How do urban mobility (geo)graph's topological properties fill a map?

Leonardo Bacelar Lima Santos[1*], Luiz Max Carvalho[2], Wilson Seron[3], Flávio C. Coelho[4], Elbert E. Macau[3,5], Marcos G. Quiles[3] and Antônio M. V. Monteiro[5]

*Correspondence:
santoslbl@gmail.com
[1]Centro Nacional de
Monitoramento e Alertas de
Desastres Naturais (Cemaden), São
José dos Campos, Brazil
Full list of author information is
available at the end of the article

Abstract

Urban mobility data are important to areas ranging from traffic engineering to the analysis of outbreaks and disasters. In this paper, we study mobility data from a major Brazilian city from a geographical viewpoint using a Complex Network approach. The case study is based on intra-urban mobility data from the Metropolitan area of Rio de Janeiro (Brazil), presenting more than 480 spatial network nodes. While for the mobility flow data a log-normal distribution outperformed the power law, we also found moderate evidence for scale-free and small word effects in the flow network's degree distribution. We employ a novel open-source GIS tool to display (geo)graph's topological properties in maps and observe a strong traffic-topology association and also a fine adjustment for hubs location for different flow threshold networks. In the central commercial area for lower thresholds and in high population residential areas for higher thresholds. This set of results, including statistical, topological and geographical analysis may represent an important tool for policymakers and stakeholders in the urban planning area, especially by the identification of zones with few but strong links in a real data-driven mobility network.

Keywords: Complex networks, Geographical information systems (GIS), (geo)graphs, Traffic-topology correlation

Introduction

Urban mobility data are important to several areas, from traffic engineering to the analysis of outbreaks and disasters. Many studies explore patterns, applicability, and limitations on urban mobility (Gonzalez et al. 2008; Song et al. 2010; Simini et al. 2012; Guo et al. 2012; Wang et al. 2012; Louail et al. 2015). Another common thread among these studies is the importance of spatial structure. In this work, the spatial structure of a actual data-based mobility complex network is explored.

There are several classical approaches to the analysis of urban mobility data, from mechanical models to statistical ones (Costa et al. 2017; Barbosa et al. 2018). According to Barat and Cattuto (2013), in many cases, urban mobility information finds a convenient representation in terms of complex networks. The complex network approach emerges as a natural mechanism to handle mobility data, taking areas as nodes and movements between origins and destinations as edges. However, there are difficulties in incorporating human mobility into models from both technical and ethical perspectives (Balcan et al. 2009). At the intra-urban scale, this difficulty is magnified due to

the complex structure of the urban territory. Thus, a general approach for handling geographical data is needed.

As presented in Barthélemy (2011), a review about spatial network, several complex systems are very often organized as networks where their elements (nodes and edges) are embedded in (geographical) space and topology alone does not contain all the information important for understanding processes and propose scientific and technological developments. A geographical approach for complex systems analysis is especially important for mobility phenomena (Barthélemy 2011).

Santos et al. (2017) proposed the (geo)graphs approach, in which a (geo)graph is defined as a graph in which the nodes have a known geographical location, and the edges have spatial dependence. (Geo)graphs provide a simple tool to manage, represent and analyze geographical complex network.

In this paper we explore the spatial structure of an actual data-based mobility complex network.

In particular, we applied a set of procedures to use origin-destination data (OD data), originally from traffic engineering, to recover useful information about mobility. OD data represent daily travels between zones on a region, especially interesting for the intra-urban scale. According to Estrada (2012), social proximity refers to actors that belong to the same space of social relations. And OD data can be seen as a "social" relation between origin and destination areas.

The central question in this paper is: how to recover useful information from urban mobility data considering its intrinsic spatial properties?

In several previous works (Song et al. 2010; De Montis et al. 2007; Chowell et al. 2003; Soh et al. 2010; Brockmann et al. 2006), the power-law behaviour of human mobility models and data were explored.

In this work, as a first step, we extend the method of Clauset et al. (2009) of distributions analysis by employing a Bayesian approach and computing Bayes factors. So, we use the mobility data to construct mobility networks and calculate topological properties. Finally, we return to the geographical domain representing and analyzing the topological measures.

The traffic-topology analysis is a traditional object of research in the mobility network literature (Chowell et al. 2003; Soh et al. 2010; De Montis et al. 2007). However, there is an open question: how is the hub's strength distributed among its links? In this work, we propose a complementary traffic-topology analysis with an explicit spatial meaning.

The network connection criteria applied in this work is similar to those in other studies in the literature (Chowell et al. 2003; De Montis et al. 2007; Soh et al. 2010), who also investigated flow weights distribution and the traffic-topology correlations. However, in contrast to those studies, we test different distributions and explore the spatial aspect of the results.

This paper is organized as follows: "Material and methods" section contains the data and methods of this investigation, in particular the fitting of power law distributions ("Power law analysis" section), power law regression ("Power law regression" section) and geographs ("(Geo)graph tools" section). Results and discussion are presented in "Result and discussion" section. Finally, some concluding remarks and perspective of future work are drawn in "Conclusions and perspectives" section.

Material and methods

Mobility data

The Metropolitan Region of Rio de Janeiro (MRRJ) encompasses 20 cities, for a total of 10,894,756 inhabitants. It is the second largest metropolitan area in Brazil, the third in South America and the 20*th* in the world.

To facilitate mobility studies, the region is divided into a set of traffic zones (TZ). For this specific work, we consider a set of 485 traffic zones, in which each TZ has appeared at least once as a source or as a destination in the set of travels of more than 99 thousands interviewed people in a Origin-Destination Survey (Companhia estadual de engenharia de transporte e logistica et al. 2010). From a network perspective, each TZ is represented by a node.

The original data (Companhia estadual de engenharia de transporte e logistica et al. 2010) consists of a list of travels, each one with an origin TZ and a destination TZ. This dataset is summarized into a flow matrix, in which each element $f(i,j)$ records the number of travels between TZs i and j, in both directions (i.e. the matrix is symmetric).

