann and K. Lehnertz, Phys. Rev. E **84** (2011).
[41] J. F. Donges, J. Heitzig, B. Beronov, M. Wiedermann, J. Runge, Q. Y. Feng, L. Tupikina, V. Stolbova, R. V. Donner, N. Marwan, H. A. Dijkstra, and J. Kurths, arXiv:1507.01571 [physics] (2015).

Appendix: Kolmogorov-Smirnoff Test

Given two cumulative distribution functions of link lengths $P(l)$ and $P'(l)$ the Kolmogorov-Smirnoff (KS) statistic κ is given as

$$\kappa = \max_{0<l<\infty} |P(l) - P'(l)|. \tag{A.1}$$

The two distributions are equal at a confidence level α if [31]

$$Q_{KS}([M_e + 0.12 + 0.11/M_e]\kappa) > \alpha. \tag{A.2}$$

Here, $M_e = M/2$ is the effective number of links constituting each distribution and Q_{KS} is given as,

$$Q_{KS}(x) = 2\sum_{j=1}^{m}(-1)^{j-1}\exp(-2j^2x^2). \tag{A.3}$$

In theory, the above sum has infinitely many entries, $m = \infty$. In this work we set $m = 100$ to obtain an acceptable approximation.