Identification of city motifs: a method based on modularity and similarity between hierarchical features of urban networks

Journal of Physics: Complexity

OPEN ACCESS

PAPER

Identification of city motifs: a method based on modularity and similarity between hierarchical features of urban networks

Guilherme S Domingues, Eric K Tokuda* and Luciano da F Costa

São Carlos Institute of Physics-DFCM, University of São Paulo, Av. Trabalhador São-carlense 400, São Carlos, S.P., 13566-590, Brazil

* Author to whom correspondence should be addressed.

E-mail: tokudaek@usp.br

Keywords: networks, motifs, cities

RECEIVED
4 August 2022

ACCEPTED FOR PUBLICATION
22 September 2022

PUBLISHED
18 October 2022

Abstract

Several natural and theoretical networks can be broken down into smaller portions, henceforth called neighborhoods. The more frequent of these can then be understood as motifs of the network, being therefore important for better characterizing and understanding of its overall structure. Several developments in network science have relied on this interesting concept, with ample applications in areas including systems biology, computational neuroscience, economy and ecology. The present work aims at reporting a methodology capable of automatically identifying motifs respective to streets networks, i.e. graphs obtained from city plans by considering street junctions and terminations as nodes while the links are defined by the streets. Interesting results are described, including the identification of nine characteristic motifs, which have been obtained by three important considerations: (i) adoption of five hierarchical measurements to locally characterize the neighborhoods of nodes in the streets networks; (ii) adoption of an effective coincidence similarity methodology for translating datasets into networks; and (iii) definition of the motifs in statistical terms by using community finding methodology. The nine identified motifs are characterized and discussed from several perspectives, including their mutual similarity, visualization, histograms of measurements, and geographical adjacency in the original cities. Also presented is the analysis of the effect of the adopted features on the obtained networks as well as a simple supervised learning method capable of assigning reference motifs to cities.

1. Introduction

Through a long period of time, cities unfolded as a means to provide resources to humans, including basic infrastructure as well as access to transportation, food, health, leisure, etc. At the same time, city planning has had to adapt to geographical and environmental constraints, including geographical and climatic characteristics. Each city can thus be understood as a solution to the specific demands and constraints at varying levels of optimization.

Given that the spatial and topological organization of cities are close and directly related to the above observed aspects, their respective study (e.g. [1–9]) provides valuable means not only for better understanding how cities are organized, but also for possibly identifying how specific topological features of a city may be related to urbanistic and transportation aspects. Respectively obtained results and insights can then be shared among municipalities as a possible subsidy for planning and improvement initiatives.

The cities to be analysed are often assumed to be represented as respective complex networks (e.g. [10–13]), which can be achieved by representing streets crossings as nodes, while the streets or avenues between two nodes are taken as a respective link.

While the overall topology of a whole city can be characterized in terms of global topological measurements, including statistical properties of blocks and streets, this type of global characterization cannot account for varying local interconnectivity taking place at different locations in a city. For instance, even if a city is found to have blocks with an average of four sides, there may still be blocks with three, five or more sides. In addition,

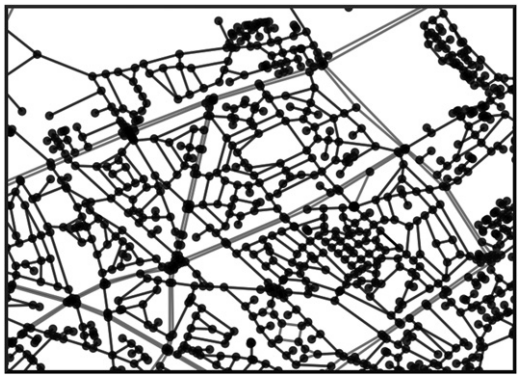

Figure 1. A small portion of Liverpool, UK (derived from Open Street Maps [14]), illustrating the potential diversity neighborhoods around each node, characterized by distinct local topological properties. The identification of recurring neighborhood types, here called *city motifs*, provides subsidies for developing several analysis aimed and better understanding and optimizing cities. The automated identification of city motifs constitutes the main objective of the present work.

some portions of a city can be more or less densely covered by streets. As a consequence, although global characterization of a city organization can provide valuable respective information, it is also of particular interest to perform studies of *local*, mesoscopic topological properties of cities. This can be achieved by focusing on a size-limited neighborhood around each of the points of interest, which are henceforth understood to correspond to each crossing between two or more streets or avenues.

Figure 1 illustrates a small portion of a city (Liverpool, UK) involving several parts characterized by specific local properties, including highly regular square blocks, varying density regions, as well as street dead ends. The identification of the recurring neighborhood types, here called *motifs*, can contribute to developing and applying enhanced approaches not only to the characterization of cities, but also their better understanding, planning and optimization.

In this work, cities are understood as corresponding to *transportation networks*. Each street intersection and dead end are represented as nodes of these networks, while the streets themselves define the respective interconnections between these nodes. Cities are typically composed by several boroughs or districts, which are not specifically taken into account in the present work. Each node i of a streets network will be associated to a respective subgraph, namely its *neighborhood*, corresponding to the nodes that are at successive topological distances, up to a maximum H, from node i. These neighborhoods will be henceforth expressed as $\eta_H(i)$. The topology of these neighborhoods are here characterized in terms of respective *hierarchical measurements*. Coincidence similarity networks are then obtained by taking into account the similarity between the hierarchical measurements.

The concept of motifs in networks (e.g. [15, 16]) has allowed several interesting results in network science, with ample applications in biochemistry, neurobiology, ecology, engineering, economy [17], transportation and infrastructure [18]. Because of the intrinsic topological variations expected to be found in streets networks, the identification of possible motifs needs to be done statistically (e.g. [19, 20]), while taking into account a set of informative local/mesoscopic topological measurements. For instance, it could be expected that highly regular, orthogonal neighborhoods would produce a respective motif, being topologically characterized by nearly constant degree distribution and small clustering coefficient values. Other possible motifs would be related to dead ended streets (degree one), triangular blocks, as well as interfaces between city regions with distinct properties.

A method for translating datasets into respective networks, reported in [21], has been successfully applied to several types of problems including datasets from the UCI database ([22]), enzymes ([23]), as well as neuronal morphology data from the Neuromorpho.org database ([24]), among other works. In the present work, we apply the coincidence similarity methodology to obtain a network expressing the similarity between the considered neighborhoods, from which it becomes possible to identify city motifs as corresponding to respectively detected communities or modules. Since the coincidence similarity concept and respective methodology have been applied before, the main contribution of the present work consists in using the coincidence similarity approach to *automatically identify the motifs in streets networks*.

Consisting of a combination of the widely employed Jaccard similarity index, adapted to real values [25, 26], and the interiority (or overlap) index (e.g. [27]), the coincidence similarity provides a particularly selective and sensitive means for translating datasets, with entries characterized by respective measurements or features, into respective networks or graphs [21, 22, 28].

In addition to its potential for obtaining particularly detailed and modular networks [26, 29], the coincidence similarity methodology can also incorporate a parameter, namely $0 \leqslant \alpha \leqslant 1$, allowing the control of the relative contribution of features with aligned or counter-aligned signs onto the resulting similarity index [21, 26]. For instance, by making $\alpha < 0.5$, the influence of positive joint variations can be relatively attenuated so that a more detailed and modular pattern of interconnectivity can be obtained. The levels of detail of the obtained coincidence similarity networks can be enhanced by selecting the α parameter.

Several works related to the characterization of cities and network motifs are briefly reviewed in section 2: related works.

Motifs should reflect the mesoscopic surroundings of each city network node, and not only its first neighbours or, at the other extreme, features that depend on the whole network. In addition, as city motifs often involve statistical variations (e.g. a roundabout may involve a varying number of radiating streets), for generalization's sake, it becomes necessary to cater for this effect while identifying the basic motifs to be taken into account in subsequent characterization and analysis of the cities. Considering the relevance of topological motifs for the study of diverse cities, it also becomes important to develop means for the respective automated identification.

The present work aims at addressing the following requirements. First, in order to obtain city motifs that reflect local, mesoscopic topological characteristics of streets networks, we characterize the motifs associated to each of the streets network nodes in terms of several respective hierarchical features (e.g. [30, 31]). These features are potentially effective because they are defined in terms of the number of neighborhoods taken around each node of interest, therefore allowing a mesoscopic characterization of the surroundings of that node. Second, in order to allow statistical generalization of the city motifs, we resource to defining each motif in terms of a respectively identified module/cluster obtained while taking into account the hierarchical features. Therefore, the intrinsic variation of the patterns within each cluster will naturally cater for their statistical variation. In addition, the clustering-based approach will also inherently allow unsupervised, automated identification of the motifs from streets networks. More specifically, the proposed methodology employs community finding on coincidence similarity networks obtained by taking into account the similarity between the hierarchical features of each of the streets network nodes.

The proposed methodology starts with a given city being represented in terms of its streets network, in which nodes correspond to crossings between two or more streets and to terminations of streets, while the streets themselves are taken as respective links. A neighborhood with a specific extension H is then obtained around each of the streets network nodes, and respective topological measurements are derived. Given that we are interested in studying varying neighborhood extensions H, node-centered topological measurements become of particular relevance for the characterization of the topological properties within the H-neighborhood around each of the streets network nodes.

The coincidence similarity methodology is here applied between the features of each possible pairwise combination of the H-neighborhood of the streets network nodes, therefore yielding a new network that, though with the same number of nodes, presents links whose strengths correspond to the similarity between the topological features of all possible pairs of H-neighborhoods. This network is henceforth called the neighborhoods network (NN).

As a consequence of the above described approach, two nodes in an NN will be strongly interconnected whenever the neighborhoods associated to those nodes have markedly similar local topological properties. Several interesting information and insights can be potentially obtained from these networks. For instance, a narrow distribution of interconnecting strengths will indicate that most of the node neighborhoods are similar, while wider strength distributions will reveal that the neighborhoods of the specific city of interest are noticeably heterogeneous. In addition, in case the obtained NN presents a well-defined modularity, community detection methods (e.g. [32–35]) can be applied in order to identify the main modules, each of which will indicate a mesoscopic region of the city, presenting particular topological properties.

Each of the communities identified in NNs will constitute a candidate for a *motif*. Therefore, in addition to studying the similarity between the topological properties of node neighborhoods across varying topological scales, the present work also aims at investigating if the motifs identified among two or more cities can be inter-related. For instance, one such module recurring between several cities can be understood as a possible shared motif. In order to develop these studies in a systematic manner, we identify the motifs respectively to a given city taken as a reference, and then compare these motifs with those of two other cities.

The obtained results revealed a surprising level of consistency and stability of the nine identified motifs, which have specific visual, topological, cross-similarity, and adjacency properties, all of which having being quantified in an objective manner in the present work.

In order to complement the discussion of the obtained results, we also performed an analysis of the influence of the adopted five hierarchical measurements on the resulting NNs, which is developed by using an approach that is also based on the coincidence similarity methodology [29].

While the above described methodology for automatically identifying the city motifs is unsupervised, the observed generality of the identified motifs motivated the proposal of a simple supervised method for assigning motifs to cities with similar characteristics. This method, which is also described and illustrated in the present work, involves using the instances of motifs identified for the reference city(ies) for assigning motif types to neighborhoods of other cities. This method was shown to perform well for the case of a two other cities.

This work starts by providing a non-exhaustive review of related works and follows by presenting the data, basic concepts and methods, including hierarchical measurements, the coincidence similarity methodology, and motifs identification. The results are then presented and discussed respectively to features interrelationship, motifs characterization, and application. The effect of the adopted features on the respectively obtained networks is also addressed, and a simple procedure for assigning the nine types of identified motifs to other cities is also presented.

2. Related works

This section revises some of the works related to the main aspects and concepts developed in the current article.

The increase of the number of cities and their population along the last century was accompanied by the development of systematic approaches aimed at discussing and better understanding urban and city spaces (e.g. [36–38]). Interesting works continued along the 60s and onward, including more quantitative approaches aimed at organizing urban scenery into several types of constituent elements [39], as well as subdividing overall cities into smaller spaces based on criteria including unimpeded visibility and navigation [40].