The driving-mode data used here takes into account: car, bus or motorcycle, representing more than 56% of the total number of travels.

Power law analysis

We extend the approach of Clauset et al. (2009) by employing a Bayesian approach to fitting and comparing distributions for the data presented in this article. Clauset et al. (2009) proposed to select the lower threshold $x_{\min}$ after which the data follows a power law regime by minimizing the Kolmogorov-Smirnov (KS) goodness-of-fit statistic. We adopt their procedure to estimate $x_{\min}$ and then proceed by assuming $x_{\min}$ is fixed and known. Parameter estimation and model selection for various distributions are made assuming the same $x_{\min}$ for all scenarios.

Let $\boldsymbol{x} = \{x_1, x_2, \dots, x_N\}$ denote the observed data. The power law distribution has probability density function (p.d.f.):

$$f(x_i|\alpha, x_{\min}) = \frac{\alpha - 1}{x_{\min}} \left(\frac{x_i}{x_{\min}} \right)^{-\alpha}.$$

We complete the model specification with a prior distribution $\pi(\alpha) \propto 1/(\alpha - 1)$, which leads to a proper posterior distribution $p(\alpha|\boldsymbol{x}, x_{\min})$.

The second distribution we analyze is the stretched exponential (Weibull, see Clauset et al. Table 1), with p.d.f

$$f(x_i|\lambda, \beta, x_{\min}) = \lambda \beta \exp \left(\lambda x_{\min}^{\beta} \right) x_i^{\beta - 1} \exp \left(-\lambda x_i^{\beta} \right).$$

We employ Gamma priors on the parameters β and λ:

$$\pi_\beta(\beta|a_1, b_1) = \frac{b_1^{a_1}}{\Gamma(a_1)} \beta^{a_1 - 1} \exp(-b_1 \beta),$$

$$\pi_\lambda(\lambda|a_2, b_2) = \frac{b_2^{a_2}}{\Gamma(a_2)} \lambda^{a_2 - 1} \exp(-b_2 \lambda).$$

The third and final distribution we consider in this study is the lower-truncated log-normal distribution:

$$f(x_i|\mu,\sigma,x_{\min}) = \sqrt{\frac{2}{\pi\sigma^2}} \frac{1}{x_i} \frac{\exp\left(-\frac{(\ln x_i - \mu)^2}{2\sigma^2}\right)}{\operatorname{erfc}\left(\frac{\ln x_{\min} - \mu}{\sqrt{2}\sigma}\right)}.$$

For the analysis of this distribution we choose a normal (Gaussian) prior for μ with mean 1 and standard deviation 5 and a Gamma prior for σ with $a = b = 1$. Notice we parametrize the normal (and log-normal) distribution in terms of mean and standard deviation.

We estimated the parameters of these distributions using the dynamic Hamiltonian Monte Carlo (HMC) algorithm implemented in the Stan probabilistic programming language (Carpenter et al. 2017) through the **rstan** package (Stan Development Team 2018) of the R programming language (R Core Team 2018), version 3.5.1. We ran four chains of 2000 iterations and checked convergence by making sure the split-Rhat statistic was below 1.01. Monte Carlo standard errors (mcse) were below 1% of the posterior standard deviations for all estimates reported in this paper.

To compare the fit of the distributions considered here to data we employ Bayes factors (Jeffreys 1935; Kass and Raftery 1995). Let $\mathcal{M}_0$ and $\mathcal{M}_1$ be two models or hypotheses one wants to test after observing data Y. The Bayes factor is defined as

$$\mathrm{BF}_{10} = \frac{p(Y|\mathcal{M}_1)}{p(Y|\mathcal{M}_0)},$$

quantifies the amount of support in favor of $\mathcal{M}_1$ compared to $\mathcal{M}_0$. For reasons of numerical stability, one usually computes $\ln \mathrm{BF}_{01}$. We employed the routines implemented in the **bridgesampling** R package (Gronau et al. 2017) to compute log-marginal likelihoods $\ln p(Y|\mathcal{M}_i)$ which were then used to compute log Bayes factors.

Power law regression

A distinct goal to fitting a power law distribution is assessing whether two variables of interest are related according to a power law. Let w_i be the weight of node i and let d_i be its degree. We say that w and d are related by a power law if the relationship $w \propto k^\beta$ holds for $\beta > 0$.

One routinely employed option to determine the exponent β from data is to fit the model:

$$w_i \sim \mathrm{Normal}(\mu_i = Kd_i^\beta, \sigma)$$

by least squares – see e.g. De Montis et al. (2007). While this approach can often lead to good estimates, it is poorly suited to strictly positive data because it allows negative predictions.

A better model for strictly positive data like that analyzed here is the Gamma regression model with an identity link function:

$$w_i \sim \mathrm{Gamma}(\mu_i = Kd_i^\beta, \kappa),$$

where we parametrize the Gamma distribution in terms of a mean μ and a shape κ, with p.d.f.

$$f(w_i|\mu_i,\kappa) = \frac{(\kappa/\mu_i)^\kappa}{\Gamma(\kappa)} w_i^{\kappa-1} \exp\left(-\frac{\kappa w_i}{\mu_i}\right).$$

This model allows for a strictly positive response variable whilst retaining directly interpretable parameters.

Finally, in the interest of completeness, we consider a log-normal model. While incorporating the positivity constraint in the data, it does not retain directly interpretability of the parameters, as estimates of the coefficients K and β pertain to the log scale. We note that an exponential transformation of the estimated parameters brings them to a comparable scale to the calculations from the two previous models.

To complete the specification of our Bayesian model, we place a Gamma(1, 1) prior on β and a Gamma(0.1, 0.01) on K. We fitted the three regression models for each weight threshold: 1, 1000 and 5000. In order to study the predictive performance of these three models (Gaussian, gamma and log-normal). We have employed leave-one-out (LOO) cross validation (Vehtari et al. 2017) in our experiments. It is important to note that in this paper we leave out individual graph nodes. In addition, we investigate model fit using the techniques described in Gabry et al. (2017) (see Supplementary figures). For these analyses we employed the **brms** package (Bürkner and et al 2017), using the same computational settings considered for the power law analysis above (four independent chains, checking split R-hat <1.01).

