epl A Letters Journal Exploring the Frontiers of Physics

October 2012

EPL, **100** (2012) 28002 www.epljournal.org
doi: 10.1209/0295-5075/100/28002

Boundary effects in network measures of spatially embedded networks

Aljoscha Rheinwalt[1,2(a)], Norbert Marwan[1], Jürgen Kurths[1,2], Peter Werner[1,2] and Friedrich-Wilhelm Gerstengarbe[1,2]

[1] *Potsdam Institute for Climate Impact Research - P.O. Box 60 12 03, D-14412 Potsdam, Germany, EU*
[2] *Humboldt-Universität zu Berlin - Unter den Linden 6, D-10099 Berlin, Germany, EU*

received 2 July 2012; accepted in final form 24 September 2012
published online 31 October 2012

PACS 89.75.Hc – Networks and genealogical trees
PACS 89.75.Fb – Structures and organization in complex systems
PACS 92.60.Ry – Climatology, climate change and variability

Abstract – In studies of spatially confined networks, network measures can lead to false conclusions since most measures are boundary affected. This is especially the case if boundaries are artificial and not inherent in the underlying system of interest (*e.g.*, borders of countries). An analytical estimation of *boundary effects* is not trivial due to the complexity of measures. The straightforward approach we propose here is to use surrogate networks that provide estimates of boundary effects in graph statistics. This is achieved by using spatially embedded random networks as surrogates that have approximately the same link probability as a function of spatial link lengths. The potential of our approach is demonstrated for an analysis of spatial patterns in characteristics of regional climate networks. As an example networks derived from daily rainfall data and restricted to the region of Germany are considered.

Introduction. – The study of spatially extended complex systems is a lively and growing field, for instance in astrophysics [1], Earth sciences [2], ecology [3], or medical image analysis [4]. In the last decades, powerful tools of time series analysis have been proposed and developed, such as singular spectrum analysis [5], wavelet analysis [6], recurrence plots [7], etc. For a spatial analysis various tools are available, such as empirical orthogonal functions [8], tools adapted from time series analysis such as spatial recurrence plots [9,10], or complex networks [11]. In many fields of research, complex networks have proven to be a successful concept for understanding complex systems, *e.g.*, resilience studies of the Internet [12], transport optimization on street networks, power grids and supply chain networks [13,14], spread of epidemics within populations [15–17], relations from structure to function in brain networks [18–25], and even in the analysis of time series [26,27]. Recently, network theory has also been utilized in climate research by the so-called *climate networks* for understanding complex climate phenomena [28–36].

Many networks are spatial networks. However, the network structure is often influenced by spatial embedding due to distance-based costs of links, *i.e.*, the link probability depends on the spatial length of links [37–39]. Although this effect is usually isotropic, it becomes anisotropic if boundaries in space are introduced to the network, as this is the case with the spatial confinement of brain networks which are embedded in three-dimensional space and confined by the area of placed electrodes [39]. Climate networks might be bounded if only a smaller region is considered [31,40]; similarly, power grids are confined by the economic region (*e.g.*, by the boundary of the European Network of Transmission System Operators for Electricity). We also call spatially confined networks *regional* networks.

Boundaries cut links which would connect the region under consideration with the outside region. Obviously, this artificially reduces node degrees and the amount of longer links in the remaining network, and hence influences corresponding network measures. Besides degree, closeness centrality and shortest-path betweenness [41] are also used as examples which measure the inverse mean topological distance from one node to all other network nodes as well as the number of shortest paths through a given node, respectively. In this study the closeness centrality of node j is defined as

$$CC_j = \frac{N}{\sum_{i=1}^{N} g_{ji}},$$

(a) E-mail: rheinwalt@pik-potsdam.de

with N being the number of nodes and the geodesic distance g_{ji} from node j to i. The shortest-path betweenness of node j is defined as

$$SB_j = \log_{10}\left(\frac{1}{2}\sum_{i,k\neq j}^{N}\frac{\sigma_{ik}(j)}{\sigma_{ik}}+1\right),$$

where σ_{ik} is the number of shortest paths from i to k and $\sigma_{ik}(j)$ are just the ones that pass through j. The effect of cut links due to boundaries is larger when the network consists of many long links, as the probability is high that such links connect the inside and the outside regions. Where and how strong boundaries affect network measures depends on the distribution of link lengths and on the network measures themselves.

