Regular Article

Node-weighted measures for complex networks with spatially embedded, sampled, or differently sized nodes

J. Heitzig[1], J.F. Donges[1,2], Y. Zou[1,3,a], N. Marwan[1], and J. Kurths[1,2,4]

[1] Potsdam Institute for Climate Impact Research, Transdisciplinary Concepts and Methods, P.O. Box 60 12 03, 14412 Potsdam, Germany
[2] Department of Physics, Humboldt University Berlin, Newtonstr. 15, 12489 Berlin, Germany
[3] Department of Electronic and Information Engineering, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
[4] Institute for Complex Systems and Mathematical Biology, University of Aberdeen, Aberdeen AB 24 UE, UK

Abstract. When network and graph theory are used in the study of complex systems, a typically finite set of nodes of the network under consideration is frequently either explicitly or implicitly considered representative of a much larger finite or infinite region or set of objects of interest. The selection procedure, e.g., formation of a subset or some kind of discretization or aggregation, typically results in individual nodes of the studied network representing quite differently sized parts of the domain of interest. This heterogeneity may induce substantial bias and artifacts in derived network statistics. To avoid this bias, we propose an axiomatic scheme based on the idea of *node splitting invariance* to derive consistently weighted variants of various commonly used statistical network measures. The practical relevance and applicability of our approach is demonstrated for a number of example networks from different fields of research, and is shown to be of fundamental importance in particular in the study of spatially embedded functional networks derived from time series as studied in, e.g., neuroscience and climatology.

1 Introduction

1.1 Motivation

In the last decades, network and graph theory have successfully been applied to various kinds of complex systems, and many different measures have been defined to study their structural and topological properties. Most of these are of a combinatorial nature, based on counts of certain nodes, links, triangles, paths, etc. (for an overview, see, e.g., [1–6]).

Often, a typically finite set of nodes of the studied network is either explicitly or implicitly considered representative of a much larger finite or infinite set of objects of interest (which we will call the *domain of interest* or in sampling contexts the *population*), either by being a somehow selected or sampled subset of this larger set or, more often, by constituting some kind of discretization, aggregation, or coarse-graining of it. Typical examples are networks of

1. functional connections between differently sized *regions of interest* (ROIs) in the human brain, as in [7–9] and Figure 1,
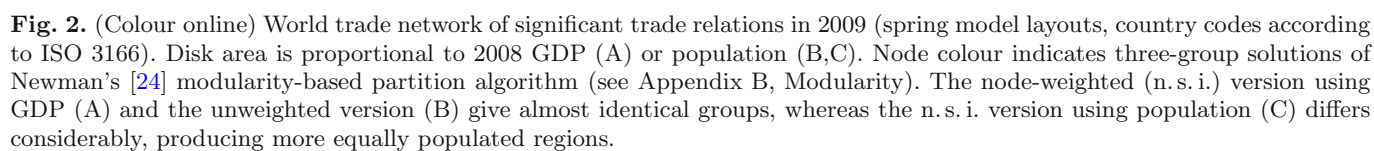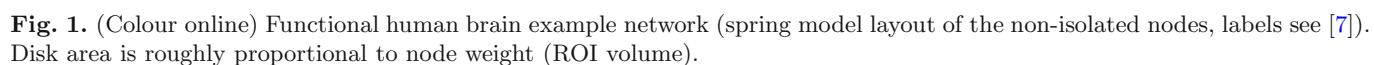2. dynamical couplings or statistical associations between time-series measured at discrete regular grid points, in irregular mesh cells, or at otherwise sampled *discrete locations* on some manifold (e.g., a climate network either using a latitude-longitude-regular grid on the Earth's surface, as in [10,11], or with meteorological stations as nodes),
3. routing connections between *autonomous systems* (AS's) in the internet, representing groups of individual servers, users or ranges of IP addresses, as in [12–15],
4. cross-references between *articles* containing different amounts of content in an online encyclopedia, as in [16,17],
5. social relationships between *households* consisting of different numbers of individuals, as in [18,19],
6. trade relationships between *countries* with differing gross domestic product and representing different numbers of consumers, as in [20] and Figure 2,
7. proximities between *sampled state vectors* in the (reconstructed) phase space of a dynamical system [21–23], sampled at irregular points in time.

Depending on the meshing, sampling, or parcellation method, the chosen level of aggregation or description, and the availability of data, some parts of the domain of interest might be represented by relatively more nodes than others, e.g.: the polar or densely populated regions on the Earth's surface in a climate network (since grid

a e-mail: yong.zou@pik-potsdam.de

Fig. 1. (Colour online) Functional human brain example network (spring model layout of the non-isolated nodes, labels see [7]). Disk area is roughly proportional to node weight (ROI volume).

Fig. 2. (Colour online) World trade network of significant trade relations in 2009 (spring model layouts, country codes according to ISO 3166). Disk area is proportional to 2008 GDP (A) or population (B,C). Node colour indicates three-group solutions of Newman's [24] modularity-based partition algorithm (see Appendix B, Modularity). The node-weighted (n.s.i.) version using GDP (A) and the unweighted version (B) give almost identical groups, whereas the n.s.i. version using population (C) differs considerably, producing more equally populated regions.

points cluster at the poles and meteorological stations cluster in populated areas); the subcortical area in the human brain (when this is parcellated into smaller regions); the younger AS's in the internet (usually having below-average numbers of users); the more technical subjects in the encyclopedia (usually being organized into shorter articles); the childless population of a village (having a higher ratio of households per people); in the trade network the industrialized world when the interest is in consumers (consisting of more countries per population) or the non-industrialized world when the interest is in GDP (consisting of more countries per GDP); the more densely sampled time periods in the dynamical system. Often, this problem of over-representing some parts of the domain of

interest is directly related to the size distribution of the objects chosen as nodes. That distribution is frequently heavy-tailed, e.g., for AS's in the internet, articles in the encyclopedia, and countries in the trade network.

The described representation bias can cause serious pitfalls in the interpretation of results obtained from the selected network, if one wants to make inferences about structural and topological properties of the domain of interest (i.e., in the above examples, about all locations on the globe or in the brain, all users of the internet, all units of content in the encyclopedia, all persons in the village, all consumers, or all time points on the trajectory, respectively). From both a statistical and an approximation-theoretical point of view it is therefore important to first decide which structural or topological properties of the domain of interest we are interested in (e.g., the connectivity distribution or amount of clustering), and then to determine what measure in the selected network could be used as a good estimate or approximation of these properties of the underlying domain of interest. Often, the network construction (that is, the choice of nodes and links) also involves some parameters like sampling density, grid origin, orientation and size, mesh size, or link inclusion thresholds, and there are often systematic influences of these parameters on the results of any measurements in the resulting network. This can lead to selective bias, as in the case of the dorsal cingulate gyrus in the brain network, whose betweenness value (a popular measure of node importance) depends very much on whether it is treated as one node DCG or as two nodes DCG.L and DCG.R (see Sect. 5.3). The effect can even lead to completely artificial features like in the climate network, as depicted in the centre of Figures 3A, 3B.

Because of the above observations, our purpose in this paper is to improve the estimation or approximation power of common network measures by introducing into their calculation a suitable kind of weighting of all individual nodes. For this, we use *node aggregation weights* based on, e.g., ROI volume, inverse grid density, mesh cell size, inverse sampling density, IP address ranges of AS's, article length, household size, or a country's population or GDP. A suitable choice of weights is sometimes difficult (e.g., for AS's or countries) and the weights may have to be estimated (e.g., in case of sampled state vectors). Our focus in this paper, however, is not on the derivation of suitable weights, but on how to make proper use of them once they are given.

To avoid confusion, we emphasize that there exists a theory of "weighted networks" [25] in which *links* instead of nodes have weights representing quantities like length or capacity. But that kind of weights and the related theory is of little help here since the type of situation we are concerned with calls for *node* weights instead, and we will see that the corresponding measures differ from those in the other theory, even if the node weights are somehow translated into link weights. Actually, some real-world networks (like the world trade network) are probably best described as having *both* node and link weights, as well as being directed. In this paper, we are treating the case
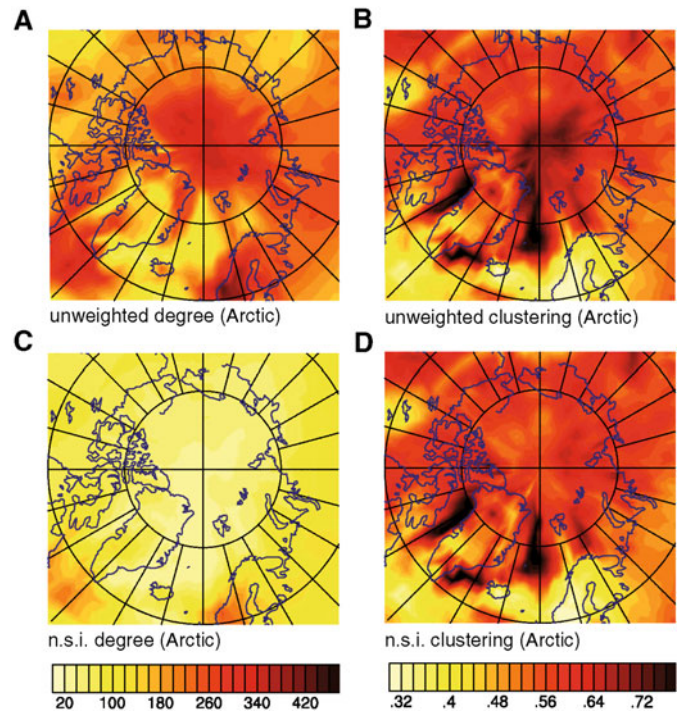

Fig. 3. (Colour online) Comparison of unweighted and weighted (n. s. i.) versions of degree (A, C) and clustering coefficient (B, D) in the northern polar region (Lambert equal area projection) of a global climate network representing correlations in temperature dynamics. The high values at the pole in (A, B) turn out to be an artefact of the increasing grid density toward the pole, as demonstrated by (C, D).

of undirected networks with node weights, but the same methodology can easily be applied to transform network measures that already make use of link weights or link direction into node-weighted versions that use node weights as well.

1.2 Climate networks

To exemplify our approach, assume that the domain of interest is the set of all points on the Earth's surface and the significant linear correlations between the surface air temperature time series at pairs of such points. This can be interpreted as a "network" (mathematically, an infinite simple graph) with uncountably many nodes and (unknown) links. However, we only observe data for a finite subset of points, say those 64 082 regular grid points that have integer latitude and longitude degrees (which is a fairly common grid type in Earth sciences). Then the set of significant linear correlations between the temperature time series at these sampled points defines a (finite) *climate network* [10,11,26–31] whose properties, as measured by common network statistics, are somehow hoped to be representative of related properties of the temperature dynamics on the whole globe. E.g., the degree of a node in the finite network corresponds to how much surface area this point is correlated to in the whole domain of interest. When degree is computed in the standard way, however,
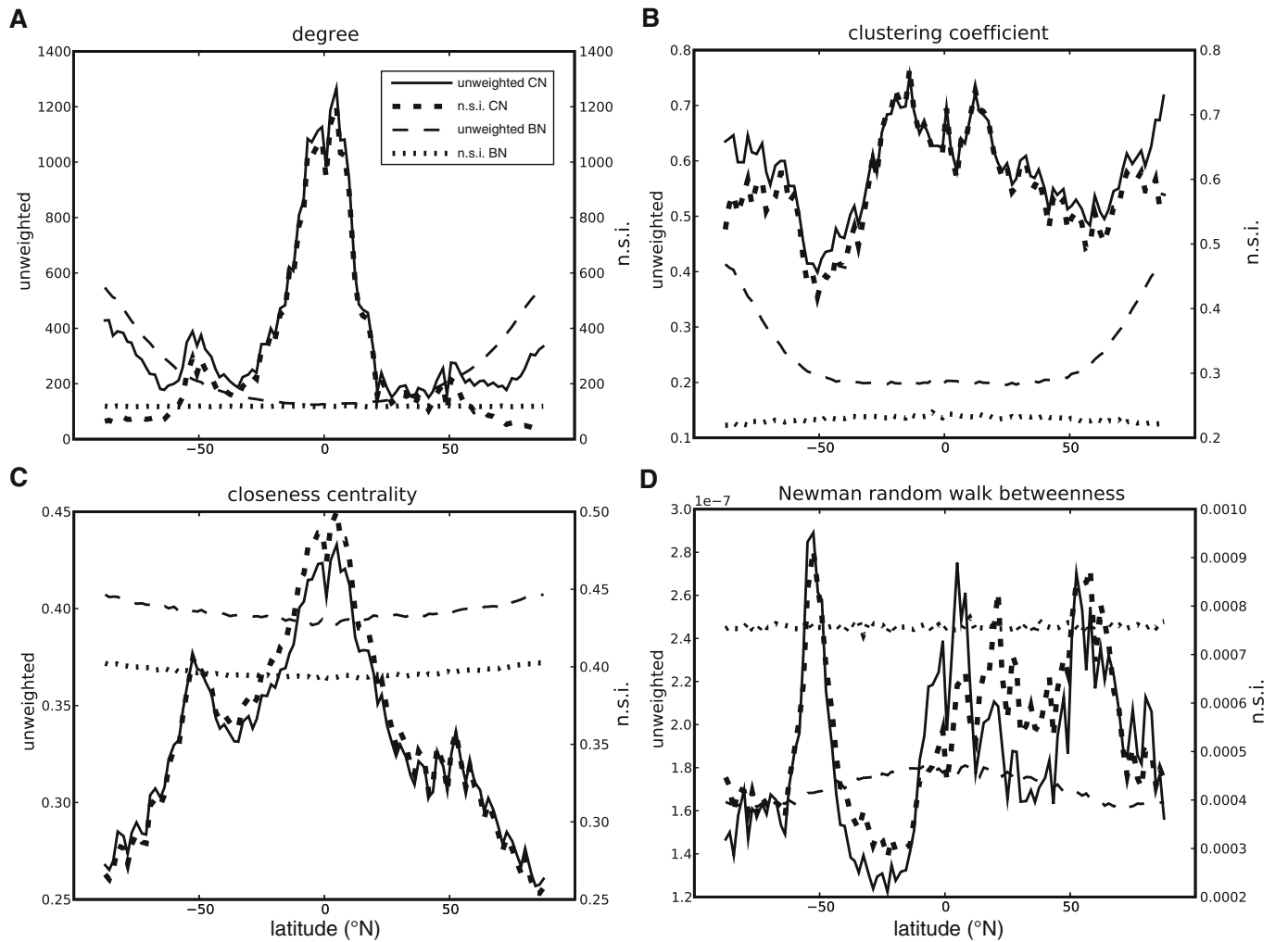
Fig. 4. Comparison of unweighted and n. s. i. versions of four common local network measures (see Sect. 5 for definitions) in a real-world climate network (CN) and a benchmark network (BN). (A) degree k_v, k_v^*, (B) clustering coefficient C_v, C_v^*, (C) closeness centrality CC_v, CC_v^*, and (D) Newman's random walk betweenness NB_v, NB_v^*. Measures are averaged along bands of equal latitude and plotted against latitude. Real-world global climate network representing correlations in surface air temperature dynamics. Benchmark network defined on the same grid with independent link probabilities depending on distance alone. The benchmark lines show that the observed increase in the unweighted degree and clustering coefficient near the poles at ±90° latitude is mainly due to the vanishing node weight of cos(latitude), whereas the effect on closeness centrality is much smaller. In case of Newman's random walk betweenness, the slight increase of the unweighted version towards the equator in the benchmark network reflects the fact that those nodes represent larger surface areas, hence a random walk on the globe will cross this area more often. This explains in part why in the unweighted version in the real-world network the central peaks are more prominent than in the n. s. i. version.

the resulting large regional differences will mostly reflect the strongly differing amounts of surface area each node represents (small area per node at the poles, large area per node on the equator), instead of indicating "real" regional differences in the connectivity of the underlying domain of interest (the temperature field).