The comparison of networks, a topic of significant interest in network science, can be implemented based on different criteria, such as the network type, the degree distribution, and the presence of communities. In [41] the similarity between the internet backbone and air transportation network is addressed by considering the hierarchy and pattern of connections among world cities.

In [42], four different standard similarity metrics (common neighbors, Jaccard, resource allocation and Leicht–Holme–Newman) are used to evaluate node similarity and reconstruct propagation networks based on the epidemics spreading dynamics. It is observed that temporal information can play an important role on the reconstruction. In [43], different samples of the urban networks of 20 different world cities are compared with basis on a set of measurements of spatial graphs, namely the meshedness, the number of short cycles of sizes three, four and five edges, the global efficiently and the cost. In particular, similarity is estimated between self-organized and planned cities.

The characterization of networks can be performed locally. In the context of spatial networks, in [44] the authors propose defining neighborhoods based on social ties as well as on physical distance. They propose four alternative manners of doing, which are applied to data of students from North Carolina schools. In [45], in the scenario where connections are susceptible to noise, the authors consider a neighborhood scheme based on shared neighbors.

Networks can also be characterized in terms of the presence of pre-defined patterns, commonly known as *motifs* (e.g. [15, 46, 47]). For instance, the distribution of triangles along a network has been used as an indicative of the tendency of the network to form clusters [48]. The distribution of motifs has also been studied respectively to its effects on specific types of dynamics on networks [49–52].

There is a variety of reported applications of network motifs. In [53, 54], the authors studied network connectivity in terms of specific types of motifs: vertices connected in a sequential way such that the inner vertices have degree equal to two. They observed highly different distributions of these motifs between real-world and artificial networks. In [55], the authors analyzed the distribution of motifs in directed networks, which they called sequential motifs. They propose a connection between sequential motifs and higher order networks, and analyze data from passenger trips through the airport network in the United States and also article navigation in Wikipedia.

Motifs have also been used to analyze data from mobile phone communication networks and related data, which can be used to study communication and human mobility patterns [56, 57]. Mobility patterns from tourists have been studied in [58]. The authors considered in their analysis temporal information, such as *when* the places were visited and semantic information (the attractions). The temporal travel motifs in this case revealed popular duration of stays in each attraction while the topological motifs the frequent travel sequences among the attractions.

Motifs have also been adopted in the study of urban networks [59, 60]. In [59], the authors analyzed how socioeconomic aspects of a city—such as mobility, market and population—can be associated to city street network patterns. They considered Greek cities and observed three distinct patterns: considering the central nodes, ring nodes and the mixture of the two. In [60] the authors studied the frequency of motifs in public transportation networks in large Chinese cities. One of the main findings regards the distribution of

Table 1. Concepts and measurements used in this work, and respective acronyms/symbols.

Measurement	Symbol
Neighborhoods network	NN
Neighborhood motifs	NM
Hierarchical ring centered around the node i	$R_h(i)$
Neighbourhood of node i for H hierarchical levels	$\eta_H(i)$
Minimum number of nodes of the detected communities	N_c
Considered maximum hierarchical level	H
Coincidence similarity index	$\mathcal{C}$
Jaccard similarity index	$\mathcal{J}$
Interiority index	$\mathcal{I}$
Hierarchical degree	hd
Hierarchical clustering coefficient	hc
Convergence ratio	cr
Hierarchical number of nodes	hn
Hierarchical number of edges	he

Table 2. The types of networks used in the current work, as well as their identification in terms of respective nodes and links. The table also incorporates references to the sections where each type of network is presented at more length.

Network	Nodes	Links
Streets network (section 3.1)	Streets crossings and dead-ends.	Interconnecting streets.
Neighborhoods network (section 3.3)	Neighborhood respective to each node in the streets network.	Coincidence similarity between the hierarchical features of the neighborhoods.
Motifs network (section 5.1)	Motif type assigned to the respective nodes in the streets network.	Same as in the streets network.
Motif types network (section 5.2)	Each identified motif.	Coincidence similarity between densities of the hierarchical features of the identified motifs.
Features network (section 7)	NN obtained for a specific feature combination.	Coincidence between the NNs.

certain three-node motifs, which seemed to be associated with the efficiency of the transportation system and robustness to failures.

3. Materials and methods

The present section describes the data, basic concepts, and methods adopted in the current work in order to automatically identify city motifs. These include the cities dataset, the hierarchical measurements, as well as the coincidence similarity-based approach for translating datasets into respective complex networks. The main concepts and measurements adopted in the present work, as well as their respective acronyms/symbols, are summarized in table 1.

Table 2 presents the types of networks employed in the present work, as well as their identification in terms of respective nodes and links. Observe that, given that this table provides only a summary description of the networks, references have been provided to the sections where the concepts are respectively presented at more length.

Figure 2 presents the main data and methods involved in the suggested approach to automated city motif identification. In the present work, the implementation of the methods adopted *Python* [61] and *igraph* [62].

3.1. Streets networks

The city data was obtained from the OpenStreetMaps [14] database. The original data contained a preliminary network delimited by a square area encompassing the geographical coordinates of the city. Additional pre-processing was required in order to obtain the streets networks. More specifically, chains of nodes [53] were identified and replaced by a single link. The identification of these chains involved finding all nodes with degree 2 while checking their adjacent.

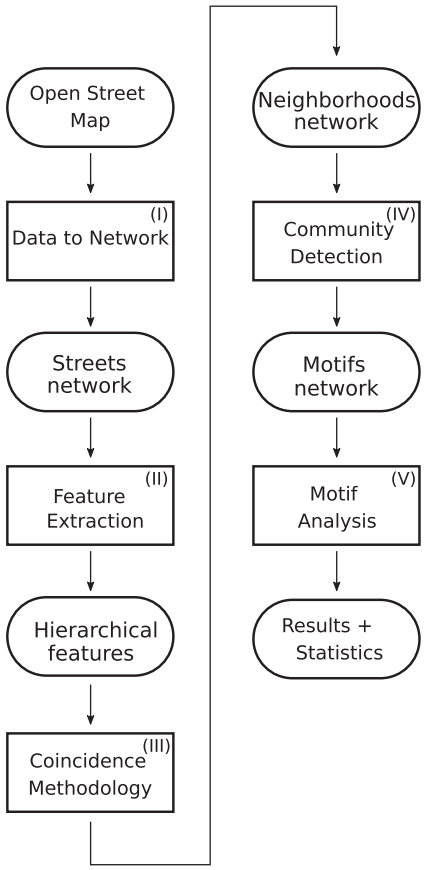

Figure 2. The main steps of the methodology for city motif identification described in the present work. Specific cities are obtained from the Open Street Map resource and transformed into respective networks (I). Hierarchical measurements are calculated (II) for each neighborhood of each node, yielding respective feature vectors. The coincidence similarity methodology is the applied (III), resulting in the respective similarity network, which has its motifs detected as corresponding to the identified communities (IV). The obtained motifs can then be analyzed in several manners (V).

After obtaining the streets networks, figure 2(I), the hierarchical measurements described in the following section were calculated (II) for each node and then used for identification of possible motifs characterizing the cities topology.

3.2. Hierarchical measurements

In this work, we considered a set of five hierarchical measurements (e.g. [30, 31, 63]) for characterizing the topology of the neighborhood around each node, which are taken as features in the coincidence methodology (section 3.3). The adoption of a H-neighborhood around the reference node i implies the hierarchical measurements to be calculated relatively to the *hierarchical ring* $R_h(i)$ with $h = H - 1$.

Figure 3 illustrates the concept of the first and second neighborhoods ($h = 1$ and $h = 2$) defined by a reference node (V), also including the calculation of the respective hierarchical measurements.

Hierarchical degree (hd). The hierarchical degree $hd_h(i)$ of node i at distance h is defined as the number of edges between the hierarchical rings $R_h(i)$ and $R_{h+1}(i)$

Hierarchical clustering coefficient (hc). The hierarchical clustering coefficient of node i at distance h is defined as

$$hc_h(i) = 2\frac{he_h(i)}{hn_h(i)(hn_h(i) - 1)} \tag{1}$$

where $e_h(i)$ is the number of edges connecting nodes of the hierarchical ring $R_h(i)$ and $hn_h(i)$ is the number of nodes in that hierarchical level.

Convergence ratio (cr). The convergence ratio of node i at hierarchical level h is defined as the ratio between $hd_h(i)$ and the number of nodes in the next hierarchical level, i.e.

$$cr_h(i) = \frac{hd_h(i)}{hn_{h+1}(i)} \tag{2}$$

Hierarchical number of nodes (hn). The hierarchical number of nodes $hn_h(i)$ in the hierarchical ring $R_h(i)$ is defined as the number of nodes inside $R_h(i)$, or the size of $R_h(i)$.

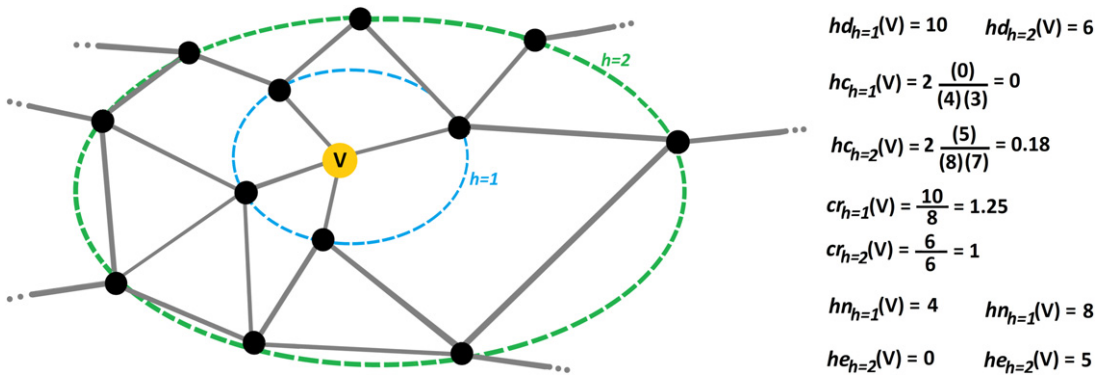

Figure 3. Illustration of the determination of the hierarchical measurements employed in this work for $h = 1$ and $h = 2$. The measurements are determined respectively to the reference node (V), which defines the two neighborhood levels $h = 1$ and $h = 2$. The respectively obtained hierarchical measurements are presented at the right-hand side of the figure.

Hierarchical number of edges (*he*). The hierarchical number of edges $he_h(i)$ among the nodes in the hierarchical level $R_h(i)$ is defined as the number of edges $he_h(i)$ between the nodes of $R_h(i)$ without considering edges connecting nodes of $R_{h+1}(i)$ or $R_{h-1}(i)$.

3.3. The coincidence similarity methodology

Several similarity indices have been considered respectively to diverse types of data and applications (e.g. [26–28, 64–66]), including cosine similarity, correlation, and the Jaccard index.

Though the *Jaccard similarity index* (e.g. [26–28, 67]) has been extensively employed as a means of quantifying the *similarity* between two non-empty sets, these applications have been mostly limited to categorical or binary data. In addition, the Jaccard index has been shown not to be able to take into account how much the two compared sets are mutually internal one another [25]. This motivated the consideration of the *coincidence similarity index* [25], corresponding to the product of the Jaccard index and the respective *interiority* or *overlap* index (e.g. [27]).

By extending multisets (e.g. [68–73]) to real-valued data [74], it has been possible to derive a respective coincidence similarity index that can be employed as a means to quantify the similarity between two non-zero, real-valued vectors or even functions. In addition, it has been shown that the Jaccard index can be decomposed into two major terms, one corresponding to the positive pairwise alignment of the signs of the compared values, and another to the anti-aligned pairs. The linear combination of these two terms, respectively weighted by α and $1 - \alpha$, yields the parametric coincidence similarity index expressed as:

$$\mathcal{C}_R(\vec{f}, \vec{g}, \alpha) = \mathcal{C}_R(\vec{g}, \vec{f}, \alpha) = \mathcal{I}_R(\vec{f}, \vec{g})\, \mathcal{J}_R(\vec{f}, \vec{g}, \alpha), \tag{3}$$

where:

$$\mathcal{J}_R(\vec{f}, \vec{g}, \alpha) = \frac{\sum_i \alpha |s_{f_i} + s_{g_i}| \min\{|f_i|, |g_i|\} - (1-\alpha)|s_{f_i} - s_{g_i}| \min\{|f_i|, |g_i|\}}{\sum_i \max\{|f_i|, |g_i|\}} \tag{4}$$

and:

$$\mathcal{I}_R(\vec{f}, \vec{g}) = \frac{\sum_i \min\{|f_i|, |g_i|\}}{\min\{\sum_i |f_i|, \sum_i |g_i|\}} \tag{5}$$

We also have that $-2(1-\alpha) \leqslant \mathcal{J}_R(\vec{f}, \vec{g}, \alpha) \leqslant 2\alpha$.

