PAPER • OPEN ACCESS

Centrality anomalies in complex networks as a result of model over-simplification

To cite this article: Luiz G A Alves *et al* 2020 *New J. Phys.* **22** 013043

New Journal of Physics
The open access journal at the forefront of physics

Deutsche Physikalische Gesellschaft DPG
IOP Institute of Physics

Published in partnership with: Deutsche Physikalische Gesellschaft and the Institute of Physics

OPEN ACCESS

RECEIVED
21 November 2019

REVISED
26 December 2019

ACCEPTED FOR PUBLICATION
7 January 2020

PUBLISHED
23 January 2020

PAPER

Centrality anomalies in complex networks as a result of model over-simplification

Luiz G A Alves[1,2], **Alberto Aleta**[3,4], **Francisco A Rodrigues**[2,5,6], **Yamir Moreno**[3,4,7] and **Luís A Nunes Amaral**[1,8,9]

[1] Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL 60208, United States of America
[2] Institute of Mathematics and Computer Science, University of São Paulo, São Carlos, SP 13566-590, Brazil
[3] Department of Theoretical Physics, University of Zaragoza, Zaragoza, E-50009, Spain
[4] Institute for Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza, Zaragoza, E-50009, Spain
[5] Mathematics Institute, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, United Kingdom
[6] Centre for Complexity Science, University of Warwick, Coventry CV4 7AL, United Kingdom
[7] ISI Foundation, Turin, I-10126, Italy
[8] Department of Physics and Astronomy, Northwestern University, Evanston, IL 60208, United States of America
[9] Northwestern Institute on Complex Systems (NICO), Northwestern University, Evanston, IL 60208, United States of America

E-mail: yamir.moreno@gmail.com, lgaalves@northwestern.edu and amaral@northwestern.edu

Keywords: complex networks, network structure, betweenness, complex systems, real data

Abstract

Tremendous advances have been made in our understanding of the properties and evolution of complex networks. These advances were initially driven by information-poor empirical networks and theoretical analysis of unweighted and undirected graphs. Recently, information-rich empirical data complex networks supported the development of more sophisticated models that include edge directionality and weight properties, and multiple layers. Many studies still focus on unweighted undirected description of networks, prompting an essential question: how to identify when a model is simpler than it must be? Here, we argue that the presence of centrality anomalies in complex networks is a result of model over-simplification. Specifically, we investigate the well-known anomaly in betweenness centrality for transportation networks, according to which highly connected nodes are not necessarily the most central. Using a broad class of network models with weights and spatial constraints and four large data sets of transportation networks, we show that the unweighted projection of the structure of these networks can exhibit a significant fraction of anomalous nodes compared to a random null model. However, the weighted projection of these networks, compared with an appropriated null model, significantly reduces the fraction of anomalies observed, suggesting that centrality anomalies are a symptom of model over-simplification. Because lack of information-rich data is a common challenge when dealing with complex networks and can cause anomalies that misestimate the role of nodes in the system, we argue that sufficiently sophisticated models be used when anomalies are detected.

Introduction

The study of complex networks produced fruitful results in many areas of knowledge, from systems biology [1, 2] and social systems [3, 4] to epidemiology [5–7] and statistical physics [8, 9]. The initial focus of complex networks and graph theory was on undirected, unweighted topologies [9, 10]. Using unweighted network projections, many properties were proved to be effective in describing complex systems [11–14]. More recently, weighted, directed, multiplexed networks have been the focus of much research attention. In many cases, these more sophisticated representations of the system are most appropriate to describe real-world networks [15–18]. Despite it, researchers still fall back on representing a system's network of interactions as if it was undirected and unweighted, many times because of the lack of information-rich data sets.

This is the case of gene regulatory networks, where usually direction, strengths, and signs of the links are overlooked because of the lack of complete data [19]. Another case where empirical studies have overlooked the details of the system is the case of multipartite networks [20]. This class of systems comprises networks with multiple groups that can only interact through nodes of different types. However, because of the lack of information-rich data sets, these systems are usually studied after projection onto networks of one single type of node. Thus, the question is how to determine when such a model is good enough to represent the system, especially in the absence of data for testing simulation predictions.

Here, we focus on the case of weighted networks projected onto unweighted networks. We propose that the presence of anomalies in the structure of the undirected and unweighted projection of the network can be a result of a situation where a model is simpler than it must be. Our starting observation is the report of betweenness centrality anomalies in transportation networks [21]. This simple measure can capture the importance of a node to connect different parts of the network [9] by the means of how often it stands between other nodes. Guimerà *et al* reported that nodes with a large degree in air transportation networks do not necessarily have the highest betweenness centrality, whereas some low degree nodes can have large betweenness centralities. The emergence of these anomalies has been attributed to the multi-community structure of the network and spatial constraints such as geopolitical boundaries [21–23]. Nevertheless, the general mechanisms governing the emergence of such anomalies remain unknown.

In order to tackle these questions, we investigate a broad class of network models with weights and spatial constraints and the structure of four transportation networks. Our analysis reveals that, like for the class of model networks, unweighted transportation networks exhibit centrality anomalies for a significant fraction of the nodes compared with an appropriate null model with the same degree distribution. However, these anomalies disappear when we consider weighted representations of the network. Our findings support the hypothesis that such centrality anomalies are a symptom of a model that is simpler than it must be.

Because model over-simplification might lead to anomalies that would misestimate the role of nodes in the system, our findings have direct implications for the modeling of dynamical processes on complex networks where betweenness centrality is used to measure the influence of nodes, such as in the modeling of human dynamics [24], the spread of diseases [25, 26], crime spreading [27], and spatial networks [22, 23]. Moreover, they also hint at the significant challenges when modeling biological [19], economic, or social phenomena because data incompleteness is so pervasive.

Results

Centrality anomalies

We collected extensive data for four large scale transportation networks: Brazil, Great Britain, and Spain bus transportation networks, and the worldwide air transportation network. We define an inter-city bus transportation network by assigning a node to each of the N municipalities (with at least one bus station) and assigning an undirected edge between two nodes if the two stops i and j are connected by at least one bus route. Throughout the period observed for each data set, the same route can be offered by more than one company and multiple times by a single company (see methods for details). This fact enables us to define the weight of the edge, w_{ij}, as the total number of buses offered by all companies over the observation period (figure 1).

In the worldwide air transportation network, each node represents a city. As a consequence, if there are multiple airports serving the same city, we assign the relevant airports to a single node. For example, JFK, La Guardia, and Newark airports are all assigned to the New York City node. We assigned undirected edges between two nodes i and j if the two cities were connected by at least one air route. Because not all air routes have daily or greater frequency, and in order not to drop less-traveled cities, we collected information on flights occurring during the week of 17 May, 2018–22 May, 2018. As for the bus transportation networks, the same route can be offered by more than one company and multiple times a day by the same company. Thus, we defined the weight of an edge, w_{ij}, as the total number of flights offered by different companies flying the route during the observation period (figure 1).

Several studies have reported that spatial networks, such as the ones we study here, can exhibit centrality anomalies [21, 22, 28, 29]—that is, the betweenness centrality of a node is not necessarily proportional to its degree squared. First, we investigate to what extent these centrality anomalies are due to the over-simplification of the networks. Specifically, we first calculate the betweenness centrality b and degree k of the nodes for an unweighted projection of the network. The betweenness centrality of node i counts the fraction of shortest paths connecting all pairs of nodes that pass through node i but do not include node i [30]. Figure 2 shows the betweenness centrality versus degree for the networks studied here.