(Geo)graph tools

Among some important tolls in literature of geoinformatics is the MovingPandas (Graser 2019) and the OSMnx (Boeing 2017). The MovingPandas (Graser 2019) is a recent library for dealing with movement data, providing the user several functions and interfaces with Geographical DataBase Management Systems and Geographical Information Systems. The OSMnx (Boeing 2017) is a tool for creating and analysis of street networks under a simple, consistent and automatable paradigm. Even more specific to handle geographic network data is the (geo)graph package (GG) (Santos et al. 2017), applied in this work.

In the (geo)graph approach, a (geo)graph is defined as a graph in which the nodes have a known geographical location, and the edges have spatial dependence. So, the GG package allows the user work with the set of nodes as a point-type shapefile and the set of edges as a line-type shapefile, very common file structures in geoinformatics.

In order to convert spatial networks to the GIS environment we propose the following workflow (Santos et al. 2017):

1. To create a shapefile for the nodes using any GIS software. A *point type shapefile* for the nodes must be created. The shapefile must have a mandatory column of type integer named *id*, representing the id's of the nodes. All the characteristics of the polygons/points will be associated to their respective points as attributes, including the geographic locations of the nodes.
2. To create an adjacency matrix (0s and 1s) representing the connections between these nodes.
3. Then, a *line type shapefile* representing the edges of the network is given as an output of our application. The point-type-shapefile and the line-type-shapefile will have topological attributes of nodes and edges respectively.
4. Then, a *line type shapefile* representing the edges of the network is given as an output of our application. The point-type-shapefile and the line-type-shapefile will have topological attributes of nodes and edges respectively.
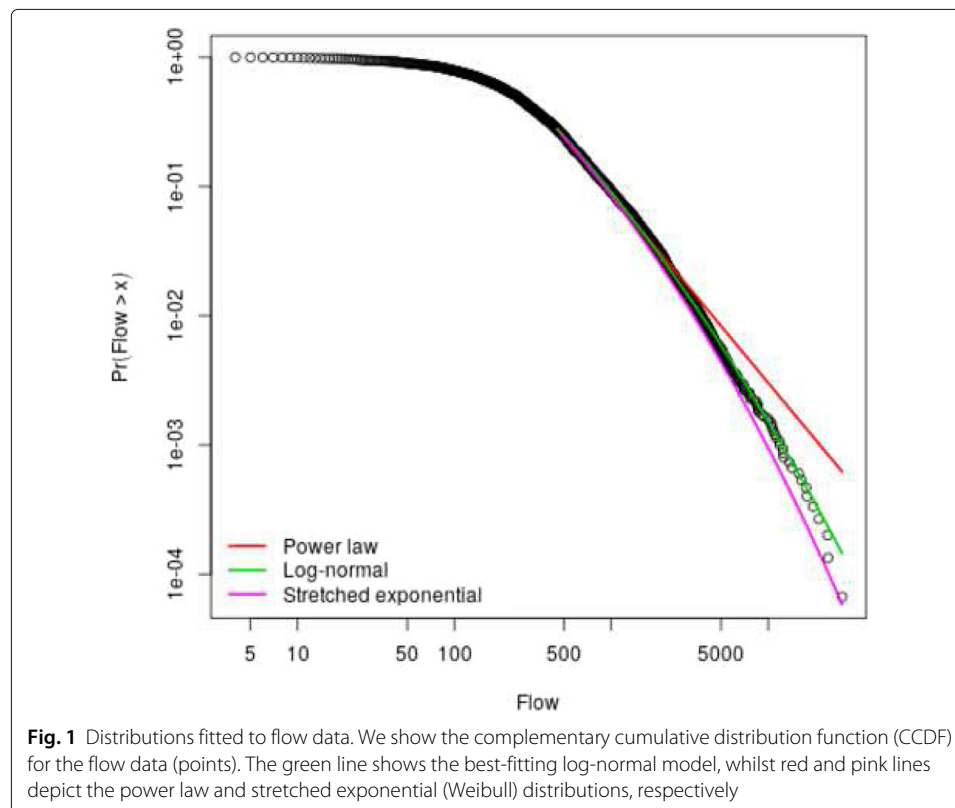
Results and discussion

Distribution of node weights

The flow values range between 0 and $w_{max} = 21228$ number of people from an origin to one destination, with an average value of $w_{ave} = 33$, which is considerably lower than w_{max}. This shows a high level of heterogeneity in the flow distribution. It is important to highlight that 12% of the travels have both origin and destination in the same node (TZ), i.e., are internal travels.

We fit three distributions to the mobility flow data (Fig. 1) and find that a truncated log-normal distribution with $x_{\min} = 452$ fits the data better than both a power law (log BF = 15) and a stretched exponential (log BF = 2). We report all marginal likelihoods and associated standard errors in the supplementary information. The estimated parameters of the log-normal are $\mu = 1.24$ with 95% credibility interval $(-1.70, 3.03)$ and $\sigma = 2.04 (1.71, 2.49)$. Interestingly, if one fails to consider other models, a power law model would be wrongly assumed to be the correct distribution of the data, as it gives an exponent firmly in the critical range $2 < \alpha < 3$, with estimated $\alpha = 2.46\ (2.42, 2.48)$.

Furthermore, while not fitting the data better than the log-normal, the stretched exponential distribution provides a better fit than the power law (log BF = 13). The posterior means of the parameters were $\beta = 0.17 (0.15, 0.18)$ and $\lambda = 3.09 (1.52, 5.45)$. Overall these results show that (i) the flow data analyzed here do not follow a power law distribution, with a log-normal providing better fit and, (ii) the importance of considering alternative distributions into consideration when analyzing real data.

The power-law behaviour of mobility data was explored in several previous work (Song et al. 2010; De Montis et al. 2007; Chowell et al. 2003; Soh et al. 2010; Brockmann et al.

