Journal of Physics: Complexity

OPEN ACCESS

PAPER

Coincidence complex networks

Luciano da Fontoura Costa*

São Carlos Institute of Physics-DFCM, University of São Paulo, PO Box 369, São Carlos, SP, 13560-970, Brazil

* Author to whom any correspondence should be addressed.

E-mail: luciano@ifsc.usp.br

RECEIVED
3 December 2021

REVISED
25 January 2022

ACCEPTED FOR PUBLICATION
14 February 2022

PUBLISHED
9 March 2022

Keywords: clustering, complex networks, data analysis, similarity indices

Abstract

Complex networks, which constitute the main subject of network science, have been wide and extensively adopted for representing, characterizing, and modeling an ample range of structures and phenomena from both theoretical and applied perspectives. The present work describes the application of the real-valued Jaccard and real-valued coincidence similarity indices for translating generic datasets into networks. More specifically, two data elements are linked whenever the similarity between their respective features, gauged by some similarity index, is greater than a given threshold. Weighted networks can also be obtained by taking these indices as weights. It is shown that the two proposed real-valued approaches can lead to enhanced performance when compared to cosine and Pearson correlation approaches, yielding a detailed description of the specific patterns of connectivity between the nodes, with enhanced modularity. In addition, a parameter α is introduced that can be used to control the contribution of positive and negative joint variations between the considered features, catering for enhanced flexibility while obtaining networks. The ability of the proposed methodology to capture detailed interconnections and emphasize the modular structure of networks is illustrated and quantified respectively to real-world networks, including handwritten letters and raisin datasets, as well as the *Caenorhabditis elegans* neuronal network. The reported methodology and results pave the way to a significant number of theoretical and applied developments.

1. Introduction

Though relatively recent, the area of network science (e.g. [1–4]) has established itself not only from the perspective of theoretical contributions, but also as a consequence of its ample range of applications to the most diverse areas and problems [4]. Basically, a complex network is a graph with structure substantially distinct from simpler, more regular, counterparts (e.g. [5]).

The representation of a structure as a complex network demands the adoption of some approach for defining the network nodes, and then the application of some objective criterion for linking these nodes. Indeed, intrinsic interrelationships have been identified between connectivity, similarities and features [6]. In particular, the potential of network science can be extended even further provided datasets can be effectively translated into respective networks, from which a wealthy of informative measurements can then be extracted (e.g. [3]).

Often employed approaches include the consideration of measurements such as the cosine similarity and Pearson correlation, as well as the adoption of set-based alternatives, such as the Jaccard index (e.g. [7–11]), which has been predominantly employed for data involving binary or categorical data. In the case of categorical data, approaches such as the traditional Jaccard index have constituted an interesting option (e.g. [12]). Though it is also possible to define connectivity with basis on differences and distances, these measurements often need to be mapped into similarities in order to be taken as respective weights.

More recently, the Jaccard index has been generalized to real-valued data [13–16] through a multiset-based approach. Multisets (e.g. [17–22]) constitute an interesting approach to set theory that allows repeated entries of elements, with their repetitions being referred to as *multiplicity*. Observe that the ability to operate

on negative values is often required in cases where the data has been preliminary normalized, e.g. through standardization.

However, the classic Jaccard index is not able to take into account the *relative interiority* of the two compared sets, which motivated the proposal [13, 16] of the *coincidence index*, corresponding to the combination, more specifically the product, between the interiority (or homogeneity) and the Jaccard indices.

Recent results [13, 16, 23] have indicated that product-based similarity indices, which include both the cosine similarity and Pearson correlation coefficient, tend to be relatively less strict regarding similarity quantification than non-bilinear indices based on combination of the minimum and maximum operations, such as the Jaccard and coincidence indices. The real-valued Jaccard, and in particular the coincidence, indices tend to allow substantially more detailed and selective similarity quantification respectively to comparison of signals, being capable of removing secondary, smaller spatial scale noise and structure while emphasizing the coincidence matches in terms of sharp, narrow matching peaks [23, 24].

Another interesting problem in network science concerns the issue of, given a network, to highlight and/or identify its respective modular structure (e.g. [2]). Though several methods have been proposed for community finding, this problem remains to a great extent open because of intrinsic difficulties. Remarkably, the real-valued Jaccard and coincidence indices-based approach reported in the current work present encouraging potential for enhancing the modular structure of existing networks accordingly to some specific set of measurements.

In this work, we aim at studying the potential of the real-valued Jaccard and the real-valued coincidence similarity indices as the basis for mapping datasets, involving one or more features corresponding to numeric values, into respective complex networks. It is shown not only that the proposed method tends to allow good levels of detail about the specific interconnecting structure of datasets, but also present potential for highlighting the network modularity. In particular, we apply the proposed method to real-world datasets of handwritten letters and raisins [25, 26]. In addition to visualizing the networks obtained by considering the cosine similarity, Pearson correlation, real-valued Jaccard and real-valued coincidence methods, we also report a quantitative comparison regarding the specificity of the indices as well as the obtained modularity.

The real-valued Jaccard and real-valued coincidence methods for translating datasets into networks are found to provide substantially better performance than the cosine similarity and Pearson correlation coefficient. In addition, as expected, by incorporating additional information about the relative interiority between the two compared datasets, the real-valued coincidence approach tended to outperform the real-valued Jaccard approach for generating networks from datasets.

The interesting issue regarding the relationship between the real-valued Jaccard and real-valued coincidence methods and network modularity was considered further by starting not from features of the elements in datasets, but from topological measurements (e.g. shortest path distances) derived from networks described by their respective adjacency matrix (0 or 1 values). Encouraging preliminary results were obtained respectively to networks derived from the raisin dataset as well as to the *Caenorhabditis elegans* neuronal network [27–29] while considering both its undirected and directed respective versions.

This article starts by presenting the main basic concepts, and follows by presenting the real-valued Jaccard and real-valued coincidence similarity indices, as well as the respective methodology for translating datasets into networks. The potential of this methodology is then illustrated respectively to real-world datasets, with the important relationship between the coincidence measurement, interconnectivity detail, and modularity being addressed subsequently.

2. Basic concepts

A *complex network* is basically a graph composed of N nodes (or vertices) and E connections (or edges, or links). Complex networks are often represented in terms of the respective adjacency matrices or edges lists. The former of these possible representations involves using a matrix A where a connection from a node j to a node i implies $A\left[i,j\right] = 1$, with $A\left[i,j\right] = 0$ being otherwise enforced. In case $A\left[i,j\right] = A\left[j,i\right]$, $\forall\ i,j$, the network is said to be *undirected*. Complex networks involving weights associated to their respective links can be represented by respective weight matrices.

The henceforth considered *datasets* are understood to incorporate N *data elements*, observations, samples, or individuals, characterized in terms of M respective measurements or *features*. Each of these datasets can be organized as a table or $N \times M$ matrix where each of the N rows corresponds to a data element, while the columns contain the respective measurements or features x_i, $i = 1, 2, \ldots, M$, each of which represented in terms of respective *feature vectors*.

In order to avoid biases typically implied by varying magnitudes between the several features, the statistic linear transformation of *standardization* (e.g. [30]) is frequently adopted, in independent manner, over each

feature f prior to similarity quantification. This transformation can be implemented as:

$$\tilde{f}_k = \frac{f_k - \mu_f}{\sigma_f}, \tag{1}$$

where μ_f and σ_f are the mean and standard deviation of the feature f taken along all data elements $k = 1, 2, \ldots, N$.

It can be shown that each of the resulting features will have zero means and unit standard deviation (e.g. [31]). In addition, most of the feature values become comprised within the interval $[-2, 2]$, depending on their respective statistical distributions. Conveniently, the newly obtained variables also become non-dimensional. Observe that the standardization ensures a bijective mapping between the original and normalized datasets, therefore preserving the features information.

The application of the standardization procedure intrinsically implies that any index or measurement adopted to quantify the similarity between data elements will need to be able to cope with negative values, which is not the case of the classic Jaccard index, or even its multiset generalization to non-negative values.

Possibly one of the most frequently adopted means for quantifying the similarity between network nodes consists of the *cosine similarity*, which relies on the inner product between two feature vectors $\vec{x}$ and $\vec{y}$, being expressed as:

$$\langle \vec{x}, \vec{y} \rangle = \sum_{i=1}^{N} x_i \, y_i = |\vec{x}||\vec{y}| \cos(\theta) \Rightarrow \cos(\theta) = \frac{\langle \vec{x}, \vec{y} \rangle}{\left|\vec{x}\right| \left|\vec{y}\right|}, \tag{2}$$

where $\langle \vec{x}, \vec{y} \rangle$ is the inner product between the vectors $\vec{x}$ and $\vec{y}$, $|\vec{x}|$ and $|\vec{y}|$ are their magnitudes, and $\cos(\theta)$ is the cosine similarity. We also have that $-1 \leqslant \cos(\theta) \leqslant 1$.

Observe that the cosine similarity does not take into account the original magnitudes of the two vectors, therefore losing part of the information about the relationship between the elements in the original dataset.