The significance of observed features can often be assessed by comparing results with those obtainable in a similar "benchmark" network in which the links have been replaced by a spatially homogeneous link distribution, which was done, e.g., in [32] to show how the underlying geometry of a cortical network influences network statistics. Similarly, for our climate network, a perfectly

homogeneous link distribution in the domain of interest (like the one in which two points are linked iff their angular distance is less than five degrees) would lead to regional differences of the degree distribution in the grid-based network, just because its node density and therefore its link density increase towards the poles (Fig. 4, dashed lines). For the same reason, the standard local clustering coefficient shows an artificial increase towards the poles. Figures 3A, 3B shows such artifacts in a real-world climate network similar to that of [11]. We will see below that this effect can be avoided by using a node weight proportional to the inverse node density (= cosine of latitude), and by using the sum of all neighbour's weights as a measure of

degree (this measure being called *area-weighted connectivity* in [26]), and the weighted proportion of interlinked pairs of neighbours of a node as a measure of clustering, instead of the classical degree measure and clustering coefficient.

Figures 3C, 3D shows the meaningful regional differences in the example real-world climate network that remain after the artifacts have been removed in this way. In that example, we can check the validity of this by comparing these results to those obtained from a climate network based not on a latitude-longitude grid but on a "geodesic" grid that has an approximately homogeneous node density all over the globe (see also [33], results not shown here for brevity). Using the classical degree and clustering measures in the latter, homogeneously sampled network gives results almost identical to those in Figures 3C, 3D rather than 3A, 3B. Such a change of grid, however, usually requires some kind of interpolation of the available data, which can introduce other problems and is obviously only possible for specific kinds of network constructions.

1.3 Outline

Surprisingly, to our best knowledge, the by now vast literature on complex networks contains almost no node-weighted measures, except for the above-mentioned area-weighted connectivity measure, whereas other techniques (e.g., finite elements) and methods of data analysis (e.g., empirical orthogonal functions, a weighted version of principal components analysis [34]) often use some weighting or other adjustment to avoid similar biases or artifacts.

In this paper, we therefore present a fairly *general strategy* for deriving node-weighted versions of network measures that can be expected to give estimates or approximations of properties of the domain of interest that are in a certain sense consistent whenever the network is of the type in which the links can be interpreted as indicating some kind of similarity or "close" relationship, as it is more or less the case in all the cited examples. The approach will not be directly useful for other kinds of networks, e.g., if links represent some kind of "complementarity" instead of similarity, like in most bipartite networks. Also, it will not be applicable when only the network as a whole can be considered "representative" of the domain of interest, but when individual nodes cannot be considered representative of some well-defined part of it, as will often be the case when network sampling methods such as random node or link sampling or snowball sampling are used (related estimation problems are treated in [6,35]).

We then apply this strategy to many of the commonly used network measures and illustrate the effects in a number of example networks from the above list. Fortunately, there is a quite simple pragmatic approach to finding useful weighted versions of network measures which allows us to postpone a more detailed analysis of the statistical estimation or numerical approximation properties for further research. This approach is *axiomatic* rather than analytic in that it requires our measures to fulfil an easily verified condition of *node splitting (or twin merging) invariance.*

After stating preliminary matter in Section 2 and giving more details on our illustrative example networks (Sect. 3), we will introduce this concept formally and shortly relate it to a statistical and approximation interpretation in Section 4. We then proceed with presenting a comprehensive set of according network measures in Sections 5, 6, illustrating each one's effect in those example networks for which the respective network measure has been considered important in the literature, but not aiming at analysing each example network with the full set of measures. We end with a more detailed description of an application to climate networks in Section 7 and a conclusion (Sect. 8). In two Appendices (online), we present some additional measures, give versions of the new measures that allow for a simpler comparison with their unweighted counterparts, and shortly discuss how the related parameter of *typical weight* can be estimated.

2 Preliminaries

Let $G = (\mathcal{N}, \mathcal{E})$ denote a finite undirected simple graph (the *network* under consideration) with known *node* or *vertex set* $\mathcal{N}$, *edge* or *link set* $\mathcal{E} \subseteq \{\{i, j\} : i \neq j \in \mathcal{N}\}$, and *adjacency matrix* $\mathsf{A} = (a_{ij})_{i,j \in \mathcal{N}}$, where $a_{ij} \in \{0, 1\}$, and $a_{ij} = 1$ iff $\{i, j\} \in \mathcal{E}$. For simplicity, we assume that $\mathcal{N} = \{1, \ldots, N\}$ for some natural number $N > 1$. The *neighbours* of a node $v \in \mathcal{N}$, i.e., the members of v's *(punctured) neighbourhood*

$$\mathcal{N}_v = \{i \in \mathcal{N} : a_{iv} = 1\} = \{i \in \mathcal{N} : a_{vi} = 1\} \quad (1)$$

are those nodes that v is directly *linked* to by an edge. We will also use the *extended adjacency matrix* $\mathsf{A}^+ = (a_{ij}^+)_{i,j \in \mathcal{N}} = \mathsf{A} + \mathsf{I}$ with

$$a_{ij}^+ = a_{ij} + \delta_{ij}, \quad (2)$$

where $\mathsf{I} = (\delta_{ij})_{ij}$ is the identity matrix and δ the Kronecker symbol, with $\delta_{ij} = 1$ for $i = j$ and $\delta_{ij} = 0$ for $i \neq j$. Moreover, we will need the *extended* or *unpunctured neighbourhood*[1]

$$\mathcal{N}_v^+ = \{i \in \mathcal{N} : a_{iv}^+ = 1\} = \mathcal{N}_v \cup \{v\}. \quad (3)$$

In addition, we will assume that each node v is assigned a positive real-valued *(aggregation) weight* w_v. As many of our measures will involve unweighted or weighted means over nodes or pairs of nodes, we also introduce the *total weight*

$$W = \sum_{i \in \mathcal{N}} w_i \quad (4)$$

---

[1] Note that in classical mathematical topology, the term "neighbourhood of a point" implies that the point itself is included, otherwise one speaks of a "punctured" neighbourhood. In the network literature, the term "neighbour" also sometimes refers to nodes not directly linked. In our terminology, a node is not a neighbour of itself but still a member of its unpunctured neighbourhood.

and a shorthand notation for (weighted) averages of functions of nodes or node pairs:

$$\langle g(v) \rangle_v = \frac{1}{N} \sum_{v \in \mathcal{N}} g(v),$$

$$\langle g(v) \rangle_v^w = \frac{1}{W} \sum_{v \in \mathcal{N}} w_v g(v),$$

$$\langle h(i,j) \rangle_{ij} = \frac{1}{N^2} \sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{N}} h(i,j), \quad \text{and}$$

$$\langle h(i,j) \rangle_{ij}^w = \frac{1}{W^2} \sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{N}} w_i h(i,j) w_j. \tag{5}$$

Most of the remaining notation will follow the reviews of Newman [1], Boccaletti [2], and da F. Costa [3]. In cases where a measure is commonly normalized using a factor of $1/(N-1)$, $1/(N-1)(N-2)$, etc., we will use instead the factors $1/N$, $1/N^2$ etc., to keep things simpler.

3 Examples of networks with nodes of different size

As stated in previous sections, node weights are ubiquitous for graph representations of complex systems. We will show the applications of our weighted measures to various example networks, ranging from the human brain, the internet, over Wikipedia to world trade. Firstly we describe the details of the constructed weighted networks. The focus application to climate networks will be presented individually in Section 7.

3.1 Human brain

Functional Magnetic Resonance Imaging (fMRI) time series have been widely used to study neural activities in the brain from a network perspective. A node is represented by a cortical region of interest, while a link is often characterized by some statistical association measuring the correlation between different regions (i.e., linear Pearson correlation, nonlinear mutual information, or frequency dependent correlation by Wavelets [7], etc.). We consider two regions are connected if their correlation exceeds a threshold, which can be either based on the correlation values or in terms of probability under some appropriate null hypothesis. The resulting functional connectivity of nervous systems has been shown to display high clustering and short path length which confers a capability for both specialized or modular processing in local neighbourhoods and distributed or integrated processing over the entire network [7,9]. The cerebral cortex is a thin folded sheet tightly confined by the skull and is thus an archetypal example of a complex network that is strongly constrained by geometry. The understanding of the properties captured by a variety of network measures (i.e., high clustering, short path length, motifs, and modularity, etc.) has been pointed out to be very limited because the role of the spatial geometry has been largely underestimated [32].

Many anatomical features show distance-dependent properties, e.g., the density of corticocortical neural connections, volume, processing steps, signal travel times, and genetic encoding needed to specify connectivity.

We re-examine a version of the network of functional connections between differently sized Region of Interests (ROIs) in the human brain as it was described in [7]. After some appropriate preprocessing on the acquired fMRI data, the network consists of 90 cortical and subcortical time series extracted from each individual. The resulting graph is shown in Figure 1, where the weight of a node is represented by its associated volume of the ROIs.

3.2 Internet

A well-known type of internet mapping is obtained by considering so-called *autonomous systems (AS's)*. On the AS level, each node represents an AS while each link between two nodes represents the existence of a peer connection among the corresponding AS's in Border Gateway Protocol (BGP) routing tables. Traditionally, much interest is in identifying a possible power law $P(k_v > x) \sim x^{1-\gamma}$ for the degree distribution [12,13,15]. Power laws have been reported by different studies of AS maps, and their exponents seem to be stable over a number of years [12,15], which could help to devise a novel class of dynamical models of the internet. However, most studies use BGP data collected by the Oregon route views project only, which may provide an incomplete picture of the internet connectivity [14].

Whereas in the literature, usually each AS is treated the same despite the considerable differences in size (three orders of magnitude), we associate here with each AS a node weight proportional to the size of the IP address space allocated to that AS in terms of Classless Inter-Domain Routing (CIDR) prefixes, a common measure of network size that can be used as an approximation of the fraction of the internet represented by that AS (although better measures might be possible). In other words, we consider the set of all IP addresses allocated by CIDR as our domain of interest G_0 which we study be means of a network G of AS's that each represent a certain part of G_0. To construct our network, we used the January 2010 BGP routing table snapshot (http://archive.routeviews.org/oix-route-views/\discretionary-2010.01/oix-full-snapshot-2010-01-27-1200.bz2) from the Oregon route views project, and a corresponding CIDR prefix allocation snapshot from (http://www.cidr-report.org/as2.0/aggr.html), giving a network of over 30 000 nodes.

3.3 Wikipedia

Wikipedia is an intriguing research object from a sociologist's point of view: nodes are articles which are published by a number of independent individuals in various languages, edges are reference hyperlinks which cover topics they consider relevant. Several models for the growth of

Wikipedia have been proposed to mimic the plausible preferential attachment mechanisms that might explain the apparent scale-freeness of the resulting networks [16,17].

As another example network, we used as nodes all 33 359 articles containing the word "physics" from the 30 July 2010 snapshot of the English language version of Wikipedia, and made an undirected link between two articles if either references the other, resulting in an average degree of 31.9. Other authors study directed Wikipedia networks [16,17], but treating it as an undirected network can be partially justified by the fact that in Wikipedia one can follow references backwards using the function "what links here". Since the individual articles represent quite different amounts of text (between one and 283 kB), it is straightforward to use their size in characters as the node weights w_v.

3.4 World trade

Finally, we also consider for illustration a network of countries where two countries are linked when they trade considerable amounts (similar to [20]), e.g., if the total value of their mutual reported imports and exports in 2009 accounted for at least 10% of the total reported foreign trade value of at least one of the two countries, based on data from comtrade.un.org. Such a network is shown in Figure 2. The topological characterization of the world trade web (WTW) is of primary interest for the modeling of crisis propagation at the global level, and it has been reported in [20] that the unweighted WTW displays some typical properties of complex networks, i.e., scale free degree distribution, small-world properties, and high clustering coefficients. As node weights, we use either population or gross domestic product in 2008 (as reported by the IMF), both showing considerable differences.

A much more realistic model of the world trade network would of course use weighted and directed links representing actual imports and exports, in addition to node weights, so one cannot attach much real-world importance to the exemplary results that we will present here for illustration with this simplified network.

4 Approaches to node weighting in network measurements

4.1 Statistical interpretation

In many cases the nodes $\mathcal{N}$ and links $\mathcal{E}$ of the studied network G are simply a subset from a larger (maybe infinite) network G_0 with nodes $\mathcal{N}_0$ and unknown links $\mathcal{E}_0$ that constitute the domain of interest whose structural and topological properties we are interested in. Since often $\mathcal{N}_0$ is a manifold like the Earth's surface, the brain volume, or a phase space, we will call the elements of $\mathcal{N}_0$ *points* here although the following consideration also applies to discrete sets $\mathcal{N}_0$ like that of all households in a society. If we can interpret the node set $\mathcal{N}$ to be a *sample* from the *population* of points $\mathcal{N}_0$, resulting from some

sampling procedure, then we might adopt a classical statistical approach and consider any measurements in the sample network G as estimates of certain statistics of the population network G_0 that we are truly interested in. E.g., the classical measure of *degree* of a node $v \in \mathcal{N}$,

$$k_v = k_v(G) = |\mathcal{N}_v| = \sum_{i \in \mathcal{N}} a_{iv}, \qquad (6)$$

could be interpreted as the simplest estimator of the number (or proportion) $k_0(v)$ of points in $\mathcal{N}_0$ to which v is linked in G_0. If, however, the sampling procedure is such that certain points $v \in \mathcal{N}_0$ are selected for the sample with a higher individual *sampling probability* p_v than others, basic statistics tells us that a much better estimator of $k_0(v)$ is a weighted sum,

$$\tilde{k}_v = \sum_{i \in \mathcal{N}_v} w_i = \sum_{i \in \mathcal{N}} w_i a_{iv} \qquad (7)$$

with suitable node weights $w_v \geqslant 0$. As in the well-known Horvitz-Thompson estimator of a sample mean, the optimal weights w_v are given by inverse probability weighting, i.e., they are inversely proportional to the sampling probabilities, $w_v \propto 1/p_v$. E.g., if G is a climate network constructed from meteorological stations and the probability p_v of having a station in location v is proportional to the local human population density, then any analysis of G should assign a node at v a weight proportional to the inverse human population density, to make sure that the climates in sparsely and densely populated areas are equally represented. In some cases, statistical considerations can also motivate more sophisticated choices of the weights w_v, like the reliability-adjusted Kriging weights used for meteorological station data in [36], equation (15).

If we were to follow this statistical approach more thoroughly, we would try to identify each property of the domain of interest we are interested in with some statistics f_0 of G_0, and then use a suitably weighted estimator $\tilde{f}$ of f_0 that is at least statistically *consistent* (i.e., converges to f_0 in a certain sense as N increases), hopefully also *unbiased* (i.e., has an average error of zero) and *efficient* (i.e., has small variance), and maybe even *robust* (i.e., is not very sensitive to only local changes in the network). Verifying these properties for a large number of different network measures is however a research program requiring much analytical effort beyond the scope of a single paper. Moreover, it would likely require some complicated continuity assumptions on G_0 that would restrict the applicability of those measures considerably, which is why we pursue a simpler approach here to find suitable weighting schemes for individual network measures.