In order to make sense of the observed values of the betweenness and their relationship with the degree, we compare the measurements for the four transportation networks to the expected values for ensembles of

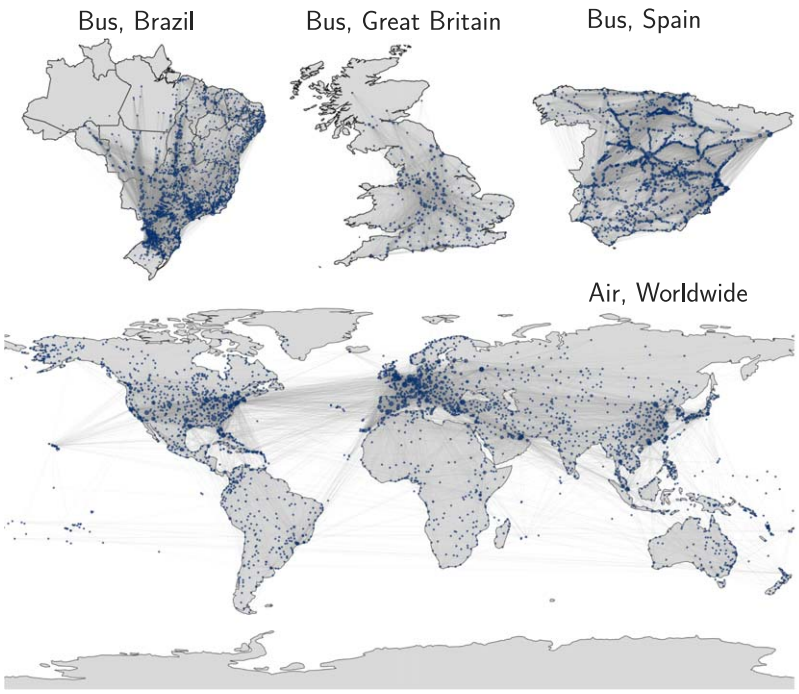

Figure 1. Illustration of the transportation network data sets. We collected four large data sets of transportation networks that include information about the number of buses or airplanes on each route. The data consist of the inter-city bus transportation networks of three countries (Brazil, Great Britain, and Spain), and worldwide air transportation network. In the plots, the node area is proportional to the degree of the node.

randomized networks with the same degree distributions. In order to provide consistency with later analyses, we do not use the typical Markov chain Monte Carlo edge switching approach, in which the structural constraints are satisfied exactly (i.e. *microcanonical ensemble*), and instead implement the undirected binary configuration model (UBCM) [31], where the constraints are met on average over the ensemble (i.e. *canonical ensemble*) [32–34]. In the UBCM, edges are placed at random following a distribution that preserves, on average, the original degree distribution observed in the data (see methods).

As has been reported earlier [21, 28], the betweennesses obtained for the randomized networks do not recapitulate those observed for the empirical networks. That is, whereas there is an approximate scaling of the betweenness with the degree squared for the randomized networks, for the empirical networks one finds many nodes with large deviations from that scaling relationship.

Model networks

It has been proposed that the existence of these centrality anomalies is due to the presence of spatial constraints and the special role, due to economic or political considerations, that some cities might have [22, 23, 28]. However, the precise factors driving the emergence of such anomalies remain unknown.

To investigate the generality of our findings, we next study a class of spatial weighted networks generated using the *strength driven preferential attachment with spatial selection* (SDPASS) model, which has been reported to produce centrality anomalies [22]. In this model, N_0 initial nodes are randomly located on a two-dimensional disc of radius L according to a uniform distribution and they are connected by links with weights w_0. At each time-step, a new node i is placed randomly on the disc, following a uniform distribution. The new node is connected to m previously existent nodes that are preferentially near and have the largest strength, according to

$$p_{ij} = \frac{s_j \, \mathrm{e}^{-d_{ij}/r_c}}{\sum_l s_l \, \mathrm{e}^{-d_{il}/r_c}} \tag{1}$$

where r_c is a desired spatial scale, s_i is the strength of the node (i.e. $s_i = \sum_j w_{ij}$), and d_{li} is the Euclidean distance between nodes l and i. The new edge (i, j) has a fixed weight w_0 and the creation of this edge perturbs the existing links attached to node j. To add this local perturbation to the model, the weights between j and its neighbors $l \in \mathcal{V}(j)$ are modified following the rule:

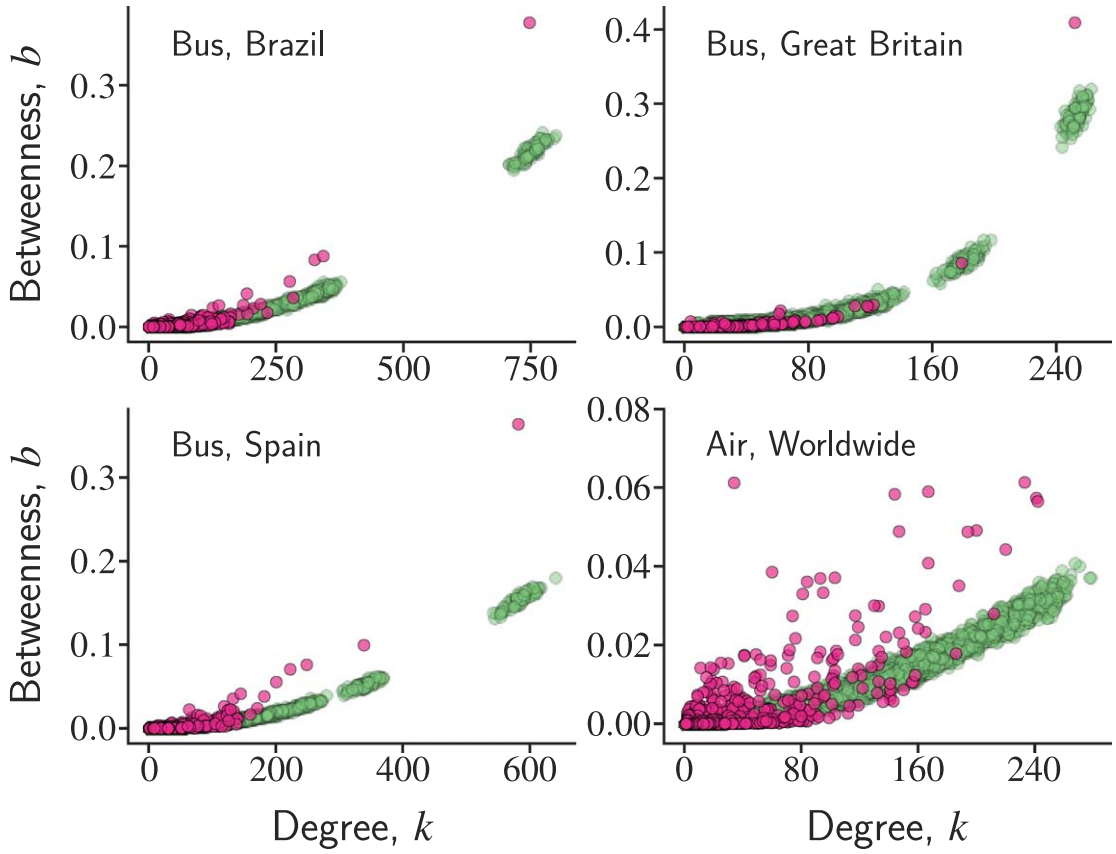

Figure 2. Investigation of centrality anomalies for the unweighted transportation networks. The panels show the comparison of real data (pink circles) with an ensemble of 10 000 networks (green circles) generated using the UBCM method [31]. As expected, for the randomized networks the betweenness centrality scales approximately with the degree squared. In contrast, for the empirical networks, the relationship between degree and betweenness is much less straightforward as there are some nodes for which the betweenness dramatically deviates from the scaling relationship.

$$w_{jl} \to w_{jl} + \delta \frac{w_{jl}}{s_j}, \tag{2}$$

where δ characterizes the susceptibility of the network to new links and $s_j = \sum_k w_{jk}$ is the strength of node j. If $\delta < w_0$, the new link has a small influence on the network. If $\delta \approx w_0$, the newly created traffic on the new edge is transferred to existing connections. If $\delta > w_0$, the traffic in the new edge generates a multiplicative effect on the traffic of the neighbors. This process is repeated until the network reaches the desired size. It is worth to note that this process generates a symmetric adjacency matrix, i.e. $w_{ij} = w_{ji}$, a necessary condition for the null models we use.