Based on the network of interest and its spatial confinement, boundary effects might be negligible, of interest, or distracting from network structure not imposed by boundaries. Neglecting boundary effects can lead to spurious conclusions, *e.g.*, for the identification of hubs in brain networks [39]. In many applications, resulting boundary effects are often not negligible and, consequently, network measures should be corrected in order to exclude them.

Here we propose a correction procedure for the network measures derived from a regional network. We will use a specific random network construction with properties similar to those in the original network, *i.e.*, it shares a similar link probability $p(\Delta_{ik})$ that two nodes, i and k that have the distance Δ_{ik} in space, are linked. However this $p(\Delta)$ is not the probability to find a link of length Δ among all. It is the probability conditioned on the number of possible links of that length due to the embedding of nodes in space.

Method. – In spatially embedded random networks (SERN) by Barnett *et al.* [38], influences of spatial embedding on network structure are quantified by a link probability that depends on the spatial length of a link in the embedding metric space. We propose this as a model for boundary effects and generate SERN for the same node positions in space as the original network and with the same link probability depending on spatial link lengths. Thus, for a network with boundaries, we consider the result of a certain network measure on such a SERN as an estimate of boundary effects in that measure. Hence, the SERN we use is a surrogate in the sense that it mimics the same length dependency in the link probability as the original network has:

- Nodes are embedded in a metric space S with the metric $\Delta\colon S\times S\to\mathbb{R}^+$; thus, Δ_{ik} is the spatial distance between node i and k.
- Nodes have been given positions X in S. These positions are the same as in the original network.
- Nodes i and k are connected with the link probability $p(\Delta_{ik})$ being the probability of finding a link of length Δ_{ik} in the original network in respect to how many links of that length could occur.

Depending on the positioning of nodes in space, a binning of spatial link length might be necessary in order to improve the link probability estimate of the original network. This can be achieved by rounding spatial link lengths to appropriate integers so that similar lengths fall into one integer length. A measurement using such a procedure can be done with the following algorithm. Here A_d is the number of possible links with integer length d and B_d is the number of actually present links with integer length d. The fraction p_d of both is an estimate for the underlying link probability $p(\Delta)$.

$A_d = B_d = p_d = 0 \quad \forall\, d \in \text{rounded } \Delta$
for $i \in nodes$ **do**
  for $k < i$ **do**
    $d \leftarrow$ rounded Δ_{ik}
    increase A_d by one
    if node i and k are linked in the original network **then**
      increase B_d by one
    end if
  end for
end for
$p \leftarrow \frac{B}{A}$.

Link probabilities p for real data derived by this algorithm can be seen in the application in fig. 4.

To improve the estimation of boundary effects in a certain network measure, such as closeness centrality, we average the result of that measure over an ensemble of surrogates. The reliability of such an estimate is derived from the distribution of ensemble measures.

If a node-based network measure that returns a value for each node is used, an estimate of boundary effects can be found for each node in space. An averaged estimate such as the median from 1000 surrogates is shown for the closeness centrality field in fig. 1 (bottom left). As a measure of reliability for that median we take the interquartile range per median. Thus, for a fixed number of surrogates we get a distribution over all nodes of interquartile ranges per median. The evolution of that distribution with increasing number of surrogates is visualized in fig. 2. As one can see we could have used only half as many surrogates and would have gotten a very similar reliability of our correction. The corresponding network to the regional closeness (bottom right) is a ripped-off part of the global network (top left). Thus, nodes in the regional network are connected if they are connected in the global network. The global network is a SERN with the link probability $p(\Delta)\propto\Delta^{-3.5}$. The scales in colorbars go from the minimum to the maximum value in all figures since we are only interested in relative quantities.