Note that in principle, the above estimator $\tilde{k}_v$ can be interpreted as a special case of the *strength* $s_v = \sum_{i \in \mathcal{N}_v} w_{vi}$ of a node v in a *directed, (link-)weighted network* in which we simply use the node weight w_i of the target node as the link weight w_{vi} of the directed link. For other network measures, this interpretation is, however, not helpful since the measure might have no counterpart for directed weighted networks, or that counterpart is unsuitable for our problem (as in case of the clustering coefficient, see below).
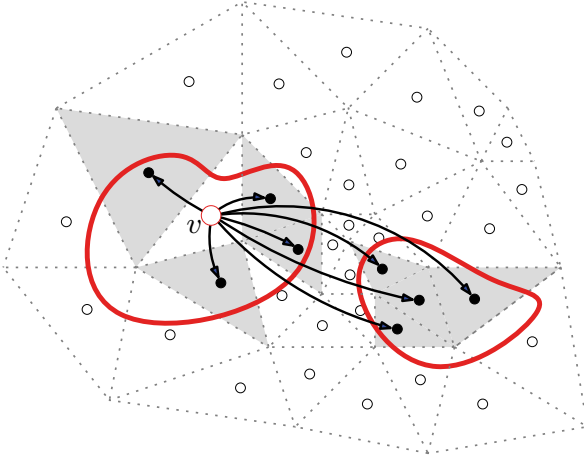
Fig. 5. (Colour online) A set $\mathcal{N}$ of nodes (circles) representing cells of different size of the domain of interest (dashed triangles). In many applications, a node v will be linked to one or more smoothly bounded regions of the domain of interest (red surrounded regions), containing v itself and some other nodes (filled circles; arrows show the resulting links in the network G). The red surrounded area of size $k_0(v)$ can then be approximated by the grey shaded area of size $\tilde{k}_v$, or more accurately by the grey shaded area plus v's own cell area, giving the estimate k_v^*. The classical degree k_v is just the number of filled circles.

4.2 Numerical approximation

If the domain of interest G_0 provides some notion of (geometric) distance (like any spatially embedded network does), an alternative approach is to consider each node $v \in \mathcal{N}$ as representative for a small cell $\mathcal{R}_v$ of points in v's geometrical vicinity in $\mathcal{N}_0$, whose size (in terms of some suitable measure, e.g., Lebesgue measure) we denote by w_v. By *geometrical vicinity* we mean those points of the underlying domain of interest that have a small geometrical distance from v, as opposed to its *neighbourhood* $\mathcal{N}_v^+$ that consists of those nodes in the network with a network-theoretic distance $\leqslant 1$. This interpretation would be adequate, e.g., if $\mathcal{N}_0$ is a continuous manifold and $\mathcal{N}$ a subset of points on a grid or derived by some meshing procedure (e.g., adaptive mesh refinement, [37]).

If, because of the continuity properties of the underlying system, it can be expected that all nodes $v' \in \mathcal{R}_v$ are linked to more or less the same nodes in G_0 as v is, then a natural approximation for an interesting statistics f_0 of G_0 would use the aggregation weight w_v wherever the formula for f_0 involves the node v.

E.g., for the above measure of degree, we could again use $\tilde{k}_v = \sum_{i \in \mathcal{N}_v} w_i$ instead of $k_v = |\mathcal{N}_v|$ to approximate $k_0(v)$, since each node i in v's G-neighbourhood represents w_i "many" nodes in G_0 of which most can be expected to be linked to v as well (Fig. 5 illustrates this idea). In the context of climate networks, each node i represents a portion of the Earth's surface of relative size $w_i = \cos$ (latitude of i), and $\tilde{k}_v$ is known as *area weighted connectivity* [26].

In many cases of continuous domains of interest G_0, a point $v \in \mathcal{N}_0$ is usually also linked to all or at least most of the points in its geometrical vicinity. More formally, in many cases the following *local connectedness* condition will hold for some suitable distance function d: for all $v \in \mathcal{N}_0$ there is $\varepsilon > 0$ such that each point $i \in \mathcal{N}_0 \backslash \{v\}$ with $d(i, v) < \varepsilon$ is linked to v. E.g., if $\mathcal{N}_0$ is the Earth's surface, and $i, j \in \mathcal{N}_0$ are linked in G_0 iff the surface temperature time series for i and j exhibit a product-moment correlation coefficient larger than some threshold value, then the smoothness of the underlying physics will imply the above.

In such a network, an alternative approximation to $k_0(v)$ would then be

$$k_v^* = \tilde{k}_v + w_v = \sum_{i \in \mathcal{N}_v^+} w_i = \sum_{i \in \mathcal{N}} w_i a_{iv}^+, \qquad (8)$$

which can be expected to be a better approximation if the mesh is fine enough. This estimator can also be interpreted as a classical numerical approximation to the integral of the indicator function of v's unpunctured neighbourhood in G_0 (see Fig. 5 again).

However, for many of the more complex network measures we will study below, e.g., random walk based measures or spectra, it will not be possible that easily to interpret our weighted versions of those measures as approximations to integrals, and a thorough analysis of their approximation qualities would require much technical effort. This is why we rather pursue a third, more pragmatic approach, which is motivated by a simple property that both the statistical and the numerical approximation interpretations have in common.

4.3 Pragmatic axiomatic approach

In both the statistical and the approximation approach, it is clear that the estimation or approximation should usually become better when the resolution of G as a description of the domain of interest G_0 is increased by replacing some or all nodes by a larger set of nodes representing smaller parts of G_0. Such *refinements* would usually change the corresponding inverse sampling densities or cell sizes that we use as our aggregation weights w_v. Let us now consider the case of sufficiently high resolution, i.e., where the sample is dense enough or the cell sizes are small enough to resolve all structural features of G_0 that are considered relevant, so that we do not expect there to be considerable inhomogeneities inside the region of G_0 represented by each individual node. Now imagine that an *elementary refinement* of $\mathcal{N}$ was performed in which only one old node $s \in \mathcal{N}$ was replaced by two new nodes s', $s'' \in \mathcal{N}_0$ which together represent more or less the same subset of $\mathcal{N}_0$ as s did. Then this would leave the aggregation weights w_i of the other nodes $i \in \mathcal{N} \backslash \{s, s', s''\}$ mostly unchanged, whereas the weights $w_{s'}$ and $w_{s''}$ of the new nodes would approximately sum up to the former w_s. Also, because the resolution was assumed to be sufficient already, s' and s'' would be linked to more or less the same nodes as s was, and most likely also to each
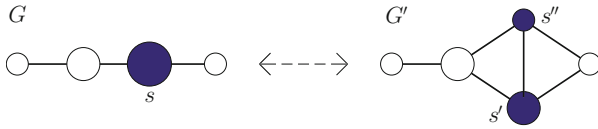
Fig. 6. (Colour online) The operation of *node splitting* replaces a node s of weight w_s in network G with two linked nodes s', s'' of weights $w_{s'} + w_{s''} = w_s$ which get the same neighbourhood as s had, giving network G'. The inverse operation of *twin merging* transforms G' back into G.

other. In that case, a good estimate or approximation $\tilde{f}$ to some statistic f_0 of G_0 should probably become somewhat more precise, but should certainly not be changed much by such an elementary refinement. This intuitive reasoning can be turned into a simple axiomatic guiding requirement when we idealize the above situation as follows.

Let $G = (\mathcal{N}, \mathcal{E})$ be a simple graph with weights $w_i > 0$ for all $i \in \mathcal{N}$, and let $s \in \mathcal{N}$ be some "old" node, $s', s'' \notin \mathcal{N}$ two "new" nodes, $w_{s'}, w_{s''} > 0$ their weights, and $w_{s'} + w_{s''} = w_s$. Then the "refined" graph $G' = (\mathcal{N}', \mathcal{E}')$ with

$$\mathcal{N}' = \mathcal{N} \backslash \{s\} \cup \{s', s''\}$$

$$\text{and} \quad \mathcal{E}' = \{\{i, j\} \in \mathcal{E} : i, j \neq s\}$$

$$\cup \{\{i, s'\}, \{i, s''\} : \{i, s\} \in \mathcal{E}\} \cup \{\{s', s''\}\} \tag{9}$$

and with weights w_i for all $i \in \mathcal{N}'$ will be called a *node splitting refinement* of G. That is, G' is derived from G by "splitting" the node s into two new interlinked nodes s', s'' with the same total weight as s and linking those two to exactly those nodes to which s was linked (see Fig. 6).

Now a (global) network measure f will be called *node splitting invariant (n. s. i.)* iff

$$f(G') = f(G) \tag{10}$$

for all pairs of networks G, G' and all weight functions w for which G' is a node splitting refinement of G. Likewise, if f_i is a network measure which is defined for nodes i, then f_i will be called *node splitting invariant* iff

$$f_i(G') = f_i(G) \quad \text{and} \quad f_{s'}(G') = f_{s''}(G') = f_s(G) \tag{11}$$

for all such G, G', w and all $i \in \mathcal{N} \backslash \{s, s', s''\}$. Similar definitions are possible for measures with more than one argument but are not needed here. In other words, a n. s. i. measure is *unaffected by any node splitting refinements.* This is, however, not to say that we actually perform any refinements or node splittings when applying a node splitting invariant measure. Rather, the notion of node splitting must be thought of as only a hypothetical, idealized operation that could be applied in principle, and the consistency requirement of node splitting invariance is only to make sure that the node weights are used in a proper way that is suitable for links which represent some kind of "similarity". In other words, to *define* a n. s. i. measure only requires us to analyse what *would* happen if nodes

were split in the suggested way. To *apply* a n. s. i. measure to a given network, one can just use the corresponding formula as one would with an unweighted measure, without any need to think about node splitting.

In the context of coarse-graining, node splitting invariance can also be called *twin merging invariance*, requiring that f should not change whenever two twins s', s'' with weights $w_{s'}$, $w_{s''}$ are merged into one node s with weight $w_s = w_{s'} + w_{s''}$, where two linked nodes s', s'' are called *twins* iff they have the same neighbourhood, $\mathcal{N}_{s'}^+ = \mathcal{N}_{s''}^+$. While node splitting corresponds to an idealized form of refinement, twin merging is an idealized form of coarse-graining and is the inverse operation of node splitting, hence both notions of invariance are equivalent.

While the above definitions of node splitting and twin merging are suitable for networks in which links represent a kind of similarity or direct connection, for which it is natural to require that twins are linked and to assume that the parts of a split node are linked, other types of networks might call for a different definition of node splitting and twin merging. If, e.g., links represent some kind of "complementarity" instead of similarity, a natural kind of splitting would leave s' and s'' unlinked, and the corresponding definition of "twin" would require that twins are not linked. Such a variant would lead to weighted network measures similar but not identical to ours, but we do not pursue this in the present paper.

An example of a n. s. i. measure is the above-defined *n. s. i. degree* $k_v^* = \sum_{i \in \mathcal{N}_v^+} w_i$, and that it is indeed n. s. i. is easily seen from the fact that

$$\mathcal{N}_{s'}^+(G') = \mathcal{N}_{s''}^+(G') = \mathcal{N}_s^+(G) \backslash \{s\} \cup \{s', s''\},$$

$$\mathcal{N}_i^+(G') = \mathcal{N}_i^+(G) \backslash \{s\} \cup \{s', s''\},$$

$$\text{and} \quad \mathcal{N}_j^+(G') = \mathcal{N}_j^+(G) \tag{12}$$

for all $i \in \mathcal{N}_s(G)$ and $j \in \mathcal{N} \backslash \mathcal{N}_s^+(G)$.

Note that the definition of node splitting invariance aka twin merging invariance does not at all rely on the formal specification of an underlying domain of interest G_0, but it depends on the network G and the weights w_i alone, which makes this tool much easier to use than estimation theory or approximation theory. Nevertheless, a conjecture and working hypothesis of this paper is that n. s. i. measures are the natural candidates for good estimation or approximation of the corresponding properties of a potentially underlying domain of interest, and that they will usually prove to be statistically consistent and exhibit good convergence properties when the domain of interest and the sampling or meshing procedures fulfil some suitable continuity or measurability properties and when the aggregation weights w_i are chosen accordingly.

In the following, we will therefore present n. s. i. versions f^* of a number of local and global network measures f that can be found in the literature, and we will refer to the possibly underlying domain of interest only when motivating some interpretations of these measures, but

without formally defining that statistics f_0 of G_0 which is supposed to be estimated or approximated by f^*.

The basic construction mechanisms we will use are

(i)   to sum up aggregation weights wherever the original measure counts nodes;
(ii) to use unpunctured neighbourhoods $\mathcal{N}_v^+$ wherever the original measure uses punctured neighbourhoods $\mathcal{N}_v$ (in other words, to consider v as linked to itself);
(iii) to also allow for equality of i, j wherever the original measure involves a sum over distinct nodes i, j; and
(iv) to "plug-in" a n. s. i. version of a measure g wherever this g is used in the definition of another measure f.

Both mechanisms (i) and (ii) were used in the definition of k_v^* above, and an example for mechanisms (iii) and (iv) will be given in the following section when we will consider the clustering coefficient.

5 Local measures

A network measure $f_v = f_v(G)$ that is defined for each node $v \in \mathcal{N}$ will be called *local* here. (Note that we always understand the terms "neighbour" and "local" as referring to the network topology, not to some possibly underlying geometry. So, local neighbours might be geometrically far apart.)

5.1 Degree

We already treated the *degree* measure $k_v = |\mathcal{N}_v|$ and defined the *n. s. i. degree* of v as

$$k_v^* = \sum_{i \in \mathcal{N}_v^+} w_i. \qquad (13)$$

Let us compare k_v and k_v^* in two example networks.

In our human brain example network, the nodes with the highest degree k_v are the right lingual gyrus and the left precuneus region (LING.R and PCUN.L, see [7] for these abbreviations), connected to 32 and 31 of the other 89 nodes, i.e., to about one third of the nodes. The volume of the 90 individual ROIs varies by a factor of 23 (Fig. 1) If it is used as a node weight for n. s. i. degree, the right lingual gyrus and the left precuneus again have the largest values, but these are now $k_v^* \approx 0.46W$, showing that in reality, both regions are functionally connected not to one third but rather to almost half the entire brain (in terms of volume). The third most connected node in terms of k_v^* (volume) is the left middle frontal gyrus (MFG.L, which seems consistent with other measures of node importance as reported in [7]), but in terms of k_v (nodes) it is the left calcarine cortex (CAL.L). Since the number of linked nodes basically depends on the level of detail in the used parcellation of the brain, k_v can change considerably for different parcellations even if the ROI represented by node v remains unchanged, and k_v seems to be influenced much more than k_v^* by the choice of parcellation.
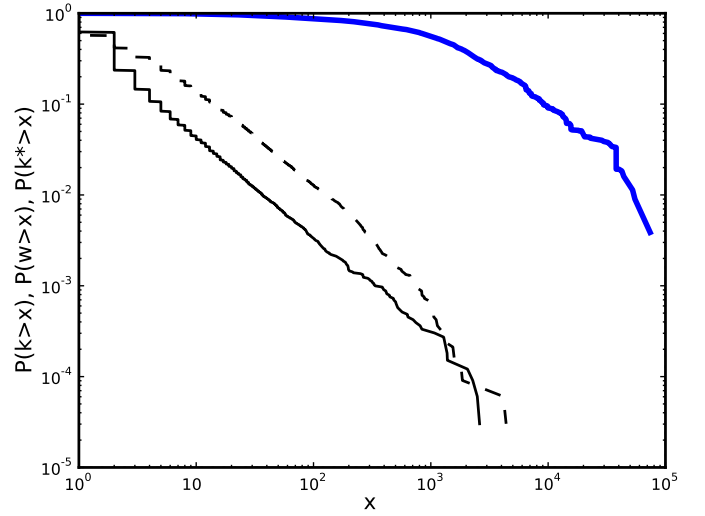

Fig. 7. (Colour online) Log-log plot of the complementary cumulative distribution function of degree k_v (thin black line), node weight w_v (dashed line), and n. s. i. degree k_v^* (thick blue line) in the routing network of autonomous systems in the internet. Power laws would appear linear.