We explore the SDPASS model for networks with $N_0 = 5$ initial nodes, $m = 4$, and size $N = 100$. We simulate all relevant limiting cases to explore how δ and the ratio $\eta = r_c/L$ affects the scaling of the betweenness centrality. For convenience, we fixed $L = 1$ to explicitly explore the dependence of the model on r_c. For each set of parameters, we generated a network using the SDPASS model, and, subsequently, we used the appropriated null models to generate an ensemble of networks to calculate the fraction of anomalous nodes in these networks.

To make the identification of centrality anomalies rigorous, we compare the observed values of the pair (k_i, b_i) of node i to the distribution of expected values for the randomized ensemble. We find that the distribution of expected values is reasonably approximated by a multivariate Gaussian, $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\mu}_i$ represents the average values of k_i and b_i for the random ensemble and $\boldsymbol{\Sigma}_i$ represents the covariance matrix. We fit a multivariate Gaussian to the random ensemble data for each node and use it to compute the line enclosing 95% of the probability mass (see methods for details).

Considering $\eta \gg 1$ the effects of distance are negligible [22] and we recover the non-spatial weighted network model of Barrat *et al* [35], which showed no anomalies in our simulations compared with an ensemble of networks generated by the UBCM model. As $\delta \to 0$, the weight effects are no longer significant and we recover the preferential attachment model [36]. The preferential attachment model does not show any anomalies in the betweenness centrality, and an ensemble of random networks generated by the UBCM model is able to predict the betweenness centrality of the nodes. For instance, using $\delta = 0.01$ and $\eta = 10$ and comparing

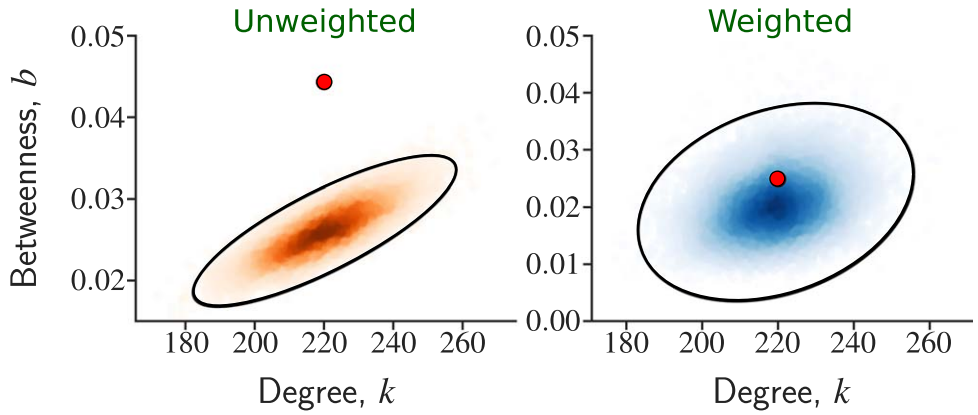

Figure 3. Illustration of centrality anomalies identification in complex networks. The red dot indicates the observed centrality and degree empirical network and the orange (blue) dots are the corresponding betweenness centrality versus degree for the 10 000 networks from the ensemble sampled using the UBCM (UECM) method. The solid black line encloses 95% of the probability mas for a multivariate Gaussian fit to the data. In the unweighted network, the observed values of betweenness centrality and degree lie outside the 95% bounds of the multivariate Gaussian adjusted to the data predicted by the synthetically generated networks. In contrast, in the weighted network, an anomaly is no longer observed.

this network with an ensemble of networks generated by the UBCM model we found that only 1% of the nodes have centrality anomalies.

Another possible scenario is $\delta \ll 1$ and $\eta \ll 1$. In this case, the effect of the link's weights is negligible and we essentially have a spatial unweighted network topology. In this case, the centrality anomalies are also not present, and our random network model (UBCM) is able to predict the betweenness centrality of the nodes. Using $\delta = 0.01$ and $\eta = 0.01$ to generate our network and comparing it with an ensemble of networks that preserves the degree distribution (UBCM), we found that only 1% of the nodes are anomalous.

Finally, we investigate the interplay between weights dynamics, i.e. $\delta \geqslant 1$, and spatial constraints, $\eta \ll 1$. In these limits, the model generates spatial weighted networks that have centrality anomalies similar to the ones observed for transportation networks. For instance, using $\delta = 10$ and $\eta = 0.01$, we found a significant fraction of nodes ($\approx$69%) that show anomalies in the unweighted projection of the network when compared to the ensemble of networks produced by the UBCM model.

Next, we compare the measurements for the model network to the expected values for an ensemble of randomized networks with the same degree and strength distributions. To this end, we use the undirected enhanced configuration model (UECM) [31, 37], which, consistently with the UBCM, preserves the constraints on average over the ensemble (i.e. *canonical ensemble*) [32–34]. In the UECM, edges and their weights are placed at random following distributions that, on average, preserve both the degree and the strength of the nodes; see methods. Note that the weights w_{ij} in our empirical networks represent the number of buses or airplanes available for the route connecting i and j. While higher values of w do reflect stronger ties, a physically appropriate calculation of the path length requires that one quantifies the length of an edge as the inverse of its weight [38]. Consistently with the transportation networks, we next consider the inverse of the weights to compute betweenness centrality of our model network. In figure 3 we show for illustration purposes the betweenness centrality data for both the unweighted and weighted randomizations. It is visually apparent that there is a centrality anomaly for one case but not the other.

Using the weighted projection of our model network and comparing it with an ensemble of networks generated by the UECM model, the fraction of centrality anomalies decrease to 18% of the nodes, a much smaller fraction than the 69% detected for the unweighted projection. Note that, because our null model does not include spatial information, our results suggest that a more sophisticated model would be a better choice for representing this network. The results of our model networks are summarized in table 1.

