A similarity approach to cities and features

Luciano da F. Costa and Eric K. Tokuda

Received 2 March 2022 / Accepted 6 September 2022 / Published online 19 September 2022
© The Author(s), under exclusive licence to EDP Sciences, SIF and Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract. Characterizing the structure of cities constitutes an important task, since the identification of similar cities can promote sharing of respective experiences. In the present work, we consider 20 European cities from 5 respective countries and with comparable populations, each of which characterized in terms of four topological as well as one geometrical feature. These cities are then mapped into respective networks by considering their pairwise similarity as gauged by the coincidence methodology, which consists of combining the Jaccard and interiority indices. The methodology incorporates a parameter alpha that can control the relative contribution of features with the same or opposite signs to the overall similarity. Interestingly, the maximum modularity cities’ network is obtained for a non-standard parameter configuration, showing that it could not be obtained were not for the adoption of the parameter alpha. The network with maximum modularity presents four communities that can be mostly related to four of the five considered countries, indicating a tendency of the cities from a same country being similar. The coincidence methodology was then applied to investigate the effect of several features combinations on the respectively obtained networks, leading to a highly modular features network containing four main communities that can be understood as the main possible models for the considered cities.

1 Introduction

The physical world is intrinsically characterized by unending diversity. Towns and cities are no exception. As a consequence of numerous influences—including geography, history, economy, age, climate, as well as traditions and culture—it is completely impossible to find two towns or cities that are identical regarding all their respective geometrical and topological characteristics, properties, or features.

The properties of cities can be derived either from direct observation of the city (e.g., photographs), respective plans, as well as by their respective representations as streets networks. The latter type of representation is typically obtained by considering intersections between streets, as well as dead-ends, as nodes, while respective segments of the streets are taken as edges. In the latter situation, which is that considered in the present work, possible features include topological measurements—such as node degree, clustering coefficient, etc.—obtained from the networks.

Given that cities are dynamic structures, adapting continuously across time and space, it becomes an important task not only to characterize them according to a representative number of features, but also to devise and apply the means for quantifying, in pairwise fashion, how much cities are similar [2,3,17,21,33,36,37].

These initiatives can lead to several valuable results. For instance, cities that are found to be similar can consider sharing their respective administrative and planning experiences. In addition, as soon as the pairwise similarity between cities has been effectively quantified, it becomes possible to generate respective networks, henceforth referred to as cities networks, in which each city becomes a node, while the interconnections reflect the respective similarities. Analysis of these cities networks using the wealthy of concepts and methods from network science (e.g., [6]) can then provide an ample understanding about types of cities corresponding to large network communities, as well as more specific types of cities.

One possible approach to study relationships between the structure of streets network involves quantifying the similarity between respective topological measurements (please refer to the Sect. 2 for additional contextual information). As a consequence of its intrinsic characteristics, the coincidence similarity index has been found [13] to lead to similarity comparisons that are more strict than those obtained by the cosine similarity and Pearson correlation alternative approaches. The coincidence similarity index has thus been applied as a means to translate datasets described by respective features, into graphs or networks presenting a marked

*e-mail: tokuda.ek@gmail.com (corresponding author)
level of interconnectivity detail and enhanced modularity [9,13].

In the present work, we understand modularity in the sense of Newman’s approach [30]. More specifically, the network modules (or communities) correspond to its subsets that yield high modularity indices, therefore indicating statistically uniform interconnections, while the respective nodes are more connected one another than with the remainder of the network.

Interestingly, the same coincidence method for translating datasets into networks can also be employed to approach another important and challenging problem in pattern recognition, namely the objective quantitative identification of the effect that different choices of features can have on the resulting networks, a problem directly related to feature analysis and selections (e.g., [10]). This type of study is essential for supplying valuable information, not only how features are interrelated, but also to select particularly suitable features then used to obtain network representations.

Given a dataset and a set of respective features, the basic idea, as described in [10], consists in deriving the features network, respectively, implied by each possible (or of interest) feature combination. More specifically, each node in these networks corresponds to a cities network obtained by a respective specific combination of measurements. The links between these nodes correspond to the similarity between the aforementioned networks as gauged by the coincidence similarity index. The obtained coincidence values are then organized into a respective weight matrix, which completely specifies the obtained city networks. Then, it is possible to consider the resulting weight matrices as features to obtain a network where each of the respective networks is associated with a node, while the pairwise interconnections between them reflect the respective similarity as gauged by the coincidence similarity index.

The present work aims at addressing the interesting problem of comparing several cities worldwide, represented by respective streets networks, in terms of cities networks obtained by the coincidence methodology [9]. The primary motivation for this study consists in harnessing more detailed and modular description of cities relationships as allowed by the intrinsic ability of the coincidence similarity index in promoting more detailed and strict information about the compared entities.

Twenty European cities with similar populations have been arbitrarily chosen, corresponding to four cities from each of 5 European countries (France, Germany, Italy, Spain, and United Kingdom) with population falling within a specific interval (see Table 2).

After obtaining the streets networks, several geometrical and topological measurements (e.g., [16]) are calculated, namely the average node degree, the standard deviation of the node degree, the standard deviation of the local node clustering coefficient, the dispersion of the node position, and the accessibility, which are considered as features characterizing each considered city.