In the internet example network, the earlier findings of power laws for the degree distribution reported in [12,13,15] are supported by the apparent linear relationship in the log-log plot in Figure 7 (thin black line). In other words, the distribution of the number k_v of linked AS's of a given AS v seems to follow a power law. However, also the distribution of the size w_v of a given AS v in terms of CIDR prefixes seems to follow a power law, as can also be seen in Figure 7 (dashed line), and the linear correlation coefficient between $\ln k_v$ and $\ln w_v$ is high (≈ 0.5), so the power law for k_v might be a consequence of the one for w_v. If we ask for the share of the internet (instead of the number of AS's) a given AS v is linked to, k_v^*/W seems a more accurate estimate of this than k_v/N, and the findings are different: the AS with highest k_v also has the highest k_v^*, but while it is linked to only 7.9% of all AS's (since $k_v = 0.079N$), it seems to be linked to approx. one fourth of all IP adresses (since $k_v^* = 0.24W$). When plotting the probability that a randomly chosen CIDR prefix belongs to an AS that is linked to other AS's with more than x total CIDR prefixes, this does no longer show a clear power law behaviour (Fig. 7, thick blue line). Also, $\ln k_v^*$ is less strongly correlated (≈ 0.3) to $\ln w_v$ than $\ln k_v$ is, and less strongly correlated (≈ 0.4) to $\ln k_v$ than $\ln w_v$ is.

5.2 Clustering coefficient

The *local clustering coefficient* of v,

$$C_v = \frac{\sum_{i \in \mathcal{N}_v} \sum_{j \in \mathcal{N}_v} a_{ij}}{k_v(k_v - 1)} = \frac{N^2 \langle a_{vi} a_{ij} a_{jv} \rangle_{ij}}{k_v(k_v - 1)} \in [0, 1], \qquad (14)$$

is the probability that two nodes drawn at random from those linked to v are linked with each other. We get
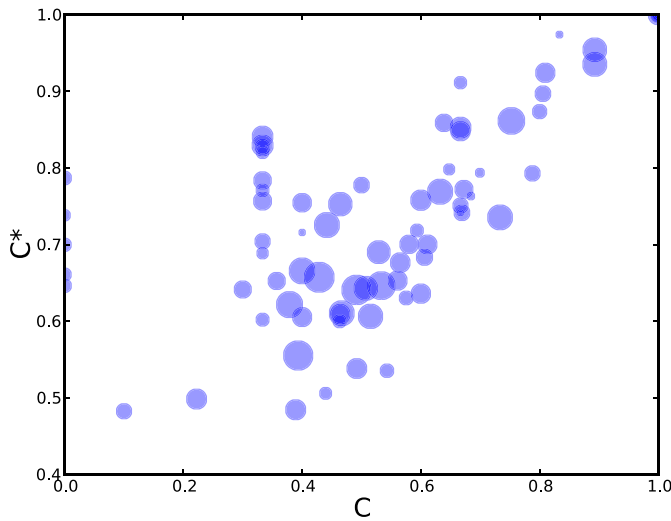
Fig. 8. (Colour online) Clustering coefficient C_v vs. n.s.i. clustering coefficient C_v^* of those nodes in the functional human brain network that have at least two neighbours, showing considerable differences (Pearson correlation is $\varrho = 0.65$). Disk area is proportional to node weight (ROI volume).

a weighted version by employing all four mechanisms (i)−(iv), giving the *n. s. i. local clustering coefficient*

$$C_v^* = \frac{W^2 \left\langle a_{vi}^+ a_{ij}^+ a_{jv}^+ \right\rangle_{ij}^w}{k_v^{*2}} \in \left[\frac{w_v \left(2k_v^* - w_v \right)}{k_v^{*2}}, 1 \right] \subseteq [0,1],$$
(15)

which estimates the probability that two weight units (or *points* in terms of G_0) drawn at random from the part of the network linked to v are linked with each other. Because we use A^+, C_v^* tends to be larger than C_v if k_v is small, and it is defined for all nodes while C_v only makes sense when $k_v > 1$. If the weights vary considerably, C_v^* and C_v can rank the nodes quite differently, although neither needs to be significantly linearly correlated with w_v. C_v^* and C_v can also differ considerably when the weights *inside* $\mathcal{N}_v$ vary strongly.

All these effects can be seen nicely in our brain example network (Fig. 8). The right thalamus region (THA.R), e.g., is in the top half (rank 37) according to its C_v of 0.54, but almost at the bottom (rank 86 out of 90) according to its C_v^* of 0.53, although the absolute values are almost equal. This is because among its 21 neighbours, it is rather the smaller ones (like THA.L and CAU.R) that are linked with many other neighbours.

One might think that similar to the case of degree, also a directed version of the formula for the local clustering coefficient from the theory of (link-)weighted networks [25], which is $c_v^w = \sum_{i \in \mathcal{N}_v} \sum_{j \in \mathcal{N}_v} a_{ij}(w_{vi} + w_{vj})/2s_v(k_v - 1)$, could be a good candidate if the directed link-weights are defined as $w_{vi} = w_i$, giving $c_v^w = \sum_{i \in \mathcal{N}_v} \sum_{j \in \mathcal{N}_v} a_{ij}(w_i + w_j)/2s_v(k_v - 1)$. But it is easy to see that this does not behave well under node splitting or twin merging since a linked pair i-j contributes $w_i + w_j$ instead of $w_i w_j$. If, e.g., $\mathcal{N}_v = \{i, j, s', s''\}$ with $w_i = w_j = w_{s'} = w_{s''} = 1$, $a_{ij} = 0$, and s', s'' are twins linked to i but not to j, then

$c_v^w = C_v = 3/6$, but after merging the twins s', s'' into one new node s with $w_s = 2$, one has $k_v = 3$, $s_v = 4$, and $c_v^w = (w_i + w_s)/2s_v(k_v - 1) = 3/16$, much smaller than before. In contrast, $C_v^* = 19/25$ before and after the twin merging.

5.3 Measures of centrality and betweenness

Many measures try to assess several aspects of "node importance". Based on the distances of one node v to all others, we consider three variants of *closeness centrality*,

$$CC_v = 1/\langle d_{vi} \rangle_i, \quad CC_v' = \left\langle 2^{-d_{vi}} \right\rangle_i, \quad CC_v'' = \langle 1/d_{vi} \rangle_i,$$
(16)

the latter also being called the *efficiency* of v, where d_{vi} is the number of links on a shortest path from v to i, or, if there is no such path, either ∞ or N, depending on the convention chosen [3,38–40]. A weighted version of CC_v should give us the inverse average distance of v from other weight units or points rather than from other nodes. But for this to become n.s.i., one has to interpret (somewhat peculiarly) each node to have unit (instead of zero) distance to itself. This is because after an imagined split $s \to s', s''$, the two parts s', s'' of s have unit not zero distance. The *n. s. i. distance function* is hence given by

$$d_{vv}^* = 1 \quad \text{and} \quad d_{vi}^* = d_{vi} \quad \text{for} \quad i \neq v,$$
(17)

i.e., the zeros on the diagonal of the ordinary distance matrix are replaced by ones to get the n.s.i. distance matrix, without changing any off-diagonal entries. Using this, we can derive v's *n. s. i. closeness centrality* measures as

$$CC_v^* = \frac{1}{\langle d_{vi}^* \rangle_i^w} = \frac{W}{\sum_{i \in \mathcal{N}} w_i d_{vi}^*} = \frac{W}{w_v + \sum_{i \in \mathcal{N}} w_i d_{vi}},$$
(18)

$CC_v'^* = \langle 2^{-d_{vi}^*} \rangle_i^w$, and $CC_v''^* = \langle 1/d_{vi}^* \rangle_i^w$. All take values in $[0,1]$.

In the internet example network, CC_v and CC_v^* are generally quite small, vary only little, and are highly correlated. Still, both measures lead to different rankings of the most central nodes (Fig. 9).

Also in our Wikipedia example network, CC_v and CC_v^* were not very discriminatory, but CC_v' and $CC_v'^*$ were. Figure 10 depicts their relationship for the most central nodes, again showing considerable differences in ranks. According to CC_v', the top ten central articles in decreasing order are those named "Physics", "Mathematics", "Chemistry", "Germany", "Science", "Physicist", "Quantum mechanics", "Albert Einstein", "Astronomy", and "Engineering", while according to $CC_v'^*$ the list is "Physics", "Mathematics", "Germany", "Chemistry", "Science", "Japan", "Italy", "Albert Einstein", "Russia", and "Astronomy". "Physicist" for instance has the sixth highest value of CC_v' but only the 21th largest value of $CC_v'^*$ because it is linked to a very large number (7.6% of the nodes) of comparatively short articles (accounting for 5.5% of the total text) on individual scientists, so that CC_v' treats them as a larger part of the network than $CC_v'^*$ does.
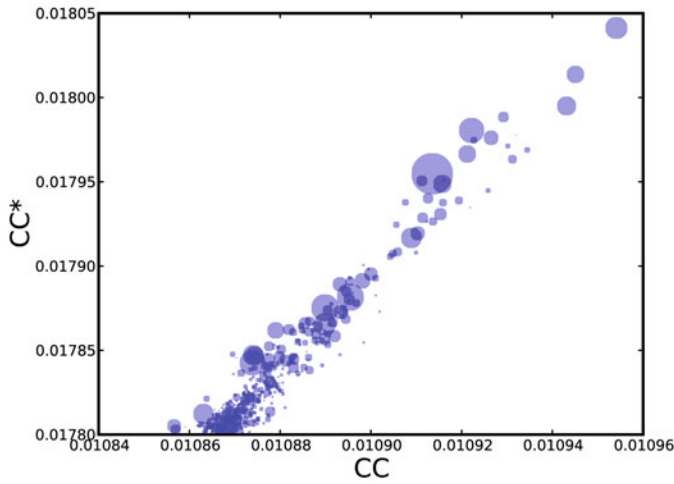
Fig. 9. (Colour online) Closeness CC_v vs. n. s. i. closeness CC_v^* of those nodes in the internet network (see text) with the highest values. Disk area is proportional to node weight (no. of CIDR prefixes).
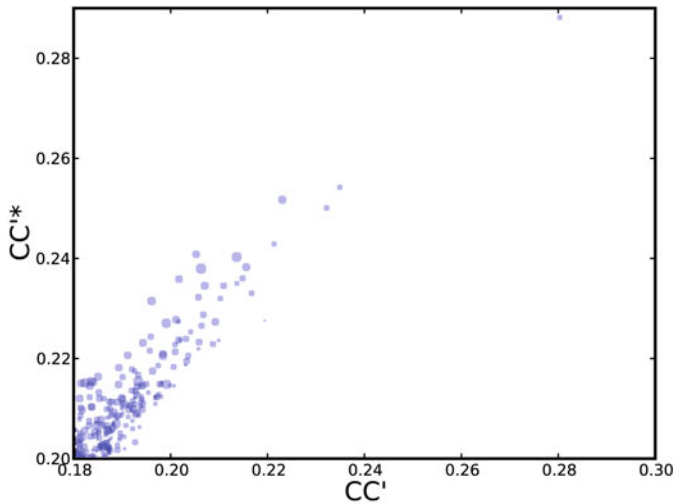

Fig. 10. (Colour online) Exponential closeness CC_v' vs. its n. s. i. version $CC_v'^*$ of those Wikipedia articles on physics (see text) with the highest values. Disk area is proportional to node weight (article size in characters).

Other, somewhat more sophisticated importance measures depend on the paths between all other nodes that lead through a given node v. The *(shortest path) betweenness* of v is the proportion of shortest paths between randomly chosen nodes a, b that lead through v:

$$BC_v = \langle n_{ab}(v)/n_{ab} \rangle_{ab} \in [0,1], \qquad (19)$$

where n_{ab} is the total number of shortest paths from a to b, and $n_{ab}(v)$ is the number of those paths that pass through v as an inner node. Formally, n_{ab} can be written as a sum over all node tuples $(t_0, \dots, t_{d_{ab}})$ with $t_0 = a$ and $t_{d_{ab}} = b$, where the summands are either zero or one, depending on whether each t_ℓ is linked to its successor $t_{\ell+1}$, for $\ell = 0, \dots, d_{ab} - 1$. As the latter condition can be
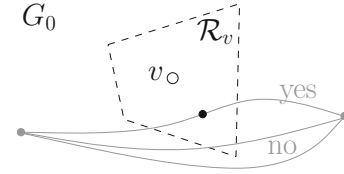

Fig. 11. Interpretation of n. s. i. shortest path betweenness as the probability density that a randomly chosen shortest path (grey curves) between two randomly chosen points (grey dots) in the domain of interest G_0 passes through a randomly chosen point (black dot) in the area $\mathcal{R}_v$ (dashed region) represented by v (black circle).

written as a product of elements of the adjacency matrix, we have:

$$n_{ab} = \sum_{\substack{(t_0, \dots, t_{d_{ab}}) \in \mathcal{N}^{d_{ab}+1} \\ t_0 = a, \; t_{d_{ab}} = b}} \prod_{\ell=1}^{d_{ab}} a_{t_{\ell-1} t_\ell}. \qquad (20)$$

A similar formula holds for $n_{ab}(v)$, only that for some m in $1 \dots d_{ab} - 1$, t_m must equal v:

$$n_{ab}(v) = \sum_{m=1}^{d_{ab}-1} \sum_{\substack{(t_0, \dots, t_{d_{ab}}) \in \mathcal{N}^{d_{ab}+1} \\ t_0 = a, \; t_m = v, \; t_{d_{ab}} = b}} \prod_{\ell=1}^{d_{ab}} a_{t_{\ell-1} t_\ell}. \qquad (21)$$

When s is hypothetically split into $s' + s''$, any shortest path through s becomes a pair of shortest paths, one through s' and the other through s''. Also, a shortest path from s'' to some $b \neq s'$ will never meet s'. Thus, to make BC_v n. s. i., it suffices to make n_{ab} and $n_{ab}(v)$ n. s. i. by making each path's contribution proportional to the weight of each inner node (that is, to the product of these weights!), where in case of $n_{ab}(v)$ we have to skip w_v in this product:

$$n_{ab}^* = \sum_{\text{(as above)}} a_{t_0 t_1} \prod_{\ell=2}^{d_{ab}} \left(w_{t_{\ell-1}} a_{t_{\ell-1} t_\ell} \right),$$

$$n_{ab}^*(v) = \frac{1}{w_v} \sum_{m=1}^{d_{ab}-1} \sum_{\text{(as above)}} \left(a_{t_0 t_1} \prod_{\ell=2}^{d_{ab}} \left(w_{t_{\ell-1}} a_{t_{\ell-1} t_\ell} \right) \right). \qquad (22)$$

The *n. s. i. shortest path betweenness*

$$BC_v^* = \langle n_{ab}^*(v)/n_{ab}^* \rangle_{ab}^w \in [0, 1/w_v] \qquad (23)$$

can then be interpreted as an estimate of the probability (or probability density) that a randomly chosen shortest path between two randomly chosen points in the underlying domain of interest G_0 passes through *a specific randomly chosen point in the area* $\mathcal{R}_v$ represented by v, as illustrated in Figure 11. The product $w_v BC_v^*$ then estimates the probability that such a path passes through *any* point in $\mathcal{R}_v$, which is not n. s. i. but is additive under node splitting: $w_{s'} BC_{s'}^* + w_{s''} BC_{s''}^* = w_s BC_s^*$.