Weighted transportation networks

To investigate the relevance of the results for networks in the real world systems, we next explore whether centrality anomalies are also present when considering the weighted representation of the transportation networks. As before, we compare the relationship between observed betweennesses and degrees to the relationship obtained for an ensemble of 10 000 randomized networks generated using the UECM (figure 4). By doing so, we observe two results. First, even for the randomized networks, there no longer exists a simple scaling relationship between betweenness and degree. Second, we no longer find systematic centrality anomalies in the

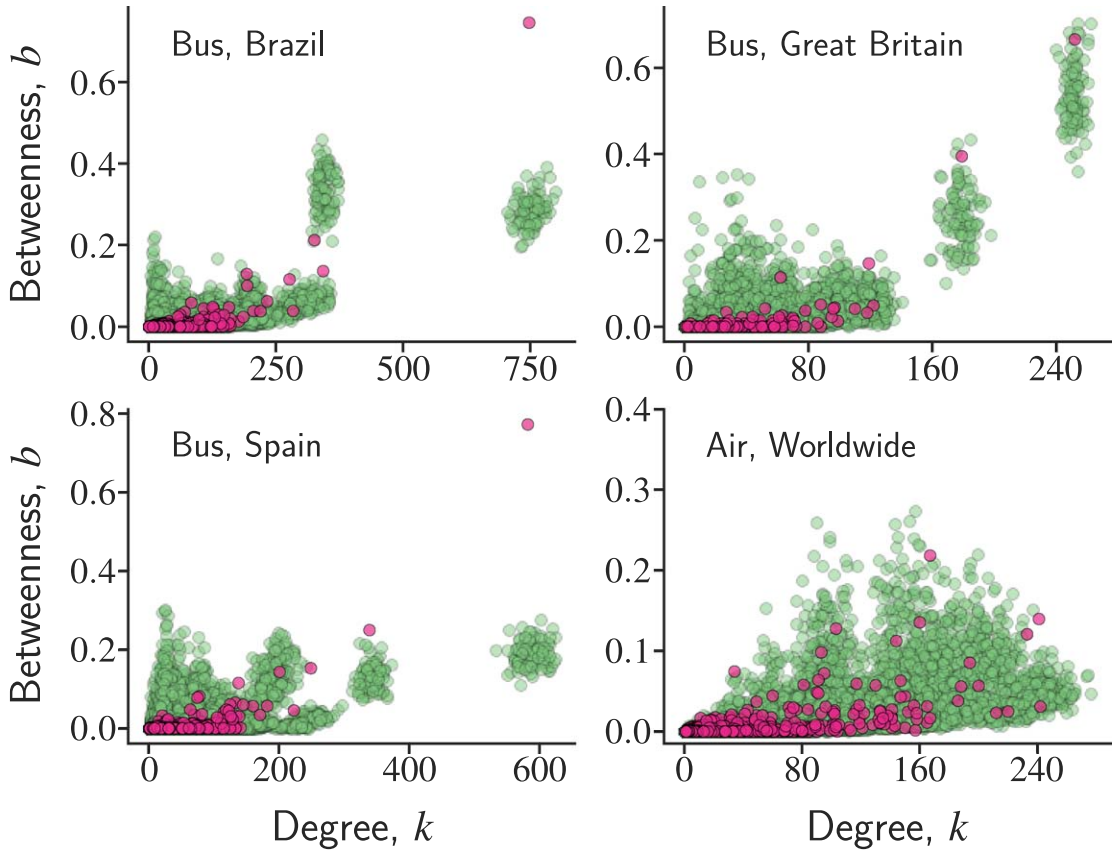

Figure 4. Investigation of centrality anomalies for the weighted transportation networks. The panels shows the comparison of real data (pink circles) with an ensemble of 10 000 networks (green circles) generated using the UECM method [31]. Notice that there no longer exists a simple scaling relationship between betweenness and degree.

Table 1. Anomalies in the SDPASS model. The first two columns show the parameters used on a simulation of networks considering $N_0 = 5$ initial nodes, $m = 4$, and size $N = 100$. The third column (unweighted network) and fourth column (weighted network) show the percentage of anomalous nodes in these networks when compared with an ensemble of networks generated by the UBCM and UECM models, respectively. The last column indicates the topology characteristics of the networks given the parameters δ and η.

δ	η	UBCM	UECM	Topology
0.01	10	1%	1%	Non-spatial unweighted
10	10	1%	1%	Non-spatial weighted
0.01	0.01	1%	1%	Spatial unweighted
10	0.01	69%	18%	Spatial weighted

data. Remarkably, only a handful of cities—Brasilia, Madrid, and Barcelona—appear to have a centrality anomaly and none of the nodes with low degree appears to have such anomalies. On the other hand, by plotting betweenness versus strength (figure 5), we uncover a simpler relationship, indicating that the strength would be a more informative measure of the nodes.

We now calculate the fraction of nodes for which we can reject the null hypothesis of no centrality anomaly (figure 6). The expectation here is that we will observe a false discovery rate of 5%. For 3 of the 4 unweighted transportation networks, we find an excess of nodes with centrality anomalies, whereas for none of the weighted networks we find such an excess. These results suggest that the existence of centrality anomalies when considering unweighted networks is a result of the neglected (but functionally crucial) role of edge weight on the evolution and performance of these networks.

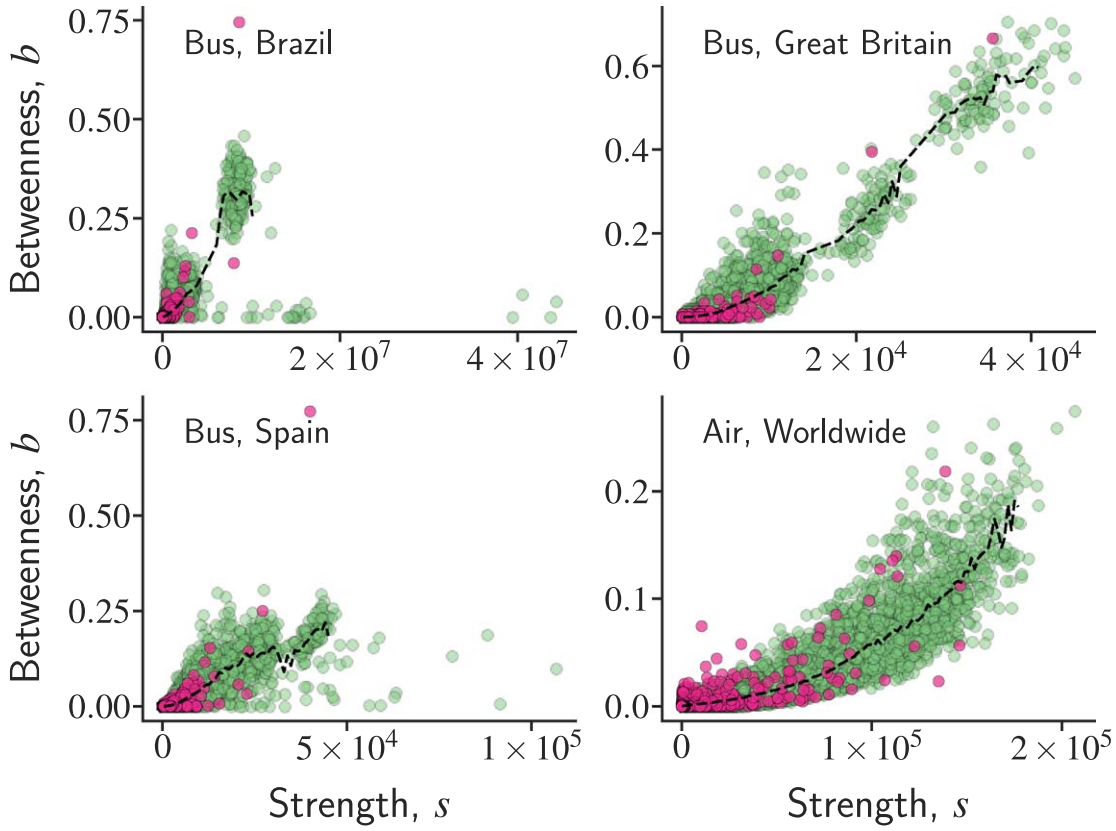

Figure 5. Investigation of betweenness centrality as a function of the strength. The pink circles show observed values, whereas the green circles show results for an ensemble of 10 000 networks generated using the UECM method [31]. The dashed black line is an average over the ensemble data. The average trend suggests that the strength is a more informative measure of the nodes than degree.