A *corrected* network measure is now calculated by subtracting the estimate of boundary effects for a certain network measure from the measure on the original network. The intrinsic spatial bias of the measure due to the artificial boundary in the regional case (bottom right) is obviously removed in the corrected measure (top right).

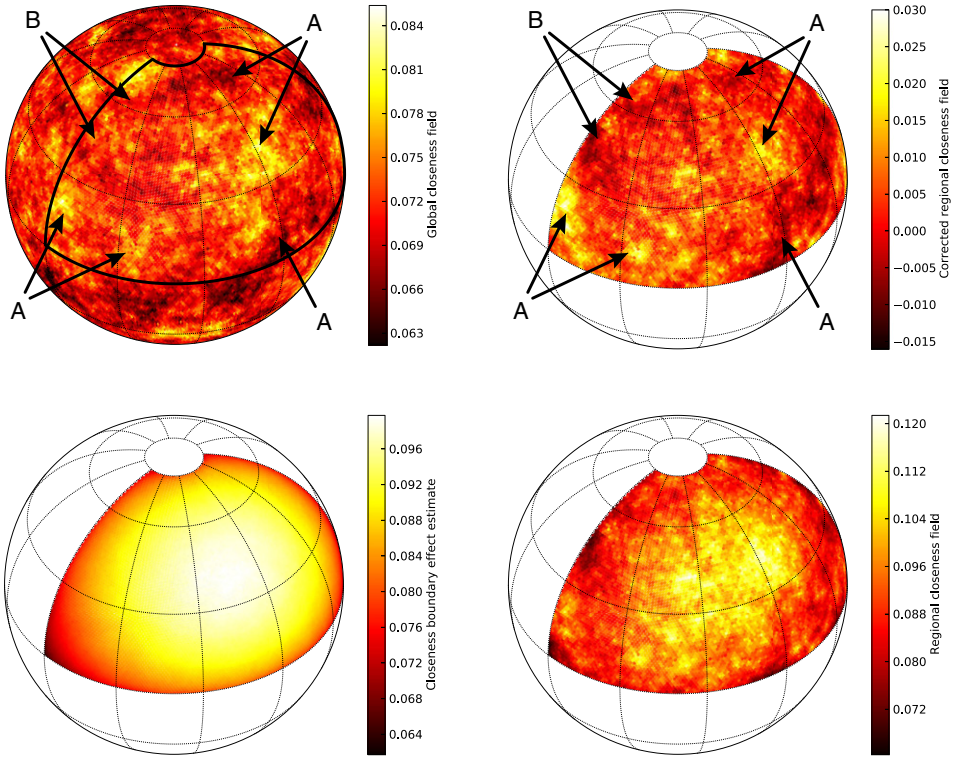

Fig. 1: (Color online) Top left: global closeness: closeness centrality of a random network on a sphere. The connection probability depends only on the spatial link length and follows a power law with the exponent −3.5. Top right: corrected regional closeness. Arrows point out areas of strong similarity (A) and dissimilarity (B) in the spatial patterns in the considered region. Bottom left: closeness boundary effects estimate, taken as the median from 1000 surrogates. Bottom right: regional closeness: closeness centrality on a part of the same network as on the whole globe (top left). Nodes in the depicted region are connected if they are connected in the global network.