Newman [41] gives an $O(|\mathcal{N}||\mathcal{E}|)$-time algorithm to compute BC_v for all v, based on Dijkstra's algorithm, and

this can easily be adapted to compute BC_v^* for all v with the same algorithmic time complexity.

In our human brain example network, according to BC_v the top ten ROIs in decreasing order are the nodes TPOsup.R, LING.L, LING.R, TPOmid.L, MFG.R, TPOsup.L, MFG.L, STG.L, CAU.R, and ORBinf.R, while according to BC_v^* the list is TPOmid.L, TPOsup.R, LING.R, LING.L, TPOsup.L, PCL.R, PHG.L, THA.R, CAU.R, and STG.L. The left and right middle frontal gyrus (MFG) are missing in the latter list (having ranks 13 and 19) because most shortest paths that lead through them are between relatively small nodes, (left and bottom region in Fig. 1), whereas the right thalamus (THA.R) has a high degree but many of its larger neighbours are not linked to each other, leading to a large value of BC_v^*. In this network one can also nicely see the effect of network design choices on network statistics. If we slightly modify the parcellation and treat the mid-sized left and right parts of the dorsal cingulate gyrus as one node s instead of two ($s' =$ DCG.L and $s'' =$ DCG.R, having quite similar neighbourhoods), leaving the rest of the network unchanged, then their BC-values increase considerably from $BC_{s'} = 0.029$ and $BC_{s''} = 0.036$ to $BC_s = 0.053$, or from 24th and 18th rank to 11th rank. Their BC^*-values, however, behave rather nicely in that the new value $BC_s^* = .0000144$ (rank 19) is approximately the average of the two old values $BC_{s'}^* = 0.0000114$ (rank 23) and $BC_{s''}^* = 0.0000171$ (rank 17). Had DCG.L/R been exact twins, BC^* would not have changed at all.

More centrality and betweenness measures will be discussed in the next subsection and in Section 6.2.

5.4 Measures based on random walks

Several network measures are based upon the idea of a random walk along the links of a network without isolated nodes, with all neighbours j of a node i having the same transition probability

$$p_{ij} = a_{ij}/k_i. \tag{24}$$

The crucial ideas in constructing n. s. i. versions of such measures are to make p_{ij} proportional to w_j and to allow the walk to stay at node i (which now also allows for the existence of isolated nodes):

$$p_{ij}^* = a_{ij}^+ w_j/k_i^*. \tag{25}$$

Such a walk can be thought to approximate a random walk in G_0 which moves to each linked point with the same probability. If G_0 is a continuous domain, this discrete random walk must not be confused with a continuous Wiener process, however. In particular, its individual steps might bridge long distances in terms of the domain's geometry.

With the above transition probabilities, the probability of visiting or staying inside $\{s', s''\}$ after a split $s \rightarrow s' + s''$ is $p_{vs'}^* + p_{vs''}^* = p_{vs}^*$ or $p_{s's'}^* + p_{s's''}^* = p_{s''s'}^* + p_{s''s''}^* = p_{ss}^*$, respectively, which means the walk

is not influenced by the split. With the original *transition matrix* $\mathsf{P} = (p_{ij})_{ij}$, the equilibrium distribution and hence the long-time average relative visiting frequencies are given by $p_v = k_v/K$, where $K = N\langle k_v \rangle_v$ is twice the number of links in G. With the *n. s. i. transition matrix* $\mathsf{P}^* = (p_{ij}^*)_{ij}$, the equilibrium distribution is $p_v^* = k_v^*/K^*$, where $K^* = W\langle k_v^* \rangle_v^w$.

The *Arenas-type random walk betweenness* of v, motivated by [42] and based on the idea of searching a target, is the expected number of visits to v on a random walk that starts and ends at some randomly chosen nodes a, b:

$$AB_v = \langle AB_{av}(b) \rangle_{ab} \tag{26}$$

with $\mathsf{AB}(b) = (AB_{ij}(b))_{ij} = \sum_{t=1}^{\infty} \mathsf{P}'(b)^t$. Since the walk is assumed to stop as soon as b is reached, the "transition" matrix is here $\mathsf{P}'(b) = (p_{ij}'(b))_{ij}$ with

$$p_{ij}'(b) = (1 - \delta_{bi}) a_{ij}/k_i. \tag{27}$$

The main problem in attaining node splitting invariance is the stopping condition introduced by the term $-\delta_{bi}$. If the target node b is split into $b' + b''$, and if b' is the new target node, the walk must not continue after reaching b'' since otherwise AB_v would increase. Hence the walk must sometimes stop earlier than before, at least when reaching a twin of the target. As exact twins are usually rare in large networks, it seems natural to adopt a somewhat more continuous stopping condition that may stop the walk with some probability as soon as it enters the neighbourhood of the target. Using a suitable *n. s. i. similarity measure* $\sigma^*(i, j) \in [0, 1]$ that equals zero for unlinked nodes and one for twins, we can then define a *n. s. i. Arenas-type random walk betweenness*:

$$AB_v^* = \langle AB_{av}^*(b) \rangle_{ab}^w / w_v \tag{28}$$

with $\mathsf{AB}^*(b) = (AB_{ij}^*(b))_{ij} = \sum_{t=1}^{\infty} \mathsf{P}'^*(b)^t$ and "transition" matrix $\mathsf{P}'^*(b) = (p_{ij}'^*(b))_{ij}$, where

$$p_{ij}'^*(b) = (1 - \sigma^*(b, i)) a_{ij}^+ w_j/k_i^*. \tag{29}$$

For the similarity measure we may, e.g., use one out of the increasing sequence

$$\sigma_I^*(i, j) = \delta_{\mathcal{N}_i^+ \mathcal{N}_j^+},$$

$$\sigma_{II}^*(i, j) = a_{ij}^+ w \left(\mathcal{N}_i^+ \cap \mathcal{N}_j^+ \right) / w \left(\mathcal{N}_i^+ \cup \mathcal{N}_j^+ \right),$$

$$\sigma_{III}^*(i, j) = a_{ij}^+ w \left(\mathcal{N}_i^+ \cap \mathcal{N}_j^+ \right) / \max \left(k_i^*, k_j^* \right),$$

$$\sigma_{IV}^*(i, j) = a_{ij}^+ w \left(\mathcal{N}_i^+ \cap \mathcal{N}_j^+ \right) 2 / \left(k_i^* + k_j^* \right),$$

$$\sigma_V^*(i, j) = a_{ij}^+ w \left(\mathcal{N}_i^+ \cap \mathcal{N}_j^+ \right) / \sqrt{k_i^* k_j^*},$$

$$\sigma_{VI}^*(i, j) = a_{ij}^+ w \left(\mathcal{N}_i^+ \cap \mathcal{N}_j^+ \right) / \min \left(k_i^*, k_j^* \right),$$

$$\sigma_{VII}^*(i, j) = a_{ij}^+ \tag{30}$$

with $w(M) = \sum_{v \in M} w_v$ for $M \subseteq N$. In all these versions, the walk may stop with some probability when a more

or less twin-like neighbour of b is reached. If nodes can be expected to "know" their neighbours with some probability, this stopping behaviour can also be interpreted as meaning that once a neighbour of the target is reached, the path to the target will more or less likely be known, hence the target can be considered to be found and the random search walk can stop.

In the brain example network, both the unweighted and n.s.i. version (using σ_{VII}^*) of Arenas-type random walk betweenness give the same set of nodes with top ten betweenness values, but in quite different order: for AB_v it is LING.R, PCUN.L, CAL.L, PCUN.R, MFG.L, MTG.R, DCG.L, DCG.R, SFGmed.L, and MFG.R, while for AB_v^* it is PCUN.L, LING.R, MFG.L, CAL.L, MTG.R, PCUN.R, SFGmed.L, DCG.L, MFG.R, and DCG.R.

In our world trade example network, when we use as node weights the considerably varying countries' gross domestic product in 2008 (as reported by the IMF), then according to AB_v the countries with the largest betweenness are CHN, USA, DEU, FRA, JPN, IND, RUS, ITA, SGP, and KOR, whereas according to AB_v^* the ranking is much different, USA, CHN, JPN, KOR, PAN, SAU, MYS, PHL, VNM, and GAB, in descending order. The last six are all connected to all three of the heavy-weight nodes USA, CHN, and JPN (or FRA in case of GAB), which explains their high AB_v^*-values. Also in the layout in Figure 2, they are more centrally located than those in the first list. Germany (DEU) and France (FRA) on the other hand, are missing from the second list mainly because they are connected to neither of those three big economies directly. A much more realistic model of the world trade network would of course use weighted and directed links representing actual imports and exports, in addition to node weights, so one cannot attach much real-world importance to the above exemplary results.

In the preceding type of random walk betweenness, each individual visit of the walk to v is counted. Newman [43] introduced a similar measure in which, however, only the "net" flow of the walk along each link is considered. A more intuitive interpretation of his measure is the expected effective current $I_{ab}(v)$ passing through v when the network is interpreted as an electric circuit with all links having unit conductance, and a unit current is sent from a random node a to a random node b. This explains the definition of *Newman's random walk betweenness*

$$NB_v = \langle I_{ab}(v) \rangle_{ab}$$

$$\text{with} \quad I_{ab}(i) = \frac{1}{2} \sum_{j \in \mathcal{N}_i} |V_j(a,b) - V_i(a,b)|, \quad (31)$$

where the "electric potential" vector $\boldsymbol{V}(a,b)$ is given by Kirchhoff's equations

$$\Lambda \boldsymbol{V}(a,b) = \boldsymbol{\delta}(a) - \boldsymbol{\delta}(b) \quad \text{with} \quad \delta_i(c) = \delta_{ic}, \quad (32)$$

where $\Lambda = \text{diag}(\boldsymbol{k}) - \mathsf{A}$ is the *Laplacian* matrix of G which will be studied more closely in Section 6.2. As with the stopping condition above, here the main problem in finding an n.s.i. version is that the current leaves the circuit at a single node b, so that when b is split into $b' + b''$ and the
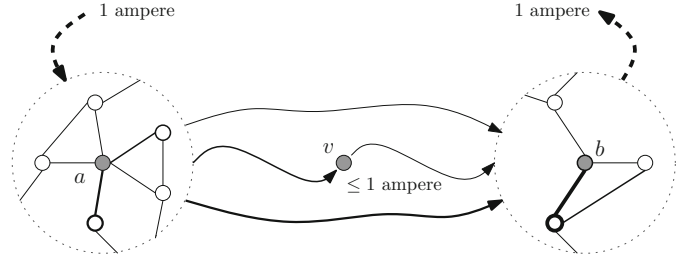

Fig. 12. Interpretation of n.s.i. Newman-type random walk betweenness as the expected (electric) current flowing through v when a unit current flows between the neighbourhoods of two randomly selected nodes a and b and each link's conductance is proportional to both its ends' weights (here represented by line thickness).

current leaves at b', the twin b'' will have a different electric potential than b'. This can be overcome in the same way as above, by using some n.s.i. similarity measure σ^* and letting a part of the current that is proportional to $w_i \sigma^*(b, i)$ leave the circuit also at each neighbour i of b. A similar thing must also be done at the a-side where the current enters. Finally, each link i-j must get conductance $w_i w_j$ to reflect the fact that the link i-j represents a bundle of links in G_0 between the w_i many points represented by i and the w_j many points represented by j. In order to avoid a dominating influence of the degree on our measure, we also restrict a and b to nodes not directly linked to v, although this would not be necessary to achieve node splitting invariance. For the choice $\sigma^* = \sigma_{VII}^*$ in which the current enters and leaves at all nodes in the neighbourhood of a and b to some extent, this approach is depicted in Figure 12.

All these considerations lead to our definition of *n.s.i. Newman-type random walk betweenness*

$$NB_v^* = \left\langle \left(1 - a_{av}^+\right) \mathsf{I}_{ab}^*(v) \left(1 - a_{vb}^+\right) \right\rangle_{ab}^w$$

$$\text{with} \quad \mathsf{I}_{ab}^*(i) = \frac{1}{2} \sum_{j \in \mathcal{N}_i^+} w_j \left| V_j^*(a,b) - V_i^*(a,b) \right|, \quad (33)$$

where $\boldsymbol{V}^*(a,b)$ is given by the new equations

$$\mathsf{D}_w \Lambda^* \boldsymbol{V}^*(a,b) = \boldsymbol{\delta}'(a) - \boldsymbol{\delta}'(b),$$

$$\delta_i'(c) = w_i \sigma^*(i,c) / \sum_{j \in \mathcal{N}_c^+} w_j \sigma^*(j,c), \quad (34)$$

and where

$$\mathsf{D}_w = \text{diag}(\boldsymbol{w}) \quad \text{and} \quad \Lambda^* = \text{diag}\left(\boldsymbol{k}^*\right) - \mathsf{A}^+ \mathsf{D}_w \quad (35)$$

are the diagonal matrix of aggregation weights and the n.s.i. Laplacian matrix (see Sect. 6.2). In this way, after a split $s \to s' + s''$, all potentials $V_i^*(a,b)$ with $i \neq s$ remain unchanged, and $V_{s'}^*(a,b)$ and $V_{s''}^*(a,b)$ equal the former $V_s^*(a,b)$. The measure NB_v^* estimates the expected effective current passing through each unit-sized part of the region $\mathcal{R}_v$ of G_0 that is represented by v, when the domain of interest G_0 is interpreted as an electric circuit with all links having unit conductance, and a unit current is sent between two random neighbourhoods of G_0.

In the brain example network, this time the top ten lists according to the unweighted and n. s. i. versions of Newman's betweenness measure differ more than for Arenas-type betweenness: we get TPOsup.R, LING.L, LING.R, ORBinf.R, MFG.R, MFG.L, TPOmid.L, CAU.R, STG.L, and ITG.R according to NB_v, but TPOmid.L, TPOsup.R, TPOsup.L, LING.L, ORBinf.R, TPOmid.R, CAU.R, THA.L, STG.L, and THA.R according to NB_v^*. For example, MFG.L/R is again missing in the second list while THA.L/R is included, and the explanation is the same as for the case of shortest path betweenness (Sect. 5.3).

Newman-type random walk betweenness is also of interest in climate networks, see Section 7.