Another closely related similarity measurement, referring to two nodes X and Y described by respective feature vectors $\vec{x}$ and $\vec{y}$, each with N components, is the *Pearson correlation coefficient*, expressed as:

$$P(X, Y) = \frac{\sum_{i=1}^{N} \left[(x_i - \bar{x}) \left(y_i - \bar{y} \right) \right]}{\sqrt{\left[\sum_{i=1}^{N} (x_i - \bar{x})^2 \sum_{i=1}^{N} \left(y_i - \bar{y} \right)^2 \right]}} \tag{3}$$

with $-1 \leqslant P(X, Y) \leqslant 1$. Observe that the Pearson correlation coefficient can be strongly influenced by outliers (e.g. [32]) as well as by small numbers of samples.

3. Real-valued Jaccard and coincidence similarity indices

The *classic Jaccard index* [7, 10, 13]) for quantifying the similarity between two sets A and B can be defined as follows:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}, \tag{4}$$

where $|A| \geqslant 0$ is the cardinality of set A. Observe that $0 \leqslant J(A, B) \leqslant 1$.

A generalization of the Jaccard index [13, 15, 16] by employing multiset concepts (e.g. [18–22, 33, 34]), henceforth referred to as *real-valued Jaccard similarity index*, allows real, possibly negative data values (multiset multiplicities) to be taken into account. The real-valued Jaccard similarity index between two multisets or feature vectors $\vec{x}$ and $\vec{y}$ can be expressed as:

$$J_{\mathrm{R}}(\vec{x}, \vec{y}) = \frac{\sum_{i \in S} s_{x_i y_i} \min \left\{ s_{x_i} x_i, s_{y_i} y_i \right\}}{\sum_{i \in S} \max \left\{ s_{x_i} x_i, s_{y_i} y_i \right\}} \tag{5}$$

with $-1 \leqslant J_{\mathrm{R}}(\vec{x}, \vec{y}) \leqslant 1$ and where x_i and y_i correspond to the multiplicities of the vectors $\vec{x}$ and $\vec{y}$, $s_x = \mathrm{sign}(x)$, $s_y = \mathrm{sign}(y)$, $s_{xy} = s_x s_y$, and S is the combined support of the multisets, which generally corresponds to the union of the abscissae in $\vec{x}$ and $\vec{y}$. Signed similarity indices, including the above, have been previously reported [35] from the perspective of the $L1$ norm.

In addition, the numerator in equation (5) has been shown to correspond to the signed intersection ($\sqcap$) between two real-valued multisets [15]:

$$x \sqcap y = \sum_{i \in S} s_{x_i, y_i} \min \left\{ s_{x_i} x_i, s_{y_i} y_i \right\} \tag{6}$$

required so that $x \sqcap \Phi = \Phi$, where Φ is the empty multiset.

Another possibility for taking into account the real values signs is as described in [36] in the context of the $L1$ norm:

$$s_+ = \sum_{i \in S} \left| s_{x_i} + s_{y_i} \right| \min \left\{ s_{x_i} x_i, s_{y_i} y_i \right\} \tag{7}$$

which allows only the cases in which the feature signs are aligned (i.e. $s_x = s_y$) to contribute to the overall sum in equation (5). The situation in which $s_x = -s_y$ can also be addressed by using the following quantity:

$$s_- = \sum_{i \in S} \left| s_{x_i} - s_{y_i} \right| \min \left\{ s_{x_i} x_i, s_{y_i} y_i \right\} \tag{8}$$

which takes into account only the reversely aligned signs among each two instances of the same measurement.

The two above terms can then be combined as:

$$s_{\pm,\alpha} = [\alpha]\, s_+ - [1 - \alpha]\, s_-, \tag{9}$$

where $0 \leqslant \alpha \leqslant 1$ is a real parameter controlling the relative weights of the aligned and anti-aligned terms s_+ and s_-.

The real-valued Jaccard index, parametrized by α, can now be expressed as:

$$J_{\mathrm{R}}(\vec{x}, \vec{y}, \alpha) = \frac{s_{\pm,\alpha}}{\sum_{i \in S} \max \left\{ s_{x_i} x_i, s_{y_i} y_i \right\}} \tag{10}$$

with $-2(1 - \alpha) \leqslant J_{\mathrm{R}}(\vec{x}, \vec{y}, \alpha) \leqslant 2\alpha$, so that the range of the real-valued Jaccard similarity index is given as:

$$\begin{aligned} &J_{\mathrm{R}}(\vec{x}, \vec{y}, \alpha)_{\max} - J_{\mathrm{R}}(\vec{x}, \vec{y}, \alpha)_{\min} \\ &\quad = 2\alpha - [-2(1 - \alpha)] = 2. \end{aligned} \tag{11}$$

It can be verified that $\alpha = 0.5$ implies that:

$$J_{\mathrm{R}}(\vec{x}, \vec{y}, \alpha = 0.5) = J_{\mathrm{R}}(\vec{x}, \vec{y}) \tag{12}$$

with $-1 \leqslant J_{\mathrm{R}}(\vec{x}, \vec{y}, \alpha = 0.5) \leqslant 1$.

There are many other ways in which the s_+ and s_- indices can be combined, and it is also interesting to consider them separately, or jointly with other indices. In addition, it is possible to separate the sign contributions in more than two cases. For instance, we could have four cases with respective selectors: $(+,+)$; $(+,-)$; $(-,+)$; and $(-,-)$, and so on. Analogue schemes can be employed respectively to other similarity measurements.

The *interiority index* (or overlap e.g. [37]) corresponds to another measurement which aims at expressing how much one of the two sets is contained in the other (commutatively). In its traditional version considering set cardinality, this index is expressed as:

$$I(A, B) = \frac{|A \cap B|}{\min \{|A|, |B|\}} \tag{13}$$

with $0 \leqslant I(A, B) \leqslant 1$.

When adapted to real values [13, 16], the interiority index between any two multisets becomes:

$$I_{\mathrm{R}}(\vec{x}, \vec{y}) = \frac{\sum_{i \in S} \min \left\{ s_{x_i} x_i, s_{y_i} y_i \right\}}{\min \left\{ \sum_{i \in S} s_{x_i} x_i, \sum_{i \in S} s_{y_i} y_i \right\}} \tag{14}$$

again with $0 \leqslant I_{\mathrm{R}}(\vec{x}, \vec{y}) \leqslant 1$.

Unlike the Jaccard index, the interiority measurement takes into account, for normalization purposes, the smallest size between the two multisets being compared. At the same time, the classic Jaccard index, as well as its generalization to multisets and real-values, have been shown not be able to take into account the relative interiority of the two sets [13].

In oder to combine the best features of each of the above two measurements, the *coincidence index* between any two multisets (or vectors) $\vec{x}$ and $\vec{y}$ has been proposed [13, 14] as corresponding to the product of the respective interiority and Jaccard indices. Therefore, we have that the *real-valued coincidence index* between any two multisets or vectors $\vec{x}$ and $\vec{y}$ can be calculated as:

$$C_R(\vec{x}, \vec{y}) = I_{\mathrm{R}}(\vec{x}, \vec{y})\; J_{\mathrm{R}}(\vec{x}, \vec{y}) \tag{15}$$

with $-1 \leqslant C(\vec{x}, \vec{y}) \leqslant 1$.

The coincidence index can also be parametrized by α as:

$$C_R(\vec{x}, \vec{y}, \alpha) = I_R(\vec{x}, \vec{y})\ J_R(\vec{x}, \vec{y}, \alpha) \tag{16}$$

with $-2(1-\alpha) \leqslant C_R(\vec{x}, \vec{y}, \alpha) \leqslant 2\alpha$.

The coincidence index tends to provide one of the most strict and detailed quantification of the similarity between two mathematical structures, having allowed enhanced performance respectively to pattern recognition [23, 24] and hierarchical clustering [38]. In the former case, the coincidence index was shown [23, 24] to be able to yield sharp, narrow peaks indicating the matches between a template and an object signal, while substantially attenuating secondary, small scale noise and otherwise unwanted background structure.

As such, the coincidence index can be understood as a binary operator (in the mathematical sense of taking two arguments) that combines low and high-pass filtering in order to best suit the pattern recognition objectives. This important feature is a direct consequence of the non-bilinearity of the operations maximum and minimum which are part of the definition of the coincidence index.

Also relevant is the fact that the normalization in equation (5) yields a similarity profile directly related to the generalized Kronecker delta function [15, 16]:

$$\delta_{i,j}^{\pm} = \text{sign}(i\ j)\ \delta_{|i|,|j|} \tag{17}$$

which implements the most strict quantification of the similarity between two numeric values.

4. Building coincidence networks

Given a dataset, we aim at obtaining a respective representation as a complex network. The commonly adopted procedure consists of understanding each data element as a node, while the links are established based on the similarity (or difference) between the features of the respective data elements.

Frequently applied methodologies rely direct or indirectly on the cosine similarity or Pearson correlation between the features of pairs of data elements. In the present work, we focus on the real-valued Jaccard and coincidence indices, especially on their versions parametrized by α.

The basic methodology proposed in this work for transforming datasets into networks consists of: (i) standardizing the original dataset; and (ii) obtaining the pairwise similarities by using the real-valued Jaccard or coincidence indices. The cosine similarity and Pearson correlation are also considered as references for comparison.