After these features are collected and standardized, the coincidence method is applied, yielding respective cities networks, whose overall connectivity can be conveniently controlled by the parameter \( \alpha \). Respective results are obtained and discussed, including the organization of cities in well-defined groups sharing properties. Interestingly, four main groups of cities have been obtained presenting the majority of cities from respective countries.

To complement our study, we then apply the coincidence methodology on the weight matrices respective to each obtained cities network, so that the whole set of weights is understood as corresponding to the features characterizing each considered network. The application of the coincidence method then allows a features network to be obtained characterizing in an accurate and objective manner the effect of the distinct possible feature combinations on the obtained networks topology.

The therefore obtained feature network was characterized by a well-defined modular structure, from which four respective communities were then identified, which can be understood as the four main models that can be obtained for the cities while considering different feature combinations. In particular, it is suggested that the hubs of each of these communities can be understood as the respective prototype, therefore summarizing each of the four models in terms of a respective reference network. It has also been found that the four obtained models share three of the five adopted features.

The present work starts by presenting how the data were obtained and then proceeds to describing the basic employed concepts and methods. The results are presented next regarding the cities networks, and then to the features network.

2 Related works

While cities networks may present common features, it is possible to group cities networks into historical non-planned cities and modern planned cities [18]. The authors of [32] employ concepts from information science to compare cities. They map the city street network to an information network and they consider a complexity measurement \( S \) for locating streets in this network. They observed that modern cities, such as Manhattan, tend to be more organized (easier to navigate) than older cities as Umeå.

In [18,33], the authors employ network centrality measurements for city comparison. They calculated four centrality measures, namely closeness, betweenness, straightness, and information, from network patches to compare cities from different countries. In [18], the result indicates that self-organized cities tend to exhibit different centrality values when compared to more planned ones. In [33], besides the differences perceived by the accessibility measurements, they also observed heterogeneity in the street length distribution across cities.

While the traditional use of networks for studying cities often takes into account their street networks, it is also possible to model the economical relation-
ships between cities (and its service firms) through a cities network [34]. In this model, differently from street networks, each node corresponds to a city and the edges correspond to the relation between cities. In [40], the authors also consider cities as nodes of a network, but here the connectivity is based on geolocated images. The authors argue that, in general, the city can be characterized by observers who look at a set of respective images. The obtained network quantifies the visual similarity among cities. A related approach was proposed by [20], where a comparative analysis of cities based on the land-use distribution was performed. The author collected data from 100 cities, computed image descriptors, and compared the images using a hierarchical clustering algorithm. In [31], the authors used venue-based data from social networks to compare cities.

Similarity indices (e.g., [1,13,28,39]) have been frequently used in scientific and technological applications. Given that the cosine similarity and Pearson correlation coefficient can cope with real-valued features (e.g., [13,15]), these two similarity measurements correspond to frequently employed approaches for transforming datasets into respective graphs (or networks, e.g., [9]).

Despite its good potential for similarity characterization, the Jaccard index (e.g., [13,39] defined in Eq. 1, respectively, to two non-empty sets $A$ and $B$, has been largely restricted to the treatment of categorical or binary data

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|},$$

where $|A| > 0$ stands for the cardinality, or number of elements, in set $A$. The Jaccard index has also found not to be able to take into account how much each of the compared sets is interior to the other [11,13].

A generalization of the Jaccard index capable of taking into account real, possibly negative valued vectors or functions, as well as incorporating the quantification of the relative interiority between the two compared vectors has been described in [11,13]. More specifically, multiset principles [12] are used to translate the intersection and union involving real values into respective multiset operations involving the minimum and maximum functions, as well as functions indicating the sign of the operands [13]. To incorporate information about the interiority between the two compared vectors, the interiority (or overlap index [39]) is also calculated using multiset representation and then multiplied by the Jaccard index, resulting in the coincidence similarity index [11,13].

### 3 Materials and methods

Table 1 presents, for reference’s sake, the three types of networks considered in the present work, as well as the nature of their respective nodes and links.

| Network type         | Nodes                                      | Links (or edges)               |
|----------------------|--------------------------------------------|---------------------------------|
| Streets network      | Streets intersections or dead-ends          | Street segments                 |
| Cities network       | Cities                                     | Similarity between cities       |
| Features network     | Cities networks respective to feature       | Similarity between weight       |
|                      | combinations                               | matrices                        |

The cities considered in this work are from five European countries (namely France, Germany, Italy, Spain, and the United Kingdom) and have populations ranging from 200,000 and 300,000 inhabitants. Other than that, the cities were chosen arbitrarily. Four cities from each country were considered, totalling $n = 20$ cities. Each city is represented by a streets network, with nodes representing street intersections, while the edges joining two nodes correspond to the streets and avenues. The graphs were obtained from OpenStreetMap, a public and crowd-sourced repository of geographical information using the OSMNx framework [7]. This repository intrinsically includes the identification of the streets. For each city, the set of nodes whose locations fall inside the administrative region of the city are considered and the resulting network obtained from these nodes is considered in our analysis. For simplicity’s sake, whenever more than one edge exists between two nodes, they were reduced to a single edge.

The following five features have been considered in this work:

1. $\langle k \rangle$: the average of the node degrees;
2. $\sigma_k$: the standard deviation of the node degree;
3. $\sigma_{CC}$: the standard deviation of the node clustering coefficient;
4. $disp_{pos}$: dispersion of the point locations;
5. $\sigma_A$: the standard deviation of the node accessibility.