6 Global measures

6.1 Aggregate measures

Popular aggregate network statistics include the *global clustering coefficient* $C = \langle C_v \rangle_v \in [0,1]$ with n. s. i. version

$$C^* = \langle C_v^* \rangle_v^w \in [0,1], \qquad (36)$$

the *global transitivity* $T = \langle a_{vi}a_{ij}a_{jv}\rangle_{vij}/\langle a_{vi}(1-\delta_{ij})a_{jv}\rangle_{vij}$ which is closely related to C^* and has the n. s. i. version

$$T^* = \frac{\langle a_{vi}^+ a_{ij}^+ a_{jv}^+ \rangle_{vij}^w}{\langle a_{iv}^+ a_{vj}^+ \rangle_{vij}^w} \in [0,1], \qquad (37)$$

the *link density* $\varrho = \langle a_{ij} \rangle_{ij} \in [0,1)$, whose n. s. i. version is

$$\varrho^* = \langle a_{ij}^+ \rangle_{ij}^w = \frac{1}{W} \langle k_v^* \rangle_v^w \in [0,1] \qquad (38)$$

the *average (geodesic) path length* or *mean geodesic distance* $L = \langle d_{ij} \rangle_{ij} \geqslant 0$ with

$$L^* = \langle d_{ij}^* \rangle_{ij}^w > 0, \qquad (39)$$

and the *global efficiency* [40] $E = \langle 1/d_{ij} \rangle_{i \neq j} \in [0,1]$ for which we can use

$$E^* = \langle 1/d_{ij}^* \rangle_{ij}^w \in [0,1]. \qquad (40)$$

Many authors, e.g., compare C and L to their values for randomly rewired graphs to assess the "small-worldness" of the network in a way that is quite sensitive to small structural changes in the network (see [44]), and using C^* and L^* for this task should at least reduce that part of this non-robustness that is related to node selection and aggregation.

In our internet example network, we get $C = 0.21$, i.e., for a randomly chosen AS v, two randomly chosen other AS's that are both linked to v are also linked with each other with probability 21%. If the AS network is constructed more meticulously than here, combining several data sources, somewhat larger values around 0.4 are found [45]. On the other hand, we have $C^* = 0.8$, meaning that for a randomly chosen IP address x, two randomly

chosen other IP addresses y, z that can be reached from x in at most one routing step, can also reach each other in at most one routing step with 80% probability. One reason for the large difference between C and C^* is the low average degree of only $\langle k_v \rangle_v = 4.24$, which means that many of the above pairs y, z belong to the same AS and therefore contribute to C^* but not to C. Also, we get $E \approx 0.27$ and $E^* \approx 0.336$, which can be interpreted as indicating that a typical distance between two randomly chosen AS's is $1/0.27 \approx 3.7$ routing steps, while between two randomly chosen IP adresses it is only $1/0.336 \approx 2.98$, which is 20% less, showing that large AS's tend to be more central than small ones.

6.2 Characteristic matrices and spectral measures

Spectral network analysis deals with the eigenvalues and eigenspaces of characteristic matrices such as the adjacency matrix A or the *Laplacian* and *normal matrices*

$$\Lambda = \mathrm{diag}(\boldsymbol{k}) - \mathsf{A}, \quad \mathsf{T} = \mathrm{diag}(\boldsymbol{k})^{-1}\mathsf{A}, \qquad (41)$$

of which we will only consider the first two here. Spectral analysis can be used to study the centrality of nodes and the community structure of the network, and to find natural partitions or node classification trees.

Depending on which matrix properties are considered essential, there are at least two different ways in which these matrices can be made n. s. i. in some sense. One way is to multiply both the rows and columns with the square root of the corresponding node weights, similar to what is done in a different context in the estimation of empirical orthogonal functions from gridded data (e.g., [34]), thereby preserving the symmetry of the matrix. For the adjacency matrix, this gives

$$\mathsf{A}^{*\prime} = \mathsf{D}_w^{1/2}\mathsf{A}^+\mathsf{D}_w^{1/2}, \qquad (42)$$

which has the property that any solution to the eigenequation $\mathsf{A}^{*\prime}\boldsymbol{x} = \lambda\boldsymbol{x}$ is n. s. i. in the sense that after a node splitting $s \to s' + s''$, the resulting new $\mathsf{A}^{*\prime}$ still has the eigenvalue λ with some new eigenvector $\boldsymbol{y}$, the entries x_v of the eigenvector $\boldsymbol{x}$ that belong to non-split nodes $v \neq s$ remain unchanged, $y_v = x_v$, and the quotient of eigenvector entry and square root of the weight is invariant also for the split nodes, $y_{s'}/\sqrt{w_{s'}} = y_{s''}/\sqrt{w_{s''}} = x_s/\sqrt{w_s}$. In particular, $\boldsymbol{y}$ has the same ℓ_2-norm as $\boldsymbol{x}$. In a sense, the splitting of node s results in a corresponding "splitting" of all the eigenvectors' s-dimension.

After the split, $\mathsf{A}^{*\prime}$ has an $(N+1)$st eigenvalue of zero, corresponding to the eigenvector $\boldsymbol{z}$ with $z_{s'} = \sqrt{w_{s''}}$, $z_{s''} = -\sqrt{w_{s'}}$, and $z_v = 0$ for $v \neq s', s''$. In particular, the *m-th moment* $\frac{1}{N}\sum_i \lambda_i^m$ of $\mathsf{A}^{*\prime}$ becomes n. s. i. if the normalization $1/N$ is replaced by the n. s. i. normalization $1/W$: $\frac{1}{W}\sum_i \lambda_i^m$. (As i does not refer to a node in this sum, we do not need to weight λ_i^m with w_i. The sum itself is already n. s. i. since all non-zero λ_i are.)

An alternative and simpler possibility is to use the node weights only for the columns and put

$$\mathsf{A}^* = \mathsf{A}^+\mathsf{D}_w, \qquad (43)$$

destroying symmetry and getting different eigenvectors $\boldsymbol{y}$ as with $\mathsf{A}^{*\prime}$, although the eigenvalues are the same. The entries of the eigenvector $\boldsymbol{x}$ that belong to non-split nodes $i \neq s$ still remain unchanged, but this time s' and s'' directly inherit their eigenvector entries from s, $y_{s'} = y_{s''} = x_s$, so the new eigenvector $\boldsymbol{y}$ has the same ℓ_∞-norm as $\boldsymbol{x}$.

As both A^* and $\mathsf{A}^{*\prime}$ are non-negative, the Perron-Frobenius theorem guarantees the existence of a non-negative eigenvalue of largest absolute value and corresponding non-negative real eigenvectors $\boldsymbol{x}^*$ and $\boldsymbol{x}^{*\prime}$. Both can therefore be used to get the n.s.i. version of another popular centrality measure: the *(adjacency) eigenvector centrality* EC_v of v is the entry x_v of the (non-negative) eigenvector $\boldsymbol{x}$ for A's largest eigenvalue, where $\boldsymbol{x}$ is usually normalized so that $\max_v x_v = 1$. The above shows that we can define *n.s.i. eigenvector centrality* as $EC_v^* = x_v^{*\prime} = x_v^*/\sqrt{w_v}$. A similar measure based on the normal matrix T is closely related to a major web search engine's page rank measure.

In the brain example network, the top ten central nodes according to EC_v are PCUN.L, LING.R, CAL.L, PCUN.R, DCG.L, DCG.R, MTG.R, SFGmed.L, CUN.L, and PCG.R, and according to EC_v^* they are PCUN.L, MTG.R, MFG.L, SFGmed.L, PCUN.R, DCG.L, LING.R, DCG.R, CAL.L, and MFG.R. Curiously, this time the second rather than the first ranking includes MFG.L/R which in this case is probably just because they have a much larger weight than CUN.L and PCG.R (which are only in the first ranking).

For the Laplacian matrix Λ, also both constructions are possible, leading to

$$\Lambda^* = \operatorname{diag}(\boldsymbol{k}^*) - \mathsf{A}^* \quad \text{and} \quad \Lambda^{*\prime} = \operatorname{diag}(\boldsymbol{k}^*) - \mathsf{A}^{*\prime}. \quad (44)$$

While the row sums of Λ^* are zero just like for Λ, those of $\Lambda^{*\prime}$ need not equal zero, but the latter matrix is symmetric. Still, both matrices have the same eigenvalues, and those and their eigenvectors behave in the same way under node splitting as those of A^* and $\mathsf{A}^{*\prime}$, except that the additional $(N+1)$st eigenvalue is now k_s^* instead of zero.

The non-symmetric version Λ^* can also be interpreted as a Laplacian matrix of a directed network in which the links instead of the nodes are weighted, with link weights $w_{ij} = w_i$. If the network is the result of the spectral coarse graining procedure described in [46], Λ^* equals the Laplacian derived in [46], equation (3).

The *spectral bisection method* of Fiedler [47] uses the signs of the eigenvector of the smallest positive eigenvalue of Λ to find the most distinguishable two groups of nodes. Using either Λ^* or $\Lambda^{*\prime}$ in the same way will provide the *n.s.i. spectral bisection*, both giving the same result since only the sign of the eigenvector entries matters.

An enhanced version of spectral bisection uses eigenvectors of Newman's [24] generalized modularity matrix to iteratively find communities. Figure 2 shows three-group solutions found by the unweighted and n.s.i. versions of that algorithm, as described in Appendix B (online only), in the world trade example network. The unweighted version and the n.s.i. version using GDP as node weight plausibly place Europe and North Africa in one group, most of
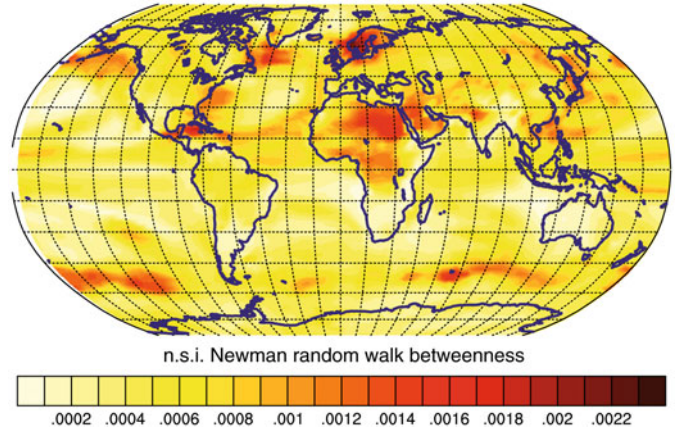

n.s.i. Newman random walk betweenness

.0002 .0004 .0006 .0008 .001 .0012 .0014 .0016 .0018 .002 .0022

Fig. 13. (Colour online) N.s.i. version NB_v^* of Newman's random walk betweenness in a global climate network representing correlations in surface air temperature dynamics (same network as in Figure 4, Robinson projection). We can clearly identify the regions of the North Pacific Subpolar Gyre, the North Atlantic Subtropical Gyre including the Gulf Stream and the Canary Current, the North and South Equatorial Currents in the Pacific, and the Antarctic Circumpolar Current. The interpretation of other regions of high values like Scandinavia and Central and North-East Africa remains unclear.

Asia in another, and the Americas in the third, with little differences in the placement of some "bordering" countries. (The placement of UGD (bottom left) in the blue group is an artifact of the algorithm due to its large distance from the network's centre.) When using population as node weight, the n.s.i. result however differs considerably, placing China and India in different groups, since the algorithm tends to produce groups of approximately equal weight.

7 Application to climate networks

Coming back to our original field of application, let us finally compare the unweighted and n.s.i. versions of a number of network measures in the case of a climate network whose node set is a latitude-longitude-regular grid on the Earth's surface, with a resolution of $2.5°$ in latitude and $3.75°$ in longitude. As in Donges et al. [11,31,33], two of these 6 816 nodes (the two poles have been excluded) were linked when the corresponding time-series of monthly averaged surface air temperature (SAT) anomalies from the 20th century reference run 20c3m of the Hadley Centre's HadCM3 model (as defined in the IPCC's Fourth Assessment Report, see [11,31] for details) showed a significant Pearson correlation coefficient. We chose the link inclusion threshold so as to achieve a relatively high link density of approx. 0.05, for which all correlations of absolute value of at least 0.25 were considered as significant.

As can be seen, e.g., from the resulting n.s.i. Newman-type random walk betweenness (Fig. 13), this network retains a number of interesting features of the global climate system. As argued in [11], increased values of Newman's random walk betweenness may be indicative of diffusive

transport processes in the climate system, e.g., turbulent eddy diffusion in the atmosphere and ocean, whereas shortest path betweenness is believed to trace advective transport processes such as strong surface ocean currents.

For comparison, we defined a second, synthetic network on the same set of nodes in which we linked each pair of nodes i, j independently with a probability of $\min(1, \exp(0.4 - 0.09\alpha_{ij}))$, where α_{ij} is the angular distance between the nodes (in degrees). The exponential relationship between link probabity and distance was fitted to the relationship between the observed link density and angular distance in the above climate network, using non-linear regression (a similar relationship was found in [27]). The resulting *benchmark network* had a slightly smaller link density of approx. 0.035 and can be interpreted as a sample from an underlying continuous network whose link distribution depends on angular distance alone and is therefore rotationally and translationally symmetric. Because of this underlying symmetry, local network measures suitable for the estimation of underlying features should not show a significant depency on the node's latitude. The thin dashed and dotted lines in Figure 4 show the longitudinally averaged values of several network measures, plotted against latitude, in this benchmark network. We can see that the n. s. i. versions (dotted lines, using cos(latitude) as node weight) fulfil this requirement of latitude independence much better than the unweighted versions (thin dashed lines), which exhibit a clear systematical increase either towards the poles (degree, clustering coefficient, and closeness centrality) or towards the equator (Newman-type random walk betweenness). For the degree, this is due to the increase in *absolute* node density, while for the clustering coefficient, this is due to the increase of the density *gradient* towards the pole.

As can be expected from this, the corresponding values in the real-world climate network also show differences between the unweighted version (solid lines) and the n. s. i. version (thick dashed lines) towards the poles or towards the equator. Unweighted Newman-type random walk betweenness, e.g., is higher in the region of the North and South Equatorial Currents at about ±10° latitude, while its n. s. i. version is higher in the region of the North Atlantic Subtropical Gyre between +15° and +60° latitude, although both show well-defined features in both regions.

The influence of the increased node density in the high latitudes on unweighted network statistics becomes even more evident when focussing on the Arctic region north of +60° (Fig. 3). While the unweighted degree and clustering coefficient are markedly increased close to the North Pole, their geographic distribution is considerably obscured by the node density induced bias further southwards (Figs. 3A, 3B). In contrast, the n. s. i. variants of degree and clustering coefficient reveal more pronounced regional structures, e.g., increased n. s. i. degree over southern Greenland and Scandinavia, or increased n. s. i. clustering coefficient surrounding Greenland (Figs. 3C, 3D). Hence, we judge n. s. i. network measures to be very promising tools for future analysis of

gridded climate data with inhomogeneous mesh cell areas or station data, particularly since the additional error that would be introduced by interpolation to equal area (geodesic) grids can thus be avoided.

8 Conclusion

To summarize, in this article we have introduced a fairly general framework to deal with biases and artifacts in complex network statistics that appear when the nodes represent differently sized parts of an underlying domain of interest. Networks of this type are routinely studied in various fields of research, including neuroscience, the Earth sciences, informatics and engineering, social sciences, economics, and dynamical systems, as our examples show, and network design choices made when describing the domain of interest by a complex network can have considerable effects on the results of network statistics.