The parameter α allows an effective control of how the aligned and anti-aligned pairwise measurements are linearly combined into the resulting overall coincidence similarity value.

In particular, when $\alpha = 0.5$, the above index becomes identical to the product between the real-valued, interiority index and the *parameterless* Jaccard index, i.e.:

$$\mathcal{C}_R(\vec{f}, \vec{g}, \alpha = 0.5) = \mathcal{I}_R(\vec{f}, \vec{g})\, \mathcal{J}_R(\vec{f}, \vec{g}), \tag{6}$$

where:

$$\mathcal{J}_R(\vec{f}, \vec{g}) = \frac{\sum_i s_{f_i} s_{g_i} \min\{|f_i|, |g_i|\}}{\sum_i \max\{|f_i|, |g_i|\}} \tag{7}$$

with $-1 \leqslant \mathcal{J}_R(\vec{f}, \vec{g}, \alpha) \leqslant 1$ (see also [66]).

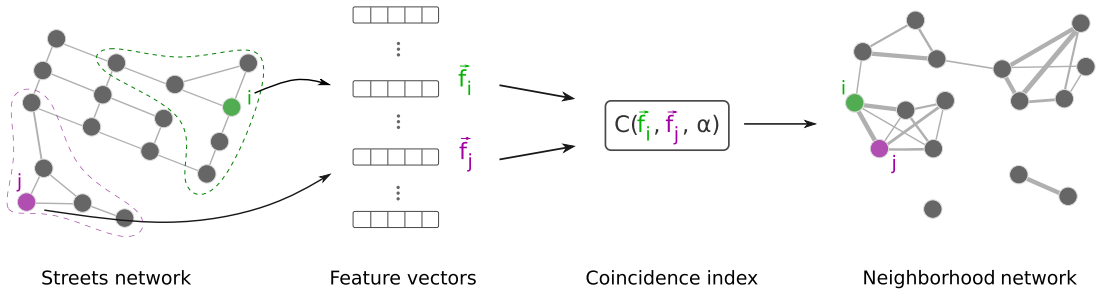

Figure 4. Diagram illustrating how a streets network is translated into the respective coincidence similarity network. The neighborhood around each node (e.g. i and j) is identified and the respective hierarchical measurements are calculated. The example in this figure considers $h = 1$, therefore including only first neighbors. The coincidence similarities between each resulting feature vectors are calculated and taken as the weights between the nodes in the coincidence similarity network. Observe that both the streets and coincidence similarity network have the same number of nodes.

The real-valued coincidence similarity index has been applied [21, 22] to translate datasets, with each data element characterized in terms of M measurements or *features*, into respective graphs or networks whose interconnecting weights between each two nodes correspond to the respective coincidence similarity values between the features of those two nodes. These coincidence similarity networks can be then thresholded by T to yield networks with weights limited to 0 and 1. However, it is also possible to preserve the values of the coincidence similarities above T while making assigning zero to the values smaller that T.

It has been shown [21, 22, 29, 75] that the interconnectivity of the resulting coincidence similarity networks strongly depends on the values of α, in the sense that higher values of α will imply more intensely interconnected networks. However, these networks may become too interconnected, to the point that the respective interconnection details and modularity are severely blurred and cluttered. This is precisely where reductions of the parameter α can contribute to limiting the overall connectivity, enhancing the level of details and modularity of the obtained networks [21, 29].

In this manner, the coincidence similarity methodology for quantifying similarity between real-valued vectors and functions (as well as other types of data) incorporates several interesting features derived from the Jaccard and interiority indices combined with the important control of the resulting overall interconnectivity by varying the parameters α.

In the current work, for each city, as illustrated in figure 4, the neighborhood around each node i is identified and the respective hierarchical measurements obtained and organized into a respective feature vector as follows:

$$\vec{f}_i = [hd(i), hc(i), cr(i), hn(i), he(i)] \tag{8}$$

The obtained features are then supplied to the above described coincidence similarity methodology in order to deriving respective coincidence similarity networks for each individual city, figure 2(III). Observe that neighborhoods defined by each node of the streets network therefore becomes associated to a single node in the NN. As a consequence, two adjacent nodes in the latter will necessarily imply some overlap between their respective neighborhoods in the streets network.

Similarity is intrinsically related to connectivity (e.g. [76]), providing a means for obtaining complex networks (e.g. [77–81]). The coincidence similarity, which has been applied as a means of translating datasets into respective complex networks [21, 22] is adopted henceforth in the present work. More specifically, after being standardized, the features describing the dataset are taken into account while calculating the coincidence similarity between every pair of data elements, yielding a coincidence similarity network in which each node corresponds to a data element while the links are determined by the respective pairwise coincidence similarity indices.

The standardization (e.g. [82]) of each of the adopted features x, respectively to data elements x_i, can be implemented as follows:

$$\tilde{x}_i = \frac{x_i - \mu_x}{\sigma_x}, \tag{9}$$

where μ_x and σ_x are the average and standard deviation of feature x taken along the whole considered dataset.

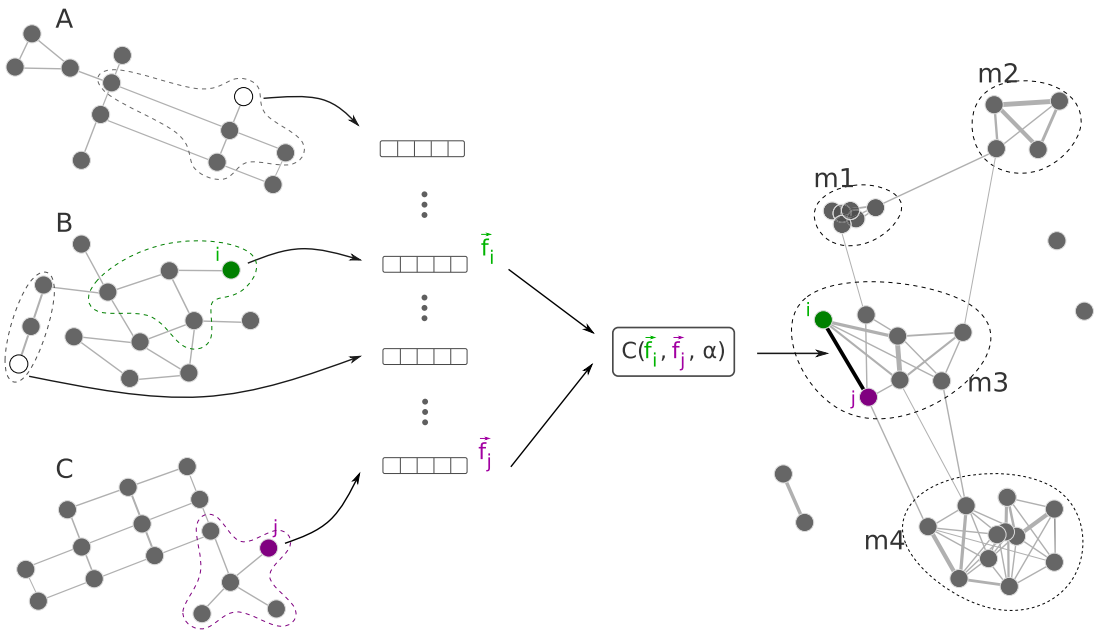

Figure 5. Given three generic cities (*A*, *B*, and *C*), all possible neighborhoods with hierarchical level *H* are identified, and the respective standardized hierarchical measurements are organized as feature vectors, four of which are shown in the figure. The coincidence similarity is then calculated between each pair of these feature vectors. By blanking these values by *T*, the combined NN CN can then be obtained. Each detected community in the latter network is then understood as a NM.

4. Motifs identification

In this section, we describe the proposed methodology for motif identification, please refer to figure 2(IV). Basically, a community finding approach is applied on the previously obtained NN, and the identified modules are understood to define the reference city motifs.

As observed in the introduction of the present work, given the diversity of interconnections typically observed in streets networks, the neighborhood motifs (NMs) to be considered here need to have a statistical nature, in the sense that each given motif type can be allowed to undergo small topological variations.

The basic hypothesis of our approach regarding the NMs is that they have some level of generality and recurrence not only within a given city, but also across other cities. Thus, the problem of motif identification as addressed in the present work can be stated as: given a city, or a set of cities, and respective neighborhoods characterized by associated topological features, to find sets of these neighborhoods that are strongly topologically similar one another while being distinct to the other neighborhoods.

The resource to be applied in order to find these groups of similar neighborhoods, which will be taken as the NMs, consists of the application of the coincidence similarity methodology [21, 22, 26, 28]. More specifically, we estimate the coincidence similarity between each pair of neighborhoods obtained from all the adopted cities. A single network is then derived from each neighborhood while the coincidence similarity between each pair of nodes corresponds to the respective link weight. The so obtained structure is henceforth called the *combined network*.

In order to simplify the resulting network, its links with coincidence similarity values smaller than a given reference T are subsequently ignored, therefore yielding a weighted network (a binary network would be otherwise obtained by standard thresholding). This operation is henceforth referred to as *blanking*.

The NMs can then be identified as being associated to the main detected communities, figure 2(IV), having at least a minimum number of nodes N_c. The nodes resulting in smaller communities are henceforth referred as *unassigned nodes*. The community detection is performed independently in the combined network and also in each considered city NN as a means to identify the correspondence among the detected communities across cities. This is implemented for each city at a time. For each community m in a given city, it is verified which among the communities in the combined network contains the largest number of the nodes in m, which is taken as the corresponding community.

The suggested methodology for identifying the NMs is illustrated in figure 5 respectively to three generic cities *A*, *B*, and *C* taken as reference. The adopted five hierarchical features are standardized (e.g. [83]) along all neighborhoods of a considered city before coincidence similarity estimation.

The motifs obtained by the suggested methodology can then be analyzed in several manners, figure 2(V).

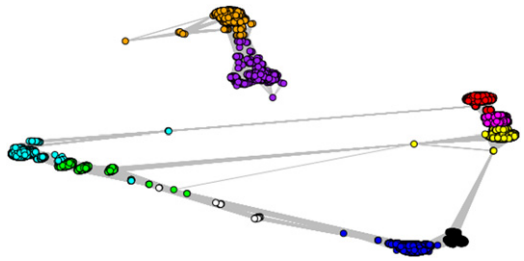

Figure 6. The NN obtained for the city of São Carlos. Each node i corresponds to the respective $\eta_2(i)$ neighborhood, while the link values are the respective pairwise coincidence similarities assuming $T = 0.08$ and $\alpha = 0.1$, yielding good separation of the nodes into clusters. The colors indicate the nine identified motifs (see also figure 9), with the unassigned neighborhoods, considering $N_c = 200$, being represented in white. This network has been visualized by using the Fruchterman–Reingold [84] approach.

Table 3. The identification and number of neighborhoods in São Carlos motifs network corresponding to each of the nine identified motifs.

Motif identif.	Motif color	No. of nodes	Rel. freq.
$m1$	Blue	1732	19.345%
$m2$	Yellow	1426	15.927%
$m3$	Orange	1377	14.554%
$m4$	Cyan	1135	12.677%
$m5$	Magenta	1103	12.320%
$m6$	Red	917	10.242%
$m7$	Green	649	7.249%
$m8$	Purple	436	4.870%
$m9$	Black	233	2.602%
Unassigned	White	19	0.212%

5. Results and discussion

This section presents the main obtained results regarding the characterization of the neighborhoods, the estimation of the NNs, the identification of the city motifs, and respective analysis.