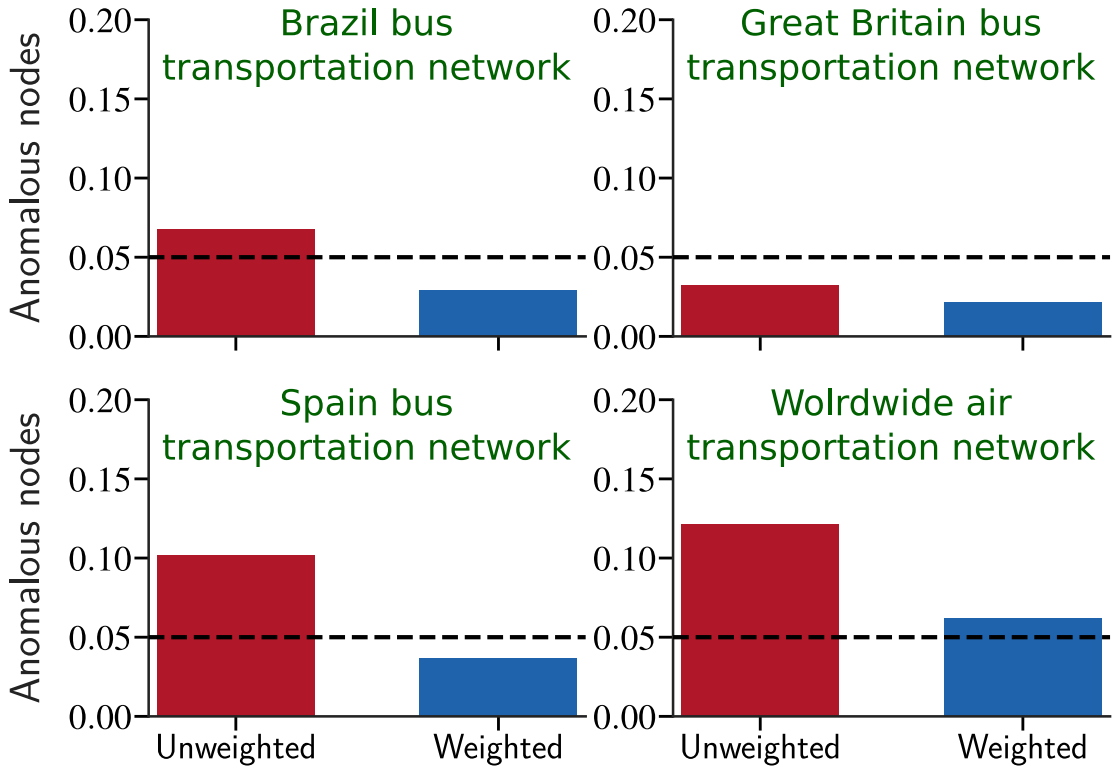

Figure 6. Quantifying the fraction of anomalous nodes. For each network, we compute the fraction of nodes that lie outside the 95% bounds of the model. The anomaly in betweenness centrality is verified for all unweighted transportation networks (except for Great Britain). In contrast, for the weighted version, the fraction of anomalous nodes is of the order of the false discovery rate, i.e. approximately 5%.

Conclusions

The findings reported here suggest that centrality anomalies present in the unweighted representation of transportation networks are masking the fact that some edges carry much larger weights than the typical edge in the network. Because of the role of spatial, temporal, and capacity constraints in real transportation networks, it is natural to expect that the degree of individual nodes cannot grow unbound, and that edge weight is a way to account for large demand. Indeed, we find that for random networks with the same degree and strength distributions the centrality structure of the network becomes indistinguishable from the observed structure.

We further extend our results to a broader class of model networks using the *SDPASS* model. Specifically, we show that when weights and spatial constraints are relevant, the centrality anomalies arise in the unweighted network projection and they cannot be predicted using a simple model that takes into account only the degree sequence as a constraint. On the other hand, when degree and strength sequences are used as a constraint for the null model, the ensemble can reproduce the betweenness centrality observed in the data, suggesting that, in the case of spatial weighted networks, more sophisticated network models are better choices for representing the system.

Our findings demonstrate that the desire to use the simplest network representations of a system carries important risks. Typically, researchers fall back on models that ignore connection directionality and weight. While this choice may be good enough in many cases, in others it could be masking important characteristics of the system. Our study shows that the presence of centrality anomalies can be an indicator that important aspects of the system are being lost in its network representation. We believe that complex systems that have nodes and edges embedded in a physical space such as spatial networks (e.g. road networks, power grids, and neural networks), might show centralities anomalies when projected onto unweighted networks. Further investigation of these systems could extend the generality of our findings to other real-world systems.

Methods

Data. We obtained data from the Brazilian inter-city bus routes for the period between January 2005 and December 2014 at a monthly time-resolution. These data are maintained and distributed by the Brazilian National Land Transportation Agency [39]. The data contains more than 19 thousand unique routes connecting 1786 cities. We gathered the geographical location of all relevant cities from the Brazilian Institute of Geography and Statistics (IBGE) [40].

We obtained data from the British inter-city bus routes for the period between 4 October, 2010 and 10 October, 2010, at an hourly resolution. These data are maintained by the National Public Transport Data Repository and distributed by the Department of Transport and licensed under the Open Government Licence. This data set was complemented with the National Coach Services Data distributed also by the Department of Transport and licensed under the Open Government Licence [41]. The total number of nodes after the aggregation into municipalities is 279 comprising almost 4 thousand unique routes.

We obtained data from the Spanish inter-city bus routes for the period between 1 January, 2017 and 31 December, 2017, at an hourly resolution. These data are maintained and distributed by the Spain Ministry of Development [42]. The data is provided as the set of routes connecting each pair of municipalities in Spain except for the province of Girona. The total number of nodes is 1435 with over 20 thousand unique routes.

The data of the worldwide air transportation network were collected in the period between 17 May, 2018 and 22 May, 2018, at an hourly resolution. These data are maintained by the website Flight Aware [43]. The data contain all flights in 2734 airports around the world, with more than 16 thousand unique routes. The geographical location of the airports was obtained from the Open Flights website [44].

Sampling of networks. To investigate the statistical properties of transportation networks we have generated 10 000 networks sampled from the ensembles for each data set and topology (non-weighted or weighted). We followed the approach proposed by Squartini *et al* [31, 33] of unbiased sampling based on maximum-entropy distributions. In this approach, the probability distributions composing the ensemble are obtained by maximizing, in sequence, the Shannon's entropy and the likelihood function subject to the desired constraints. In particular, for the non-weighted networks case we used the 'UBCM', where the constraint is the degree sequence $\{k_i\}_{i=1}^{N}$. Notice that the constraints in the *canonical ensemble* are met on average over the network samples, differently from the *microcanonical ensemble*, i.e. Markov Chain Monte Carlo edge switching approach, where the constraints are satisfied exactly [32–34]. With the UBCM model the probability of having a link between nodes i and j, p_{ij} is given by