Fig. 1 Distributions fitted to flow data. We show the complementary cumulative distribution function (CCDF) for the flow data (points). The green line shows the best-fitting log-normal model, whilst red and pink lines depict the power law and stretched exponential (Weibull) distributions, respectively

2006). In this work, for this case study, we have shown that for the mobility flow data a log-normal distribution outperformed the power law.

Traffic-topology correlation

We also studied the traffic-topology correlation, i.e., the relationship between a node's strength (total weight), w_i, and its degree, d_i. In Table 1 we present the parameter estimates for each weight (flow) threshold.

We have found β values between the previous found in literature: 0.94 in Chowell et al. (2003) and 1.8 in De Montis et al. (2007).

The results show that there is a significant correlation between the weights and degrees, especially for higher thresholds. Also, the weights of the nodes grow slightly faster than their degrees (Barrat et al. 2004; De Montis et al. 2007). For the non-zero threshold (flow > 1), however, we find that the posterior distribution for β includes the "null" hypothesis $\beta = 1$. See Fig. 2 for the fitted regression lines, along with prediction intervals using the Gamma error structure.

From a statistical perspective, we find that a Gamma regression model provides better predictive ability (better LOO scores) compared to the usually employed Gaussian regression model. For example, for flow threshold 1, difference in LOOIC scores was 735 with standard error of 99, while for threshold 1000 the difference was 53 with standard error of 21. While in some settings the log-normal model yielded better fit, the differences in predictive ability are small enough to justify employing the Gamma model, for which estimates of K and β can obtained directly with no need for transformation.

Network properties

In Fig. 3 we present several properties (statistical indices) for flow networks considering distinct connection threshold values, i.e. minimum flow value for connecting a pair of zones.

The index $< k >$ represents the expected value for the number of connections of a node for a specific connection threshold value. This value can be viewed as the network's average connectivity. The minimum connection threshold value is 0, in this case we have a fully connected network (complete graph). For small non-null connection threshold values, almost all zones are connect to the others, while for high values just a few pairs of zones remain. The index $< c >$ is associated to the network's transitivity: measures the average (over all nodes) connection probability of a pair of nodes connect to a same other node. On the other hand, the index $< l >$ represents the average (over all nodes) number of edges in the smallest path between a pair of nodes, while the index D is the greatest shortest path length. A detailed description of complex network's indexes can be founded in da F. Costa et al. (2007).

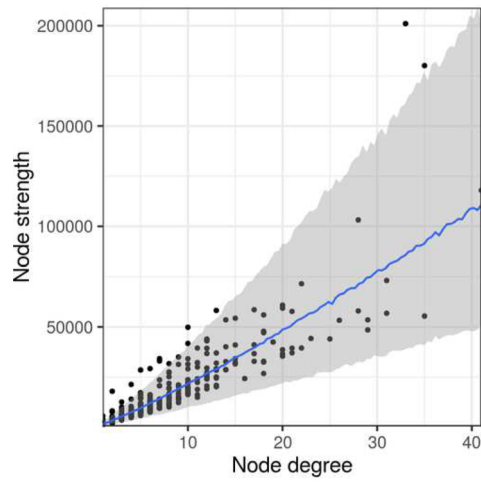**Table 1** Power law regression results

	K (BCI)	β (BCI)
Threshold 1	402 (284, 546)	1.04 (0.96, 1.13)
Threshold 1000	1561 (1462, 1663)	1.16 (1.12, 1.20)
Threshold 5000	6887 (6117, 7617)	1.18 (1.03, 1.35)

We report posterior means and 95% Bayesian credibility intervals (BCI) of parameters under the Gamma regression model
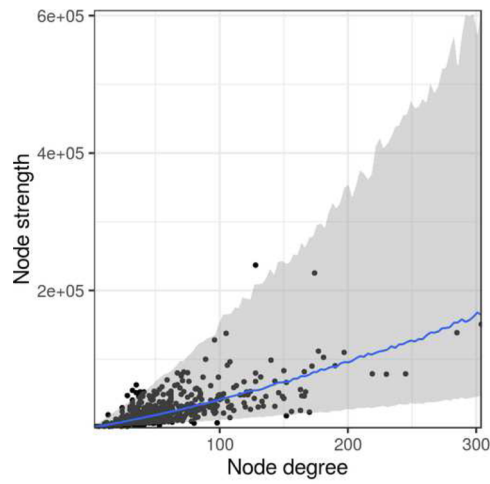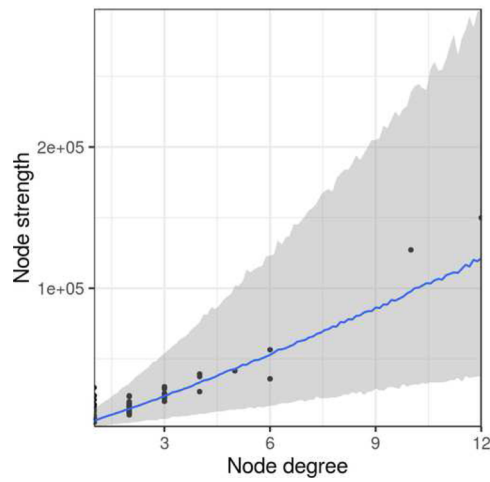
(a) Threshold = 1   4.7E-3 percent of the maximum flow

(b) Threshold = 1000   4.7 percent of the maximum flow

(c) Threshold = 5000   24 percent of the maximum flow

Fig. 2 Fitted regression lines and 95% prediction intervals for power law regression of weight versus degree. We present results for each threshold. **a**) Threshold = 1   (4.7E-3 percent of the maximum flow), **b**) Threshold = 1000   (4.7 percent of the maximum flow), **c**) Threshold = 5000   (24 percent of the maximum flow)