We considered the Brazilian city of São Carlos (SP), with population between approximately 250 000 inhabitants, whose streets were represented by complex networks as described in section 3.1, with each node representing streets crossing or termination, while the corresponding street as link between that pair of nodes. The obtained network has the $N = 8953$ (São Carlos) nodes.

5.1. Neighborhoods and motifs networks

As a first step in our approach, we calculated the five hierarchical measurements for each neighborhood $\eta_2(i)$ ($H = 2$) associated to each node i, which were used to characterize locally the topological properties of the streets network.

The NNs obtained respectively to the city of São Carlos are presented in figure 6. These visualizations were obtained by using the Fruchterman–Reingold [84] method.

Of particular interest is the relatively high modularity of all obtained NNs, which was mostly allowed by the strict similarity quantification implemented by the coincidence similarity methodology, as well as the mutual coherence between the neighborhoods. The network in figure 6 provides the basis for identifying the city motifs, which was done by detecting the respective communities using the Infomap methodology (e.g. [85]). Each of the nine identified community with at least $N_c = 200$ nodes (neighborhoods) were understood as identified motifs, with the smaller communities remaining unassigned.

Table 3 presents the number of motifs of each type, from 1 to 9, identified in the São Carlos motifs network. Interestingly, the relative frequency of the occurrence of motifs decreases in an almost linear manner. Observe that only 19 nodes resulted unassigned.

Once the motif types have been identified, all the corresponding nodes in the respective NNs can be labeled with the respective motif type, resulting in a *motifs network*. Observe that the latter network is identical to the original streets network of the city being considered, except for the labelling of the nodes with the respectively obtained motifs.

Figure 7 presents the motifs network obtained for the city of São Carlos by using the described methodology.

Figure 7. The geographical distribution of motif types in the city of São Carlos (SP, Brazil).

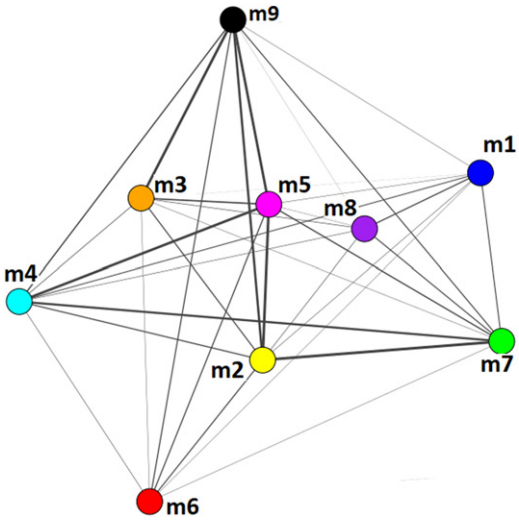

Figure 8. Coincidence similarity network obtained for the nine identified city motifs, referred to as *motif types network*. Each node corresponds to one of the motifs, while the width of the links is proportional to the respective pairwise coincidence similarity between the densities of the five features adopted for characterizing each neighborhood $\eta_2(i)$, assuming no blanking ($T = 0$) and $\alpha = 0.1$. This network was visualized by using the distributed recursive layout algorithm. [86].

5.2. Motif types network

It is interesting to construct a network where each node corresponds to one of the identified motifs, while the width of the link between two motifs i and j reflects the value of the respective coincidence similarity between the respective features densities. These densities, shown in the first column in figures 17 and 18 of appendix A, correspond to the histograms of relative frequency of each of the hierarchical measurements characterizing each of the nine motifs obtained for the city of São Carlos. The so obtained *motif types network* is presented in figure 8 illustrates respectively to the nine identified motifs.

The node with the largest strength (sum of coincidence similarities respective to its links) in figure 8 corresponds to the city motif $m4$, which can thus be understood as being more mutually similar to several of the remainder motifs. Also worth noticing is the relatively stronger relationship between motifs $m2$–$m5$–$m9$, $m2$–$m7$, $m4$–$m7$, $m9$–$m3$, as well as $m4$ and $m5$, meaning that they are intrinsically more similar one another.

Motifs $m3$, $m6$ and $m8$ have the smallest coincidence similarity strengths, being therefore relatively more distinct to the remainder motifs.

5.3. Motifs characterization

In this section we discuss the nine identified city motifs in terms of three particularly important respective perspectives: (i) visual appearance; (ii) relative frequency histograms (densities) of features; and (iii) geographical adjacency between motifs.

Figure 9 depicts five samples of each of the nine identified motifs. The reference node has been shown as circles (red), while its first and second neighborhoods are shown as squares (green) and triangles (blue), respectively.

Of particular relevance is the high level of similarity observed among samples from the same type of motif. In addition, despite intrinsic statistical variations, the samples from motifs of different types resulted with marked topological differences. For instance, motifs $m1$, $m8$ and $m9$ are characterized by reference nodes with degrees that are, with just one exception, all equal to four, in contrast, for instance, to the degrees one or two observed for the reference nodes of $m6$. Motifs $m2$, $m3$, $m4$, $m5$ and $m7$ have reference nodes with degree equal to three. The distinction between these motifs having the same reference node degree is accounted for by the other considered hierarchical features, which cannot be straightforwardly discerned by visual analysis.

In order to characterize the identified motifs in a more comprehensive manner, it is necessary to resource to the relative frequency histograms (densities) of the adopted five hierarchical measurements obtained for each of the nine identified motifs respectively to the São Carlos motifs network (please see first column of figures 17 and 18 in appendix A).

Figure 10 presents the relative frequency histograms (densities) of the hierarchical node degree $hd_{h=1}$ for the São Carlos motifs network for each of the nine identified motifs. Interestingly, most of the obtained histograms are mutually distinct, except for the cases $m4/m5$ and $m2/m7$. However, these two pairs of motifs that have similar hierarchical degree have been verified to differ regarding the distribution of the other adopted measurements.

These histograms provide an objective characterization of the distribution of the feature $hd_{h=1}$ with in the identified motifs, therefore complementing the preliminary visual analysis.

A more comprehensive characterization of the identified motifs taking into account not only the distributions of all the five adopted hierarchical features, but also additional aspects including the motif shapes (figure 9) and the adjacency between motifs, is presented in section 6.

Another important property of the city motifs concerns their geographical relationships in the original streets network (see examples in figure 11).

Indeed, it could be expected that some types of motifs tend to appear adjacent one another as one moves from more uniform to less uniform, or from more central to more periphery regions of a city. In order to verify this possibility in a more systematic and quantitative manner, figure 12 depicts the histograms of city motif adjacencies for the city of São Carlos.

The understanding of the spatial relationships between motif types can be complemented by visualizing their distribution within the considered city, as shown in figure 7.

It is interesting to keep in mind that, given a NN, where each node is associated to a respective neighborhood, the fact that two nodes i and j are adjacent implies overlap between their respective neighborhoods. As a consequence, two adjacent neighborhoods tend to have similar local topological properties. That is one of the reasons why each of the motif types tends to present specific adjacency preferences.

Regarding the predominant adjacencies observed for the nine identified motifs, we have that each of motif types $m1$ and from $m3$ to $m5$ tend to be most adjacent to itself. This transitive property is of particular importance as it leads to patches of neighborhoods sharing the same motif type. For instance, motif $m1$ tends to form extensive regions of almost perfect orthogonality, and therefore regularity, in cities. Interestingly, this type of motif tends to be adjacent also to $m2$ or $m7$, frequently appearing at the border of the regular patches corresponding to $m1$. Motif $m6$ is mostly adjacent to motif type $m5$. Given that $m6$ often corresponds to streets dead-ends, we also have that the motif type $m5$ also tends to occur near the geographical borders between the communities within cities. In addition, motif $m9$ is adjacent to $m2$, with the former tending to appear surrounded by the latter type. Also, motif $m8$ is mostly adjacent to $m3$, both of them frequently composing triangular topologies in the city.

It should be observed that having similar topological properties contributes to making a pair of motif types to appear geographically adjacent, but this is not always the case. Take, for instance, motif types $m2$ and $m7$. As indicated from figure 12, they are not adjacent. At the same time, as illustrated by the strong respective connection in figure 8), they are significantly similar one another. Thus, though topologically similar, these two motif types are highly unlikely to be found geographically adjacent in the considered city.

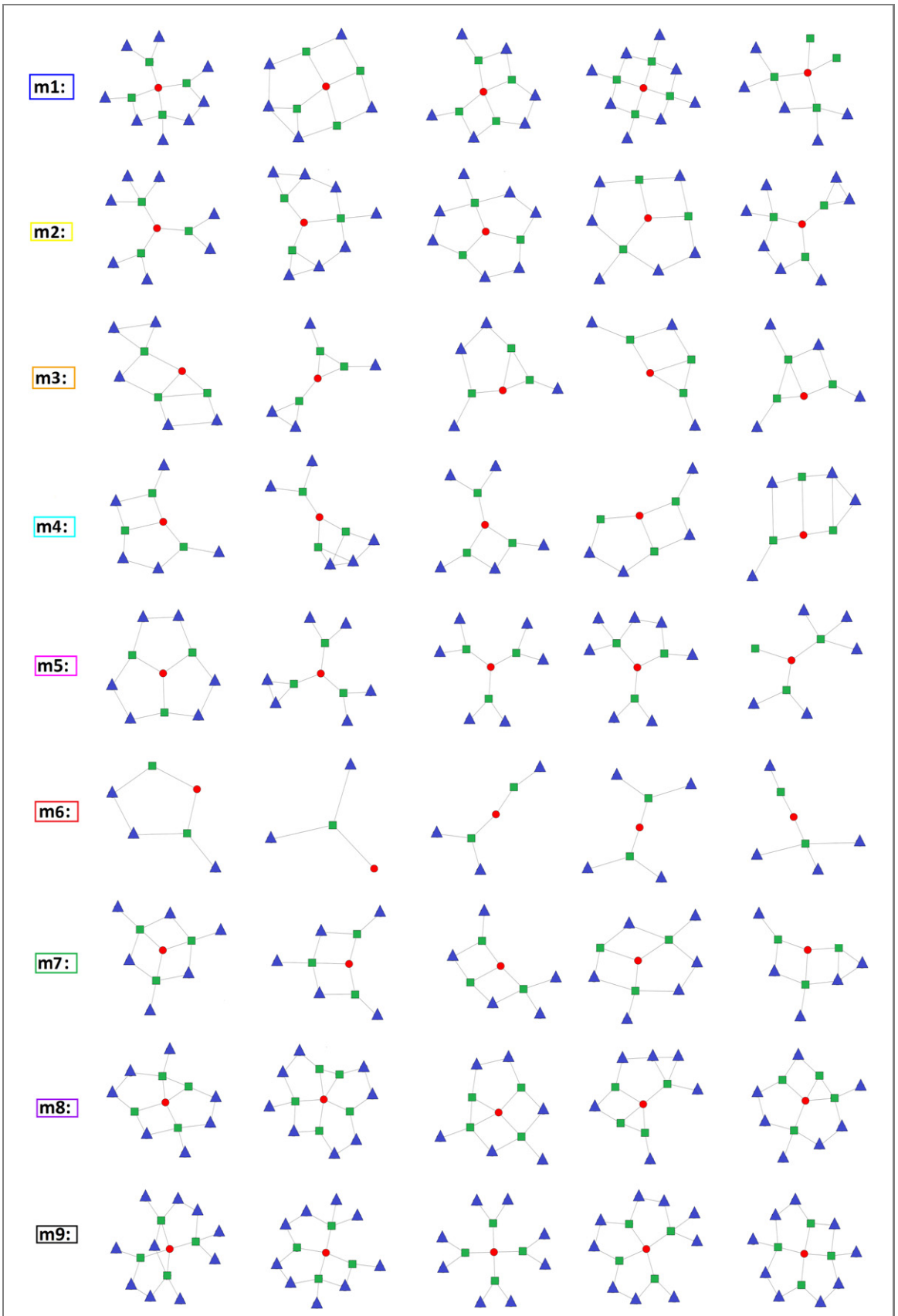

Figure 9. Examples of the nine motifs identified for the city of São Carlos, with the reference nodes shown as circles, while the respective first and second neighborhoods are depicted as squares and triangles, respectively. From top to bottom, the types of motifs are presented in decreasing order of respective frequency. As expected, small variations can be observed among motifs of the same type, which justifies the adopted statistical approach for motif identification. The networks shown in this figure were visualized by using the Kamada–Kawai method [87].