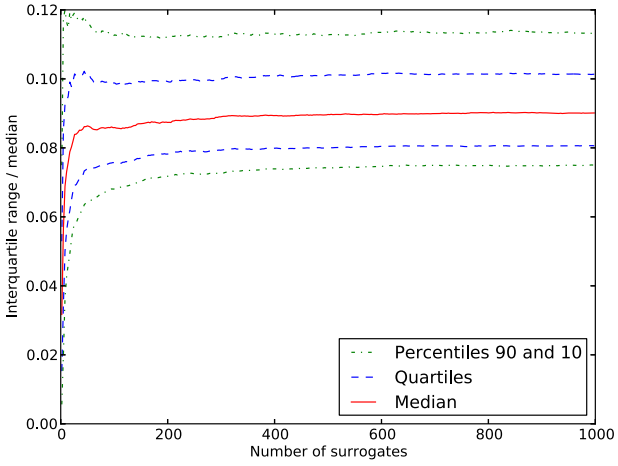

Fig. 2: (Color online) Evolution of surrogate reliability with increasing number of surrogates for the example shown in fig. 1. Shown are important quantiles of the distribution of node-wise interquartile ranges per median. After 400 surrogates reliability does not improve much further.

Interestingly, too, the spatial structure of the corrected closeness centrality field resembles the one of the global network in the corresponding region (fig. 1). Strong similarities are denoted by A and dissimilarities by B. Quantitatively the similarity between corrected closeness centrality and the closeness centrality values in the corresponding region of the global network can be expressed by a Spearman's rank correlation coefficient of 0.661. Respectively, this can be compared to a coefficient of 0.575 if the uncorrected measure is used. Figure 1 is just a visual example, thus if we generate an ensemble of 1000 such examples we get the distributions of correlation coefficients as shown in in fig. 3. The distribution of coefficients corresponding to the corrected closeness is not only shifted to higher similarity, but is also narrower. The difference as well as the absolute value in similarity vary strongly with measures and link probabilities used. However similarities for corrected measures seem to be always higher than for regional measures.

The similarity in network measures is due to the removed boundary effects as well as to the similarity in network structure. All links that connect nodes within the specified region are the same in both networks. The global network has more links as well as links that connect nodes in the region with nodes that are not in the region; in particular, links that reach deep into the region are rare due to the power-law dependency in the link probability $p(\Delta)$. Note that the degree is not as strongly affected by these additional links in the global in comparison to the regional case as is the path-based measure closeness

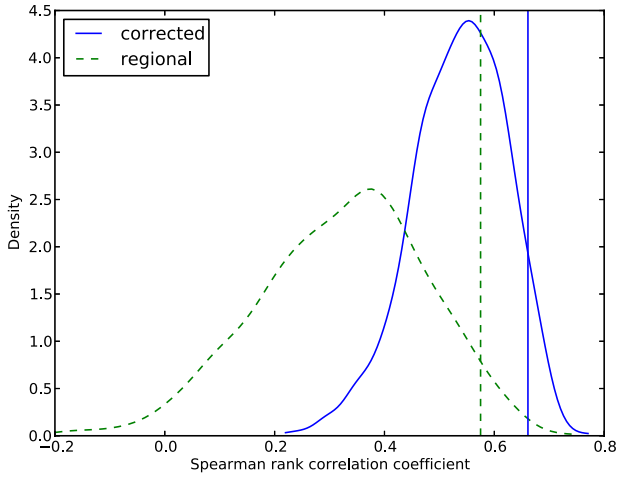

Fig. 3: (Color online) Gaussian kernel density estimate from 1000 samples for the distribution of Spearman's rank correlation coefficients between regional and global closenesses compared to between corrected and global closenesses. Vertical lines correspond to the example shown in fig. 1.