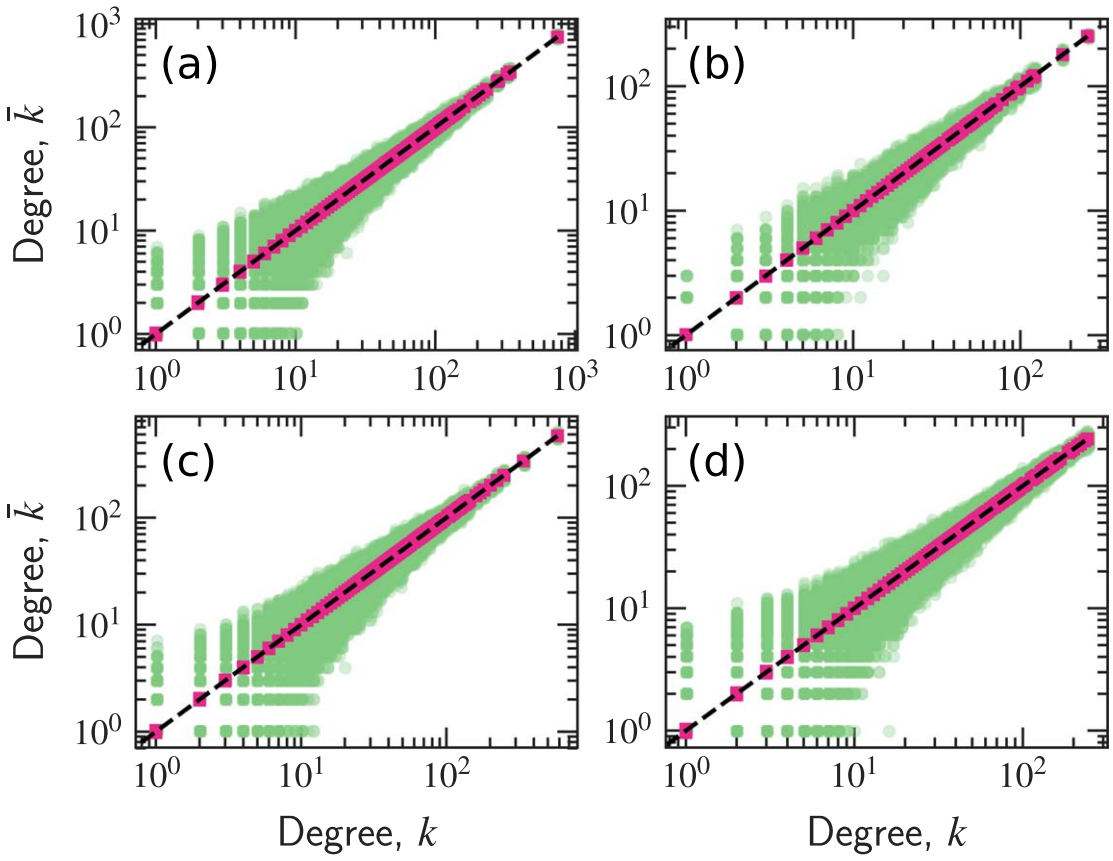

Figure 7. Samples of networks using the undirected binary configuration model (UBCM) given their degree sequence. In each plot, the green dots show the measured degree $\bar{k}_i$ in each sample versus the degree k_i in the observed network. The pink squared-dots represents the average degree $\langle k_i \rangle$ over the 10 000 networks of the ensemble versus the empirical degree k_i. The dashed line is a straight line with slope 1. (a) Brazilian buses transportation network, (b) British buses transportation network, (c) Spanish buses transportation network, and (d) worldwide air transportation network.

$$p_{ij} \equiv \frac{x_i x_j}{1 + x_i x_j}, \tag{3}$$

where the vector $\mathbf{x}$ of N unknown parameters can be determined by either maximizing the log-likelihood function

$$\lambda(\mathbf{x}) = \sum_i k_i(\mathbf{A}) \ln x_i - \sum_i \sum_{i<j} \ln(1 + x_i x_j), \tag{4}$$

where $\mathbf{A}$ refers to the adjacency matrix of the observed graph, or by solving the system of N equations:

$$\langle k_i \rangle = \sum_{j \neq i} \frac{x_i x_j}{1 + x_i x_j} = k_i(\mathbf{A}) \quad \forall i, \tag{5}$$

where $k_i(\mathbf{A})$ is the observed degree of node i and $\langle k_i \rangle$ is the ensemble average. Once the values of the p_{ij} have been determined, we can extract a sample graph from the ensemble by running a Bernoulli trial for each pair of vertices to connect i and j with probability $p_{ij}\,(a_{ij} = 1)$ and not connect with probability $1 - p_{ij}\,(a_{ij} = 0)$. Repeating this last step, we can generate any desired number of networks that, on average, have the same degree sequence as the observed one. Figure 7 shows a good agreement between the average degree versus the empirical ones.

Similarly, for the weighted network we have considered the 'UECM', where the constraints are the degree and strength sequences. Again, the constraints are met on average over the network samples (i.e. *canonical ensemble*). In this case, the probability p_{ij} is given by

$$p_{ij} \equiv \frac{x_i x_j y_i y_j}{1 - y_i y_j + x_i x_j y_i y_j} \tag{6}$$

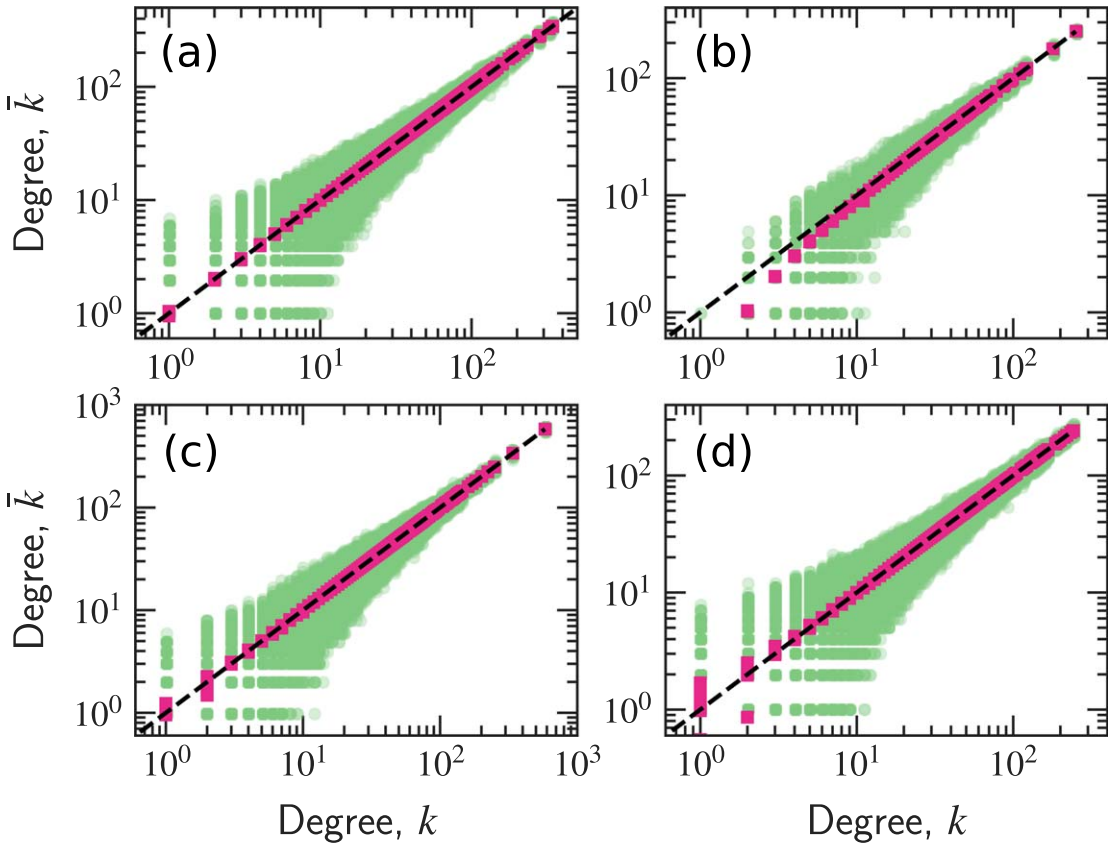

Figure 8. Samples of networks using undirected enhanced configuration model (UECM) given their degree and strength sequences. In each plot, the green dots show the measured degree $\bar{k_i}$ in each sample versus the degree k_i in the observed network. The pink squared-dots represents the average degree $\langle k_i \rangle$ over the 10 000 networks of the ensemble versus the empirical degree k_i. The dashed line is a straight line with slope 1. (a) Brazilian buses transportation network, (b) British buses transportation network, (c) Spanish buses transportation network, and (d) worldwide air transportation network.