These five measurements have been chosen so as to provide a characterization of complementary topological and geometrical properties of the respective networks, as described in the following. While the average degree $\langle k \rangle$ provides a direct quantification of the local connectivity of the networks, the respective standard deviation $\sigma_k$ reflects the heterogeneity of the local topology of the networks (for instance, a completely regular network has null node degree standard deviation). The standard deviation of the clustering coefficient $\sigma_{CC}$ helps to complement the characterization of the interconnectivity of the network while considering the links between the neighbors of each node. The
Table 2 List of the analyzed cities, grouped by country, with corresponding populations. Adapted from [29]  

| City         | Country     | Population |
|--------------|-------------|------------|
| Nantes       | France      | 280,000    |
| Bordeaux     | France      | 230,000    |
| Lille        | France      | 230,000    |
| Rennes       | France      | 210,000    |
| Brunswick    | Germany     | 240,000    |
| Freiburg     | Germany     | 220,000    |
| Kiel         | Germany     | 230,000    |
| Augsburg     | Germany     | 260,000    |
| Bari         | Italy       | 280,000    |
| Messina      | Italy       | 220,000    |
| Verona       | Italy       | 220,000    |
| Padova       | Italy       | 200,000    |
| Vigo         | Spain       | 200,000    |
| Granada      | Spain       | 230,000    |
| Oviedo       | Spain       | 220,000    |
| Mostoles     | Spain       | 210,000    |
| Bradford     | U.K.        | 300,000    |
| Derby        | U.K.        | 270,000    |
| Luton        | U.K.        | 260,000    |
| Southampton  | U.K.        | 240,000    |

dispersion of the positions (\(\text{disp}_{\text{pos}}\)) of the nodes provides indication about the spreading, around the center of mass, of the geographical position of the network nodes. Therefore, this measurement is related to the spatial density of the nodes. The accessibility measurement can be understood as a generalization of the node degree [4] at successive topological distances while taking into account transition probabilities between each pair of network nodes. As such, the standard deviation of this measurement can provide indication about the regularity of a network at mesoscopic topological scales.

The node degree distribution is a simple yet rich graph measurement, providing information about the connectivity of the nodes [16]. We consider two statistics from this distribution: the average and the standard deviation (\(\langle k \rangle\) and \(\sigma_k\)).

The clustering coefficient accounts for the cyclic structure of the graph by considering the number of 3-cliques (triangles) present [16]. This local measurement in particular takes into account the number of triangles considering just the neighborhood of the reference node

\[
CC(i) = \frac{N_\Delta(i)}{N_3(i)},
\]

where \(N_\Delta(i)\) corresponds to the number of triangles formed by the nodes in the neighborhood of \(i\), and \(N_3(i)\) corresponds to the possible number of edges among them.

The geographical distribution of nodes is also a factor of similarity among city graphs, so that the dispersion of the node positions is in the current work taken as a network feature. It is computed by Eq. 3, where \(\vec{p}_i\) represent the position for each node, \(m\) is the number of nodes, and \(\vec{c}\) is the centroid of the positions. Here, we adopted the the \(L_2\) norm \(|| \cdot ||\) as measurement of position difference, yielding

\[
\text{disp}_{\text{pos}} = \frac{\sum_i^m ||\vec{p}_i - \vec{c}||}{m}.
\]  

Another feature that has been successfully employed for city analysis is the network accessibility [19,35,38], defined respectively to transition probabilities between pairs of nodes. In the present case, these transition probabilities refer to uniform random walks performed on the network topology. A parameter \(h\) defines the topological scale of the neighborhood (e.g., \(h = 2\) indicates the nodes that are at topological distance 2 from the reference node). The accessibility can be calculated as

\[
A_h(i) = \exp \left[ -\sum_{j=1}^{m} p^h_{i,j} \log(p^h_{i,j}) \right],
\]

with \((p^h_{i,j})\) corresponding to the transition probability from node \(i\) to node \(j\), and \(m\) corresponds to the number of nodes in the network.

Figure 1 depicts the main steps adopted for obtaining the cities network. It starts by representing each city in terms of a respective street network, in which the nodes correspond to crossings of two or more streets, while the links stand for respective streets. Network measurements (topological and geometric) are obtained for each of the street network, and the similarity between each possible pair of cities is estimated in terms of the respective coincidence similarity index. The cities network is obtained by thresholding the coincidence values to remove the smallest coincidence links. This threshold can be chosen while taking into account the intended overall level interconnections (see also Sect. 4).

Figure 2 illustrates the estimation of the features network, starting from the weight matrices \(W_i\) corresponding to the networks obtained for each possible combination \(i\) \((i = 1 \ldots p = 31)\) of the adopted features. The coincidence method is then applied to quantify the similarity between each of these matrices, yielding the features network. Each node in the latter network corresponds to a specific feature combination, while the links between these nodes reflect the respective pairwise similarity. Community detection can then be applied on the obtained features networks to identify the possible models respective to the original data.

Another important characteristic of graphs is that the interconnectivity around each node can vary significantly. One popular respective measurement is the clustering coefficient (or transitivity) [16]. In this work, the standard deviation of the node transitivity is considered.

Each of the 20 considered cities was characterized in terms of the above-discussed measurements. To ensure more commensurate values between these five features, they are, respectively, standardized (e.g., [9]) so as to
Fig. 1 Diagram indicating how the cities networks is obtained. The $n = 20$ original cities, represented as maps here, are converted into respective street networks where each node corresponds to crossings of two or more streets, while the links stand for the streets themselves. Features, corresponding to five topological measurements, are then obtained for each of the street networks. The similarity between the cities is then inferred by using the coincidence index, leading to the respective cities network.

