On the Self-Coincidence Structure of Networks

Luciano da Fontoura Costa

On the Self-Coincidence Structure of Networks

Luciano da Fontoura Costa
luciano@ifsc.usp.br

São Carlos Institute of Physics – DFCM/USP

15th May 2022

Abstract

The characterization of the topological properties of complex networks remains a challenging subject, as it is difficult to define sets of measurements capable of comprehensive characterization of their structure. In the present work, we approach this important and interesting issue from the perspective of the similarity between the nodes of network, while considering several local and global respective topological node properties. Capable of implementing a particularly selective and robust quantification of the similarity between two mathematical structures, the coincidence similarity is adopted in this work for that finality. Given a network, and having each of its nodes described in terms of a set of topological features, the coincidence methodology can then be applied in order to reveal a respective similarity network, which is here called the self-similarity structure of the original network. Several interesting results are obtained by applying this method to model-theoretic as well as real-world networks, including the identification of an opposite relationship between topological and coincidence links in some of the analyzed cases as well, small average node similarity tending to result at hubs and nodes at the core of communities, the understanding of the node(s) with maximum average node similarity as hubs of similarities, as well as the finding that the topological links do not tend to be related to the respective coincidence similarity between the involved nodes.

1 Introduction

Thanks to the ability of graphs and networks to represent virtually any discrete structure and system, the area of network science established itself as a dynamic and important area along a relatively short period of time (e.g. [1, 2, 3, 4]).

Though a great deal of initiatives have addressed the characterization of the topological properties of complex networks (e.g. [1, 2, 3]), there is still no definitive answer regarding sets of these measurements capable of providing particularly comprehensive representations of the network topology, eventually to the level of establishing a fully invertible mapping between the original network and its representation by a set of features. Thus, there is still motivation for new approaches to the characterization of the topological properties of complex networks.

In the present work, we resource to the *coincidence similarity index* [5, 6, 7] for that finality. Corresponding to the product between the real-valued Jaccard [5, 6] and the interiority (or overlap, e.g. [8]) indices, the coincidence similarity has been found to allow a particularly selective, sensitive and accurate quantification of the similarity between pairs of mathematical structures including vectors, matrices, functions and fields (e.g. [6, 7, 9]), therefore paving the way to several successful applications to a diversity of fields (e.g. [6, 7, 10]).

Here, we address the promising perspective of quantifying graphs and networks in terms of the respective distribution of the *similarity* between several topological properties of the involved nodes and/or subgraphs while simultaneously considering several topological measurements.

The derivation of coincidence networks from original topological networks was brief and preliminary addressed in [7], considering topological measurements taken for each of the nodes in the original network as features to be translated, by using the coincidence methodology suggested in that same work [7], into a respective coincidence network whose interconnections not necessarily correspond to the original links. In this work, we address this interesting possibility in a more systematic and comprehensive manner, taking into account additional topological measurements as well several model-theoretical and real-world networks to be analyzed from the perspective of the distribution of respective similarities.

More specifically, the suggested quantification of the network topological properties in terms of pair-wise similarity values suggested involves the following two-steps: (i) first, the similarity quantification is performed between

all possible pairs of nodes in a network, including a node with itself; and (ii) the overall similarity of the network is then gauged in terms of the average and standard deviation of the values obtained in (i).

In addition, the step (i) above immediately yields a respective coincidence network, which can be respectively visualized and studied in several manners. In the present work, emphasis is given to comparing this resulting links in the obtained coincidence network with the topological links in the original network. In this way, it becomes possible, among other possibilities, to investigate if higher coincidence values are eventually associated to the presence of an actual respective topological link in the original network.

Thus, after presenting the proposed method for quantifying the complexity of a network at its successive topological scales, the present work applies this approach to three theoretical models, namely Erdős -Rényi, Barabási-Albert, and Watts-Strogatz, as well as to two real-world data. Several noticeable results are reported, including the tendency of hubs and nodes at the center of communities to have small average node similarity, the identification of an opposite relationship between topological and coincidence links, the possibility to understand node(s) with maximum average node similarity as hubs/prototypes of a network, as well as the verification of lack of relationship between topological connectivity and respective coincidence similarity, therefore suggesting that the latter measurements has substantial potential for complementing the characterization of networks.

This work is organized as follows. We start by briefly presenting the involves basic concepts – including multiset concepts and the coincidence similarity, and follows by describing the application of the similarity-based approach to the quantification of model-theoretical as well as some networks derived from real-world data.

2 Basic Concepts

Three model-theoretical networks (e.g. [1, 2, 3]) are considered in the present work: Erdős-Rényi – ER, Barabási-Albert – BA, and Watts-Strogatz – WS in its ring configuration. These network types are respectively characterized by statistical uniformity, scale-free node distribution, and small world property combined with geographical adjacency, and geographical regularity. The ER and BA models are characterized by small average shortest path distances, and the WS model can present this property depending on the rewiring probability. The present work focuses on undirected networks.