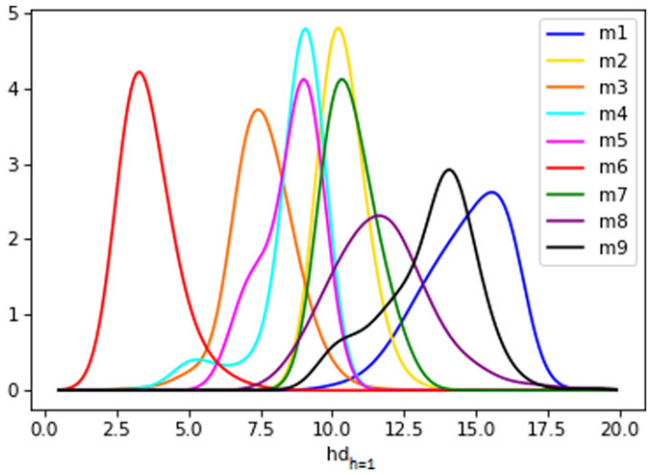

Figure 10. Relative frequency histograms of $hd_{h=1}$ obtained for the nine identified city motifs respectively to the city of São Carlos. Interestingly, distinct histograms have been obtained for most of the motifs, except for $m4$ being similar to $m5$ and $m2$ being similar to $m7$.

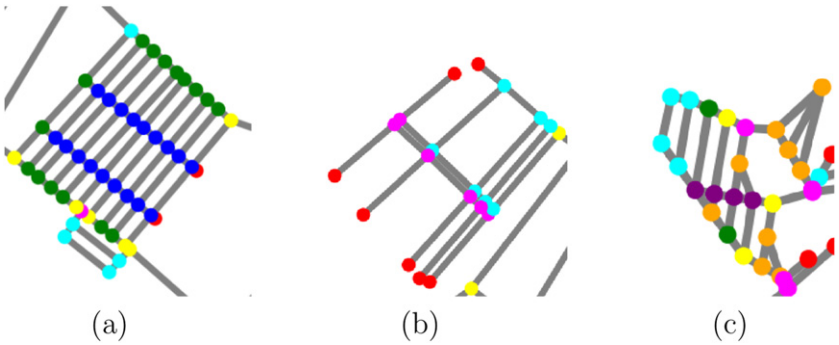

Figure 11. Visualization of some instances of the identified motifs in the streets networks of the city of São Carlos. (a) $m1$ (blue) nodes appear as representing highly regular and rectangular blocks, with predominance of $m7$ (green) on their borders. (b) $m6$ (red) are end nodes, generally linked to the network through $m4$ (cyan) and $m5$ (magenta). (c) The occurrence of $m3$ (orange) and $m8$ (purple), as part of triangular blocks.

6. The nine identified motifs

By referring to figures 8–10 (complemented by figures 17 and 18 in appendix A), and figure 12, we can now typify each of the nine identified motifs as follows:

m1, blue. As it can be observed in figure 9, this motif type tends to have its reference node with degree 4. In addition, we have from figures 10, 17, and 18 that this motif is characterized (together with $m8$ and $m9$) by the highest hierarchical degree and hierarchical number of nodes. As it can be discerned from figure 9 as well as the geographical distributions (figure 7), this motif is intrinsically associated to highly regular patches of square blocks. This motif type also tends to appear adjacent to itself as well as to $m2$ and $m7$.

m2, yellow. This motif type, whose reference nodes tend to have degree 3, being also similar to $m5$ and $m7$. However, $m2$ and $m5$ have the hd histograms at different positions. At the same time, the cr densities are at different positions in $m2$ and $m7$, and the latter motif has a wider dispersion of hd. Motif $m2$ tends to appear adjacent to itself, $m1$, and $m5$. This motif, which tends to have a relatively large number of second neighbors, often corresponds to irregular neighborhoods inside the patches of $m1$ motifs.

m3, orange. This motif type tends to have reference node characterized by node degree equal to 3, as well as by relatively low hd values. As it can be readily inferred from figure 7, this motif type is characterized by having its reference node as corresponding to one of the vertices of a triangular block. Motif type $m5$ tends to be adjacent to itself as well as to $m8$.

m4, cyan. This motif, which often has reference node with degree 3, is similar to $m5$, but it tends to have cr larger than that of $m5$. Interestingly, this motif appears adjacent mostly to itself, and then with $m2$.

m5, magenta. This motif, with reference node tending to have degree 3, is similar to motifs $m2$, $m4$, and $m9$. However, the hd histograms are different among these three motifs. In particular, $m9$ tends to have larger hn

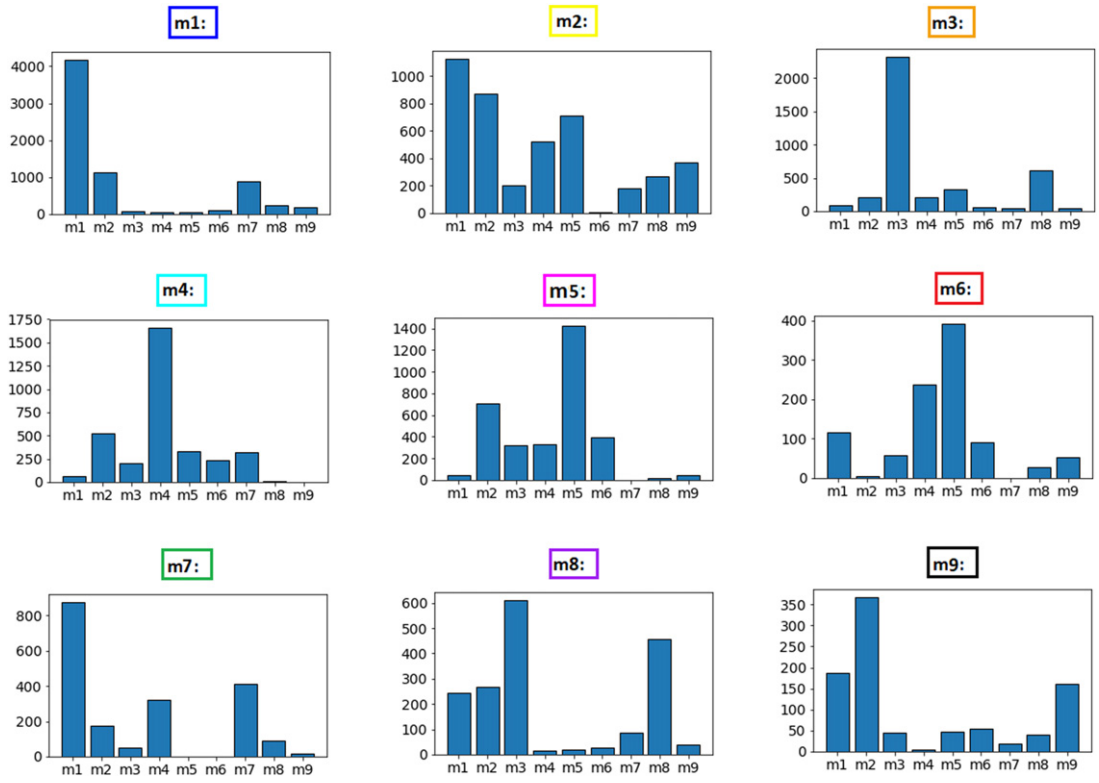

Figure 12. Histograms of the geographical adjacency between the nine identified motifs respectively to the city of São Carlos. These results corroborate the tendency of certain motif types presenting higher probability of being near one another.

than $m5$, and $m4$ has cr larger than $m5$. This motif tends to be adjacent to itself and to $m2$. Generally speaking, motifs $m2$, $m4$, and $m5$ are typically found at the interfaces or transitions between the more highly regular patches of $m1$ motifs.

m6, red. This motif type tends to have reference node with degree 1 or 2. In addition, we have from figures 10, 17, and 18 that this motif has the smallest hierarchical degree (hd) and hierarchical number of nodes (hn). Unlike many other motifs, this motif does not tend to appear adjacent to itself, being predominantly adjacency with $m4$. Figure 7 indicates that this motif type tends to correspond to street dead-ends, being therefore expected to appear mostly near the city borders.

m7, green. The reference node associated to this type of motif tends to have node degree equal to 3. Its hierarchical measurements are mostly similar to those of $m2$, though presenting cr larger. This motif type tends to be predominantly adjacent to $m1$ and itself. Figure 7 indicates that this type of motif tends to correspond to borders of the highly regular patches of $m1$ motifs.

m8, purple. This is the second least frequently observed type of motif, with only 436 occurrences in the city of São Carlos. It is most similar to $m1$, but the latter has larger hd. As with motif type $m3$, the reference node of $m8$ tends to correspond to one of the vertices of a triangular block. Motif type $m8$ tends to be adjacent to itself and $m3$.

m9, black. The reference node of this motif type tends to have degree equal to 4. It is most similar to $m2$ and $m5$. However, $m9$ tends to have he distinct from $m3$, and hd distinct from $m5$. This type of motif tends to be adjacent to itself, $m1$, and $m2$.

7. Analysis of the influence of the adopted features

Almost invariably, the results obtained from comparisons and classifications depend substantially on the adopted measurements or features used to characterize each data element. Even though the five selected features (see section 3.2) allowed remarkable results regarding the identification of city motifs, it is still interesting to study the effect of each of them on the obtained motifs networks. The present section focuses on this aspect.

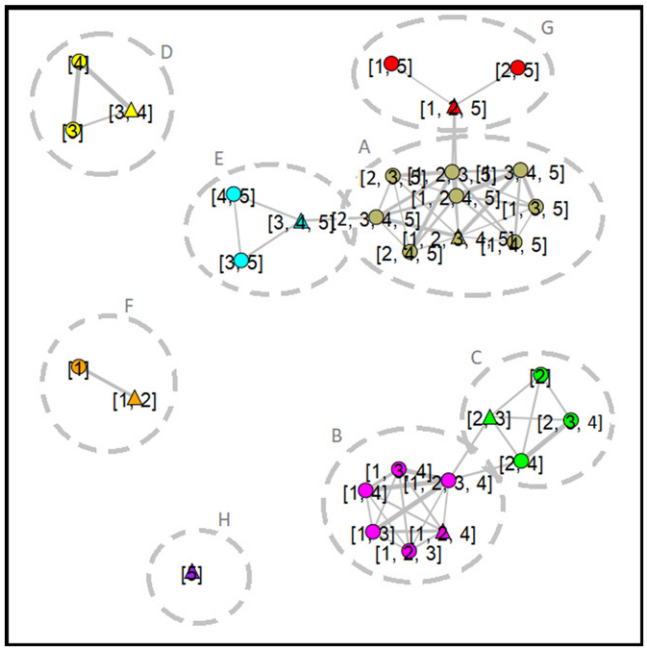

Figure 13. The *features network* obtained for the motifs identified for the city of São Carlos. Each node corresponds to one of the 31 possible combinations of the five adopted features (hierarchical measurements), while the links width is proportional to the value of the coincidence similarity between the NNs obtained for the respective combinations and assuming $T = 0.3$ and $\alpha = 0.1$. Eight communities, discriminate by respective colors, have been identified, which can be understood as the main possible models of the effect of the features on the obtained motifs network. The hubs within each identified community are shown as triangles. The colors in this figure were used only for highlighting the eight models, bearing no relationship whatsoever with the identified motifs. This network was visualized by using the Kamada–Kawai method [87].

In order to do so, in the present work we apply the feature analysis methodology described in [29]. More specifically, NNs are obtained considering all possible combinations of the adopted features. Each of these networks is then represented by the respective weight matrix, whose entries correspond to the obtained coincidence similarity values. Then, the coincidence similarities are obtained between every pair of respective weight matrices, yielding a respective *features network*. Each node in the latter therefore corresponds to a coincidence similarity network respective to some features combination, while the link weights indicate the respective coincidence similarities.

Given that five hierarchical measurements (features) have been adopted, the resulting features network will necessarily have 31 nodes, each corresponding to a possible combination, except for the null case. Figure 13 depicts the therefore obtained features network.