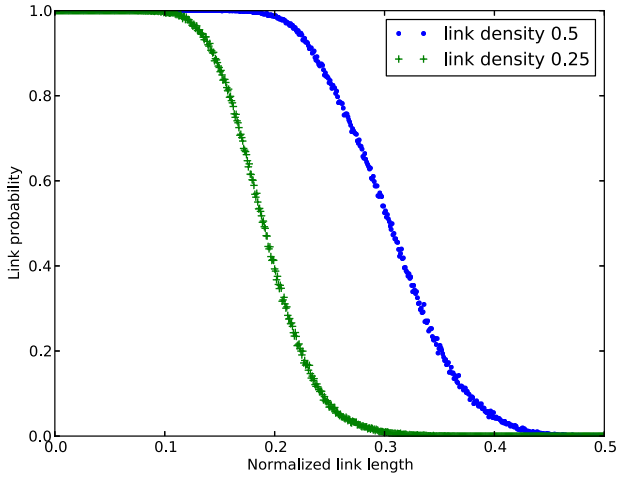

Fig. 4: (Color online) Link probability $p(\Delta)$ for both precipitation networks. Link lengths normalized to 1 for the longest possible link in the region.

centrality. For instance, correlation coefficients for degree are higher.

However, due to the lack of information in the regional in comparison to the global network, *i.e.*, border-crossing links cannot be resolved, the corrected values of the network measures of the regional network can still differ from those derived from ripped-off parts of a global network. Additional links in the global in comparison to the regional network can have an additional effect on network measures —especially if they are long.

These examples show the potential of the method. The described method removes boundary effects in network measures, but does not predict how these measures would be different if the network grows. However corrected measures for different boundaries are comparable, whereas the uncorrected measures are not.

Application. – Previous studies of climate networks have mostly considered a global network [28–30,32–36]. Such networks are spatially embedded networks in a two-dimensional space without boundaries (*e.g.*, fig. 1, top left). As soon as we focus on a smaller region, artificial boundaries in the embedding space are introduced to the network (*e.g.*, fig. 1, bottom right). Here we restrict rainfall networks to the region of Germany; thus, boundaries are purely artificial and not part of the underlying system.

The here considered regional climate networks are constructed from precipitation data for the region of Germany. We use daily meteorological weather station data (precipitation in millimeters) from 1951 to 2007 provided by the German national weather service (Deutscher Wetterdienst) for over 2000 stations. From these time series we construct a simple, undirected and unweighted network with a number of nodes equal to the number of time series. Nodes are connected by the nonlinear similarity measure event synchronization [40,42]. Thus, networks are constructed by thresholding. However events are generated in a way that reduces seasonality at each station. This is done by defining events only for days where the daily precipitation sum exceeds the 75% percentile threshold of all similar days in the years from 1951 to 2007.

Link densities are either 25% or 50%, which corresponds to similarity values above 0.619 or 0.558, respectively. Such values are very unlikely to arise by chance. According to the distribution of random events in time with the same event density, the probability for a similarity above 0.413 drops below 1%.

Our main results are summarized in the following. The link probability $p(\Delta)$ for both networks follows roughly a Gaussian decay. Thus, the connection probability for two nodes is strongly distance dependent. This dependence is shifted to longer distances for the network with the link density of 50% (fig. 4). This is why we show results for both networks. Boundary effect estimates for degree are stronger for the link density of 50% due to the higher probability of longer links (fig. 5). For the link density of 25%, degree boundary effects are strongest close to boundaries and become more constant towards the spatial center because longer links become more unlikely (fig. 6). For the path-based measures, closeness centrality and shortest-path betweenness, this is not the case (figs. 8 and 7).

Conclusion. – We have proposed an approach to estimate boundary effects in network measures from spatially embedded networks based on surrogate ensembles of spatially embedded random networks for a given link probability $p(\Delta)$. These boundary effects can vary depending on $p(\Delta)$ and on the network measures used. Using boundary effects estimates, network measures can be corrected in order to reduce the influence of boundaries. We have demonstrated the potential of the approach on an artificial random network and on two regional networks constructed