Fig. 2 Diagram illustrating the estimation of the features network. Weight matrices are obtained, using the coincidence methodology, for each of the possible feature combinations ($p = 31$). The coincidence index is then calculated for each pair of weight matrices, yielding the features network, in which each node correspond to the network with respective feature combination. The communities of the features network can then be detected and understood as possible models of the original data. In particular, the hub of each of the detected communities can be understood as respective prototypes.

Features have null mean and unit variance. The standardization of each of the features $f_i$ can be performed as

$$\tilde{f}_i = \frac{f_i - \mu_f}{\sigma_f},$$

where $\mu_f$ and $\sigma_f$ correspond to the mean and standard deviation of $f_i$, respectively.

The coincidence similarity between two non-zero, real-valued feature vectors $\vec{f}$ and $\vec{g}$ can be expressed \cite{9,13} as

$$C_R(\vec{f}, \vec{g}) = \mathcal{I}_R(\vec{f}, \vec{g}) \cdot \mathcal{J}_R(\vec{f}, \vec{g}),$$

where $\mathcal{I}_R(\vec{f}, \vec{g})$ is the interiority index between $\vec{f}$ and $\vec{g}$, calculated as

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

[9,13] Springer
and \( J_R(\vec{f}, \vec{g}) \) is the Jaccard index between the two non-zero, real-valued vectors, given as

\[
J_R(\vec{f}, \vec{g}) = \frac{\sum_i \text{sign}(f_i g_i) \min \{|f_i|, |g_i|\}}{\sum_i \max \{|f_i|, |g_i|\}}.
\] (8)

It is of particular interest to adopt the \( \alpha \) (0 ≤ \( \alpha \) ≤ 1) parameter \([9,13]\) as a means to control the contributions of the pairwise aligned and anti-aligned signs of the involved feature values on the overall coincidence result. This can be immediately implemented \([9,13]\) as

\[
C_R(\vec{f}, \vec{g}, \alpha) = I_R(\vec{f}, \vec{g}) J_R(\vec{f}, \vec{g}, \alpha),
\] (9)

where

\[
J_R(\vec{f}, \vec{g}, \alpha) = \frac{\sum_i \alpha |s_{f_i} + s_{g_i}| \min \{|f_i|, |g_i|\} - (1 - \alpha)|s_{f_i} - s_{g_i}| \min \{|f_i|, |g_i|\}}{\sum_i \max \{|f_i|, |g_i|\}}.
\] (10)

where \( s_{f_i} = \text{sign}(f_i) \). We also have that \(-2(1 - \alpha) ≤ J_R(\vec{x}, \vec{y}, \alpha) ≤ 2\alpha\).

When \( \alpha = 0.5 \) we have that \( J_R(\vec{f}, \vec{g}, \alpha) = J_R(\vec{f}, \vec{g}) \). For \( \alpha > 0.5 \), the pairwise features having the same sign will have greater contribution than those with opposite signs. The contrary effect is obtained when \( \alpha < 0.5 \). As a consequence, the overall degree of interconnection between the nodes of the resulting networks can be controlled by varying \( \alpha \), with more interconnected structures being obtained for larger values of \( \alpha \). It has been observed that \( \alpha < 0.5 \) tends to significantly enhance the modularity and levels of interconnectivity detail of the, respectively, obtained networks \([9,13]\).

### 4 Topological Study of European Cities

As a first step, we calculated the five features obtained from the street graph of each city. They are shown in Table 3, along with the the number of nodes and edges. As a preliminary visual inspection of this table will indicate, all adopted measurements contribute to the separation between groups, though with distinct effectiveness. For instance, the measurement \( \sigma_A \) resulted mostly similar within cities from the same country, while tending to be distinct between cities from different countries. In addition, \( \sigma_k \) tends to single out the British cities, while \( \langle k \rangle \) resulted more distinct for the German cities. The combination of the several obtained differentiating measurements adopted in the suggested coincidence-based approach can contribute to further enhancing the distinction between cities from different countries.

### Table 3

Features from each city considered in this work. For each city, the number of nodes (\( m \)), the number of edges (\( \text{ecount} \)), the average node degree (\( \langle k \rangle \)), the standard deviation of the node degrees (\( \sigma_k \)), the standard deviation of the clustering coefficient (\( \sigma_{CC} \)), the dispersion of the node positions (\( \text{disp}_{pos} \)), and the standard deviation of the node accessibilities (\( \sigma_A \)).