A total of eight communities have been found by using the Infomap methodology (e.g. [85]), each of which corresponding to a respective *putative model* of the NNs that can be obtained for different features combinations. Interestingly, the network obtained while considering all the five features resulted strongly interconnected to other nodes.

Further understanding of the influence of the features can be derived by taking into account the histograms of features to be found within each detected community. These histograms are shown in figure 14.

All features contributed almost equally to the NNs in model A. The motifs in model B employ features 1 to 4, while the feature 5 is not found in this model. The model C involves features 2 to 4. All in all, we have that the adopted features led to eight main putative models of the networks, with the model A corresponding to being the most interconnected within itself as well as with some other models, therefore having special relevance deriving from its centrality.

8. A simple supervised method for assigning motifs

Given that the nine identified city motifs depend exclusively on local measurements, namely only the two neighborhood levels around each reference node, they are not influenced by the remainder of the streets networks, also tending to be invariant to border effects. In addition, it is likely that the local city topology is shared betwen similar cities, as required to cater for similar demands, such as transportation, mobility, access to resources, etc. Yet another important aspect possibly supporting the possible generality of the identified

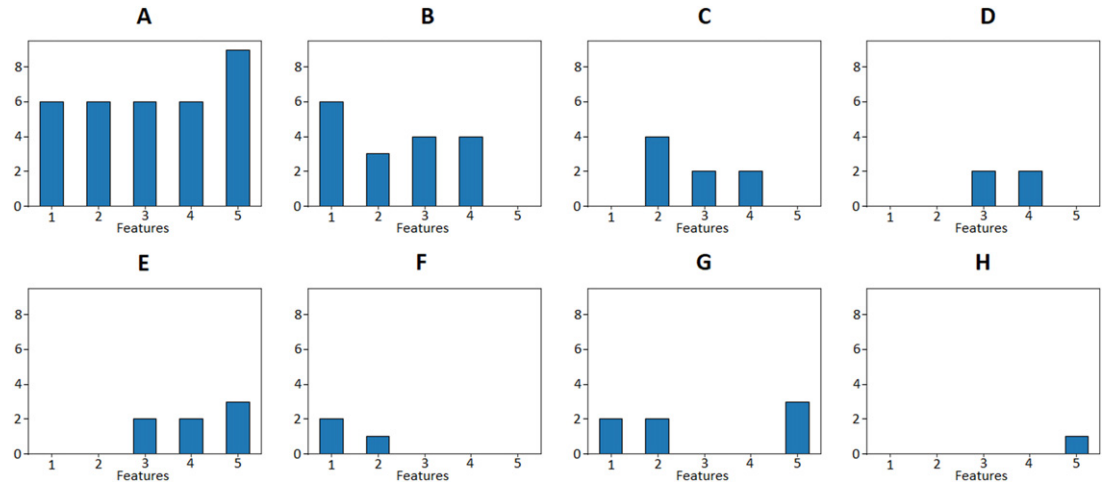

Figure 14. Histograms of the features adopted within each of the eight obtained models describing the effect of features combinations, respectively to the communities A–H.

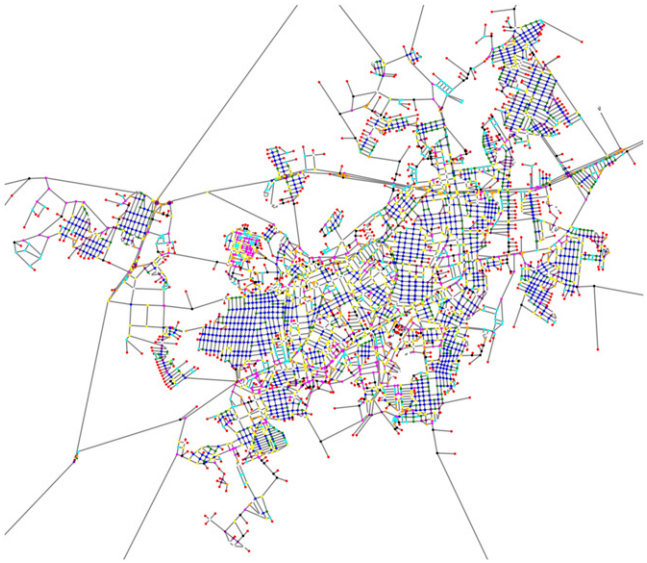

Figure 15. The motifs associated to the neighborhoods of Lages by using the suggested supervised motif assignment method while considering the motif reference table obtained for the city of São Carlos as reference for motif identification.

motifs is the fact that streets networks are largely geographical networks with scant long range connections.

In the light of the above discussion, it is reasonable to posit that the identified motifs can be mostly shared by cities that are reasonably similar. In other words, the reference motifs are henceforth considered among a given set of cities with similar topology.

Under this assumption, it becomes possible to consider transferring the motifs learned in unsupervised manner respective to some reference cities to other cities, which can be done in a relatively simple manner. First, some cities are taken as models, and their combined NN is respectively obtained as described in the current work. Then, a table is derived in which each line corresponds to one of the neighborhoods of the combined network that have been identified as motifs, followed by its respective motif type, as well as its five hierarchical measurements. Now, given a neighborhood from another city to be classified, its standardized features can be compared to those in the reference table and, in case the maximum coincidence similarity is larger than a given threshold, the motif type of the respective entry in the table is assigned to the new neighborhood.

In order to illustrate the above suggested supervised methodology for estimating motifs to nodes of other cities, we consider the motifs identified for the city of São Carlos as described in the previous sections

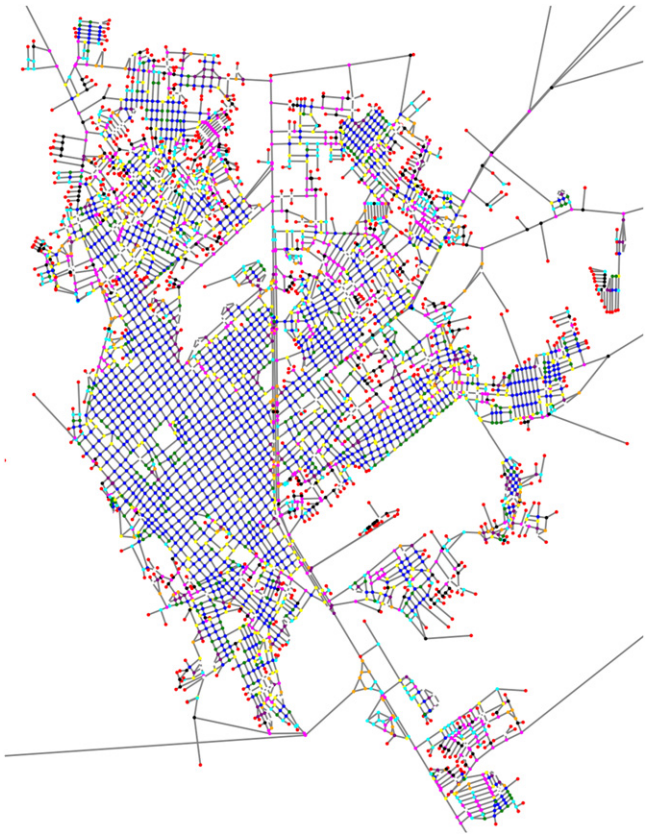

Figure 16. The motifs associated to the neighborhoods of Imperatriz by using the supervised suggested motif assignment method while considering the motif reference table obtained for the city of São Carlos as reference for motif identification.

Table 4. The identification, number and relative frequency of neighborhoods assigned to each of the nine identified motifs obtained for the city of Lages.

Motif identif.	Motif color	No. of nodes	Rel. freq.
*m*1	Blue	926	18.494%
*m*2	Yellow	853	17.036%
*m*3	Orange	377	7.529%
*m*4	Cyan	484	9.666%
*m*5	Magenta	513	10.246%
*m*6	Red	924	18.454%
*m*7	Green	394	7.869%
*m*8	Purple	246	4.913%
*m*9	Black	290	5.792%

as templates for assigning motifs to two other Brazilian cities, namely Lages (SC) and Imperatriz (MA). The respective results are presented in figures 15 and 16. It can be observed that the obtained motifs have characteristics and roles markedly similar to those obtained in the case of São Carlos. The densities of the five hierarchical features of the nine motif types obtained in supervised manner for the cities of Lages and Imperatriz are presented in figures 17 and 18 of appendix A.

Tables 4 and 5 present the relative frequency of the motifs obtained for the cities of Lages and Imperatriz, respectively. It could be expected that many of the motifs result with similar frequencies, though some of the frequencies obtained for different cities could be expected to result distinct, reflecting intrinsic geographical and other types of environmental, geographical and other types of constraints and specificities.

The above tendencies can be observed by comparing the obtained relative frequencies in tables 3–5. We have a markedly large number of motifs *m*1 in all cases, suggesting a predominant orthogonal organization of the cities. Similar relative frequencies were obtained also for motif *m*2. At the same time, a particularly large relative frequency of motif *m*3, as well as relatively small frequency of motif *m*6, were observed for the city of São Carlos.

Table 5. The identification, number and relative frequency of neighborhoods assigned to each of the nine identified motifs obtained for the city of Imperatriz.

Motif identif.	Motif color	No. of nodes	Rel. freq.
$m1$	Blue	1340	24.192%
$m2$	Yellow	888	16.032%
$m3$	Orange	330	5.958%
$m4$	Cyan	417	7.528%
$m5$	Magenta	755	8.214%
$m6$	Red	1165	21.033%
$m7$	Green	392	7.077%
$m8$	Purple	194	3.502%
$m9$	Black	358	6.463%

9. Concluding remarks

The study and characterization of cities have constituted the focus of significant attention along the last decades, especially given the potential of such analysis for enhancing urban aspects and better understanding relationships between the city topology and socioeconomic factors, among several others possibilities.

In network science, the concept of network motifs has been applied with particular effectiveness for characterizing and better understanding the network topology. Here, we approached the interesting topic of city characterization in terms of statistical motifs identified from network representations of cities, i.e. streets networks. More specifically, we adopted a local characterization of the topological properties of neighborhoods around each of the streets networks nodes. This has been accomplished by using five hierarchical measurements considering two neighborhood levels around each reference node, thus allowing a mesoscopic characterization of the respective topological properties.

The pairwise similarity between the topological properties of the neighborhoods was then quantified by using the coincidence similarity methodology, which implements a particularly strict similarity quantification, therefore contributing to enhanced levels of interconnection details and network modularity.

A NN was obtained for a Brazilian city (São Carlos), which then had its communities detected by the Infomap approach. The properties of the identified motifs were then characterized and discussed based on four main perspectives, namely the motifs similarities, visualizations of samples of each motif, distributions of the five adopted hierarchical measurements, as well as histograms of adjacency between the nine motifs.

The obtained city motifs can be understood from both the perspective of homogeneity, complexity, as well as centrality, with one of the motifs ($m1$) corresponding to the prototypical square block organization characterizing full orthogonal street plans. This type of motif tends to be the most regular and central among the identified types. Another particularly interesting motif, namely $m6$, tended to be related to streets dead ends. Motifs $m3$ and $m8$ both have their reference nodes corresponding to one of the vertices of a triangular block, but they distinguish one another respectively to other hierarchical measurements. Motifs of type $m2$, $m4$ and $m5$ tended to be particularly irregular, frequently appearing as an interface or transitions between more regular patches. These three motifs, however, have distinct hierarchical degrees.

As a complement to the reported approach to city motifs identification, we also performed an analysis of the influence of the adopted hierarchical features on the respectively obtained NNs. This was accomplished by using the coincidence similarity, leading to the identification of eight possible models (communities) of NNs that can be obtained by combining the five adopted features. The most cohesive model involves all the five adopted hierarchical measurements.

Although the proposed methodology to identify city motifs involves several concepts and steps, a simple supervised method has been also suggested and illustrated in this work for estimating motif types in a given streets network. This procedure is based on a reference table containing several instances of neighborhoods and their motifs identified respective to a set of reference cities used for training. Then, given a new city with similar characteristics represented in terms of its streets network, motif types can be assigned to its neighborhoods by taking into account the motif of the table entry with the features that are most similar to those of each of the nodes in the new city.