and the **x** and **y** vectors can be computed, again, by either maximizing the log-likelihood

$$\lambda(\mathbf{x}, \mathbf{y}) \equiv \sum_i [k_i(\mathbf{W}) \ln x_i + s_i(\mathbf{W}) \ln y_i] + \sum_i \sum_{j<i} \ln \frac{1 - y_i y_j}{1 - y_i y_j + x_i x_j y_i y_j}, \tag{7}$$

where **W** represents in this case the adjacency matrix of the weighted graph, or by solving the $2N$ equations

$$\langle k_i \rangle = \sum_{j \neq i} p_{ij} = k_i(\mathbf{W}) \quad \forall\, i, \tag{8}$$

$$\langle s_i \rangle = \sum_{j \neq i} \frac{p_{ij}}{1 - y_i y_j} = s_i(\mathbf{W}) \quad \forall\, i, \tag{9}$$

where $k_i(\mathbf{W})$ and $s_i(\mathbf{W})$ are, respectively, the observed degree and strength of node i and $\langle k_i \rangle$ and $\langle s_i \rangle$ are the ensemble averages.

Thus, solving the above equations, the probabilities of generating a link of weight w between any pair of nodes i and j is given by

$$q_{ij} = \begin{cases} 1 - p_{ij}, & \text{if } w = 0, \\ p_{ij}(y_i y_j)^{w-1}(1 - y_i y_j), & \text{if } w > 0. \end{cases} \tag{10}$$

Figures 8 and 9 show, respectively, the average degree and strength over the ensemble generated by the UBCM method compared to the empirical observations.

Detecting anomalies. To detect the anomaly in betweenness centrality versus degree, we have calculated these quantities for each node over a 10 000 ensemble of synthetic networks considering the appropriate null models. For every node, we approximated the distribution of k and b by a multivariate Gaussian distribution and computed the fraction of nodes that lie outside the 95% confidence interval for the null model.

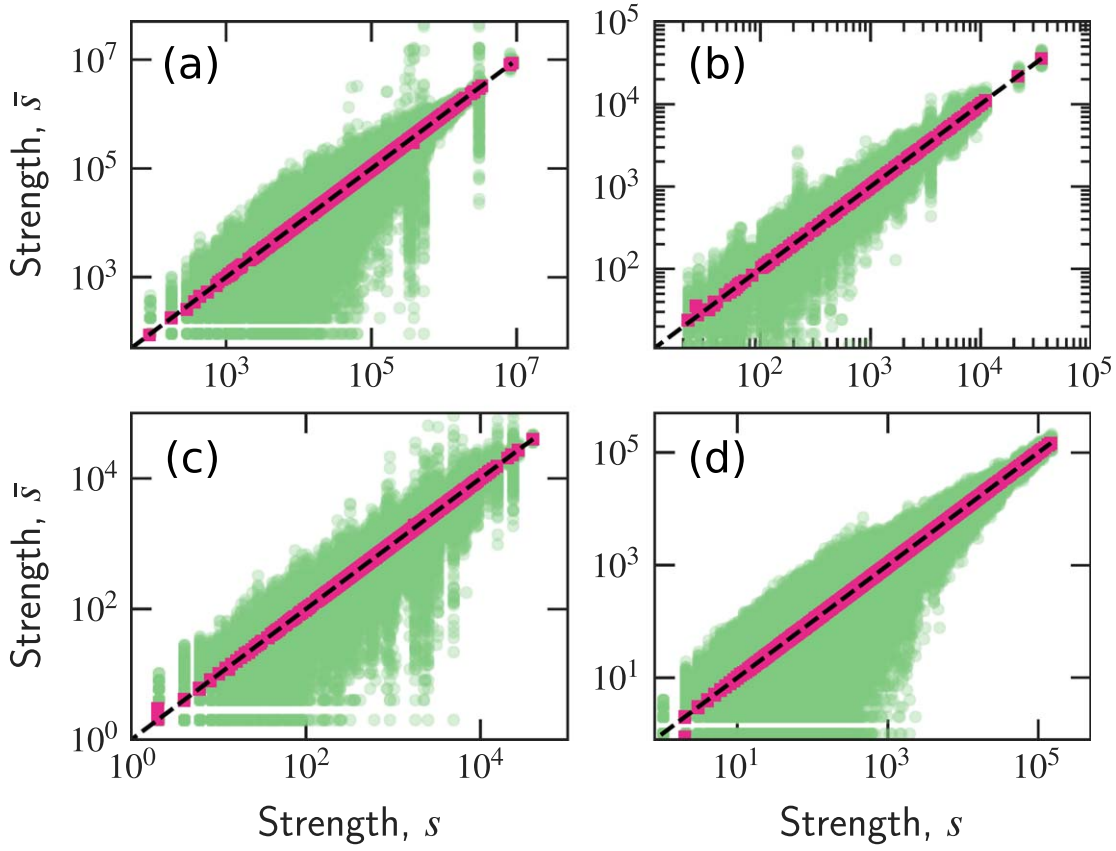

Figure 9. Samples of networks using undirected enhanced configuration model (UECM) given their degree and strength sequences. In each plot, the green dots show the measured strength $\bar{s_i}$ in each sample versus the strength s_i in the observed network. The pink squared-dots represents the average strength $\langle s_i \rangle$ over the 10 000 networks of the ensemble versus the empirical strength s_i. The dashed line is a straight line with slope 1. (a) Brazilian buses transportation network, (b) British buses transportation network, (c) Spanish buses transportation network, and (d) worldwide air transportation network.

Multivariate Gaussian fitting. For each node, we approximated the joint distribution of betweenness centrality and degree (or strength) by a multivariate Gaussian, that is

$$\mathcal{N}(\mathbf{x}, \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)}{\sqrt{(2\pi)^2|\boldsymbol{\Sigma}|}}, \tag{11}$$

where $\mathbf{x} = (k, b)^T$,

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_k \\ \mu_b \end{pmatrix}, \tag{12}$$

is the mean, and

$$\boldsymbol{\Sigma} = \rho \begin{pmatrix} \sigma_{kk} & \sigma_{kb} \\ \sigma_{kb} & \sigma_{bb} \end{pmatrix}, \tag{13}$$

is the covariance matrix, where ρ is the correlation between k and b. Thus, the line enclosing 95% of the probability mass for the null model is a ellipsoid (under a rotated coordinate system) with radii given by the eigenvalues $\sqrt{\lambda_1}$ and $\sqrt{\lambda_2}$ of the scaled covariance matrix $s\dot{\Sigma}$, where $s = -2\log(1 - p)$ and p is the confidence probability that the null hypothesis is true.