The adjacency matrix of the resulting network can then be obtained by thresholding the respective similarity matrix (understood as a weight matrix) by a value T. It is also possible to preserve the weights of the thresholded links, resulting in the representation of the original dataset in terms of a respective weight network, in which case thresholding will not be necessary.

A welcomed characteristic of the proposed methodology to translate datasets into networks is that it involves only two parameters, namely α and the threshold T.

Given that α controls the relative contribution of the joint signs of the features, it will typically have a major effect on the respectively obtained distribution of similarity values. Therefore, it is interesting to consider several values of α so as to identify the respective setting that is more suitable for each problem. For instance, several values of α can be checked in order to identify which can lead to the largest modularity (e.g. [39, 40]). In addition, the consideration of sequences of progressively increasing values of α, as illustrated in section 5, can provide interesting insights about how the topology of the obtained networks progressively changes as α is increased. Observe that larger values of α tend to imply more connected networks.

Regarding the threshold T, it influences the results in the sense that larger values will yield overall less connected networks. Observe that, unlike α, the threshold T does not control the relative contribution of the pairwise feature signs, but only acts on the overall obtained similarity values. As with α, several values of T can be compared so as to identify those more suitable for each specific situation. Interestingly, the same value of $T = 0.25$ allowed good results to be obtained for all results in the present work.

5. Real-world datasets

In this section, we illustrate the potential of the real-valued Jaccard and coincidence indices respectively to real-world datasets. In particular, we consider handwritten *letters* and the *raisin* (e.g. [25, 26]) datasets. All network visualizations have been obtained from the respective adjacency matrices. In addition, every visualization presented in this section were obtained by using the Kamada–Kawai method [41], which was found to allow particularly effective results respectively to the considered data.

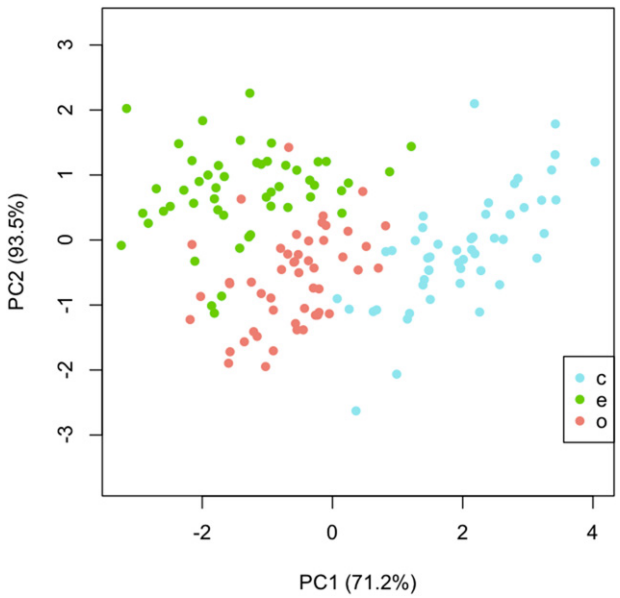

Figure 1. PCA of the letters dataset considering all four original features. The percentages in the axes indicate the relative explanation of the data variance considering 1 and 2 principal components, from which it can be verified that the PCA projection accounts for 93.5% of the data variation. Each of the handwritten letters has been shown in specific respective colors: 'c' in blue, 'o' a in salmon, and 'e' in green. There are 50 samples of each handwritten letter. Overlaps can be observed between every pair of categories.

The letters dataset contains 50 samples handwritten letters 'c', 'e' and 'o', which have been chosen because of their mutual similarity. Each of the data elements is characterized in terms of four respective numeric features, namely the written area, the width and height of each character, as well as the arc length of their external contour.

Figure 1 depicts the principal component analysis (PCA) [30, 31] of the *letters* data set considering all four original features. In addition, in order to avoid effects of the varying magnitudes of the four original measurements, which could therefore bias the network representations, we consider the respectively standardized dataset. All networks obtained from the letters dataset take into account all the four original features. We can see from figure 1 that overlaps are found between each of the respective three categories.

The application of the cosine similarity and Pearson correlation approaches to obtain the weights defining the respective complex networks resulted in the graphs shown in figure 2. One of the categories (blue) resulted relatively well separated in the case of the cosine similarity. However, the other two categories (nodes in green and red) present substantial overlap in both cases.

In order to illustrate the interiority index, we obtained the respective complex network for the letters dataset. The result is shown in figure 3. All the three categories present substantial overlaps, indicating that, despite its ability in providing valuable information about the relative interiority between the datasets, the interiority index has limited potential for accurately reflecting the interconnectivity structure and modularity of the original dataset, at least for the cases and configurations considered here.

Now, we proceed to applying the recently introduced real-valued Jaccard index parametrized by $\alpha = 0.25, 0.32, 0.39, 0.46, 0.53, 0.60$ to the letters dataset, yielding the networks depicted in figure 4.

The resulting networks reveals a surprising separation between the groups for the smaller values of α. Not only the groups have coalesced into a relatively compact community (especially for the two smallest adopted values of α), but the blue group resulted well-separated while a substantial portion of the red group has also resulted separated from the green group. In addition, when compared to the results obtained by using the cosine similarity and Pearson correlation, these results are characterized by a substantially more detailed and specific representation of the original dataset.

Figure 5 shows the *network* obtained for the letters data by adopting the coincidence index with $\alpha = 0.25, 0.32, 0.39, 0.46, 0.53, 0.60$ and threshold $T = 0.25$.

The enhanced discriminative potential of the coincidence index [14, 23] has allowed a network representation of the letters dataset that is even more detailed and structured regarding the interrelationship between the data elements, as well as the overall data modular structure, presenting compact communities (especially for the two smallest values of α) and little overlap between the three groups even in the cases where the green and red groups are adjacent (contrast with the networks in figure 2).

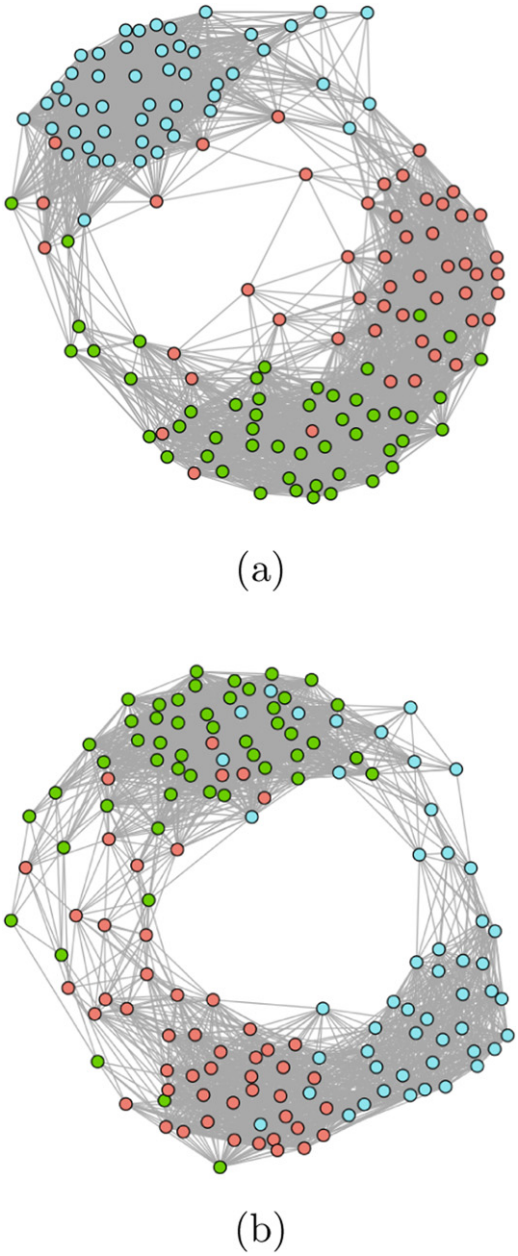

Figure 2. Complex networks of the letters dataset obtained by using the *cosine similarity* (a) and *Pearson correlation* (b) methods while considering all four original features in standardized version. The adopted thresholds were 0.60 and 0.70, respectively. Although the blue group resulted moderately separated in both cases, the other two categories (red and green) present a substantial overlap.

A particularly detailed network has obtained respectively to $\alpha = 0.32$, in which not only the blue cluster resulted significantly separated from the others, but substantial separation can be observed also regarding the green and a substantial part of the red groups.

It is interesting to compare the networks obtained for the letters dataset by using the Jaccard (figure 4) and coincidence (figure 5) methods. Substantially less compact networks, with less-separated communities, were obtained for the Jaccard method as a consequence of the respective similarity index being less strict than the coincidence index, the latter also reflecting the interiority between the two multisets. By imposing less demanding conditions, larger similarity values are obtained for the real-valued Jaccard index that are related to the respectively obtained networks being less separated.

Given that both the real-valued Jaccard and the real-valued coincidence indices convey separated information about the sign of the joint variation of the two multisets, it becomes possible to derive respective complex networks considering exclusively the negative coincidence similarities. In the case of the letters dataset, the obtained network considering only negative coincidence values (i.e. $\alpha = 0$) is shown in figure 6.