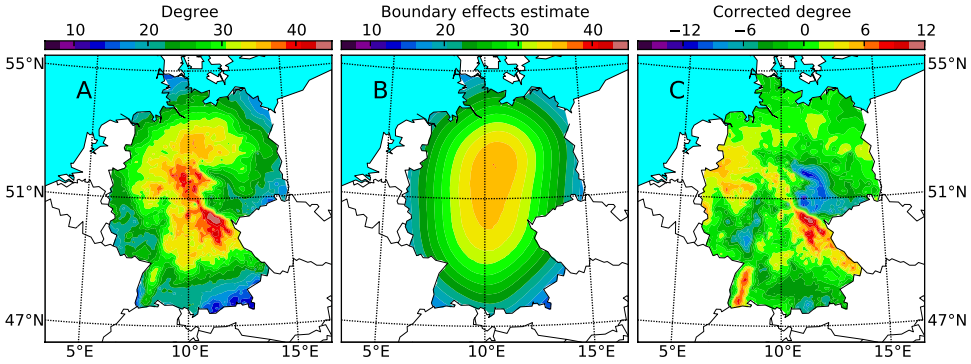

Fig. 5: (Color online) Uncorrected degree field for the network with 50% link density (A), the corresponding boundary effects estimate (B) and the corrected measure (C).

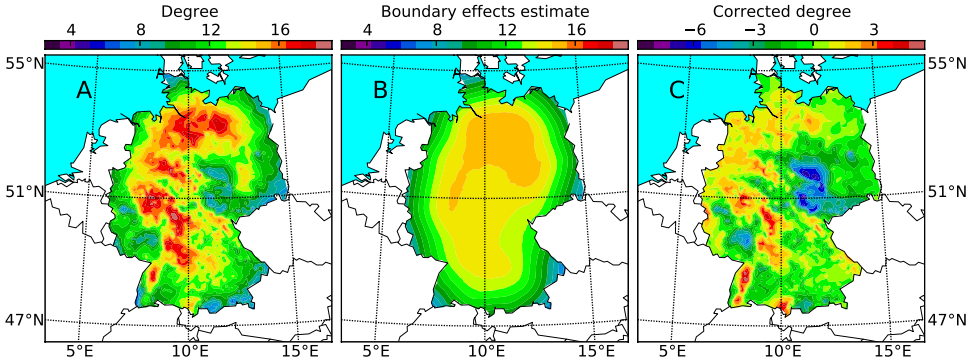

Fig. 6: (Color online) Uncorrected degree field for the network with 25% link density (A), the corresponding boundary effects estimate (B) and the corrected measure (C).

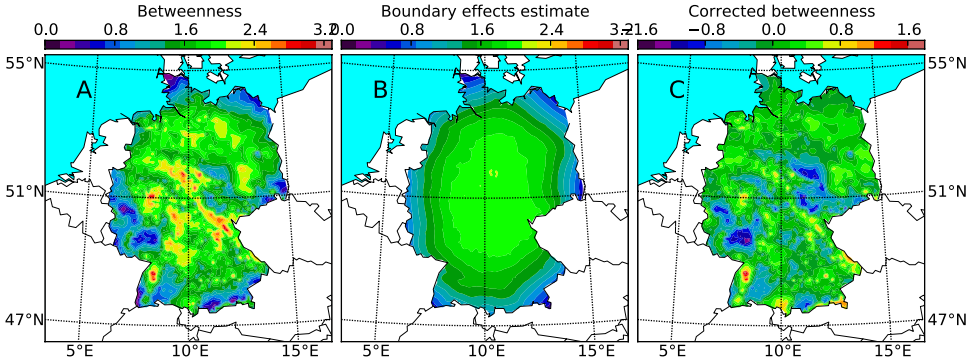

Fig. 7: (Color online) Uncorrected betweenness field for the network with 25% link density (A), the corresponding boundary effects estimate (B) and the corrected measure (C).