The central axiomatic notion of node splitting (or twin merging) invariance provides an elegant means to tackle the problem of how to use information on the size of (the part of the domain of interest represented by) individual nodes in a way that is robust against different choices of node selection, grid, meshing, parcellation, aggregation, or coarse-graining. Our framework allowed us to derive consistently weighted versions of a large representative set of commonly used statistical network measures that quantify different aspects of networks in which links represent some kind of similarity or closeness. Despite the diversity of these n. s. i. measures, a simple set of design rules (given in Sect. 4.3) guided the introduction of node weights into their definitions. The resulting formulas are in most cases computationally no more demanding than the original ones, and slightly modified versions of standard algorithms with the same complexity can usually be used to apply them in practice. Also, since the construction of all our measures is based on the same first principles, they work better together and allow for an easier interpretation than alternative ad hoc approaches might.

Most importantly, n. s. i. measures reflect the features of the domain of interest more accurately than classical unweighted measures. This was demonstrated in particular in the case of a synthetic and a real-world climate network, for which it was possible to compare the results of n. s. i. measures on a non-homogeneous grid with those based on a homogeneous (geodesic) grid, both showing exactly the same features and avoiding the artifacts that were produced by unweighted measures on the non-homogeneous grid. We have further illustrated the applicability and practical relevance of our axiomatic approach by qualitatively showing their effects in a number of semi-realistic example networks. In particular, we showed how in many of these examples, the judgement of which parts of the network are the most central or otherwise structurally important can change when node weights are used.

Our results also indicate that the topological properties of network representations of technical infrastructure such as the internet depend on whether the sometimes considerably varying size of the individual subsystems

chosen as nodes in the network representation is taken into account or not, and we conjecture that n. s. i. measures will prove highly relevant and beneficial for the consistent analysis of the vulnerability of distributed technical systems. When analyzing how the connectivity of the internet decreases due to targeted attacks [48], the size of both the attacked and the affected autonomous systems should be of obvious interest. A more thorough study of the AS network used for illustration here is of course beyond the scope of this methodological paper. It would have to construct the links much more carefully by taking additional data sources into account, as described in [14], and verify that the number of CIDR prefixes is indeed a suitable size measure for autonomous systems. We also leave for future research the general question of how one should choose node weights suitable for a given network and research question.

Appendix A: Measures corrected for a typical weight

Since we introduced n. s. i. measures f^* as weighted versions of existing network measures f, it is natural to consider the case where all node weights w_v are equal to some constant value ω. Usually, because of the construction mechanisms (i) and (ii) given in Section 4.3, f^* will not exactly equal f in that case but rather be some (usually simple) transformation of it, and only for $N \to \infty$, f^* is often asymptotic to f, e.g., $k_v^* = (k_v + 1)\omega = \omega k_v + |o(k_v)| \sim k_v$. When one wants to compare values of f^* with those of f, this behaviour presents some difficulties, hence we will present here for most of the treated measures a second, somewhat more complex, *corrected n. s. i.* version of f^*, denoted by $f^{*\omega}$, which is also n. s. i., but which involves a parameter $\omega > 0$ and has the property that

$$f^{*\omega} = f \quad \text{whenever} \quad w_v = \omega \text{ for all } v \in \mathcal{N}, \quad (A.1)$$

so that $f^{*\omega}$ can be compared to f more easily than f^*. The parameter ω is then called the *typical weight* and can be thought of as a kind of resolution or scale on which the analysis focuses. The question of how a suitable value for ω can be estimated from G and w if the weights are not all equal is addressed later, in the next subsection.

In many cases, the following construction mechanisms are helpful in addition to (i)–(iv) given in Section 4.3 to derive $f^{*\omega}$ from f^*:

(v) Correct for the effect of (i) by replacing each occurrence of w_i by w_i/ω, and

(vi) correct for the effects of (ii) and (iii) by subtracting suitable terms (often constants) from sums over nodes.

It will be convenient to express the corrections in terms of the corrected version of W, which can be called the *corrected n. s. i. number of nodes,*

$$N^{*\omega} = W/\omega. \quad (A.2)$$

In the example of *degree,* we apply both (v) and (vi) and put

$$k_v^{*\omega} = \sum_{i \in \mathcal{N}_v^+} w_i/\omega - 1 = k_v^*/\omega - 1 \quad (A.3)$$

which obviously reduces to k_v if $w_v \equiv \omega$. In case of non-constant weights, it can happen that $k_v^{*\omega}$ turns out to be negative for some v if ω is chosen too large. The same effect can happen for corrected n. s. i. versions of other measures, as will be obvious from their definitions presented below. In general, negative values can be avoided by lowering ω or by replacing them by zero.

For the *local clustering coefficient,* we have

$$C_v^* (w_i \equiv \omega) = \frac{C_v k_v (k_v - 1) + 3k_v + 1}{(k_v + 1)^2} = C_v + \left| O\left(\frac{1}{k_v}\right) \right|.$$

A *corrected n. s. i. local clustering coefficient* can only be defined for the case that $k_v^{*\omega} > 1$,

$$C_v^{*\omega} = \frac{(N^{*\omega})^2 \langle a_{vi}^+ a_{ij}^+ a_{jv}^+ \rangle_{ij}^w - 3k_v^{*\omega} - 1}{k_v^{*\omega} (k_v^{*\omega} - 1)} \leqslant 1, \quad (A.4)$$

where, following (vi), we subtract $(3k_v^{*\omega} + 1)$ since in $\langle a_{vi}^+ a_{ij}^+ a_{jv}^+ \rangle_{ij}^w$, the nodes i and j can be equal or equal to v. For $\omega \leqslant w_v/2$, one can prove that $C_v^{*\omega} \geqslant 0$.

Also *corrected n. s. i. closeness centrality* measures can be derived via (v) and (vi):

$$CC_v^{*\omega} = \frac{1}{1/CC_v^* - 1/N^{*\omega}} \leqslant 1,$$

$$CC_v'^{*\omega} = CC_v'^* - 1/2N^{*\omega} \leqslant 1,$$

$$CC_v''^{*\omega} = CC_v''^* - 1/N^{*\omega} \leqslant 1. \quad (A.5)$$

In case of *(shortest path) betweenness centrality* $BC_v^* = \langle n_{ab}^*(v)/n_{ab}^* \rangle_{ab}^w$, we only have to divide n_{ab}^* and $n_{ab}^*(v)$ by a suitable power of ω, according to (v). Subtractions as in (vi) are unnecessary since we did not extend any sums to derive BC_v^* from BC_v. Hence the corrected n. s. i. versions are

$$n_{ab}^{*\omega} = \omega^{1-d_{ab}} n_{ab}^*, \quad n_{ab}^{*\omega}(v) = \omega^{2-d_{ab}} n_{ab}^*(v),$$

$$BC_v^{*\omega} = \langle n_{ab}^{*\omega}(v)/n_{ab}^{*\omega} \rangle_{ab}^w = \omega BC_v^* \in [0, \omega/w_v]. \quad (A.6)$$

For *random walk-based measures,* a corrected n. s. i. version of the transition probabilities p_{ij} is only obvious in the pathological case in which $\omega \leqslant \min_v w_v$ and when no isolated nodes exist. We could then put

$$p_{ij}^{*\omega} = a_{ij}^+ \left(w_j/\omega - \delta_{ij} \right)/k_i^{*\omega}. \qquad (A.7)$$

For larger, more realistic choices of typical weight, this definition of $p_{ii}^{*\omega}$ would result in negative values for many i and could thus not be interpreted as a transition probability. We will therefore not present corrected versions of random-walk based measures.

The *aggregate statistics* we presented also allow for corrected n. s. i. versions: for the *global clustering coefficient* it is

$$C^{*\omega} = \langle C_v^{*\omega} \rangle_v^w, \qquad (A.8)$$

while for *transitivity* it is

$$T^{*\omega} = \frac{(N^{*\omega})^3 \langle a_{vi}a_{ij}a_{jv} \rangle_{vij} - N^{*\omega} \left(3 \langle k_v^{*\omega} \rangle_v^w + 1 \right)}{(N^{*\omega})^3 \langle a_{iv}a_{vj} \rangle_{vij} - N^{*\omega} \left(3 \langle k_v^{*\omega} \rangle_v^w + 1 \right)} \leqslant 1. \qquad (A.9)$$

For the *link density, average (geodesic) path length,* and *global efficiency,* it is just

$$\varrho^{*\omega} = \varrho^* - \frac{1}{N^{*\omega}}, \quad L^{*\omega} = L^* - \frac{1}{N^{*\omega}}, \quad E^{*\omega} = E^* - \frac{1}{N^{*\omega}}. \qquad (A.10)$$

For *spectral analysis,* the corrected n. s. i. versions of *adjacency* A and *Laplacian* Λ are

$$\mathsf{A}^{*\omega} = \frac{1}{\omega}\mathsf{A}^* - \mathsf{I}, \qquad \Lambda^{*\omega} = \operatorname{diag}\left(\boldsymbol{k}^{*\omega} \right) - \mathsf{A}^{*\omega} = \frac{1}{\omega}\Lambda^*,$$

$$\mathsf{A}^{*\prime\omega} = \frac{1}{\omega}\mathsf{A}^{*\prime} - \mathsf{I}, \qquad \Lambda^{*\prime\omega} = \operatorname{diag}\left(\boldsymbol{k}^{*\omega} \right) - \mathsf{A}^{*\prime\omega} = \frac{1}{\omega}\Lambda^{*\prime}. \qquad (A.11)$$

In particular, $EC_v^*(w_i \equiv \omega) = EC_v$, hence n. s. i. eigenvector centrality needs no correction, and also unweighted and n. s. i. spectral bisections are identical if $w_i \equiv \omega$.

A.1 Estimation of typical weight

The usefulness of our corrected versions of the n. s. i. measures depends on a suitable choice of the typical weight ω. In applications in which the underlying domain of interest does not provide a natural choice for ω, we can try to determine a suitable ω from the network itself. Such an "estimate" $\tilde{\omega}$ of ω should ideally (i) depend monotonically on the node weights w_v, (ii) lie in $[\min_v w_v, \max_v w_v]$, (iii) be n. s. i. itself, and (iv) be small enough so that measures such as $k_v^{*\omega}$ and $C_v^{*\omega}$ are defined and non-negative for all or at least almost all nodes. In addition, it would be nice if (v) $\tilde{\omega}$ is statistically robust, i.e., cannot change unboundedly by only local changes to the network.

To fulfil (i), (ii), and (v), the most natural choice seems to be the median node weight. With a small adjustment,

also (iii) is satisfied: define the *n. s. i. twin-adjusted weight* of v as

$$w_v'^* = \sum_{i \in \mathcal{N}_v^+} w_i \delta_{\mathcal{N}_i^+ \mathcal{N}_v^+} \in [w_v, k_v^*] \qquad (A.12)$$

and let $\tilde{\omega}_I$ be the w-weighted median of $w_v'^*$. This fulfils (i)–(iii) and (v), though not necessarily (iv). Only when there are pathologically many twins, $\tilde{\omega}_I$ might exceed $\max_v w_v$.

A different approach is to address (iv) first by putting

$$w_v''^* = \frac{3}{4}k_v^* - \sqrt{\frac{9}{16}k_v^{*2} - \frac{1}{2}\Delta_v^*} \in \left(0, \frac{1}{2}k_v^* \right]. \qquad (A.13)$$

Using $\omega = \min_v w_v''^*$ would then ensure that for all v, $k_v^{*\omega} \geqslant 1$ and both $C_v^{*\omega}$ and $C_v'^{*\omega}$ are defined and non-negative except maybe for those few v where the above minimum is attained. When v is not isolated, $w_v''^* \geqslant \min_v w_v$, but it may easily exceed $\max_v w_v$. Hence a good choice that fulfils (i)–(iv), though not (v), is

$$\tilde{\omega}_{II} = \min\left(\tilde{\omega}_I, \min_v w_v''^* \right). \qquad (A.14)$$

Finally, a trade-off between (iv) and (v) can be made by using in this definition not $\min_v w_v''^*$ but a small quantile, say the first percentile $P_1^w(w_v''^*)$, of the w-weighted distribution of $w_v''^*$:

$$\tilde{\omega}_{III} = \min\left(\tilde{\omega}_I, P_1^w \left(w_v''^* \right) \right). \qquad (A.15)$$

Appendix B: Some additional measures

The *average (nearest) neighbours' degree* of v,

$$k_{nn,v} = \sum_{i \in \mathcal{N}_v} k_i/k_v, \qquad (B.1)$$

represents the average size of the region a point linked to v is linked to. Using the plug-in mechanism (iv) and the correction mechanism (v) and (vi), we define its weighted versions, the *n. s. i. average (nearest) neighbours' degree* and the *corrected n. s. i. average (nearest) neighbours' degree,* as

$$k_{nn,v}^* = \sum_{i \in \mathcal{N}_v^+} w_i k_i^*/k_v^*,$$

$$k_{nn,v}^{*\omega} = \sum_{i \in \mathcal{N}_v^+} w_i k_i^{*\omega}/\omega k_v^{*\omega} - 1.$$

These are n. s. i. since their ingredients k_i^*, k_v^*, $k_i^{*\omega}$, and $k_v^{*\omega}$ are. In case of constant weights $w_i \equiv \omega$, we have $k_{nn,v}^{*\omega} = k_{nn,v}$, whereas the uncorrected version is then

$$k_{nn,v}^* \left(w_i \equiv \omega \right) = \frac{k_{nn,v}k_v + 2k_v + 1}{k_v + 1}\omega$$

$$= \left(k_{nn,v} + 2 \right)\omega + O\left(\frac{1}{k_v} \right).$$

The Pearson product-moment correlation coefficient between the degrees of the two end nodes of each link in G

is called *degree correlation* or *assortativity* and can be computed from the degrees and average neighbours' degrees as follows:

$$r = \frac{\langle k_v^2 k_{nn,v}\rangle_v \langle k_v\rangle_v - \left(\langle k_v^2\rangle_v\right)^2}{\langle k_v^3\rangle_v \langle k_v\rangle_v - \left(\langle k_v^2\rangle_v\right)^2}. \tag{B.2}$$

A *[corrected] n. s. i. degree correlation* is most easily found using the plug-in mechanism (iv):

$$r^* = \frac{\langle k_v^{*2} k_{nn,v}^*\rangle_v^w \langle k_v^*\rangle_v^w - \left(\langle k_v^{*2}\rangle_v^w\right)^2}{\langle k_v^{*3}\rangle_v^w \langle k_v^*\rangle_v^w - \left(\langle k_v^{*2}\rangle_v^w\right)^2},$$

$$r^{*\omega} = \frac{\langle k_v^{*\omega 2} k_{nn,v}^{*\omega}\rangle_v^w \langle k_v^{*\omega}\rangle_v^w - \left(\langle k_v^{*\omega 2}\rangle_v^w\right)^2}{\langle k_v^{*\omega 3}\rangle_v^w \langle k_v^{*\omega}\rangle_v^w - \left(\langle k_v^{*\omega 2}\rangle_v^w\right)^2}.$$

Note that r^* is the common Pearson correlation coefficient between k_i^* and k_j^* in the probabilistic model in which the link or self-loop i-j is drawn with relative probability $w_i w_j$. The latter measures the number of links in G_0 between the regions represented by i and j, hence r^* estimates the correlation coefficient between the degrees of the two end points of all links in G_0.