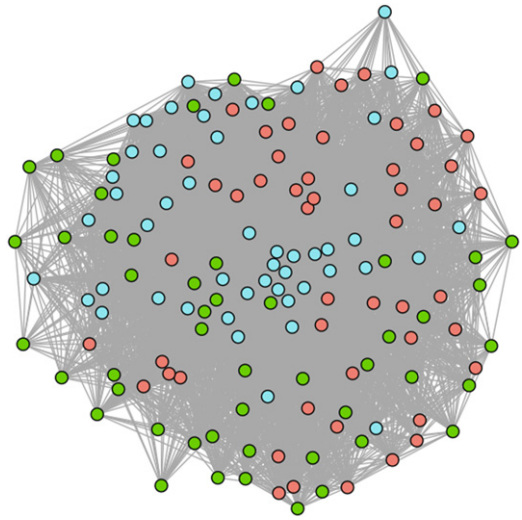

Figure 3. Complex network of the letters dataset obtained by taking into account the *real-valued interiority index*, adopting 0.9 as threshold. The obtained result corroborates that, despite the ability of the interiority index to quantify the relative interiority between two datasets, it has limited potential for providing a detailed representation of the interconnections and respective modularity.

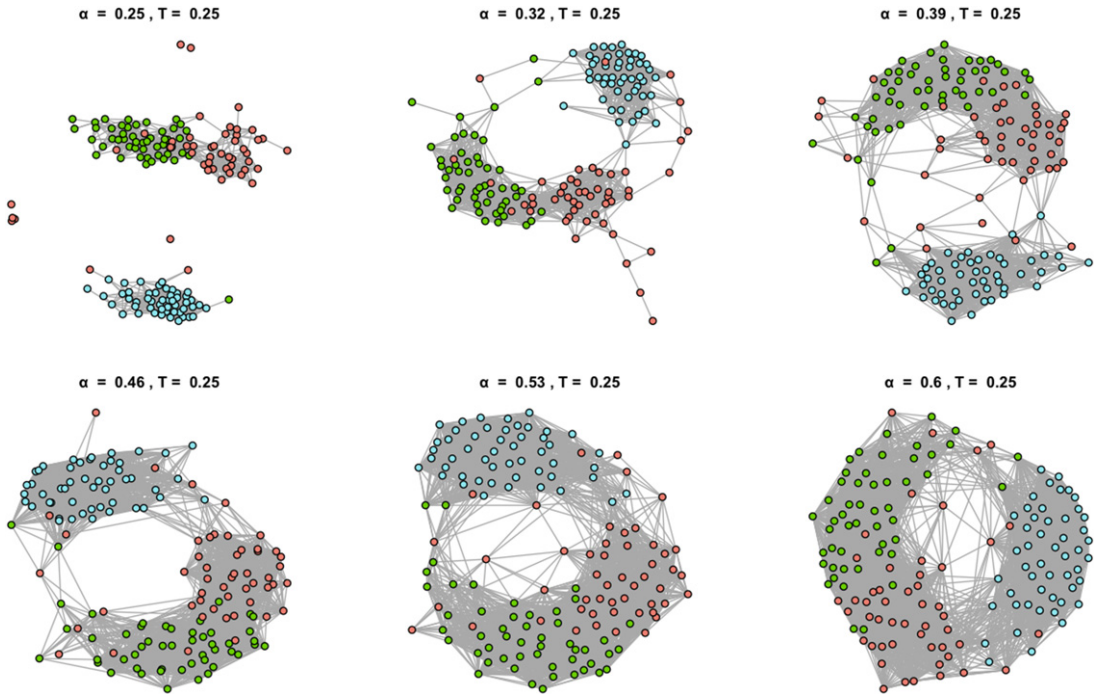

Figure 4. Complex network of the letters dataset obtained by using the *real-value Jaccard index* and the four original features, for $\alpha = 0.25, 0.32, 0.39, 0.46, 0.53, 0.60$ and threshold $T = 0.25$.

Interestingly, two hierarchical levels, or modules, can be observed. The most central one corresponds to the blue group. The other two groups, which are more intertwined and dispersed, resulted at the outer community (network border).

In order to illustrate the potential of the coincidence methodology respectively to a larger dataset, we considered the *raisin* dataset [25, 26], containing 900 individual raisin samples from two varieties grown in Turkey, *Kecimen* and *Besni*. Each of the data elements is characterized in terms of seven measurements corresponding to area, perimeter, major axis length, minor axis length, eccentricity, convex area, and extent. In the original dataset, each of the two varieties is represented in terms of 450 respective samples. The visualizations of the respectively obtained networks were obtained by using the Fruchterman–Reingolg (e.g. [42]) approach.

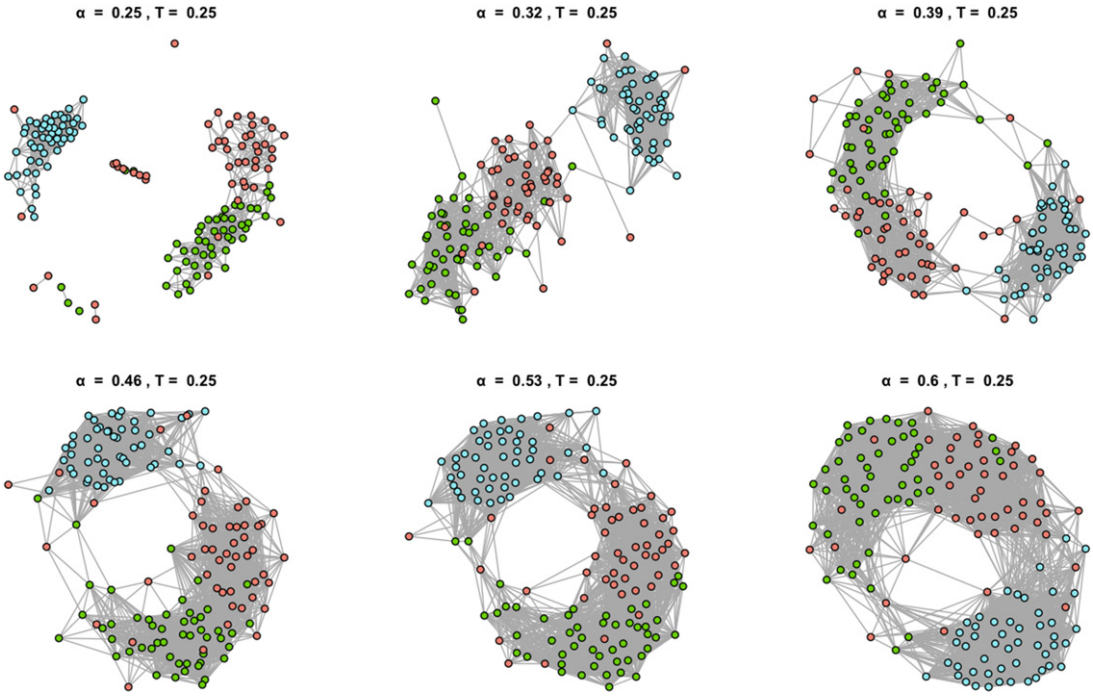

Figure 5. Complex network of the letters dataset obtained by using the *coincidence index* on the four original features for $\alpha = 0.25, 0.32, 0.39, 0.46, 0.53, 0.60$ and threshold $T = 0.25$.

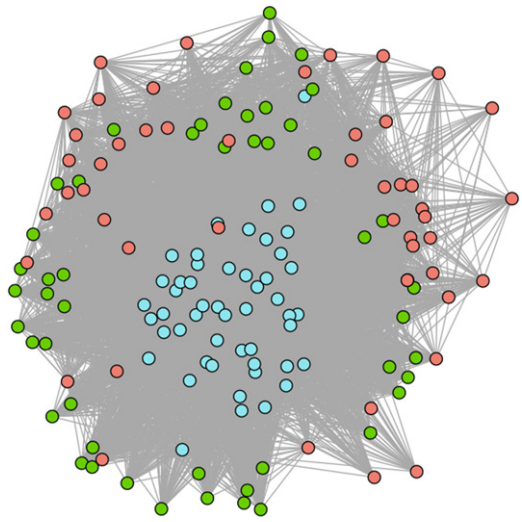

Figure 6. Network obtained for coincidence values smaller or equal than -0.30, assuming $\alpha = 0$. The connections are therefore established between features having opposite signals. The most well-separated original category (blue) resulted as the most internal module in the negative coincidence network, with the two other categories found at the respective border.

Figure 7 illustrates the coincidence network obtained for the raisin dataset considering all measurements and $T = 0.25$, and $\alpha = 0.15, 0.18, 0.21, 0.24, 0.27, 0.3$. For the sake of enhanced visualization, only 450 alternated (every other one) samples have been taken into account.

The raisin example illustrates the *potential* of the coincidence methodology for identifying groups (clusters) and communities in relatively large datasets. As α increases, two separated groups can be observed to progressively appear, grow and then merge into two respective communities interconnected in a ring-like manner. The two obtained groups and communities correspond very closely to the two varieties or raisins, though a number of *Besni* raisins having been progressively incorporated into the *Kecimen* group as a consequence of these specific samples being more similar to the other group.