| City     | \( m \)  | \( \text{ecount} \) | \( \langle k \rangle \) | \( \sigma_k \) | \( \sigma_{CC} \) | \( \text{disp}_{pos} \) | \( \sigma_A \) |
|----------|----------|---------------------|---------------------|--------------|-----------------|---------------------|--------------|
| Nantes   | 6732     | 14622               | 4.295               | 1.688        | 0.142           | 0.274               | 10.249       |
| Bordeaux | 4332     | 8825                | 4.056               | 1.344        | 0.133           | 0.263               | 11.351       |
| Lille    | 3424     | 7054                | 4.092               | 1.358        | 0.151           | 0.286               | 11.821       |
| Rennes   | 4216     | 8984                | 4.216               | 1.615        | 0.153           | 0.294               | 11.597       |
| Lille    | 3424     | 7054                | 4.092               | 1.358        | 0.151           | 0.286               | 11.821       |
| Rennes   | 4216     | 8984                | 4.216               | 1.615        | 0.153           | 0.294               | 11.597       |
| Lille    | 3424     | 7054                | 4.092               | 1.358        | 0.151           | 0.286               | 11.821       |
| Rennes   | 4216     | 8984                | 4.216               | 1.615        | 0.153           | 0.294               | 11.597       |
Subsequently, aimed at obtaining a preliminary comparison reference for our studies, we obtained the principal component analysis (PCA, e.g., [24]) of the 20 cities, which is shown in Fig. 4, while taking into account all the respective five measurements. The total variance explanation accounted by the first two principal axes is relatively low (68%), indicating that the adopted measurements are little correlated one another, therefore effectively complementing the characterization of the city structures.

Three main clusters, shown in Fig. 3, were identified from the topological features of the considered cities by using the single-linkage agglomerative hierarchical clustering method (e.g., [14,22]). The three main clusters were obtained by cutting the obtained dendrogram at linkage distance 0.022.

The resulting distribution of the cities in the obtained PCA is not uniform, with a concentration being observed on the right-hand side. The three main obtained groups (A, B, and C) are also shown. Group A contains 4 British cities. Group B contains cities from Spain, Germany, and France, while Group C includes cities Spain, Germany and Italy. Groups B and C are, therefore, largely heterogeneous.

The transformation of the cities dataset into respective networks involves only two parameters: the overall threshold $T$, and the parameter $0 \leq \alpha \leq 1$ controlling the relative contribution of aligned and anti-aligned features signs (two types of joint variations). By testing several combinations of these two parameters, we identified that $T = 0.10$ constitutes a particularly adequate choice in the sense of yielding modular structures and enhancing the interconnectivity details for several values of $\alpha$. In particular, smaller values of $T$ will imply less interconnected networks. Once the small value of $T = 0.10$ is set, the greatest modularity can be obtained by varying the parameter $\alpha$ can then be. As discussed in Sect. 3, the parameter $\alpha$ can be used to control the overall level of interconnections between the nodes in the obtained networks, in the sense that the larger the value of $\alpha$, the more connected the obtained networks will be. Interestingly, the adoption of smaller values of $\alpha$ generally contributes to enhancing substantially the overall modularity and details in the obtained networks [9].

Another important characteristic observed from the adopted multi-alpha analysis is that the connections established for a given $\alpha$ will be necessarily preserved for larger values of that parameter. Therefore, ‘early’ connections obtained for a relatively small value $\alpha_1$ can be understood as being stronger and more stable in more interconnected networks obtained for larger values of $\alpha > \alpha_1$.

Figure 5 shows the cities networks obtained for $T = 0.10$ and $\alpha = 0.25, 0.32, 0.39, 0.46, 0.53, 0.6$. The five node colors identify the respective countries to which the cities belong. As expected, little connected networks have been obtained for the two smallest values $\alpha_1$ (a),(b). However, these two networks also contain four connected components presenting a manifest homogeneity, in the sense of including cities mostly from the same country, as can be appreciated from the colors within each component. Observe that connected components are particular cases of modules, obtained when the respective modularity is particularly high.

As $\alpha$ is increased (Fig. 5c), three of the four groups in (b) merge into a major community, with the remainder group corresponding mostly to the British cities. After increasing $\alpha$ further (d), all groups coalesce into a single component, with the community containing mostly British cities connected through a single link to the remainder of the network. As $\alpha$ is then increased...
Fig. 4  Principal component analysis of all (5) features considered. The axes shown in the figure correspond to the first two principal components. The percentage of variance explained by each axis is shown, respectively (a) $\alpha = 0.25$, (b) $\alpha = 0.32$, (c) $\alpha = 0.39$, (d) $\alpha = 0.46$, (e) $\alpha = 0.53$, (f) $\alpha = 0.60$.

Fig. 5  Cities networks obtained for different values of $\alpha$ by the coincidence index calculation. The same edge threshold of $T = 0.10$ was considered for all graphs. As expected, the overall connectivity tends to increase with $\alpha$. Particularly detailed networks were obtained for the smaller values of $\alpha$. 

France
Germany
Italy
Spain
UK
to obtain the networks in (e) and (f), the respective networks become more and more interconnected at the expense of the respective modularity and level of details, which are both substantially decreased. Even so, many neighboring cities in (e) and (f) tend to be from the same country.

It is also interesting to keep in mind that substantially more information and insights can be obtained by considering the multi-alpha analysis such as that shown in Fig. 5, than by considering a single value of \( \alpha \). Indeed, in addition to the already observed possibility of identifying the strongest links appearing for the first values of \( \alpha \), the multi-alpha analysis also plainly indicates how the initial modules progressively merge, while their relative adjacencies are mostly maintained as \( \alpha \) increases.

The obtained results corroborate the critical role of the parameter \( \alpha \), respectively, to the obtained highly modular networks with detailed interconnectivity. Indeed, had only the standard configuration of \( \alpha = 0.5 \) been used, the obtained result would be characterized by almost no modularity and little level of details. It was only thanks to the possibility to vary \( \alpha \) to smaller values that the identification of highly modular and detailed networks has become possible.