Regarding the choice of measurements to be employed for characterizing the topological properties of each node, there are basically two main approaches that can be taken: (a) to incorporate as many measurements as necessary to provide a comprehensive representation of the topological properties; and (b) to adopt a set of measurements that are of particular interest respectively to each research and applications. Here, we chose an intermediate possibility, in the sense that we adopt a set of measurements that is reasonably comprehensive and still convey information about topological properties of special interest regarding the distance between nodes, as well as the local interconnectivity around the nodes. More specifically, the following node-related measurements are employed in the present work (e.g. [3]):

[1] - Node degree: The node degree, corresponding to the number of connections existing between that node and the remainder of the network, constitutes arguably the most important topological property of a node in a graph or network. Here, we consider its average k_av and standard deviation k_sd;

[2] - Clustering coefficient: The clustering coefficient, or transitivity, of a node measures how much the neighbors of that node are interconnected. Here, we take into account its average cc_av and standard deviation cc_sd;

[3] - Average and standard deviation of the shortest path lengths: Given a node, it is possible to identify its shortest paths to all other nodes in the network, each with its respective number of links, which is understood as the length of the path. The average sh_av and standard deviation sh_sd of this measurement taken respectively to all networks nodes is are adopted here as topological measurements, therefore providing valuable information about the overall proximity between the network nodes;

[4] - Betweeness centrality: This interesting measurement quantifies the participation of a given node respectively to the overall structure of shortest paths in the respective network. We adopt its average bt_av and standard deviation br_sd in order to reflect this interesting property of the nodes in the considered networks.

Thus, we have a total of five features characterizing each of the network nodes according to a mixture of local and global properties.

It is interesting to observe that the degree and clustering coefficient measurements can be understood as being *local*, as they reflect the topological properties around the most immediate neighborhood of each node. The other two measurements, namely the average shortest path length and the betweeness centrality can be said to be *global* because they convey information about the overall topology of the network. Therefore, the set of adopted

topological measurements constitute a good balance between local and global measurements.

Similarity indices have been often considered in order to compare sets and other mathematical structure (e.g. [11, 12, 13, 14, 15, 16, 17]). The coincidence similarity index was suggested [5, 6] as an enhancement of the widely applied Jaccard index (e.g. [18, 19, 16, 20, 21, 22]) with two main motivations in mind: (i) to complement the Jaccard index in order to incorporate information about the relative interiority between the two compared structures; and (ii) to allow consideration of real-valued real values. The coincidence similarity is based on multiset theory (e.g. [23, 24, 25, 26, 27, 28]).

Given that all adopted measurements are non-negative, we can limit our attention to the multiset coincidence between two non-negative real vectores $\vec{x}$ and $\vec{y}$, which can be defined as:

$$\mathcal{C}(\vec{x},\vec{y}) = \frac{\sum \min\{\vec{x},\vec{y}\}}{\sum \max\{\vec{x},\vec{y}\}} \frac{\sum \min\{\vec{x},\vec{y}\}}{\min\{\sum \vec{x}, \sum \vec{y}\}} \quad (1)$$

where $\sum \vec{v}$ indicates the sum of the elements of vector $\vec{v}$. The two terms in the above equation correspond to the Jaccard and interiority indices, respectively. Conveniently, we have that $0 \leq \mathcal{C}(\vec{x},\vec{y}) \leq 1$.

In the present work, the features to be compared by the coincidence similarity are preliminary normalized by dividing each value by the respective average of each type of feature. This normalization is related to normalizing the features to result in densities with unit area, but larger feature values are obtained by using the adopted normalization.

The coincidence methodology to translate datasets where each element is characterized in terms of M features (or measurements) into a respective network [7] involves representing each data element as a node, while the weights of the links between two nodes correspond to the respective pair-wise coincidence values between those two nodes. Therefore, the higher the coincidence value, the stronger the connection in the obtained coincidence network.

3 Average Node Similarity and the Self-Similarity Structure

In the present work, we transform the original network of interest into a dataset of elements, corresponding to the nodes, characterized in terms of respective feature vectors involving the five topological measurements described in Section 2. The coincidence methodology is then applied in order to transform this dataset into a respective coincidence network. The original network has N nodes.

The *average similarity* of a specific node i can now be defined as corresponding to:

$$\langle \mathcal{C}_i \rangle = \frac{1}{N} \sum_{k=1}^{N} \mathcal{C}(\vec{v}_i, \vec{v}_k) \quad (2)$$

where $\vec{v}_i$ and $\vec{v}_i$ are the feature vectors describing the topological properties of nodes i and k, respectively.