Fig. 3 Topological properties for different threshold connection. Log-log plot, base 10
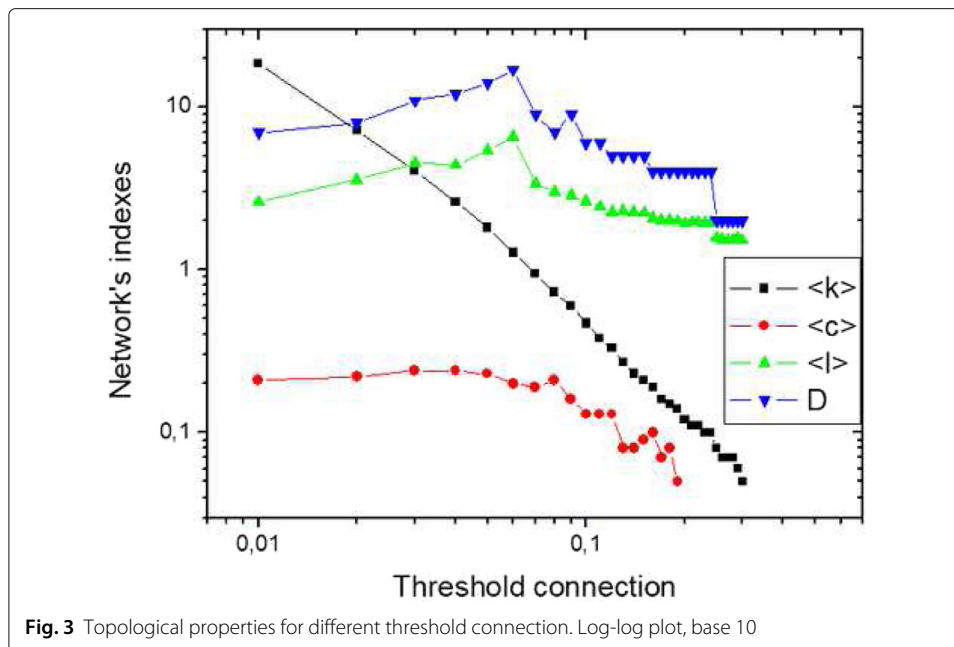
The minimum connection threshold value in order to get a non-complete connected network is $4.7E - 5$: we connect every pair of nodes (origin-destination) with at least one person going from this origin to this destination. In this case, there is only one connected component, with 485 nodes and 14.155 edges.

Then, we looked for the connection threshold associated to the greatest network's diameter, in order to balance the weak and the strong edges for a single connection threshold: for this connection threshold we have not disconnected the largest connect component yet. We call this specific value as the Critical Connection Threshold (CCT). This critical connection threshold (CCT) is 0.06. In the CCT is 133 nodes in the largest connected component. After this value, the network's diameter decreases when we increase the connection threshold. The CCT is between the thresholds 4.7E-5 (Fig. 4) and 0.06 (Fig. 5).

In Table 2, we present the network's properties for three selected connection thresholds: the smallest non-null connection threshold, the CCT and one that delivers a fragmented (disconnected) network.

For the CCT, we highlight that for a random network with the same number of nodes and edges, the expected value for $< c_{rand} >$ is 0.01 and $< l_{rand} >$ is 20.5. Thus, there is a statistical small world effect in this network: $< c >$ is greater than $< c_{rand} >$ and $< l >$ is smaller than $< l_{rand} >$ (Watts and Strogatz 1998).

This means that it is often possible to find connections between pairs of zones already connected to a common zone (creating triangles in the network structure) and even for a pair of nodes where the directly flow between them is not so high there are some "shortcuts" on the mobility network, getting these zones closer (in a topological point of view).