The ring-like observed in the raisin example illustrates an interesting situation in which the data elements are gradually similar, defining a chain of transitive interconnections that, in the case of this particular dataset,

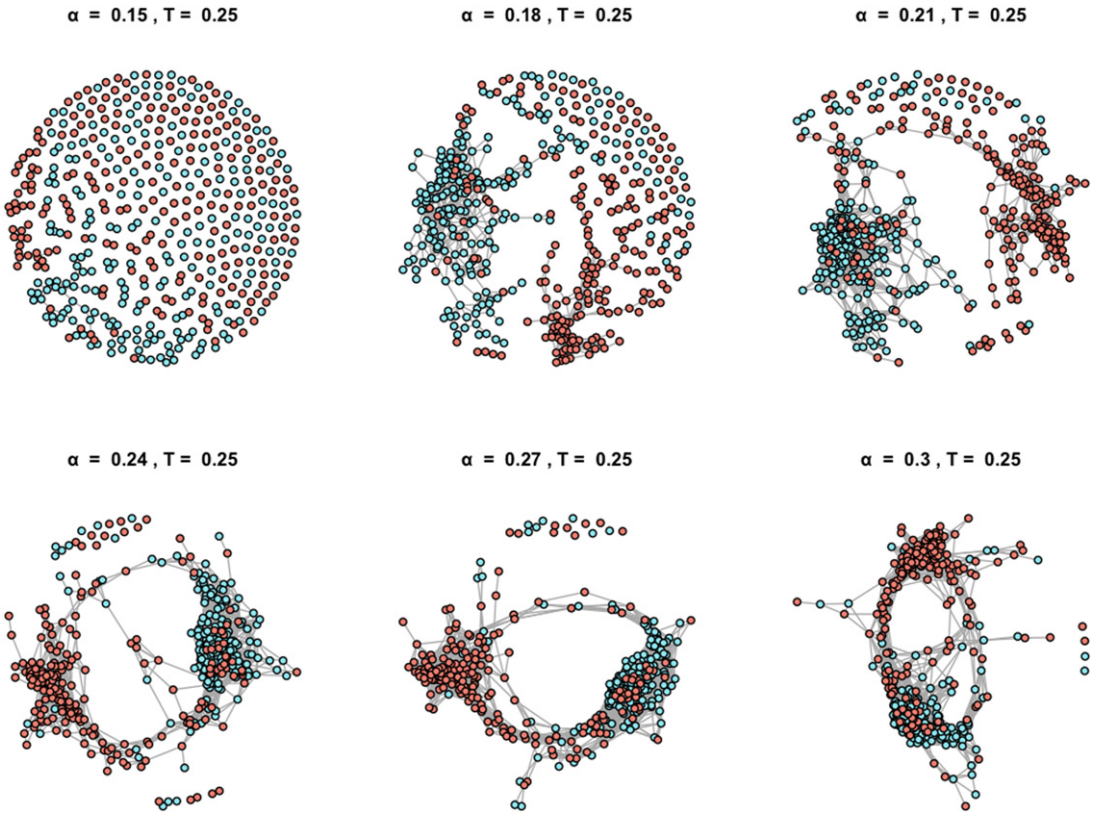

Figure 7. Coincidence complex networks obtained for the raisin dataset [25, 26] considering all seven original features, $T = 0.25$, and $\alpha = 0.15, 0.18, 0.21, 0.24, 0.27, 0.3$. As α is increased, two separated groups, corresponding mostly to the two main varieties of raisins (*Kecimen*, in blue, and *Besni*, in red), progressively appear and then grow until coalescing into two respective modules or communities at $\alpha \approx 0.21$. A ring-like topology can be observed for the three largest values of α. This example illustrates the potential of the coincidence methodology for identifying groups and communities in datasets. Relatively few misplacements can be observed, which are mostly limited to cases of *Besni* raisins appear within the *Kecimen* group.

ends up by establishing a major cycle. Several additional examples of varying topologies found in datasets visualized as networks by using the proposed methodology, as well as complementary information on data and other material related to the present work, can be found at [43].

6. Coincidence networks and modularity

When visually compared to the network construction approaches based on the cosine and Pearson correlation indices, the real-valued Jaccard and real-valued coincidence methods tend to yield networks characterized not only by an impressive level of connectivity details, but also enhanced modular structure. Then, the coincidence method typically allows improved results respectively to the real-valued Jaccard approach. These characteristics have been observed not only respectively to the networks presented in section 5, but also in other studies considering these similarity indices for template matching between two signals [23, 24].

Possible reasons accounting for the relative limitations of the cosine and Pearson methods are as follows. Regarding the cosine similarity, we have that this measurement does not take into account the magnitude of the two compared vectors, therefore overlooking potentially important information. In addition, as a consequence of its own mathematical properties, the cosine index frequently tends to result higher similarity values [16, 23]. In the case of the Pearson correlation, we have that it is known to be particularly sensitive to outliers (e.g. [32]), and also to provide biased estimations when applied to relatively few samples. For instance, maximum Pearson correlation (equal to 1) will be always be obtained between any two points. In addition, as with the cosine similarity, the Pearson correlation has also been observed to provide relatively high similarity values (e.g. [23, 24]).

The real-valued Jaccard and real-valued coincidence indices not only take into account the magnitude of the vectors, but also implement a more strict quantification of their similarity. This is mainly a consequence of the close relationship between the Jaccard and the generalized Kronecker delta function [15, 16], which accounts for the most strict possible characterization of the similarity between two numeric values. The even more strict quantification of similarity implemented by the real-valued coincidence index relatively to the

Table 1. The average and standard deviation of the similarity values obtained by the cosine, Pearson correlation, Jaccard and coincidence indices respective to the *letters* dataset, with $\alpha = 0.5$. The more strict quantification of similarity implemented by the two latter approaches is implied by the respectively smaller obtained standard deviations.

Similarity	Average	St. dev.
Cosine	0.000	0.651
Pearson	−0.005	0.682
Jaccard	0.011	0.387
Coincidence	0.011	0.349

Table 2. The average and standard deviation of the similarity values obtained by the cosine, Pearson correlation, Jaccard and coincidence indices respective to the *raisin* dataset, for $\alpha = 0.5$. As with the letters dataset, the real-valued Jaccard and real-valued coincidence again led to a more strict similarity quantification.

Similarity	Average	St. dev.
Cosine	0.0181	0.636
Pearson	0.0174	0.649
Jaccard	0.0273	0.354
Coincidence	0.0249	0.319

real-valued Jaccard index stems from the fact that the former corresponds to a combination of the real-valued Jaccard and interiority index, therefore complementing the latter index regarding its inability to take the relative interiority between multisets into account [13].

In order to illustrate and complement the aforementioned discussion on the relative characteristic performance of the considered similarity methods, in this section, we consider a quantitative approach to characterizing the interconnectivity detail and modularity allowed by the several considered similarity methods with respect to the letters and raisin datasets, as well as by the *C. elegans* neuronal network [27–29]. In addition, we apply the modularity index (e.g. [39, 40]) for characterizing the separation of the obtained network respectively to the original data elements category. Observe that the modularity is closely related to the obtained level of details, in the sense that an enhanced level of details tends to contribute to a better definition of eventual communities.

First, we quantify how much each of these methods is strict regarding the resulting connectivity in terms of the respectively obtained similarity values. Similarity indices leading to broader histograms, i.e. having larger standard deviations, are understood to be less strict that indices resulting in narrowed histograms, given that less strict indices will favor larger similarity values. The standard deviation of the cosine similarity, Pearson correlation, real-valued Jaccard and real-valued coincidence indices can be directly compared in terms of the respective standard deviations because, in all cases with $\alpha = 0.5$ their average is close to zero while their values are distributed within the same interval extension (i.e. $[-1, 1]$).

Table 1 presents the mean and standard deviation of the similarity values obtained by the cosine, Pearson correlation, real-valued Jaccard and real-valued coincidence indices respectively to the letters datasets. The two latter indices adopt $\alpha = 0.5$.

The nearly null obtained average values are a consequence of the standardization of the original features, which implies the respective new random variables to result centered at the coordinate axis, and to become uncorrelated.

As expected, the real-valued Jaccard and real-valued coincidence yielded substantially smaller standard deviations of the similarity values than obtained for the cosine and Pearson correlation, corroborating the ability of the two real-valued Jaccard and real-valued coincidence approaches to implement a more demanding quantification of the similarity between feature vectors.

Table 2 shows the mean and standard deviation of the similarity values of the cosine, Pearson correlation, real-valued Jaccard and real-valued coincidence applied to the raisin dataset. The two latter indices consider $\alpha = 0.5$. The respective results again substantiate the more strict quantification of similarity implied by the real-valued Jaccard and real-valued coincidence indices.