Acknowledgments

LGAA and AA contributed equally to this work. LGAA acknowledges FAPESP (2016/16987-7) for partial financial support. AA acknowledges the support of the FPI doctoral fellowship from MINECO and its mobility scheme (FIS2014-55867-P). FAR acknowledges the Leverhulme Trust, CNPq (305940/2010-4) and FAPESP (2016/25682-5 and 2013/07375-0) for the financial support given to this research. YM acknowledges partial support from the Government of Aragón, Spain through grant E36-17R (FENOL), and by MINECO and FEDER funds (FIS2017-87519-P). LANA thanks the John and Leslie McQuown Gift.

ORCID iDs

Luiz G A Alves https://orcid.org/0000-0001-6204-5552
Alberto Aleta https://orcid.org/0000-0002-1192-8707
Francisco A Rodrigues https://orcid.org/0000-0002-0145-5571
Yamir Moreno https://orcid.org/0000-0002-0895-1893
Luís A Nunes Amaral https://orcid.org/0000-0002-3762-789X

References

[1] Guimera R and Amaral L A N 2005 Functional cartography of complex metabolic networks *Nature* **433** 895
[2] Park H-J and Friston K 2013 Structural and functional brain networks: from connections to cognition *Science* **342** 1238411
[3] Girvan M and Newman M E J 2002 Community structure in social and biological networks *Proc. Natl Acad. Sci.* **99** 7821–6
[4] Wang Z, Szolnoki A and Perc M 2013 Interdependent network reciprocity in evolutionary games *Sci. Rep.* **3** 1183
[5] Cohen R and Havlin S 2010 *Complex Networks: Structure, Robustness and Function* (Cambridge: Cambridge University Press)
[6] Helbing D *et al* 2015 Saving human lives: what complexity science and information systems can contribute *J. Stat. Phys.* **158** 735–81
[7] Moreno Y, Pastor-Satorras R and Vespignani A 2002 Epidemic outbreaks in complex heterogeneous networks *Eur. Phys. J.* B **26** 521–9
[8] Pastor-Satorras R, Rubi M and Diaz-Guilera A 2003 *Statistical Mechanics of Complex Networks* vol 625 (New York: Springer)
[9] Newman M 2018 *Networks* (Oxford: Oxford University Press)
[10] Barrat A, Barthelemy M and Vespignani A 2008 *Dynamical Processes on Complex Networks* (Cambridge: Cambridge University Press)
[11] Strogatz S H 2001 Exploring complex networks *Nature* **410** 268
[12] Newman M E J 2003 The structure and function of complex networks *SIAM Rev.* **45** 167–256
[13] Amaral L A N and Ottino J M 2004 Complex networks *Eur. Phys. J.* B **38** 147–62
[14] Boccaletti S, Latora V, Moreno Y, Chavez M and Hwang D-U 2006 Complex networks: structure and dynamics *Phys. Rep.* **424** 175–308
[15] Barrat A, Barthélemy M, Pastor-Satorras R and Vespignani A 2004 The architecture of complex weighted networks *Proc. Natl Acad. Sci.* **101** 3747–52
[16] Buldyrev S V, Parshani R, Paul G, Stanley H E and Havlin S 2010 Catastrophic cascade of failures in interdependent networks *Nature* **464** 1025–8
[17] Kivelä M, Arenas A, Barthelemy M, Gleeson J P, Moreno Y and Porter M A 2014 Multilayer networks *J. Complex Netw.* **2** 203–71
[18] Boccaletti S, Bianconi G, Criado R, Del Genio C I, Gómez-Gardenes J, Romance M, Sendina-Nadal I, Wang Z and Zanin M 2014 The structure and dynamics of multilayer networks *Phys. Rep.* **544** 1–122
[19] Sanz J, Cozzo E, Borge-Holthoefer J and Moreno Y 2012 Topological effects of data incompleteness of gene regulatory networks *BMC Syst. Biol.* **6** 110
[20] Benson A R, Abebe R, Schaub M T, Jadbabaie A and Kleinberg J 2018 Simplicial closure and higher-order link prediction *Proc. Natl Acad. Sci.* **115** E11221–30
[21] Guimera R, Mossa S, Turtschi A and Amaral L A N 2005 The worldwide air transportation network: anomalous centrality, community structure, and cities' global roles *Proc. Natl Acad. Sci.* **102** 7794–9
[22] Barrat A, Barthélemy M and Vespignani A 2005 The effects of spatial constraints on the evolution of weighted complex networks *J. Stat. Mech.* P05003
[23] Barthélemy M 2011 Spatial networks *Phys. Rep.* **499** 1–101
[24] Barbosa H, Barthelemy M, Ghoshal G, James C R, Lenormand M, Louail T, Menezes R, Ramasco J J, Simini F and Tomasini M 2018 Human mobility: models and applications *Phys. Rep.* **734** 1–74
[25] Meloni S, Arenas A and Moreno Y 2009 Traffic-driven epidemic spreading in finite-size scale-free networks *Proc. Natl Acad. Sci.* **106** 16897–902
[26] Meloni S, Perra N, Arenas A, Gómez S, Moreno Y and Vespignani A 2011 Modeling human mobility responses to the large-scale spreading of infectious diseases *Sci. Rep.* **1** 62
[27] Caminha C, Furtado V, Pequeno T H C, Ponte C, Melo H P M, Oliveira E A and Andrade J S Jr 2017 Human mobility in large cities as a proxy for crime *PLoS One* **12** e0171609
[28] Guimera R and Nunes Amaral L A 2004 Modeling the world-wide airport network *Eur. Phys. J.* B **38** 381–5
[29] Mukherjee S 2012 Statistical analysis of the road network of India *Pramana-J. Phys.* **79** 483–91
[30] Freeman L C 1977 A set of measures of centrality based on betweenness *Sociometry* **40** 35–41
[31] Squartini T, Mastrandrea R and Garlaschelli D 2015 Unbiased sampling of network ensembles *New J. Phys.* **17** 023052
[32] Bianconi G 2007 The entropy of randomized network ensembles *Europhys. Lett.* **81** 28005
[33] Squartini T and Garlaschelli D 2011 Analytical maximum-likelihood method to detect patterns in real networks *New J. Phys.* **13** 083001
[34] Gabrielli A, Mastrandrea R, Caldarelli G and Cimini G 2019 Grand canonical ensemble of weighted networks *Phys. Rev.* E **99** 030301
[35] Barrat A, Barthélemy M and Vespignani A 2004 Weighted evolving networks: coupling topology and weight dynamics *Phys. Rev. Lett.* **92** 228701
[36] Simkin M V and Roychowdhury V P 2011 Re-inventing Willis *Phys. Rep.* **502** 1–35
[37] Mastrandrea R, Squartini T, Fagiolo G and Garlaschelli D 2014 Enhanced reconstruction of weighted networks from strengths and degrees *New J. Phys.* **16** 043022
[38] Brandes U 2008 On variants of shortest-path betweenness centrality and their generic computation *Soc. Netw.* **30** 136–45
[39] National Land Transport Agency—ANTT 2017 Statistics and Road Studies—Operational Data (Accessed: 1 September, 2017) (http://antt.gov.br/)
[40] Brazilian Institute of Geography and Statistics (IBGE) 2017 Cartography (Accessed: 1 September, 2017) (http://ibge.gov.br/)
[41] National Public Transport Data Repository 2017 (Accessed: 1 September, 2017) (https://data.gov.uk/)
[42] bus.es Autobuses España 2017 (Accessed: 1 September, 2017) (http://bus.es/)
[43] Flight Aware 2018 (Accessed: July 2018) (https://flightaware.com/)
[44] Open Flights 2018 (Accessed: July 2018) (https://openflights.org/)