Considering both the smallest non-null and the critical connection threshold, we analyzed the degree distribution (Figs. 4 and 5). For the non-null threshold (flow of at least one travel), the distribution presented in Fig. 4 is not well approximated by any of tested distributions, with the power law providing a marginally better fit (log Bayes factor against
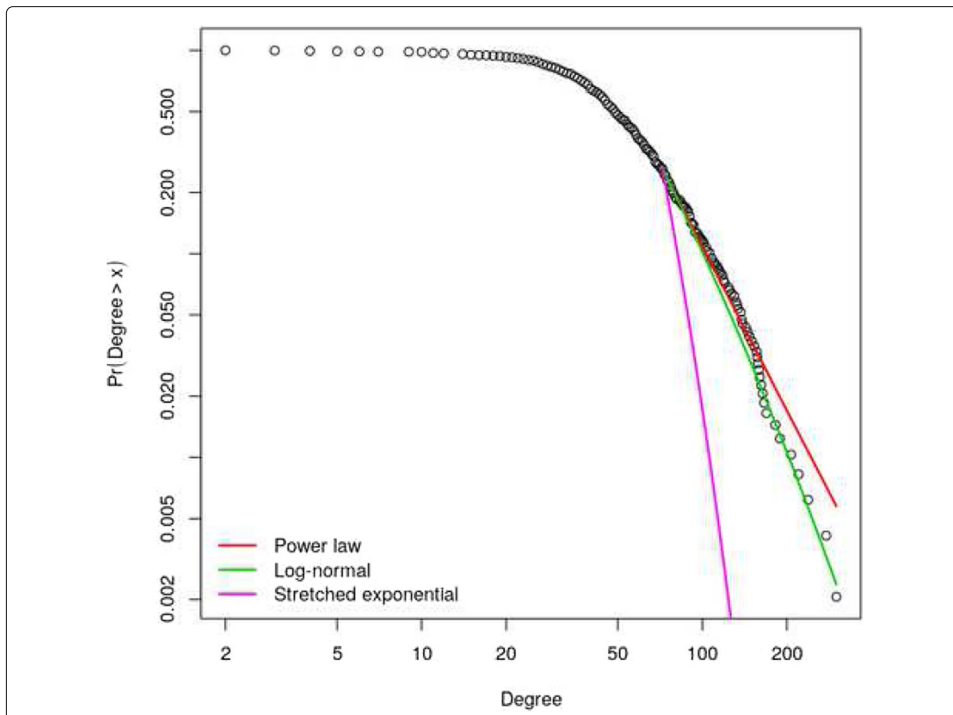
Fig. 4 Degree distribution for threshold 1 (4.7E-3%). We show the fitted complementary c.d.f.s for a power law, a log-normal and stretched exponential (Weibull) distributions. We estimated $x_{min} = 72$ using the method of Clauset et al. (2009). For this threshold, the log-normal provides the best fit to data (see text)
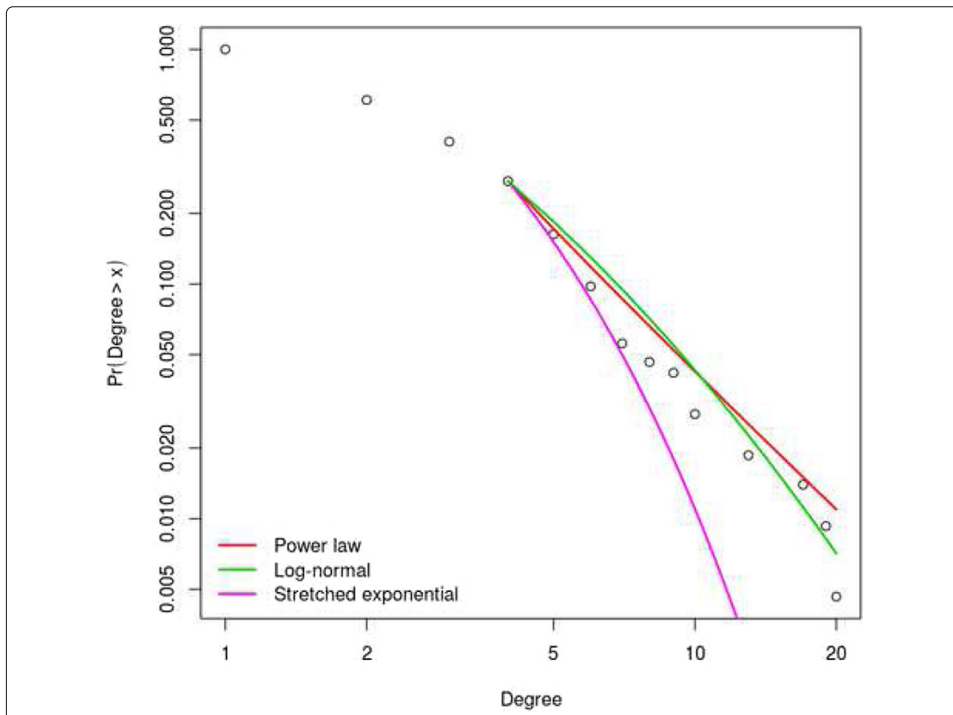
Fig. 5 Degree distribution for threshold 1273 (6%). We show the fitted complementary c.d.f.s for a power law, a log-normal and stretched exponential (Weibull) distributions. We estimated $x_{min} = 4$ using the method of Clauset et al. (2009). For this threshold, the power law provides the best fit to data (see text)

Table 2 Statistical properties of Rio de Janeiro's mobility network

	< k >	< c >	< l >	D
4.7E-5	58.4	0.28	1.94	4
4.7E-2	2.00	0.23	5.63	14
0.06	1.27	0.2	6.6	17
0.2	0.12	0	1.93	4
0.24	0.09	0	1.93	4

Each collum represents a topological measure: $< k >$ is the average degree, $< c >$ is the average clustering coeficiente, $< l >$ is the average shortest path length and D is the diameter of the network; and each line brings a threshold: from smaller to higher values

log-normal: 0.42). A bootstrap test of adequacy of the power law yielded a p-value of 0.08, which means the power law is not suitable to describe the degree distribution for this threshold.

On the other hand, for the critical connection threshold (1273 trips between locations) shown in Fig. 5, the power law provides adequate fit (bootstrap p-value: 0.745), although the support for the power law against a log-normal is weak (log Bayes factor against log-normal: 0.25).

These findings are in agreement with those of Alessandretti et al. (2017), who find that when the whole distribution of displacements is considered, i.e., when both small and large values are used in the fitting procedure, the log-normal outperforms the (Pareto) power law. Conversely, when restricting attention to large values, the power law provides better fit.

The connection threshold plays a central hole also on the (geographical) space. For small connection thresholds the networks' hubs are in the central region of Rio de Janeiro and Niterói (the two most important cities in this metropolitan region) - Fig. 6, and for the
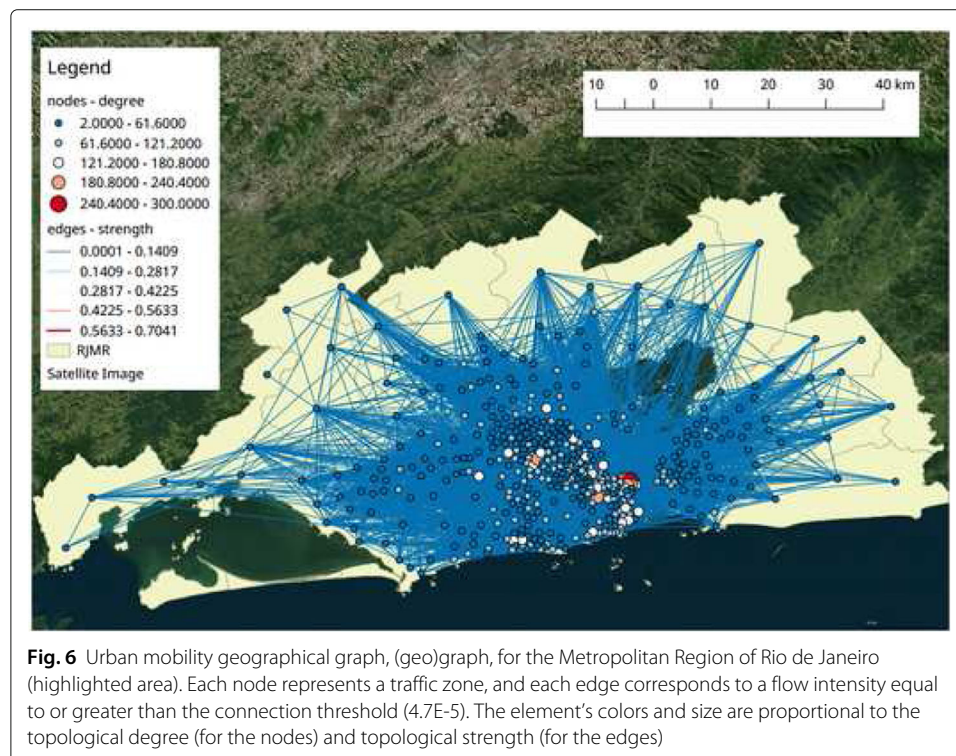
Fig. 6 Urban mobility geographical graph, (geo)graph, for the Metropolitan Region of Rio de Janeiro (highlighted area). Each node represents a traffic zone, and each edge corresponds to a flow intensity equal to or greater than the connection threshold (4.7E-5). The element's colors and size are proportional to the topological degree (for the nodes) and topological strength (for the edges)