Figure 8 presents the histograms of similarity values obtained by the real-valued Jaccard and real-valued coincidence indices obtained respectively to the letters and raisin datasets, for $\alpha = 0.5$. In addition to being consistent with the average and standard deviation values discussed above, the results in this figure allow us to directly observe the narrower similarity values distribution implied by the real-valued coincidence index relatively to the real-valued Jaccard measurement. The obtained differences are a direct consequence of the

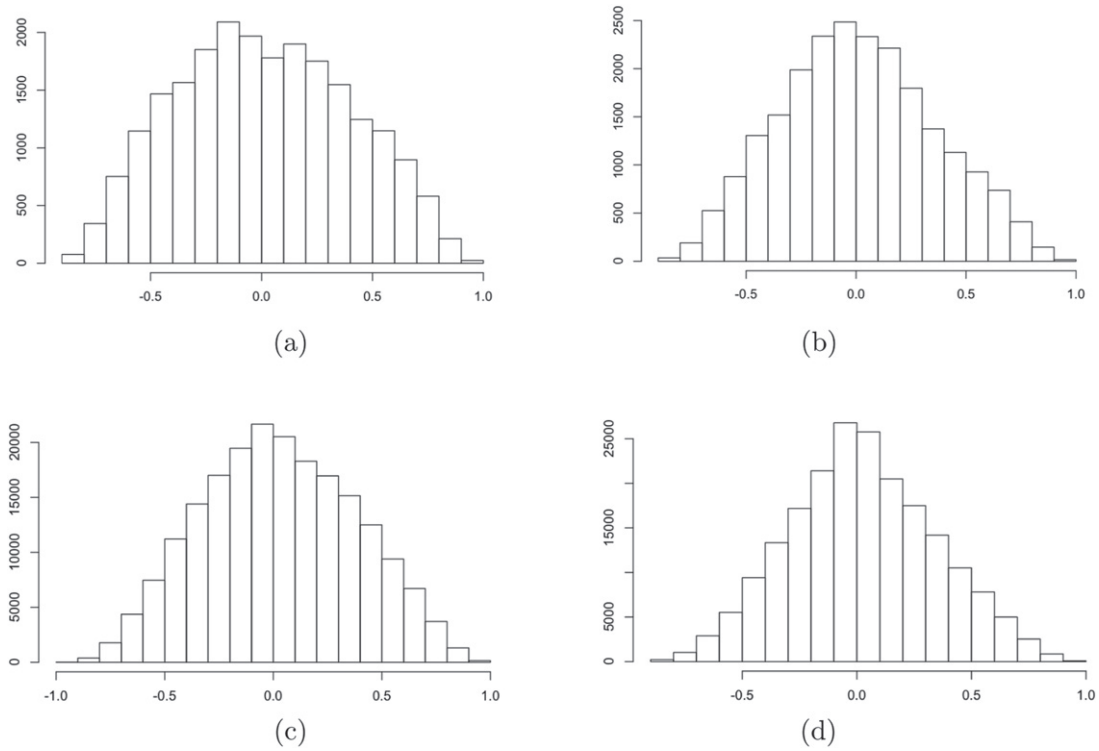

Figure 8. Histograms of the similarity values obtained by applying the real-valued Jaccard and real-valued coincidence, with $\alpha = 0.5$, for the letters and raisin datasets: (a) real-valued Jaccard applied to letters; (b) real-valued coincidence applied to letters; (c) real-valued Jaccard applied to raisin; (d) real-valued coincidence applied to raisin. A narrower distribution of similarity values obtained by the coincidence index can be observed in both cases.

Table 3. Modularities of the considered similarity indices respectively to the networks obtained for the letters and raisin datasets. Observe the larger modularity values resulting from the Jaccard and coincidence methods respectively to both datasets.

Similarity	Modul. *letters*	Modul. *raisin*
Cosine	0.438	0.252
Pearson	0.348	0.151
Jaccard	0.436	0.266
Coincidence	0.467	0.271

fact that the latter index incorporates the quantification of the relative interiority between the two compared vectors, being therefore even more strict and complete.

Having substantiated in a quantitative manner the enhanced ability of the real-valued Jaccard and real-valued coincidence indices for implementing a more strict quantification of the similarity between two feature vectors, therefore contributing to more detailed respective networks to be achieved, we now approach the *modularity* of the networks obtained by using the adopted similarity methods.

Table 3 presents the modularity (e.g. [39, 40]) of the letters and raisin networks obtained in section 5, assuming $\alpha = 0.5$ in the case of the Jaccard and coincidence indices. The modularity was calculated respectively to the original categories, three letters in the case of the letters dataset, and the two varieties of raisins in the raisin dataset.

It can be readily observed that the Pearson correlation yielded the smallest modularity, followed by the cosine similarity or the real-valued Jaccard, and then the real-valued coincidence. These results mirror the same characteristics already observed with respect to the standard deviation of the similarity values allowed by each similarity index, substantiating the tendency of more strict and purposive similarity quantification to contribute to enhanced modularity characterization.

In order to study the effect of the parameters α, modularity values were also calculated regarding all respective configurations in section 5. Table 4 presents the modularity values obtained for the application of the real-valued Jaccard and real-valued coincidence methods to the networks respectively derived from the letters dataset.

Table 4. Modularity values resulting from the letters network by application of the using the real-valued Jaccard and real-valued coincidence index parametrized by α. The latter method allowed larger modularity values for each considered instance of α, except for the case $\alpha = 0.25$. Observe that, in the case of this particular dataset, the maximum modularity is not obtained for the reference value $\alpha = 0.5$ in neither of the two considered methods.

Similarity	$\alpha = 0.25$	$\alpha = 0.32$	$\alpha = 0.39$	$\alpha = 0.46$	$\alpha = 0.53$	$\alpha = 0.60$
Jaccard	0.537	0.535	0.505	0.458	0.415	0.381
Coincidence	0.533	0.546	0.521	0.491	0.452	0.410

Table 5. Modularity values resulting from the raisin networks by the two multiset methods parametrized by α. The real-valued coincidence method allowed larger modularity values to be obtained for each considered instance of α.

Similarity	$\alpha = 0.15$	$\alpha = 0.18$	$\alpha = 0.21$	$\alpha = 0.24$	$\alpha = 0.27$	$\alpha = 0.30$
Jaccard	0.323	0.300	0.297	0.280	0.268	0.264
Coincidence	0.327	0.301	0.303	0.289	0.278	0.273

It is of particular interest to observe that, were not for the additional degree of freedom allowed by the incorporation of the parameter α into the real-valued Jaccard and real-valued coincidence methods, the highest observed modularity could not have been obtained. In fact, the parameterless versions of both these two indices, which correspond to taking $\alpha = 0.5$, can be found to lead to modularity values that are smaller than most of those obtained for the other values of α.

Table 5 presents the modularity values obtained respectively to the raisin dataset. The obtained results substantiate the potential of the parametrized real-values coincidence method, respectively to the real-valued Jaccard, for obtaining detailed interconnections and enhanced modularity.

Having approached in a quantitative manner the potential of the considered methods respectively to implementing more strict quantification of similarity and reflecting the modular structure of the obtained networks, we complement our studies by applying the multiset similarity-based network generation approaches not on dataset features, but on *topological measurements* (e.g. [3]) derived from networks represented by respective adjacency matrix involving only 0s and 1s as entries.

The rationale here is that, provided the topological measurements of the networks interconnectivity reflect to some significant extent the modular structure of the original networks, the similarity-based methods of deriving respective networks may become capable of highlighting the respective modularity accordingly to some specific criterion.

Community finding methods have been proposed (e.g. [12]) based on the application of the traditional Jaccard index for categorical data (sets) to quantify the similarity between the sets of neighbors of each node in the original network. The basic idea is that nodes in the same community will tend to have similar neighbors. Once these similarities are obtained, some respective clustering approach can then be applied in order to identify the communities. Here, analogously to [44], we associate to each of the nodes of the network of interest the relative shortest distance values between that node and other network nodes.

Figure 9 depicts the networks obtained from the raisin network derived with $T = 0.25$ and $\alpha = 0.24$ (figure 7). This network has been chosen among the others in this work for presenting two well-defined main modules, mostly corresponding to the original data elements categories, while also incorporating additional details.

The respective adjacency matrix (0 or 1 values) was obtained, and the shortest distances between each node and the remainder nodes in the original network were adopted as feature vectors to be supplied to the real-valued Jaccard and real-valued coincidence methods of network generation from data. The derivation of the new network assumed $T = 0.25$ and $\alpha = 0.5$. The shortest path lengths have been obtained by using the Floyd–Warshall method [45, 46].

Though starting with only the binary interconnectivity between the nodes of the original network, both the real-valued Jaccard and real-valued coincidence methods were capable of yielding networks with two well separated modules or communities. As a consequence of its more strict and complete characterization of similarity, the real-valued coincidence approach allowed an even more separated and detailed resulting network. The additional small scale structures in the network in figure 9(b) reflects the interconnections between the two communities originally present in figure 7. The larger separation allowed by the real-valued coincidence is indicated by the fact that thinner interconnections between the two main communities can be observed, with fewer nodes along them, and also by the respectively more compact (smaller) obtained modules.