The potential of this simple supervised methodology has been illustrated by applying a motif table derived from the city of São Carlos to two other cities, with suitable results. In particular, the nine city motifs therefore identified were found to be remarkably consistent not only within a same city, but also across the considered cities. However, it should be kept in mind that this approach requires the new cities to have characteristics similar to those used in the training stage, and incorrect or biased results can be otherwise obtained.

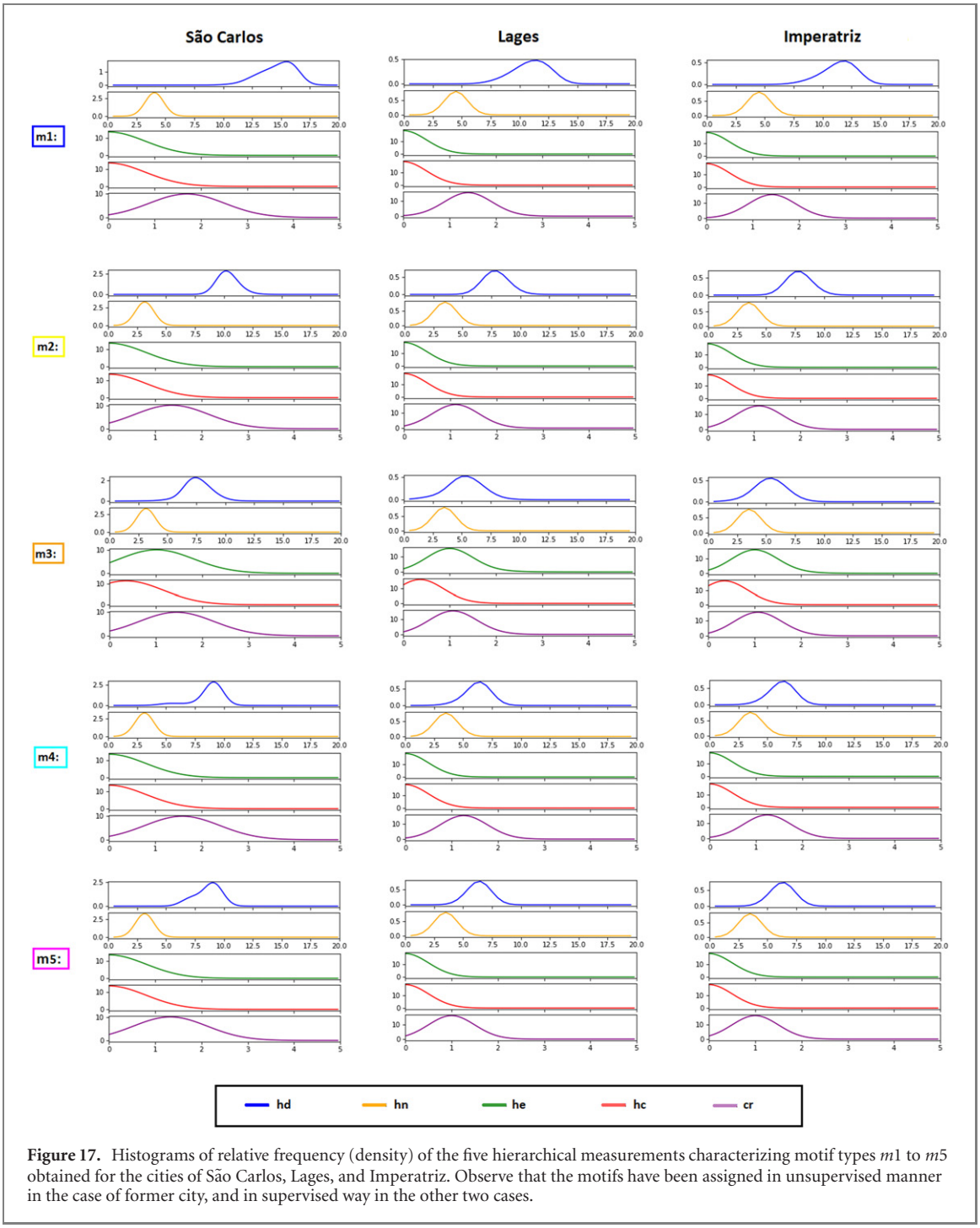

Figure 17. Histograms of relative frequency (density) of the five hierarchical measurements characterizing motif types $m1$ to $m5$ obtained for the cities of São Carlos, Lages, and Imperatriz. Observe that the motifs have been assigned in unsupervised manner in the case of former city, and in supervised way in the other two cases.

Additional information about this work can be found at https://github.com/ericktokuda/city-motifs-supp.

The encouraging results reported in the present work respectively to concepts, methodology and results, pave the way to a large number of future possible developments. For instance, it would be interesting to investigate the effect of larger neighborhood extensions (H) on the resulting motifs. It would also be interesting to compare a substantial number cities based on their respective distribution of motifs, as well as the adjacency between them.

Given the inherently hierarchical nature of the accessibility (e.g. [88–90]), this measurement could also be considered instead, or as a complement to the hierarchical measurements currently adopted. Another particularly promising perspective regards the incorporation of geometrical features as a complement of the topological features. For instance, even more strict identification of motifs belonging to highly orthogonal portions of a city can be obtained by also taking into account the lengths of each of the block sides.

In addition, given that motifs can be expected in a wide range of real-world and theoretical networks, it would be of particular interest to extend the concepts and methodology proposed in the present work to

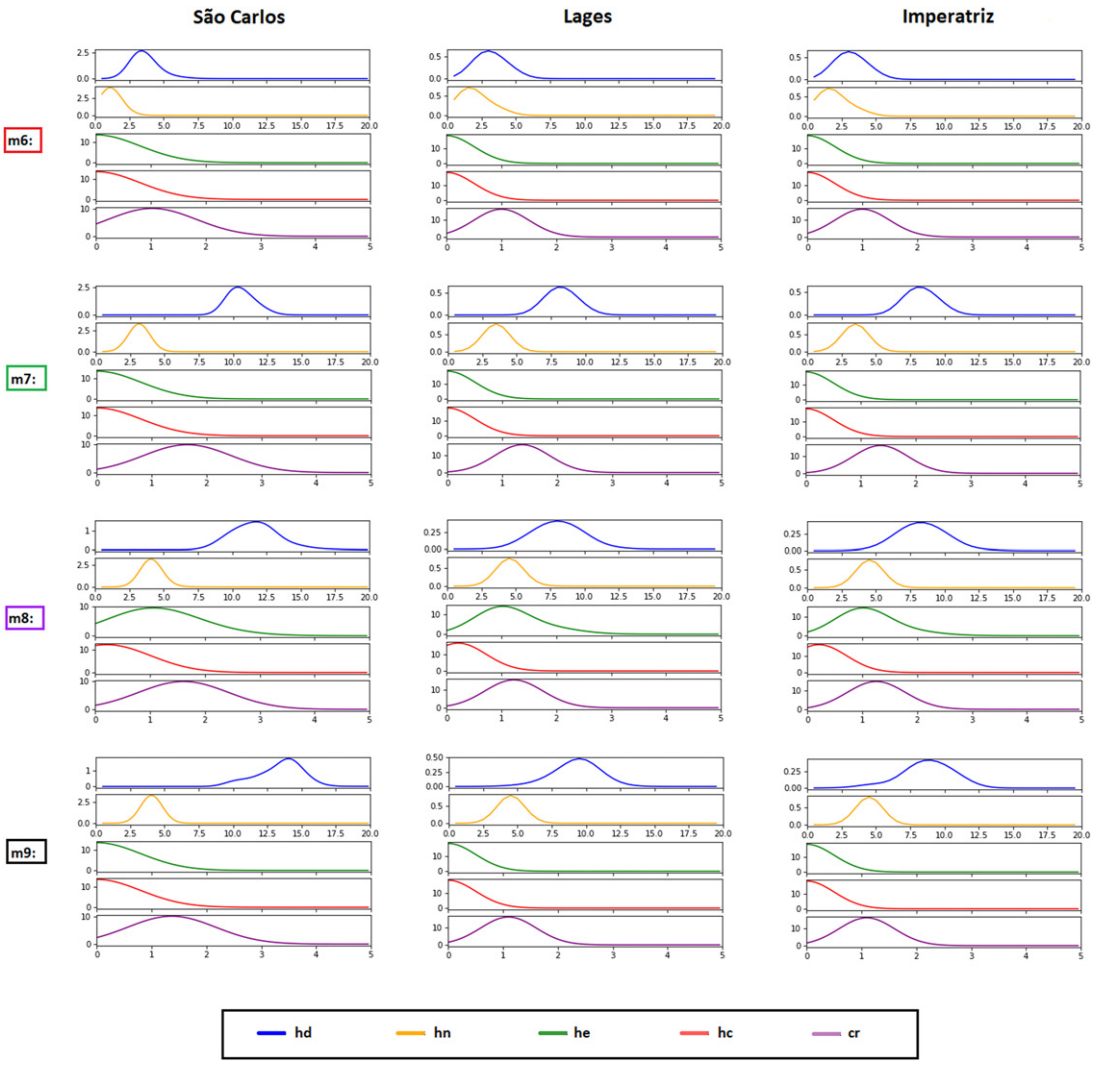

Figure 18. Histograms of relative frequency (density) of the five hierarchical measurements characterizing motif types $m6$ to $m9$ obtained for the cities of São Carlos, Lages, and Imperatriz. Observe that the motifs have been assigned in unsupervised manner in the case of former city, and in supervised way in the other two cases.

other types of networks, such as roads and airport routes, energy distribution, Internet and WWW, protein interaction, scientific collaboration, text and citations networks, among many other possibilities.

Acknowledgments

G S Domingues thanks CAPES (88887.601529/2021-00) for financial support. E K Tokuda thanks FAPESP (2019/01077-3) for financial support. L da F Costa thanks CNPq (307085/2018-0) and FAPESP (2015/22308-2) for support. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001.

Data availability statement

No new data were created or analysed in this study. https://github.com/ericktokuda/city-motifs-supp

Appendix A. Comparison of all measurements

Shown in this appendix are the relative frequency histograms (densities) of the nine types of motifs obtained in unsupervised manner in the case of the city of São Carlos, and in supervised manner respectively to the cities of Lages and Imperatriz, which are presented respective in figure 17 (from motif $m1$ to $m5$) and figure 18 (from motif $m6$ to $m9$).

Of particular importance is the fact that the histograms obtained for each of the three NNs have similar shapes, suggesting potential consistency and generality of the identified motifs. At the same time, markedly distinct histogram shapes can be observed between distinct motif types. To a considerable extension, these important results have been allowed by the choice not only of the informative hierarchical measurements, but also by the strict similarity characterization implemented by the coincidence similarity methodology.

Interesting relationships can be observed among different motif types. For instance, the histograms of *hierarchical degree hd* obtained for distinct motifs tended to appear at different positions along the horizontal axes and with varying shapes. Consequently, this feature can be deemed to be of particular importance for distinguishing between the nine identified city motif types. Particularly noticeable variation has also been observed for the *convergence ratio cr* hierarchical feature. In addition, observe that the histogram of the *hierarchical number of edges ne* resulted mostly null for several motif types.