A particularly remarkable result observed in Fig. 5 is the noticeable uniformity of the communities obtained for small values of \( \alpha \) resulting from networks (a) and (b). This result indicates that the adopted features and network construction methodology were accurate enough to reveal impressive levels of topological homogeneity between cities from a same country. This important result motivated us to proceed further in the sense of trying to identify the value of \( \alpha \) leading to the greatest possible modularity given the adopted features. To do so, we considered several values of \( \alpha \) between 0.25 and 0.6 (with resolution 0.07) and calculated the respective modularity (e.g., [23]), so that its maximum could be identified. The modularity was calculated using the country membership as reference, in the sense that the maximum modularity would correspond to obtaining completely homogeneous communities.

Figure 6 depicts the modularity obtained for the considered cities and features in terms of the parameter \( \alpha \).

Figure 7a shows the relative frequency histogram of the coincidence similarity values obtained for the maximally modular network (\( \alpha_M = 0.29 \)). Given that \( \alpha < 0.5 \), implying the positive joint feature signs to be penalized, the obtained average becomes negative which, at least for this case, results in a more detailed network. The relative frequency histogram obtained for \( \alpha = 0.5 \), shown in Fig. 7b, has mean close to zero and standard deviation comparable to that obtained for the maximally modular network, therefore indicating the presence of higher coincidence values, which tends to yield a more densely connected network with less details and smaller modularity. Observe also the markedly distinct shapes of the two obtained relative frequency histograms, implied by the two distinct values of \( \alpha \). This corroborates the important fact that the variation of the parameter \( \alpha \) has an effect that goes beyond transforming the coincidence values in a trivial manner (e.g., shifting or scaling).

As could be expected from our previous experiment involving 6 values of \( \alpha \), the modularity decreases substantially for the larger values of \( \alpha \), with a respective maximum being observed at \( \alpha_M = 0.29 \). The respectively defined network, characterized by the maximum possible overall modularity for the considered cities and features, is shown in Fig. 8.

It should be observed that the fact that all the considered cities have been chosen from 5 European countries actually tends to contribute to the uniformity between the cities. In addition, the fact that the groups in Fig. 8 were obtained while maximizing the respective modularity by changing two parameters would by no means \( a \ priori \) imply that a relatively high modularity value could be achieved, because this result would require not
The cities network obtained using the coincidence methodology with $T = 0.10$ and $\alpha = 0.29$, which implied a modularity of 0.31. The nodes represent the cities and the connections widths are proportional to the values of the coincidence index between the features of the two respective cities. Four separated components have been obtained, which can be mostly associated with the United Kingdom (I), Germany (II), France (III), and Italy (IV) only the original cities obtained from a same country to be similar while differing from the other groups, but also that these relationships were properly captured by the adopted measurements.

To provide some indication about the significance of the obtained maximum modularity, the country own-<sup>1</sup>erships, i.e., the set of nationalities of the cities, were assigned across the cities in uniformly random manner, and the respective modularities were quantified. Figure 9 depicts the histogram of the values of modularities obtained in this experiment together with a red vertical line indicating the maximum modularity obtained for the coincidence similarity. The marked difference between the maximum and shuffled modularity values supports the fact that the observed countries uniformity is unlikely to be obtained by chance.

The maximally modular network in Fig. 8 can now be compared to the PCA in Fig. 4. Other than the group A containing only British cities, the marked modularity obtained by the coincidence method cannot be directly inferred from PCA results, which substantiates the potential of the reported methodology for characterizing the interrelationship between the data elements (cities). Indeed, though the adopted PCA approach considers all the five original features, it is intrinsically limited to being a two-dimensional projection of the original data elements that cannot preserve all the original information. Conversely, the links in the obtained maximally modular network take into account all the five original features while not involving any projection or information loss other than those implied by the adopted (optional) thresholding.

To conclude this section, it is interesting to make some additional considerations regarding the remarkable obtained cities network presenting enhanced modularity and uniformity regarding respective countries. Indeed, there are several conditions necessary for reaching this result, including an appropriate selection of features capable of effectively describing the distinctive properties of the cities, as well as the adoption of

**Fig. 8** The cities network obtained using the coincidence methodology with $T = 0.10$ and $\alpha = 0.29$, which implied a modularity of 0.31. The nodes represent the cities and the connections widths are proportional to the values of the coincidence index between the features of the two respective cities. Four separated components have been obtained, which can be mostly associated with the United Kingdom (I), Germany (II), France (III), and Italy (IV).

**Fig. 9** Distribution of the modularity valued for 3000 permutations of the cities among the considered countries. The red line indicates the maximum modularity value (0.31) attained by using the coincidence-based approach described in the present work.

The network in Fig. 8 presents four major connected components, which have been labeled as I, II, III, and IV by decreasing interconnectivity (total strength). Four of the cities in group I are British, with the Spanish city of Vigo being also included. The cities in group II are predominantly German, also encompassing two Italian cities. Three French cities, plus the Spanish city of Granada, compose group III. The last group, IV, contains two Italian and one Spanish city. Three cities from distinct countries remained isolated for this value of $\alpha$: Augsburg, Nantes, and Mostoles. Therefore, these four obtained groups can be associated with respective countries as: I ↔ Britain; II ↔ Germany; III ↔ France, IV ↔ Italy.

The varying uniformity of the obtained groups, respectively, to country membership is expected as a consequence of the fact that, even if the cities of each country were indeed similar one another while differing between countries, they would still likely present structural differences within each country. Therefore, distinct samples of cities by country are likely to be characterized by intrinsic variations, which could explain why some groups resulted more uniform than others.