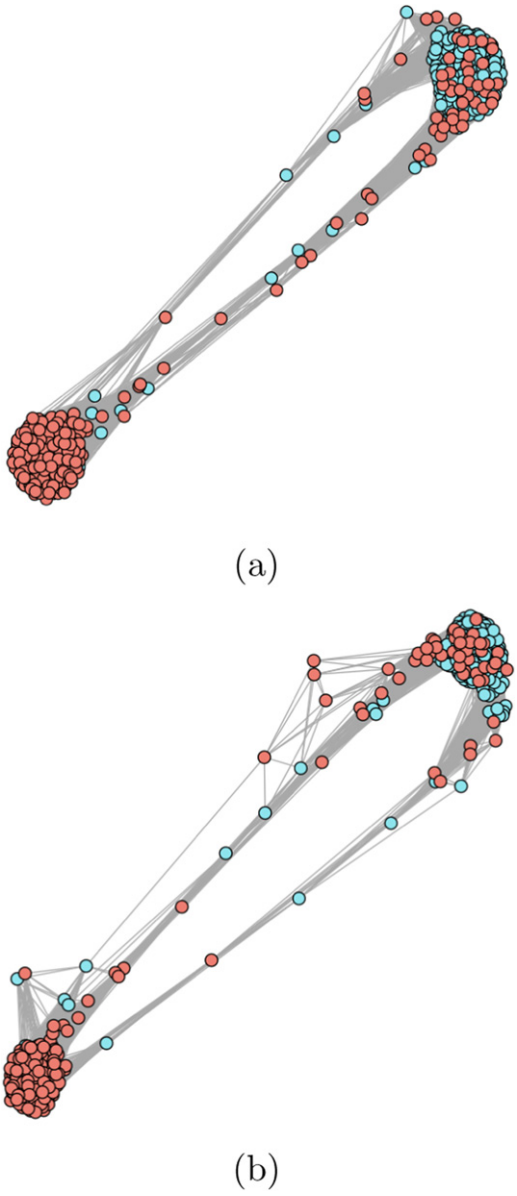

Figure 9. The networks obtained from the adjacency matrix of the raisin network with $T = 0.25$ and $\alpha = 0.24$ (figure 7) by using the real-valued Jaccard (a) and real-valued coincidence (b) methods. The shortest path distances between each pair of nodes were adopted as features. Observe that the more strict and complete quantification of the similarities allowed by the real-valued coincidence index led to more detailed interconnectivity and enhanced separation of the two main modules (fewer nodes along the two main long interconnections between the two communities and more compact modules).

To complement our analysis, we also applied the real-valued coincidence approach to the binary adjacency matrix of the *C. elegans* network [27–29]. Figure 10 illustrates the original *C. elegans* network visualized from its respective binary adjacency matrix by using the Fruchterman–Reingold methodology (e.g. [42]).

As in the previous example, the feature vector assigned to each data element corresponds to the shortest path distances from the respective node to the remainder nodes. As before, the shortest paths distances have been calculated by using the Floyd–Warshall method [45, 46]. The generation of the new network from the adjacency matrix of the previous network assumed $T = 0.25$ and $\alpha = 0.4$.

Given that the *C. elegans* neuronal network is a directed graph, here we consider both its undirected and directed versions. The undirected network is obtained by the symmetrization of the original network, i.e. by adding the original adjacency matrix with its transpose and considering the obtained elements that are larger than zero as 1.

The obtained network respectively to the *undirected C. elegans* distances, shown in figure 11, presents an enhanced modular structure with more intricate and detailed interconnectivity than the original network shown in figure 10.

Figure 12 depicts the network obtained while taking into account the respective *directed* adjacency matrix. As expected, the consideration of the edges directionality impacted considerably on the respectively obtained

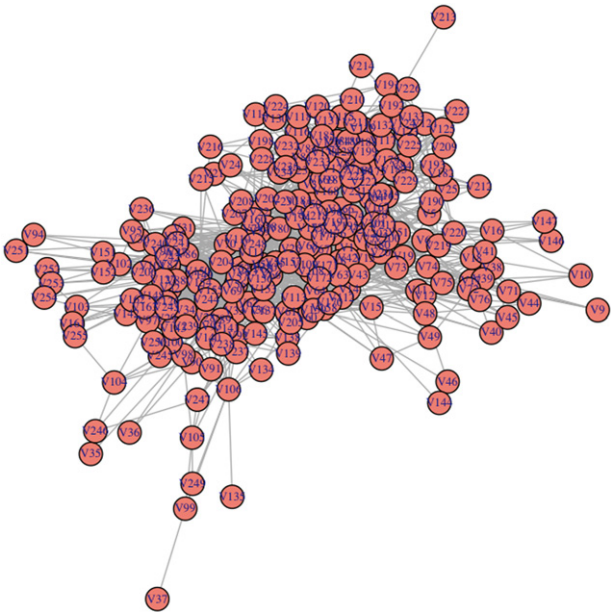

Figure 10. Visualization of the *C. elegans* [27–29] from its original adjacency matrix (0 and 1 values). No particular interconnectivity pattern or modularity can be observed in this network.

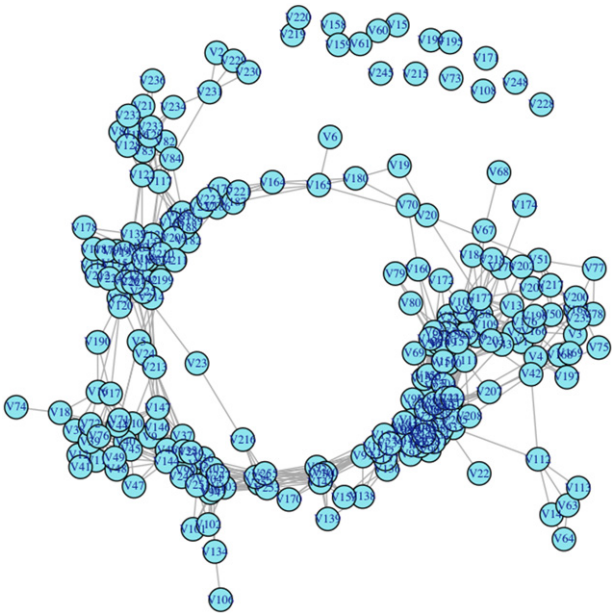

Figure 11. The coincidence network obtained from the unweighted *undirected* adjacency matrix of the *C. elegans* neural network, adopting $T = 0.25$ and $\alpha = 0.4$. The resulting network presents a well-defined modular structure, as well as intricate interconnectivity details.

modularity. More specifically, the reduced number of edges in this case implied two separated groups, or connected components, to appear.

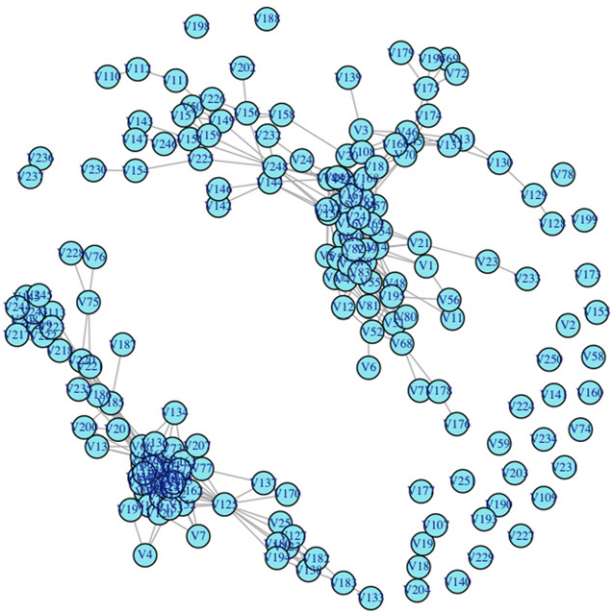

Figure 12. The network obtained from the unweighted *directed* adjacency matrix of the *C. elegans* network, adopting $T = 0.25$ and $\alpha = 0.4$ by using the real-valued coincidence methodology. As could be expected, the consideration of the edges directionality strongly impacted on the obtained modularity, which is now characterized by two completely separated groups.

7. Concluding remarks

Network science has consolidated itself as an important area, with ample theoretical and applied contributions. Much of this success stems from the ability of complex networks to *represent* virtually every possible discrete structure or system, while continuous counterparts can always be discretized to a given resolution.

One of the remaining challenges in network science concerns how to effectively transform datasets, having elements characterized by respective features, into networks. Though measurements such as the cosine similarity and Pearson correlation coefficient have been frequently applied, other interesting measurements such as the traditional Jaccard index have been mostly constrained to categorical or binary data.

The present work set out at addressing the possibility of using the real-valued Jaccard and real-valued coincidence indices as the basis for deriving complex network representations from numeric (e.g. integer, rational, real) data. The motivation for this stems from the potential of these two multiset-based indices for providing more strict and complete quantification of the similarity between two multisets or vectors [13, 16, 23].

The results obtained with respect to the letters and raisin real-world datasets substantiated the enhanced precision and discrimination potential of the real-valued Jaccard and real-valued coincidence indices. In particular, when compared to results obtained by using the cosine similarity and Pearson correlation coefficient, the networks obtained by the real-valued Jaccard and real-valued coincidence indices were systematically found to present enhanced levels of interconnection detail and modularity structure.

In addition, the incorporation of the parameter α into these indices was found to provide a flexible and valuable means for selecting the overall level of interconnectivity of the obtained networks, with less densely connected structures being obtained for smaller values of α. Thanks to this parameter, networks have been achieved that present modularity higher than could be otherwise obtained by using the respective parameterless indices (i.e. $\alpha = 0.5$).