ORCID iDs

Guilherme S Domingues https://orcid.org/0000-0001-8872-545X
Eric K Tokuda https://orcid.org/0000-0002-6159-2500
Luciano da F Costa https://orcid.org/0000-0001-5203-4366

References

[1] Rosvall M, Trusina A, Minnhagen P and Sneppen K 2005 Networks and cities: an information perspective *Phys. Rev. Lett.* **94** 028701
[2] Strano E, Viana M, da Fontoura Costa L, Cardillo A, Porta S and Latora V 2013 Urban street networks, a comparative analysis of ten European cities *Environ. Plan.* B **40** 1071–86
[3] Porta S, Crucitti P and Latora V 2006 The network analysis of urban streets: a dual approach *Physica* A **369** 853–66
[4] Buhl J, Gautrais J, Reeves N, Solé R V, Valverde S, Kuntz P and Theraulaz G 2006 Topological patterns in street networks of self-organized urban settlements *Eur. Phys. J.* B **49** 513–22
[5] Louf R and Barthelemy M 2014 A typology of street patterns *J. R. Soc. Interface* **11** 20140924
[6] Barthelemy M 2016 *The Structure and Dynamics of Cities* (Cambridge: Cambridge University Press)
[7] Batty M and Longley P A 1994 *Fractal Cities: A Geometry of Form and Function* (New York: Academic)
[8] Batty M 2007 *Cities and Complexity: Understanding Cities with Cellular Automata, Agent-Based Models, and Fractals*
[9] Batty M 2013 *The New Science of Cities* (Cambridge, MA: MIT press)
[10] Barabási A L and Pósfai M 2016 *Network Science* (Cambridge: Cambridge University Press)
[11] Newman M 2010 *Networks: An Introduction* (Oxford: Oxford University Press)
[12] da Fontoura Costa L, Oliveira O Jr, Travieso G, Rodrigues F, Boas P R V, Antiqueira L, Viana M and Correa Rocha L 2011 Analyzing and modeling real-world phenomena with complex networks: a survey of applications *Adv. Phys.* **60** 329–412
[13] Boccaletti S, Latora V, Moreno Y, Chavez M and Hwang D 2006 Complex networks: structure and dynamics *Phys. Rep.* **424** 175–308
[14] Open Street Map 2021 Open Street Map www.openstreetmap.org/ (accessed July 2021)
[15] Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D and Alon U 2002 Network motifs: simple building blocks of complex networks *Science* **298** 824–7
[16] Budach S and Marsico A 2018 pysster: classification of biological sequences by learning sequence and structure motifs with convolutional neural networks *Bioinformatics* **34** 3035–7
[17] Xie W, Yong Y, Wei N, Yue P and Zhou W 2021 Identifying states of global financial market based on information flow network motifs *North Am. J. Econ. Finance* **58** 101459
[18] Jin Y, Wei Y, Xiu C, Song W and Yang K 2019 Study on structural characteristics of China's passenger airline network based on network motifs analysis *Sustainability* **11** 2484
[19] Nissen P, Ippolito J, Ban N, Moore P and Steitz T 2001 RNA tertiary interactions in the large ribosomal subunit: the a-minor motif *Proc. Natl Acad. Sci. USA* **98** 4899–903
[20] D'Haeseleer P 2006 What are DNA sequence motifs? *Nat. Biotechnol.* **24** 423–5
[21] da Fontoura Costa L 2022 Coincidence complex networks *J. Phys. Complex.* **3** 015012
[22] da Fontoura Costa L 2021 A caleidoscope of datasets represented as networks by the coincidence methodology Accessed 01 June 2022
[23] Reis R and da Fontoura Costa L 2022 Enzyme similarity networks (arXiv:2205.05163)
[24] Benatti A, Arruda H and da Fontoura Costa L 2022 Neuromorphic networks as revealed by features similarity Accessed 01 July 2022 arXiv
[25] da Fontoura Costa L 2021 Further generalizations of the Jaccard index (arXiv:2110.09619)
[26] da Fontoura Costa L 2022 On similarity *Physica* A **599** 127456
[27] Vijaymeena M and Kavitha K 2016 A survey on similarity measures in text mining *Mach. Learn. Appl.: Int. J.* **3** 19–28
[28] da Fontoura Costa L 2021 Multiset neurons
[29] da Fontoura Costa L 2021 Elementary particles networks as revealed by their spin, charge and mass
[30] Travençolo B and da Fontoura Costa L 2008 Hierarchical spatial organization of geographical networks *J. Phys. A: Math. Theor.* **41** 224004
[31] da Fontoura Costa L and Silva F N 2006 Hierarchical characterization of complex networks *J. Stat. Phys.* **125** 841–72
[32] Newman M 2006 Modularity and community structure in networks *Proc. Natl Acad. Sci. USA* **103** 8577–82
[33] Fortunato S 2010 Community detection in graphs *Phys. Rep.* **486** 75–174
[34] Reichardt J and Bornholdt S 2006 Statistical mechanics of community detection *Phys. Rev.* E **74** 016110
[35] Fortunato S and Hric D 2016 Community detection in networks: a user guide *Phys. Rep.* **659** 1–44

[36] Geddes P 1915 *Cities in Evolution* (London: Routledge)
[37] Mumford L 1937 *What Is a City?* (Athens, GA: University of Georgia Press)
[38] Wirth L 1938 Urbanism as a way of life *Am. J. Sociol.* **44** 1–24
[39] Lynch K 1964 *The Image of the City* (Cambridge, MA: MIT press)
[40] Hillier B and Hanson J 1989 *The Social Logic of Space* (Cambridge: Cambridge University Press)
[41] Choi J, Barnett G and Chon B 2006 Comparing world city networks: a network analysis of internet backbone and air transport intercity linkages *Glob. Netw.* **6** 81–99
[42] Liao H and Zeng A 2015 Reconstructing propagation networks with temporal similarity *Sci. Rep.* **5** 11404
[43] Cardillo A, Scellato S and Latora V 2006 Structural properties of planar graphs of urban street patterns *Phys. Rev.* E **73** 066107
[44] Hipp J R, Faris R W and Boessen A 2012 Measuring 'neighborhood': constructing network neighborhoods *Soc. Netw.* **34** 128–40
[45] Li A and Horvath S 2007 Network neighborhood analysis with the multi-node topological overlap measure *Bioinformatics* **23** 222–31
[46] Alon U 2007 Network motifs: theory and experimental approaches *Nat. Rev. Genet.* **8** 450–61
[47] Stone L, Simberloff D and Artzy-Randrup Y 2019 Network motifs and their origins *PLoS Comput. Biol.* **15** e1006749
[48] da Fontoura Costa L, Rodrigues F, Travieso G and Boas P R V 2007 Characterization of complex networks: a survey of measurements *Adv. Phys.* **56** 167–242
[49] Lodato I, Boccaletti S and Latora V 2007 Synchronization properties of network motifs *Europhys. Lett.* **78** 28001
[50] Ciriello G and Guerra C 2008 A review on models and algorithms for motif discovery in protein–protein interaction networks *Brief. Funct. Genom. Proteomics* **7** 147–56
[51] Sporns O, Kötter R and Friston K 2004 Motifs in brain networks *PLoS Biol.* **2** e369
[52] Liu P, Masuda N, Kito T and Sariyüce A 2022 Temporal motifs in patent opposition and collaboration networks *Sci. Rep.* **12** 1917
[53] Boas P R V, Rodrigues F, Travieso G and da Fontoura Costa L 2008 Chain motifs: the tails and handles of complex networks *Phys. Rev.* E **77** 026106
[54] Boas P R V, Rodrigues F A, Travieso G and da Fontoura Costa L 2007 Border trees of complex networks *J. Phys. A: Math. Theor.* **41** 224005
[55] LaRock T, Scholtes I and Eliassi-Rad T 2021 Sequential motifs in observed walks (arXiv:2112.05642)
[56] Schneider C M, Belik V, Couronné T, Smoreda Z and González M C 2013 Unravelling daily human mobility motifs *J. R. Soc. Interface* **10** 20130246
[57] Stoica A and Prieur C 2009 Structure of neighborhoods in a large social network *2009 Int. Conf. Computational Science and Engineering* vol 4 (IEEE) pp 26–33
[58] Yang L, Wu L, Liu Y and Kang C 2017 Quantifying tourist behavior patterns by travel motifs and geo-tagged photos from Flickr *ISPRS Int. J. Geo-Inf.* **6** 345
[59] Tsiotas D and Polyzos S 2017 The topology of urban road networks and its role to urban mobility *Transport. Res. Proc.* **24** 482–90
[60] Ping L, Xing X, Zhong-Liang Q, Gang-Qiang Y, Xing S and Bing-Hong W 2006 Topological properties of urban public traffic networks in Chinese top-ten biggest cities *Chin. Phys. Lett.* **23** 3384
[61] Rossum G and Drake F Jr 1995 *Python Reference Manual* (Amsterdam: Centrum voor Wiskunde en Informatica)
[62] Csardi G and Nepusz T 2006 The igraph software package for complex network research *InterJournal, Complex Systems* **1695**
[63] Ahnert S, Travencolo B and da Fontoura Costa L 2009 Connectivity and dynamics of neuronal networks as defined by the shape of individual neurons *New J. Phys.* **11** 103053
[64] Mirkin B 1996 *Mathematical Classification and Clustering* vol 11 (Berlin: Springer)
[65] Arruda G F, da Fontoura Costa L and Rodrigues F A 2012 A complex networks approach for data clustering *Physica* A **391** 6174–83
[66] Akbas C, Bozkurt A, Arslan M, Aslanoglu H and Cetin E 2014 L1 norm based multiplication-free cosine similarity measures for big data analysis *2014 Int. Workshop Computational Intelligence for Multimedia Understanding (IWCIM)* (IEEE) pp 1–5
[67] Wikipedia 2004 Jaccard index—wikipedia, the free encyclopedia https://en.wikipedia.org/wiki/Jaccard_index (accessed 10 October 2021)
[68] Singh D, Ibrahim M, Yohana T and Singh J 2011 Complementation in multiset theory *Int. Math. Forum* **38** 1877–84
[69] Mahalakshmi P and Thangavelu P 2019 Properties of multisets *Int. J. Innov. Technol. Explor. Eng.* **8** 1–4
[70] Knuth D 1998 *The Art of Computing* (Reading, MA: Addison-Wesley)
[71] Heinz S 2011 *Mathematical Modeling* (Berlin: Springer)
[72] Blizard W 1989 Multiset theory *Notre Dame J. Form. Log.* **30** 36–66
[73] Blizard W 1991 The development of multiset theory *Mod. Logic* **4** 319–52
[74] da Fontoura Costa L 2021 Multisets ResearchGate
[75] da Fontoura Costa L and Tokuda E 2022 A similarity approach to cities and features (arXiv:2202.08301)
[76] Comin C, Peron T, Silva F, Amancio D, Rodrigues F and da Fontoura Costa L 2020 Complex systems: features, similarity and connectivity *Phys. Rep.* **861** 1–41
[77] da Fontoura Costa L 2004 Complex networks, simple vision (arXiv:cond-mat/0403346)
[78] Onnela J, Kaski K and Kertész J 2004 Clustering and information in correlation based financial networks *Eur. Phys. J.* B **38** 353–62
[79] Yang H, Cheng J, Yang Z, Zhang H, Zhang W, Yang K and Chen X 2021 A node similarity and community link strength-based community discovery algorithm *Complexity* **2021** 8848566
[80] Putra J and Tokunaga T 2017 Evaluating text coherence based on semantic similarity graph *Proc. TextGraphs-11: The Workshop on Graph-Based Methods for Natural Language Processing* pp 76–85
[81] Backes A and Bruno O 2010 Shape classification using complex network and multi-scale fractal dimension *Pattern Recognit. Lett.* **31** 44–51
[82] da Fontoura Costa L and Cesar R Jr 2010 *Shape Analysis and Classification: Theory and Practice* (Boca Raton, FL: CRC Press)
[83] Gewers F, Ferreira G, Arruda H, Silva F, Comin C, Amancio D and da Fontoura Costa L 2021 Principal component analysis: a natural approach to data exploration *ACM Comput. Surv.* **54** 1–34
[84] Fruchterman T and Reingold E 1991 Graph drawing by force-directed placement *Softw. - Pract. Exp.* **21** 1129–64
[85] Rosvall M and Bergstrom C 2008 Maps of random walks on complex networks reveal community structure *Proc. Natl Acad. Sci. USA* **105** 1118–23
[86] Martin S, Brown W M and Wylie B N 2007 Dr. l: distributed recursive (graph) layout *Technical Report* (Albuquerque, NM (United States) Sandia National Lab.(SNL-NM))
[87] Kamada T and Kawai S 1989 An algorithm for drawing general undirected graphs *Inf. Process. Lett.* **31** 7–15
[88] Travençolo B and da Fontoura Costa L 2008 Accessibility in complex networks *Phys. Lett.* A **373** 89–95

[89] Arruda G F, Barbieri A, Rodriguez P, Rodrigues F, Moreno Y and da Fontoura Costa L 2014 Role of centrality for the identification of influential spreaders in complex networks *Phys. Rev.* E **90** 032812
[90] Viana M, Fourcassié V, Perna A, da Fontoura Costa L and Jost C 2013 Accessibility in networks: a useful measure for understanding social insect nest architecture *Chaos Solitons Fractals* **46** 38–45