It is also interesting to observe that the cities appearing in minority in the four main groups very probably present values of the adopted features that make them indeed more similar to the groups to which they have been, respectively, assigned.

To conclude this section, it is interesting to make some additional considerations regarding the remarkable obtained cities network presenting enhanced modularity and uniformity regarding respective countries. Indeed, there are several conditions necessary for reaching this result, including an appropriate selection of features capable of effectively describing the distinctive properties of the cities, as well as the adoption of
a sound and strict methodology for translating from these features to a respective network with high uniformity and homogeneity. Plainly, these two conditions have been mostly met in the reported approach: (i) the adopted five features, despite their relatively small number, provided a specific characterization of the cities; and (ii) the coincidence method confirmed its tendency to provide sound and strict characterization of the interrelationships between the original data elements (cities), yielding a highly modular and detailed network representation of the relationships between the considered cities.

However, there is a third condition for obtaining the remarkable results in Fig. 8, and this consists in the fact that the adopted cities need to present, from the outset, distinctive characteristics between cities of different countries while presenting homogeneity regarding a respective country. The obtained results corroborate this interesting possibility, suggesting that European cities (at least those considered here) tend to present surprising homogeneity within a same country, while being relatively different between countries. This interesting tendency may be related to intrinsic climates, distinct planning traditions, as well as intrinsic socio-economic characteristics.

5 Features interrelationship

It is well known (e.g., [14, 26]) that the features adopted for characterizing data elements can greatly influence and even determine the results of respective classifications. This problem is so important that whole areas, including feature selection (e.g., [8, 25, 27]), are to a large extent dedicated to studying how the choice of different types of features can impact on pattern recognition and machine learning.

Having obtained remarkable results while translating the 20 considered cities into a respective modular and detailed network, it remains an equally interesting problem to study to which an extent the adopted features contributed to the reported results. Interestingly, it is possible to apply the same coincidence methodology employed to obtain the networks also for this finality [10]. More specifically, networks are obtained, respectively, to several combinations of features, and the coincidence method is then applied over the obtained weight matrices so as obtain a new network in which each node corresponds to one of the feature combination network, while the links correspond to the pairwise similarity between the weights of respective matrices. This interesting issue constitutes the main subject of the present section.

First, we obtained networks for each of the $p = 31$ possible combinations (2^5, except for the null combination) of the adopted features. Then, the coincidence method with $T = 0.10$ and $\alpha = 0.29$ was applied considering the obtained weight matrices as features. The thus obtained features network is shown in Fig. 10.

![Fig. 10](image-url)

Network of feature combinations. Each node corresponds to a configuration of features, while the edges reflect the coincidence index between the two coincidence graphs. The widths of the links are proportional to the respective pairwise coincidence values. The communities were detected using the approach described in [5]. The hub within each identified community is shown as a triangle.

![Fig. 11](image-url)

Cities networks corresponding (from left to right) to the hub, as well as the nodes with the second and third highest strengths in each of the communities identified in the features network. Each of the four rows corresponds to the respective communities I–IV in Fig. 10. The cities networks along each of the rows present a marked mutual similarity.
Interestingly, the resulting network, henceforth called features network, also resulted strongly modular (modularity equal to 0.448), with four well-defined communities (labeled W, X, Y, and Z) having been detected by the multilevel community finding method [5].

In a similarity network with well-defined modularity, each of the nodes belonging to any of its communities will tend to be similar to the other nodes in that same community. Indeed, this corresponds to one of the possible rationales behind the very concept of network modularity. In this sense, each of the nodes in each of the obtained four communities in Fig. 8 corresponds to respectively similar cities. This important property allows us to conclude that, in the case of the 20 adopted cities, there are four possible main respective representations or models of the cities as networks, corresponding to each of the four modules, identified as W, X, Y, and Z. Given that the hub within each of the detected communities is the node most intensely interconnected to the others in the same community, it becomes possible to consider that hub as a prototype of the respective cities networks represented by that community. The hubs identified, respectively, to each of the four communities in Fig. 8 can be identified as having triangular shape. Interestingly, the four obtained hubs can be found to be interconnected along a square motif corresponding to the intersection between the four detected communities.

Each of the rows in Fig. 11 presents the cities networks corresponding to the respective prototype (hub), as well as the nodes with the second and third highest strengths within each of the identified four communities. For simplicity’s sake, the networks have been shown in circular format, with the nodes following the same order as in Table 2, while the countries are identified by respective colors. Therefore, most of the adjacent nodes tend to be from the same respective country. The four obtained hubs correspond to the features combinations (1, 2, 3), (1, 2, 3, 4), (1, 2, 3, 5), and (1, 2, 3, 4, 5), corresponding in the case of the specific example to the nodes with maximum number of features within each, respectively, obtained community. In addition, as it can be readily appreciated from Fig. 11, the cities networks obtained for each of the four modules (rows) are remarkably similar one another, confirming the effectiveness of the coincidence methodology for representing the similarity between the structures of the considered networks.

Figure 12 presents histograms of how many times each feature appeared within each of the four identified communities. These results indicate that the main difference between the four modules relates to features 4 and 5, namely the dispersion of point locations and standard deviation of the accessibility. Indeed, each of the four modules are characterized by the four combinations while taking features 4 and 5. The histograms also indicate that the three other features, namely the average and standard deviation of the degree and the local transitivity, are shared by all the four main obtained modules.