Soffer and Vázquez [50] justify a version of the *clustering coefficient C_v* which is partially adjusted for degree correlations:

$$C_v' = \frac{N^2 \langle a_{vi} a_{ij} a_{jv}\rangle_{ij}}{\sum_{i\in\mathcal{N}_v}(\min(k_i, k_v) - 1)} \in [C_v, 1]. \tag{B.3}$$

Again using all four mechanisms (i)–(iv), we get their n. s. i. versions:

$$C_v'^* = \frac{W^2 \langle a_{vi}^+ a_{ij}^+ a_{jv}^+\rangle_{ij}^w}{\sum_{i\in\mathcal{N}_v^+} w_i \min(k_i^*, k_v^*)} \in [C_v^*, 1] \quad \text{and}$$

$$C_v'^{*\omega} = \frac{(N^{*\omega})^2 \langle a_{vi}^+ a_{ij}^+ a_{jv}^+\rangle_{ij}^w - 3k_v^{*\omega} - 1}{\sum_{i\in\mathcal{N}_v^+} \frac{w_i}{\omega}\min(k_i^{*\omega}, k_v^{*\omega}) - 2k_v^{*\omega}} \in [C_v^{*\omega}, 1],$$

where we define the latter only for the case where $\sum_{i\in\mathcal{N}_v^+}\frac{w_i}{\omega}\min(k_i^{*\omega}, k_v^{*\omega}) > 2k_v^{*\omega}$.

Bonacich [51] defined a measure of v's *power centrality* based on the idea that a node's "power" in a network should be the sum of a linear function of each of its neighbours' powers:

$$PC_v = \sum_{i\in\mathcal{N}} a_{vi}(\alpha + \beta PC_i) \tag{B.4}$$

for some parameters $\alpha, \beta > 0$. This implicit definition is solved by the power centrality vector

$$\boldsymbol{PC} = (PC_i)_{i\in\mathcal{N}} = \alpha\left(\mathsf{I} - \beta\mathsf{A}\right)^{-1}\boldsymbol{k}, \tag{B.5}$$

where $\boldsymbol{k} = (k_i)_{i\in\mathcal{N}}$ is the degree vector, assuming that $\mathsf{I} - \beta\mathsf{A}$ is invertible (which will be true for a general choice of β). To find the n. s. i. version, we make the defining

equation n. s. i. and require that v's power is the weighted sum of the linear function of its neighbours' powers:

$$PC_v^* = \sum_{i\in\mathcal{N}} a_{vi}^+ w_i(\alpha + \beta PC_i^*). \tag{B.6}$$

Whenever $\mathsf{I} - \beta\mathsf{A}^+\mathsf{D}_w$ is invertible, the solution

$$\boldsymbol{PC}^* = \alpha\left(\mathsf{I} - \beta\mathsf{A}^*\right)^{-1}\boldsymbol{k}^* \tag{B.7}$$

is then the sought *n. s. i. power centrality* vector, where $\boldsymbol{k}^*$ is the n. s. i. degree vector, $\boldsymbol{k}^* = (k_i^*)_{i\in\mathcal{N}}$.

Following (v) and (vi) again, the corrected n. s. i. equation is

$$PC_v^{*\omega} = \sum_{i\in\mathcal{N}} a_{vi}^+ \frac{w_i}{\omega}(\alpha + \beta PC_i^{*\omega}) - (\alpha + \beta PC_v^{*\omega}), \tag{B.8}$$

and its solution is the *corrected n. s. i. power centrality* vector

$$\boldsymbol{PC}^{*\omega} = \alpha\left(\mathsf{I} - \beta\mathsf{A}^{*\omega}\right)^{-1}\boldsymbol{k}^{*\omega}. \tag{B.9}$$

The *random walk centrality* of v as defined by Noh and Rieger [52] measures how fast a walk starting at a random node reaches v:

$$RWC_v = p_v \Big/ \sum_{t=0}^{\infty}\left((P^t)_{vv} - p_v\right). \tag{B.10}$$

The obvious n. s. i. version of this is

$$RWC_v^* = p_v^* \Big/ \sum_{t=0}^{\infty}\left(\left((P^*)^t\right)_{vv} - p_v^*\right).$$

The *normal matrix* $\mathsf{T} = \mathrm{diag}(\boldsymbol{k})^{-1}\mathsf{A}$ is also often used as the basis of a centrality measure and has the n. s. i. versions

$$\mathsf{T}^* = \mathrm{diag}\left(\boldsymbol{k}^*\right)^{-1}\mathsf{A}^*, \quad \mathsf{T}^{*\omega} = \mathrm{diag}\left(\boldsymbol{k}^{*\omega}\right)^{-1}\mathsf{A}^{*\omega} \tag{B.11}$$

which both retain T's property of having row sums equal to one.

B.1 Modularity

Following Newman [24], the *modularity* of a partition $\mathcal{P}$ of $\mathcal{N}$ expresses how well the groups defined by $\mathcal{P}$ are internally connected and separated from each other. It is defined as the observed within-group link density minus its expected value given the observed degree distribution:

$$\mathsf{Q} = \frac{\langle\delta_{\mathcal{P}(i)\mathcal{P}(j)}B_{ij}\rangle_{ij}}{\langle a_{ij}\rangle_{ij}} \quad \text{with} \quad B_{ij} = a_{ij} - \frac{k_i k_j}{N\langle k_v\rangle_v}, \tag{B.12}$$

where $\mathcal{P}(i)$ is that set in $\mathcal{P}$ which contains i. The (corrected) n. s. i. versions of these are:

$$\mathsf{Q}^* = \frac{\langle\delta_{\mathcal{P}(i)\mathcal{P}(j)}B_{ij}^+\rangle_{ij}^w}{\langle a_{ij}^+\rangle_{ij}^w},$$

$$B_{ij}^+ = a_{ij}^+ - \frac{k_i^* k_j^*}{W\langle k_v^*\rangle_v^w},$$

$$Q^{*\omega} = \frac{\left\langle \delta_{\mathcal{P}(i)\mathcal{P}(j)} B_{ij}^{+\omega} \right\rangle_{ij}^{w} - 1/N^{*\omega}}{\left\langle a_{ij}^{+} \right\rangle_{ij}^{w} - 1/N^{*\omega}},$$

$$B_{ij}^{+\omega} = a_{ij}^{+} - \frac{k_i^{*\omega} k_j^{*\omega}}{N^{*\omega} \left\langle k_v^{*\omega} \right\rangle_v^{w}}.$$

Q^* estimates the within-group link density in G_0 minus its expected value given the degree distribution in G_0, for the partition of $\mathcal{N}_0$ induced by the partition $\mathcal{P}$ of $\mathcal{N}$.

Similar to the matrices A, $\mathsf{\Lambda}$, and T, also the *spectrum of the modularity matrix* $\mathsf{B} = (B_{ij})_{ij}$ can be made n. s. i., using the matrices $\mathsf{B}^{+} = (B_{ij}^{+})_{ij}$ and $\mathsf{B}^{+\omega} = (B_{ij}^{+\omega})_{ij}$:

$$\mathsf{B}^* = \mathsf{D}_w^{1/2} \mathsf{B}^{+} \mathsf{D}_w^{1/2}, \qquad \mathsf{B}^{*\omega} = \frac{1}{\omega} \mathsf{D}_w^{1/2} \mathsf{B}^{+\omega} \mathsf{D}_w^{1/2} - \mathsf{I}.$$

Given a subset $g \subseteq N$ of the nodes, the *generalized modularity matrix* is the $|g| \times |g|$ matrix $\mathsf{B}^{(g)} = (B_{ij}^{(g)})_{ij}$ with

$$B_{ij}^{(g)} = B_{ij} - \delta_{ij} \sum_{v \in g} B_{iv}. \tag{B.13}$$

Newman [24] uses the signs in the eigenvector of its largest positive eigenvalue in an efficient iterative *network dividing algorithm* similar to Fiedler's spectral bisection method, in which a hierarchical clustering tree is constructed top-down by starting with $g_0 = \mathcal{N}$, bisecting into two groups g_1, g_2 according to the eigenvector signs, and then repeating with each g_i thus obtained. The following versions of $\mathsf{B}^{(g)}$ can be used to derive similar n. s. i. network divisions, as exemplified in Figure 2:

$$B_{ij}^{(g)*} = B_{ij}^* - \delta_{ij} \sum_{v \in g} w_v B_{iv}^+,$$
$$B_{ij}^{(g)*\omega} = B_{ij}^{*\omega} - \delta_{ij} \sum_{v \in g} \left(\frac{w_v}{\omega} B_{iv}^{+\omega} - \delta_{iv} \right).$$

If we assume that when splitting $s \to s' + s''$ with $s \in g$, both s' and s'' are put into the new g, then the eigenvectors of these matrices have the same n. s. i. properties as those of B^* above. In particular, the eigenvector entries for s' and s'' have the same sign as that of s, whereas all other signs are unchanged, hence the division of g will be the same after the split except that s is replaced by s' and s'' in its subgroup.

References

1. M.E.J. Newman, SIAM Rev. **45**, 167 (2003)
2. S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D. Hwang, Phys. Rep. **424**, 175 (2006)
3. L. da F Costa, F.A. Rodrigues, G. Travieso, P.R. Villas Boas, Adv. Phys. **56**, 167 (2007)
4. M.E.J. Newman, *Networks: An Introduction* (Oxford University Press, 2010)
5. R. Cohen, S. Havlin, *Complex Networks: Structure, Robustness and Function* (Cambridge University Press, 2010)
6. E. Kolaczyk, *Statistical Analysis of Network Data: Methods and Models,* Springer Series in Statistics (Springer, 2009)
7. S. Achard, R. Salvador, B. Whitcher, J. Suckling, E. Bullmore, J. Neurosci. **26**, 63 (2006)
8. C. Zhou, L. Zemanová, G. Zamora, C. Hilgetag, J. Kurths, Phys. Rev. Lett. **97**, 238103 (2006)
9. P. Hagmann, L. Cammoun, X. Gigandet, R. Meuli, C.J. Honey, V.J. Wedeen, O. Sporns, PLOS Biol. **6**, e159 (2008)
10. A.A. Tsonis, P.J. Roebber, Physica A **333**, 497 (2004)
11. J.F. Donges, Y. Zou, N. Marwan, J. Kurths, Europhys. Lett. **87**, 48007 (2009)
12. R. Pastor-Satorras, A. Vázquez, A. Vespignani, Phys. Rev. Lett. **87**, 258701 (2001)
13. R. Pastor-Satorras, A. Vázquez, A. Vespignani, Topology, hierarchy, and correlations in Internet graphs, In *Complex Networks*, edited by E. Ben-Naim and Z. Toroczkai (2004), Vol. 650, pp. 425–440.
14. Q. Chen, H. Chang, R. Govindan, S. Jamin, S. Shenker, W. Willinger, The origin of power-laws in internet topologies revisited, in *IEEE INFOCOM* (2002), Vol. 2, pp. 608–617
15. G. Siganos, M. Faloutsos, P. Faloutsos, C. Faloutsos, IEEE/ACM Trans. Netw. **11**, 514 (2003)
16. A. Capocci, V. Servedio, F. Colaiori, L. Buriol, D. Donato, S. Leonardi, G. Caldarelli, Phys. Rev. E **74**, 036116 (2006)
17. V. Zlatić, H. Štefančić, Europhys. Lett. **93**, 58005 (2011)
18. M. Comola, M. Fafchamps, CEPR Discussion Papers 7406, C.E.P.R. Discussion Papers, 2009
19. M. Comola, M. Fafchamps, CSAE Working Paper Series 2010-20, Paris-Jourdan Sciences Économiques, 2010
20. M.A. Serrano, M. Boguñá, Phys. Rev. E **68**, 015101 (2003)
21. N. Marwan, J.F. Donges, Y. Zou, R.V. Donner, J. Kurths, Phys. Lett. A **373**, 4246 (2009)
22. R.V. Donner, Y. Zou, J.F. Donges, N. Marwan, J. Kurths, New. J. Phys. **12**, 033025 (2010)
23. R.V. Donner, Y. Zou, J.F. Donges, N. Marwan, J. Kurths, Phys. Rev. E **81**, 015101 (2010)
24. M.E.J. Newman, Proc. Natl. Acad. Sci. USA **103**, 8577 (2006)
25. A. Barrat, M. Barthelemy, Proc. Natl. Acad. Sci. USA **101**, 3747 (2004)
26. A.A. Tsonis, K.L. Swanson, P.J. Roebber, B. Am. Meteorol. Soc. **87**, 585 (2006)
27. A.A. Tsonis, K.L. Swanson, Phys. Rev. Lett. **100**, 228502 (2008)
28. K. Yamasaki, A. Gozolchiani, S. Havlin, Phys. Rev. Lett. **100**, 228501 (2008)
29. A. Gozolchiani, K. Yamasaki, O. Gazit, S. Havlin, Europhys. Lett. **83**, 28005 (2008)
30. R.V. Donner, T. Sakamoto, N. Tanizuka, in *Nonlinear Time Series Analysis in the Geosciences: Applications in Climatology, Geodynamics and Solar-Terrestrial Physics*, edited by R.V. Donner, S.M. Barbosa (Springer, 2008), pp. 125–154
31. J.F. Donges, Y. Zou, N. Marwan, J. Kurths, Eur. Phys. J. Special Top. **174**, 157 (2009)
32. J.A. Henderson, P.A. Robinson, Phys. Rev. Lett. **107**, 018102 (2011)
33. J.F. Donges, H.C.H. Schultz, N. Marwan, Y. Zou, J. Kurths, Eur. Phys. J. B **84**, 635 (2011)
34. G.R. North, T.L. Bell, R.F. Cahalan, Mon. Weather Rev. **110**, 699 (1982)
35. S.H. Lee, P.J. Kim, H. Jeong, Phys. Rev. E **73**, 016102 (2006)
36. R. Rohde, J. Curry, D. Groom, R. Jacobsen, R.A. Muller, S. Perlmutter, A. Rosenfeld, C. Wickham, J. Wurtele, *Under Review* (2011), pp. 1–39

37. T. Plewa, T.J. Linde, V.G. Weirs, in *Adaptive mesh refinement, theory and applications, Lecture notes in computational science and engineering* (Springer, 2005), Vol. 41
38. C. Dangalchev, Physica A **365**, 556 (2006)
39. K.A. Stephenson, M. Zelen, Soc. Networks **11**, 1 (1989)
40. V. Latora, M. Marchiori, Physica A **314**, 109 (2002)
41. M.E.J. Newman, Phys. Rev. E **64**, 016132 (2001)
42. A. Arenas, A. Cabrales, A. Díaz-Guilera, R. Guimerà, F. Vega-Redondo, in *Statistical Mechanics of Complex Networks, Lecture Notes in Physics*, edited by R. Pastor-Satorras, M. Rubi, A. Díaz-Guilera (Springer Berlin, Heidelberg, 2003), Vol. 625, pp. 175–194
43. M.E.J. Newman, Soc. Networks **27**, 39 (2005)
44. S. Bialonski, M.T. Horstmann, K. Lehnertz, Chaos **20**, 013134 (2010)
45. J. Barcelo, Comput. Netw. **45**, 333 (2004)
46. D. Gfeller, P. De Los Rios, Phys. Rev. Lett. **100**, 174104 (2008)
47. M. Fiedler, Czech. Math. J. **23**, 298 (1973)
48. R. Albert, H. Jeong, A.L. Barabási, Nature **406**, 378 (2000)
49. G. Csárdi, T. Nepusz, InterJournal Complex Systems **CX. 18**, 1695 (2006), http://igraph.sf.net
50. S.N. Soffer, A. Vázquez, Phys. Rev. E **71**, 057101 (2005)
51. P. Bonacich, Am. J. Sociol. **92**, 1170 (1987)
52. J.D. Noh, H. Rieger, Phys. Rev. Lett. **92**, 118701 (2004)