critical connection threshold the hubs are located in zones belonging to Nova Iguaçu and Duque de Caxias (two of the most populous cities and with a large number of commuters) - Fig. 7.
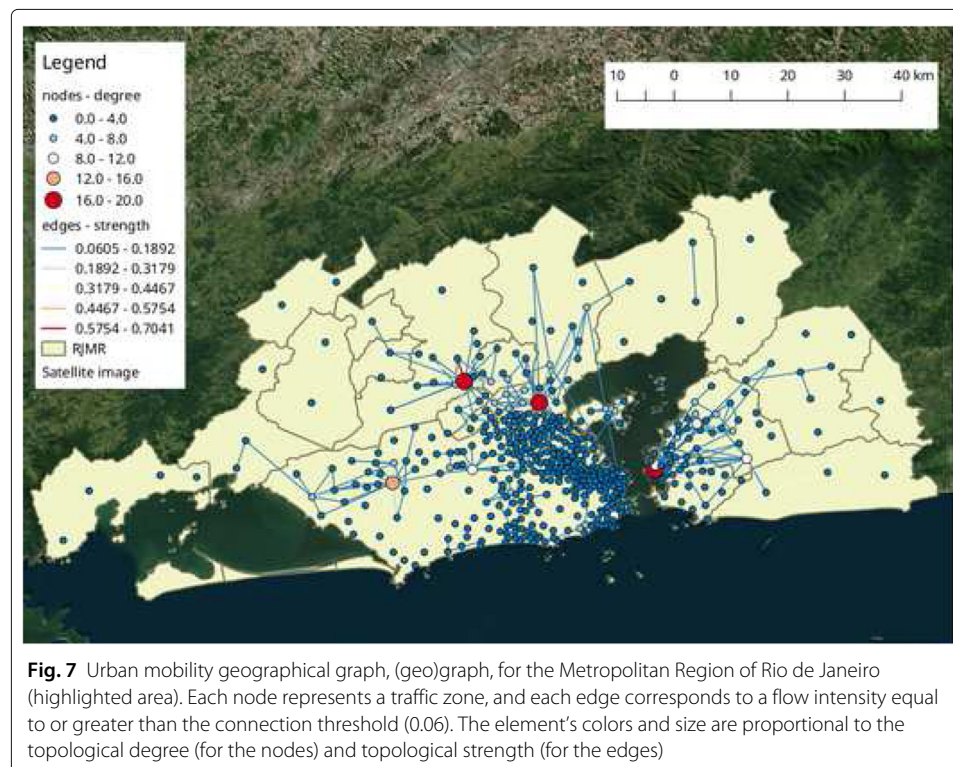
When we compare our topological findings with some previous similar works, especially Chowell et al. (2003) and De Montis et al. (2007), it is possible to note that the RJ's mobility network is bigger (in terms of number of edges) and presents high diameter.

These are expected results, as the study area is a Metropolitan Region, with several municipalities combining, daily, intense inter-city travels (mainly for work and study) with intra-city travels (for general purposes). Conforming the Origin-Destination Survey for the Metropolitan Region of Rio de Janeiro, 21% of the travel were due working motivation, 16% for studying activities and 50% starts or ends at the traveler residence (Companhia estadual de engenharia de transporte e logistica et al. 2010).

On the other hand, a non-intuitive and important result is related to how the hub's strength is distributed among its links. In this case study, the central area of MRRJ is connected with many other areas, however most of its connections are weak - associated to small-flow values. Even though Nova Iguaçu and Duque de Caxias do not connect to as many zones as does the central region, these zones have a set of a few pairs of nodes with high flow - a few but strong links.

Conclusions and perspectives

In this paper we address some aspects of the urban mobility phenomenon under a geographical point of view using a Complex Network approach. We applied the (geo)graph approach and tolls in order to recovery useful information from urban mobility data considering its intrinsic spatial proprieties. We found a high level of heterogeneity in

Fig. 7 Urban mobility geographical graph, (geo)graph, for the Metropolitan Region of Rio de Janeiro (highlighted area). Each node represents a traffic zone, and each edge corresponds to a flow intensity equal to or greater than the connection threshold (0.06). The element's colors and size are proportional to the topological degree (for the nodes) and topological strength (for the edges)

the Metropolitan Region of Rio de Janeiro's flow distribution. Considering the complex network analysis, we showed a statistical small world effect behaviour around the Critical Connection Threshold - set of connection thresholds that provide a network with maximum diameter.

An important point in the traffic-topology analysis is how the hub's strength is distributed among its links. We have shown that, in our case study, the central area of the region is connected with many other areas, however most of its connections are weak - associated to small-flow values. On the other hand, there are some zones do not connect to as many other zones as does the central region, but these zones have a few but strong (high flow) links.

From a methodological point of view, we also extended the method of Clauset et al. (2009) of distributions analysis by employing a Bayesian approach and computing Bayes factors. In addition, we show that a Gamma model for power law regression analysis leads to better fit, provides directly interpretable parameters and respects non-negativity constraints in the data.

Among our perspectives is replicating our analysis on other datasets (from different cities) and compare the results, in attempt to capture some regional behaviour.