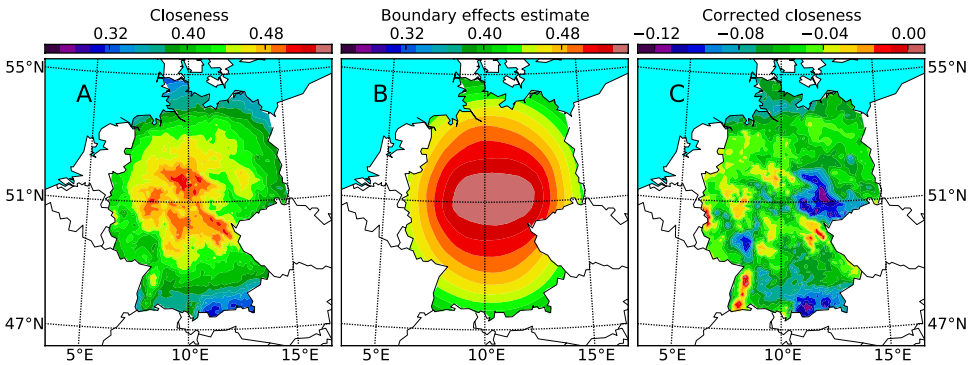

Fig. 8: (Color online) Uncorrected closeness centrality field for the network with 25% link density (A), the corresponding boundary effects estimate (B) and the corrected measure (C).

from rainfall data —all three with different link probabilities $p(\Delta)$.

Spatial patterns in network measures may look very different if corrected measures are used instead of uncorrected measures (figs. 5–8). This is the case if boundaries of a network break the isotropy of the link probability $p(\Delta)$. The effect of this on network measures is quantified by our boundary effects estimates.

Depending on the network structure and network measures used, the effects of boundaries have a bearing on the entire network. For instance, degree is a local measure in the network, but depending on the distribution of link lengths it is non-local in the embedding space. The clustering coefficient is not even local in the network, *i.e.*, it depends on topological paths of length three, thus the spread of boundary effects becomes more complex. Path-based measures such as closeness centrality and shortest-path betweenness are extremely non-local in this sense. This explains why boundaries affect network measures in the entire network and not only close to them.

* * *

This work was supported by the German Research Foundation by the DFG project "Interactions and complex structures in the dynamics of changing climate" and the German Federal Ministry for Education and Research (BMBF) via the Potsdam Research Cluster for Georisk Analysis, Environmental Change and Sustainability (PROGRESS). The authors would as well like to thank Jonathan Donges and Jobst Heitzig for inspiring discussions and ideas.