All in all, the interesting features provided by the reported methodology, and especially by the real-valued coincidence index, can be summarized as: (i) effectiveness for revealing more detailed patterns of interconnections between the original data elements; (ii) potential for enhancing the modular structure of the original data; (iii) can be applied on datasets involving real-valued, integer, categorical and binary types, including hybrid cases: (iv) based on non-bilinear operations (including the maximum and minimum) that can substantially attenuate noise and secondary structures (low-pass filtering) simultaneously as more meaningful structures are enhanced (high-pass filter), an effect that cannot be achieved through bilinear operations such as Pearson correlation and cosine similarity: (v) conceptual and computational simplicity, relying on the maximum, minimum, sign and division operations; (vi) employ an improved version of the well-known and extensively used (mostly for categorical and binary data) Jaccard index; (vii) integrates verification of the relative interiority

between the datasets (in the case of the coincidence index); (viii) founded on a formal framework corresponding to the generalization of multisets to real, possibly negative values; (ix) incorporate a parameter α allowing the control of the contributions of the positive and negative pairwise combinations of sign alignments between the original features; (x) the features remain in their original domain, except for their eventual standardization, not being mapped into other less intuitive spaces; (xi) can be readily adapted to take into account similarity between three or more data elements [13]; and (xii) the implemented similarity comparisons can have a probabilistic interpretation reflecting the Jaccard formulation and normalization.

It should be nevertheless observed that the choice of the similarity measurement, as well as the frequently adopted standardization, can have significant respective effects on the obtained results. The adoption of these approaches therefore needs to take into account demands and requirements specific to each application and problem. It is also important to keep in mind that data analysis methods and results should be always understood only as hypotheses to be further explained, validated, and verified.

The proposed methodology bears significant potential for ample application in complex network research, allowing an enhanced manner of obtaining representations of the most diverse type of data and features, including cases of mixed types of measurements. Several prospects for further developments are therefore established that includes but are not restricted to the following possibilities.

It would be interesting to have the original datasets to be pre-processed so as to enhance the separation between their elements, e.g. by using linear discriminant analysis [31], before being fed into the considered similarity-based methods for network construction from datasets. Another promising possibility consists in extending other similarity indices, such as those described in [13], to real-valued data, including those capable of taking into account higher order combinations of data elements (e.g. three feature vectors instead of just a pair). Along a related line, it would also be of interest to take integer powers of the numerator in the Jaccard index [13], with potential for more strict characterization of similarity.

Acknowledgments

Luciano da Fontoura Costa thanks CNPq (Grant No. 307085/2018-0) and FAPESP (Grant 15/22308-2).

Data availability statement

No new data were created or analysed in this study.

ORCID iDs

Luciano da Fontoura Costa https://orcid.org/0000-0001-5203-4366

References

[1] Barabási A L and Pósfai M 2016 *Network Science* (Cambridge: Cambridge University Press)
[2] Newman M 2010 *Networks: An Introduction* (Oxford: Oxford University Press)
[3] da Fontoura Costa L, Rodrigues F A, Travieso G and Villas Boas P R 2007 Characterization of complex networks: a survey of measurements *Adv. Phys.* **56** 167–242
[4] da Fontoura Costa L, Oliveira O N, Travieso G, Rodrigues F A, Villas Boas P R, Antiqueira L, Viana M P and Correa Rocha L E 2011 Analyzing and modeling real-world phenomena with complex networks: a survey of applications *Adv. Phys.* **60** 329–412
[5] da Fontoura Costa L 2018 What is a complex network? https://researchgate.net/publication/324312765_What_is_a_Complex_Network_CDT-2
[6] Comin C H, Peron T, Silva F N, Amancio D R, Rodrigues F A and da Fontoura Costa L 2020 Complex systems: features, similarity and connectivity *Phys. Rep.* **861** 1–41
[7] Jaccard P 1901 Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines *Bull. Soc. Vaudoise Sci. Nat.* **37** 241–72
[8] Jaccard P 1901 Étude comparative de la distribution florale dans une portion des alpes et des jura *Bull. Soc. Vaudoise Sci. Nat.* **37** 547–9
[9] Samanthula B K and Jiang W 1989 Secure multiset intersection cardinality and its application to Jaccard coefficient *IEEE Trans. Dependable Secure Comput.* **13** 591–604
[10] Wikipedia 2021 Jaccard index https://en.wikipedia.org/wiki/Jaccard_index (accessed 10 October 2021)
[11] Schubert A and Telcs A 2014 A note on the Jaccardized Czekanowski similarity index *Scientometrics* **98** 1397–9
[12] 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
[13] da Fontoura Costa L 2021 Further generalizations of the Jaccard index https://researchgate.net/publication/355381945_Further_Generalizations_of_the_Jaccard_Index
[14] da Fontoura Costa L 2021 Multisets https://researchgate.net/publication/355437006_Multisets

[15] da Fontoura Costa L 2021 Generalized multiset operations https://researchgate.net/publication/356191988_Generalized_Multiset_Operations (accessed 10 November 2021)
[16] da Fontoura Costa L 2021 On similarity https://researchgate.net/publication/355792673_On_Similarity
[17] Hein J 2003 *Discrete Mathematics* (Sudbury, MA: Jones and Bartlett Publishers)
[18] Knuth D E 1998 *The Art of Computing* (Reading, MA: Addison-Wesley)
[19] Blizard W D 1989 Multiset theory *Notre Dame J. Form. Log.* **30** 36–66
[20] Blizard W D 1991 The development of multiset theory *Mod. Logic* **4** 319–52
[21] Mahalakshmi P M and Thangavelu P 2019 Properties of multisets *Int. J. Innovative Technol. Explor. Eng.* **8** 1–4
[22] Singh D, Ibrahim M, Yohana T and Singh J N 2011 Complementation in multiset theory *Int. Math. Forum* **38** 1877–84
[23] da Fontoura Costa L 2021 Comparing cross correlation-based similarities https://researchgate.net/publication/355546016_Comparing_Cross_Correlation-Based_Similarities
[24] da Fontoura Costa L 2021 Multiset neurons https://researchgate.net/publication/356042155_Common_Product_Neurons
[25] Dua D and Graff C 2017 UCI machine learning repository http://archive.ics.uci.edu/ml
[26] Cinar I, Koklu M and Tasdemir S 2022 Classification of raisin grains using machine vision and artificial intelligence methods *Gazi J. Eng. Sci.* **6** 1–34
[27] Kaiser M and Hilgetag C C 2006 Nonoptimal component placement, but short processing paths, due to long-distance projections in neural systems *PLoS Comput. Biol.* **2** e95
[28] Choe Y, McCormick B H and Koh W 2004 Network connectivity analysis on the temporally augmented *C. elegans* web: a pilot study *Soc. of Neurosc. Abstracts* **30**
[29] Cherniak C 1994 Component placement optimization in the brain *J. Neurosci.* **14** 2418–27
[30] Gewers F, Ferreira G R, Arruda H F, Silva F N, Comin C H, Amancio D R and da Fontoura Costa L 2020 Principal component analysis: a natural approach to data exploration *ACM Comput. Surv.* **54** 200–19
[31] Johnson R A and Wichern D W 2002 *Applied Multivariate Analysis* (Englewood Cliffs, NJ: Prentice-Hall)
[32] Kim Y, Kim T-H and Ergün T 2015 The instability of the Pearson correlation coefficient in the presence of coincidental outliers *Finance Res. Lett.* **13** 243–57
[33] Blizard W D 1989 Real-valued multisets and fuzzy sets *Fuzzy Sets Syst.* **33** 77–97
[34] Blizard W D 1990 Negative membership *Notre Dame J. Form. Log.* **31** 346–68
[35] Akbas C E, Bozkurt A, Arslan M T, Aslanoglu H and Cetin A E 2014 *L*1 norm based multiplication-free cosine similarity measures for big data analysis *IEEE IWCIM* (France)
[36] Mirkin B 1996 *Mathematical Classification and Clustering* (Dordrecth: Kluwer)
[37] Vijaymeena M K and Kavitha K 2016 A survey on similarity measures in text mining *Mach. Learn. Appl.* **3** 19–28
[38] da Fontoura Costa L 2021 Real-valued Jaccard and coincidence based hierarchical clustering https://researchgate.net/publication/355820021_Real-Valued_Jaccard_and_Coincidence_Based_Hierarchical_Clustering
[39] Newman M E 2006 Finding community structure in networks using the eigenvectors of matrices *Phys. Rev.* E **74** 036104
[40] Fletcher R J Jr, Revell A, Reichert B E, Kitchens W M, Dixon J D and Austin J D 2013 Network modularity reveals critical scales for connectivity in ecology and evolution *Nat. Commun.* **4** 2572
[41] Kamada T and Kawai S 1989 An algorithm for drawing general undirected graphs *Inf. Process. Lett.* **31** 7–15
[42] Fruchterman T M J and Reingold E M 1991 Graph drawing by force-directed placement *Softw. Pract. Exp.* **21** 1129–64
[43] da Fontoura Costa L 2021 A caleidoscope of datasets represented as networks by the coincidence methodology https://researchgate.net/publication/356392287_A_Caleidoscope_of_Datasets_Represented_as_Networks_by_the_Coincidence_Methodology
[44] de Souza P J P, Comin C H and da Fontoura Costa L 2019 Distance-based network partitioning (arxiv:1911.01775)
[45] Floyd R W 1962 Algorithm 97: shortest path *Commun. ACM* **5** 345
[46] Warshall S 1962 A theorem on Boolean matrices *J. ACM* **9** 11–2