Author details
[1]Centro Nacional de Monitoramento e Alertas de Desastres Naturais (Cemaden), São José dos Campos, Brazil. [2]Programa de Computação Científica (PROCC), Fundação Oswaldo Cruz, Rio de Janeiro, Brazil. [3]Universidade Federal de São Paulo (UNIFESP), São José dos Campos, Brazil. [4]Fundação Getúlio Vargas, Rio de Janeiro, Brazil. [5]Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, Brazil.

References
Alessandretti L, Sapiezynski P, Lehmann S, Baronchelli A (2017) Multi-scale spatio-temporal analysis of human mobility. PloS ONE 12(2):0171686

Balcan D, Colizza V, Goncalces B, Hud H, Ramascob J, Vespignani A (2009) Multiscale mobility networks and the spatial spreading of infectious diseases. PNAS 106(51):21487

Barat A, Cattuto C (2013) Empirical temporal networks of face-to-face human interactions. Eur Phys J Spec Top 222:1295–1309

Barrat A, Barthelemy M, Pastor-Satorras R, Vespignani A (2004) The architecture of complex weighted networks. Proc Natl Acad Sci 101(11):3747–3752

Barbosa H, Barthelemy M, Ghoshal G, James CR, Lenormand M, Louail T, Menezes R, Ramasco JJ, Simini F, Tomasini M (2018) Human mobility: Models and applications. Phys Rep 734:1–74. https://doi.org/10.1016/j.physrep.2018.01.001. Human mobility: Models and applications

Barthélemy M (2011) Spatial networks. Phys Rep 499(1):1–101

Boeing G (2017) Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. Comput Environ Urban Syst 65:126–139

Brockmann D, Hufnagel L, Geisel T (2006) The scaling laws of human travel. Nature 439:462–465

Bürkner P-C, et al (2017) brms: An r package for bayesian multilevel models using stan. J Stat Softw 80(1):1–28

Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A (2017) Stan: A probabilistic programming language. J Stat Softw 76(1)

Chowell G, Hyman JM, Eubank S, Castillo-Chavez C (2003) Scaling laws for the movement of people between locations in a large city. Phys Rev E 68(6):066102

Clauset A, Shalizi CR, Newman ME (2009) Power-law distributions in empirical data. SIAM Rev 51(4):661–703

Companhia estadual de engenharia de transporte e logistica, Secretaria de estado de transporte, Governo do Estado do Rio de Janeiro (2010) Resultado da pesquisa origem/destino. http://setrerj.org.br/wp-content/uploads/2017/07/175_pdtu.pdf. Accessed 20 Oct

Costa PB, Neto GCM, Bertolde AI (2017) Urban mobility indexes: A brief review of the literature. Transp Res Procedia 25:3645–3655. https://doi.org/10.1016/j.trpro.2017.05.330. World Conference on Transport Research - WCTR 2016 Shanghai. 10-15 July 2016

da F. Costa L, Rodrigues F, Travieso G, Villas Boas P (2007) Characterization of complex networks: A survey of measurements. Adv Phys 56:167–242. https://doi.org/10.1080/00018730601170527

De Montis A, Barthélemy M, Chessa A, Vespignani A (2007) The structure of interurban traffic: a weighted network analysis. Environ Plan B Plan Des 34(5):905–924

Estrada E (2012) Epidemic spreading induced by diversity of agents mobility. Phys Rev E 84:036110

Gabry J, Simpson D, Vehtari A, Betancourt M, Gelman A (2017) Visualization in bayesian workflow. arXiv preprint arXiv:1709.01449. https://arxiv.org/abs/1711.05879

Gonzalez MC, Hidalgo CA, Barabasi A-L (2008) Understanding individual human mobility patterns. Nature 453(7196):779

Guo D, Zhu X, Jin H, Gao P, Andris C (2012) Discovering spatial patterns in origin-destination mobility data. Trans GIS 16(3):411–429

Graser A (2019) Movingpandas: Efficient structures for movement data in python. GIForum 1:54–68

Gronau QF, Singmann H, Wagenmakers E-J (2017) Bridgesampling: an r package for estimating normalizing constants. arXiv preprint arXiv:1710.08162. https://arxiv.org/abs/1711.05879

Jeffreys H (1935) Some tests of significance, treated by the theory of probability. In: Mathematical Proceedings of the Cambridge Philosophical Society. Cambridge University Press Vol. 31. pp 203–222

Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90(430):773–795

Louail T, Lenormand M, Picornell M, Cant'u OG, Herranz R, Frias-Martinez E, Ramasco JJ, Barthelemy M (2015) Uncovering the spatial structure of mobility networks. Nat Commun 6:6007

R Core Team (2018) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Santos LB, Jorge AA, Rossato M, Santos JD, Candido OA, Seron W, de Santana CN (2017) (geo) graphs-complex networks as a shapefile of nodes and a shapefile of edges for different applications. arXiv preprint arXiv:1711.05879. https://arxiv.org/abs/1711.05879

Simini F, González MC, Maritan A, Barabási A-L (2012) A universal model for mobility and migration patterns. Nature 484(7392):96

Soh H, Lim S, Zhang T, Fu X, Lee GKK, Hung TGG, Di P, Prakasam S, Wong L (2010) Weighted complex network analysis of travel routes on the singapore public transportation system. Phys A Stat Mech Appl 389(24):5852–5863

Song C, Koren T, Wang P, Barabasi AL (2010) Modelling the scaling properties of human mobility. Nat Phys 6:818–823

Stan Development Team (2018) RStan: the R interface to Stan. R package version 2.18.2. http://mc-stan.org/. Accessed 20 Oct

Vehtari A, Gelman A, Gabry J (2017) Practical bayesian model evaluation using leave-one-out cross-validation and waic. Stat Comput 27(5):1413–1432

Wang P, Hunter T, Bayen AM, Schechtner K, González MC (2012) Understanding road usage patterns in urban areas. Sci Rep 2:1001

Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks,. Nature 393(6684):409–10

Publisher's Note

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