REFERENCES

[1] Amelino-Camelia G., *Nature*, **478** (2011) 466.
[2] Davidsen J. and Paczuski M., *Phys. Rev. Lett.*, **94** (2005) 48501.
[3] Jones K. E., Patel N. G., Levy M. A., Storeygard A., Balk D., Gittleman J. L. and Daszak P., *Nature*, **451** (2008) 990.
[4] Schmah T., Marwan N., Thomsen J. S. and Saparin P., *Med. Phys.*, **38** (2011) 5003.
[5] Ghil M., Allen M., Dettinger M., Ide K., Kondrashov D., Mann M., Robertson A., Saunders A., Tian Y. and Varadi F. *et al.*, *Rev. Geophys.*, **40** (2002) 1003.
[6] Torrence C. and Compo G. P., *Bull. Am. Met. Soc.*, **79** (1998) 61.
[7] Marwan N., Romano M. C., Thiel M. and Kurths J., *Phys. Rep.*, **438** (2007) 237.
[8] Hannachi A., Jolliffe I. T. and Stephenson D. B., *Int. J. Clim.*, **27** (2007) 1119.
[9] Marwan N., Kurths J. and Saparin P., *Phys. Lett. A*, **360** (2007) 545.
[10] Agustí P., Traver V. J., Marin-Jimenez M. J. and Pla F., *Lect. Notes Comput. Sci.*, **6855** (2011) 364.
[11] Newman M., *SIAM Rev.*, **45** (2003) 167.
[12] Cohen R., Erez K., ben Avraham D. and Havlin S., *Phys. Rev. Lett.*, **85** (2000) 4626.
[13] Danila B., Yu Y., Marsh J. A. and Bassler K. E., *Phys. Rev. E*, **74** (2006) 046106.
[14] Chen Y.-H., Wang B.-H., Zhao L.-C., Zhou C. and Zhou T., *Phys. Rev. E*, **81** (2010) 066105.
[15] Klovdahl A., Potterat J., Woodhouse D., Muth J., Muth S. and Darrow W., *Soc. Sci. Med.*, **38** (1994) 79.
[16] Tuckwell H. C., Toubiana L. and Vibert J.-F., *Phys. Rev. E*, **57** (1998) 2163.
[17] Ferguson N. M. G. P. G., *Sex. Transm. Dis.*, **27** (2000) 600.
[18] Sporns O., Chialvo D., Kaiser M. and Hilgetag C., *Trends Cognit. Sci.*, **8** (2004) 418.
[19] Zhou C., Zemanová L., Zamora G., Hilgetag C. and Kurths J., *Phys. Rev. Lett.*, **97** (2006) 238103.
[20] Sporns O., Honey C. and Kötter R., *PLoS ONE*, **2** (2007) e1049.
[21] Stam C. and Reijneveld J. *et al.*, *Nonlinear Biomed. Phys.*, **1** (2007) 3.
[22] Bullmore E. and Sporns O., *Nat. Rev. Neurosci.*, **10** (2009) 186.
[23] Zamora-López G., Zhou C. and Kurths J., *Front. Neuroinform.*, **4** (2010) 1.
[24] Bashan A., Bartsch R., Kantelhardt J., Havlin S. and Ivanov P., *Nat. Commun.*, **3** (2012) 702.
[25] Gallos L., Makse H. and Sigman M., *Proc. Natl. Acad. Sci. U.S.A.*, **109** (2012) 2825.
[26] Marwan N., Donges J. F., Zou Y., Donner R. V. and Kurths J., *Phys. Lett. A*, **373** (2009) 4246.
[27] Donner R. V., Small M., Donges J. F., Marwan N., Zou Y., Xiang R. and Kurths J., *Int. J. Bifurcat. Chaos*, **21** (2011) 1019.
[28] Tsonis A. and Roebber P., *Physica A*, **333** (2004) 497.
[29] Tsonis A., Swanson K. and Roebber P., *Bull. Am. Meteorol. Soc.*, **87** (2006) 585.
[30] Tsonis A., Swanson K. and Wang G., *J. Clim.*, **21** (2008) 2990.
[31] Yamasaki K., Gozolchiani A. and Havlin S., *Phys. Rev. Lett.*, **100** (2008) 228501.
[32] Donges J., Zou Y., Marwan N. and Kurths J., *EPL*, **87** (2009) 48007.
[33] Donges J., Zou Y., Marwan N. and Kurths J., *Eur. Phys. J. ST*, **174** (2009) 157.
[34] Steinhaeuser K., Chawla N. and Ganguly A., *ACM SIGKDD Explor. Newslett.*, **12** (2010) 25.
[35] Steinhaeuser K., Chawla N. and Ganguly A., *Stat. Anal. Data Min.*, **4** (2011) 497.
[36] Paluš M., Hartman D., Hlinka J. and Vejmelka M., *Nonlinear Processes Geophys.*, **18** (2011) 751.
[37] Gastner M. and Newman M., *Eur. Phys. J. B*, **49** (2006) 247.
[38] Barnett L., Di Paolo E. and Bullock S., *Phys. Rev. E*, **76** (2007) 056115.
[39] Henderson J. and Robinson P., *Phys. Rev. Lett.*, **107** (2011) 18102.
[40] Malik N., Bookhagen B., Marwan N. and Kurths J., *Clim. Dyn.*, **39** (2012) 971.
[41] Newman M., *Networks: An Introduction* (Oxford University Press, Inc.) 2010.
[42] Quiroga R., Kreuz T. and Grassberger P., *Phys. Rev. E*, **66** (2002) 041904.