6 Concluding remarks

Cities can be understood as organic entities, in the sense that they are born and then keep adapting to the environment, as well as to intrinsic demands including effective transportation, basic resources and infrastructure, etc. Given that these two factors can vary from city to city, or even from country to country, it becomes an interesting research subject to characterize their properties while trying to establish similarity interrelationships. In addition to contributing to a better knowledge about cities, the identification of similarities paves the way for sharing urbanistic, administrative, and planning experiences.

The present work applied the recently introduced concept of coincidence similarity, consisting of a combination of the interiority and Jaccard index, as the means not only for transforming sets of cities characterized by respective features into respective networks but also for studying the effect of choice of these features on the obtained results. Twenty European cities with comparable populations were selected and characterized in terms of five respective features, four topological and one geometric. Except for the group of British cities, no well-defined grouping could be observed from the traditional PCA methodology.

The coincidence methodology was then applied, and the decisive effect of its parameter α in obtaining detailed networks illustrated, respectively, to six uniformly distributed values between 0.25 and 0.60. Then, by taking into account the countries of the cities as cat-
egories, we found the value of $\alpha$ that optimized the overall modularity. Interestingly, this value (equal to 0.29) resulted markedly distinct from the reference value of 0.5 that would be otherwise implied in case the parameter $\alpha$ had not been taken into account. This result corroborates the fact that modular and detailed representations of the cities could not have been obtained were not for the possibility to try different values of $\alpha$. The maximally modular cities network obtained was characterized by four completely separated components, each of which presenting majority of cities from a same respective country.

To complement our analysis, we applied the coincidence methodology on the weight matrices associated to the obtained cities networks while considering all possible combinations of the five adopted features. This procedure resulted in a network whose each node corresponds to a cities network obtained by each of the possible feature combination. Interestingly, this network resulted markedly modular, containing four well-defined communities which can be understood as the main possible data models given the adopted features.

The obtained features network allowed interesting insights regarding the effect of the features on the cities networks. In particular, all networks in a same of these communities will by construction have similar topology, allowing all these possible cases to be represented in terms of a model or prototype network, which was chosen to correspond to the respective hub. Consequently, the four identified hubs can be understood as representing the main four models representing the considered cities. In addition, by comparing the features in each of the four obtained communities, it has been possible to infer their respective influence on the obtained results. In particular, we observed that all main four models share the use of the first three features, which seem to be directly related to the obtained country-specific modularity. Interestingly, the consideration of the two remaining features then allowed a respective subdivision into the four obtained models.

In addition to their specific contributions related to cities characterization, the reported study and results also corroborated the potential of the coincidence methodology for yielding particularly detailed and modular networks when mapping datasets into networks. It also confirmed the advantage of being able to control, through the parameter $\alpha$, the contributions of sign aligned and anti-aligned pairs of features, as a critical resource for allowing the identification of maximal modularity. The potential of the application of the coincidence methodology for studying the effect of feature combinations was also further substantiated by the described results.

The reported results paved the way to a number of related further studies. For instance, it would be interesting to consider other possible topological and geometrical features. It would also be of particular interest to apply the described similarity approach to characterize parts of cities instead of their whole. The interesting finding that European cities from a same country seem to present some uniformity could also be further investigated by considering not only additional European cities, but also samples from other continents. Another particularly promising possibility would be to identify which combination among the adopted features yields the most homogeneous modules regarding the respective countries. This could be achieved by systematic consideration of the combinations of features or by applying some optimization approach such as gradient descent.

Acknowledgments E.K.T. acknowledges FAPESP Grant No. 2019/01077-3 and L. da F.C. thanks CNPq Grant No. 307085/2018-0 and FAPESP 2015/22308-2 for sponsorship. The authors also thank Filipi N. Silva, for publicly providing (GitHub) the source code for the accessibility computation.

Author contributions All authors have contributed equally to the paper.

Data availability statement This manuscript has no associated data or the data will not be deposited. [Authors’ comment: Original data from OpenStreetMap.]

References

1. C.E. Akbas, A. Bozkurt, M.T. Arslan, H. Aslanoglu, A.E. Cetin, L1 norm based multiplication-free cosine similarity measures for big data analysis, in 2014 International Workshop on Computational Intelligence for Multimedia Understanding (IW CIM), IEEE, pp. 1–5 (2014)

2. M. Barthelmy, The Structure and Dynamics of Cities (Cambridge University Press, Cambridge, 2016)

3. M. Batty, The New Science of Cities (MIT press, New York, 2013)

4. A. Benatti, L.d.F. Costa, Accessibility: A Generalization of the Node Degree (A Tutorial) (2021) (arXiv)

5. V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. 2008(10), P10008 (2008)

6. S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Complex networks: structure and dynamics. Phys. Rep. 424(4–5), 175–308 (2006)

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

8. G. Chandrashekar, F. Sahin, A survey on feature selection methods. Comput. Electric. Eng. 40(1), 16–28 (2014)

9. L.d.F. Costa, Coincidence complex networks. ResearchGate (2021). https://doi.org/10.13140/RG.2.2.13469.23521/2

10. L.d.F. Costa, Elementary particles networks as revealed by their spin, charge and mass. ResearchGate (2021). https://doi.org/10.13140/RG.2.2.20598.73282/3