The *self-similarity* structure of the original is henceforth understood as the coincidence network derived from the original network as described above. As such, it provides comprehensive information about the degree of similarity, as revealed by the coincidence index, between all pairs of nodes in the network of interest, including between the node and itself. This structure can be characterized by the distribution of the average node similarity of its nodes, as well as by the respective average $\langle\langle \mathcal{C} \rangle\rangle$ and standard deviation $\sigma_{\mathcal{C}}$.

Interestingly, several interesting questions arise in consequence of the introduction of the above measurements. For instance, given a network or one of its subgraphs, the respective node with the highest average node similarity can be understood as a *hub* of similarity [29], therefore representing a possible prototype for that network. Another interesting point regards if it is possible to have networks specific distributions (e.g. scale free) of average node similarity, also motivating the problem of designing networks with given distributions of this measurement. In a sense, all concepts and methods that have been developed in terms of the node degree and average node degree can be extended or revisited from the perspective of the average node similarity.

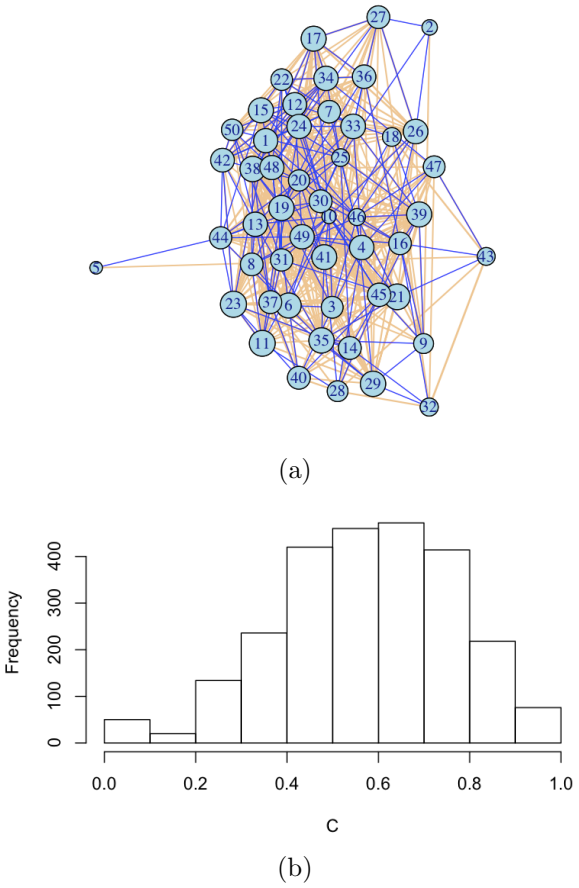
4 Model-Theoretic Networks

Having presented the basic concepts and methods to be employed in the present work, we now proceed to applying them to study model-theoretical networks, more specifically ER, BA and WS. All networks presented in the subsequent figures have been threshold for the sake of improved visualization.

Figure 1(a) shows the original networks (in blue) superimposed on the coincidence networks obtained for the ER example, with $N = 50$ nodes and uniform interconnection probability $p = 0.15$. The average node similarity is indicated by the size of the nodes in the presented network. Of particular relevance is the complete lack of relationship between the original and coincidence links.

Figure 1(b) shows the respective histogram (a) of average node similarities (coincidences). Interestingly, the average of the average node similarity resulted equal to 0.5733, which is not far from the middle of the possible excursion of similarity values, which are constrained in the interval $[0, 1]$. A small standard deviation of 0.1937

has been obtained, corroborating the regularity expected from ER networks.



(a)

(b)

Figure 1: The ER network (a) and respective histogram of the average node similarity (b) considering each of the $N = 50$ nodes. Each node was characterized by five features: node degree, clustering coefficient, betweenness centrality, as well as the average and standard deviation of the the shortest path lengths. The overall average is 0.5733, with standard deviation 0.1937.

One aspect of particular interest regards possible relationship between the average node similarity and node degree. Figure 2 illustrates the respective scatterplot obtained for the above ER network, which indicates very small joint variation for this type of uniformly random structure.

The considered BA network with $m = 2$ and $N = 50$ and respective histogram of average node similarities are depicted in Figure 3, resulting in overall average of 0.4615 and standard deviation 0.2449. Interestingly, the average node similarities resulted much smaller than for the ER, reflecting the more heterogeneous structure of this network respectively to the several considered topological measurements taken as features. It is interesting to observe the small average similarity of the hubs.

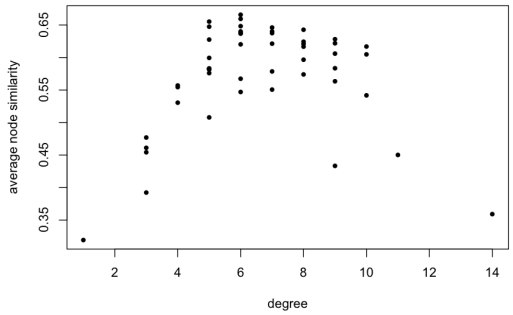
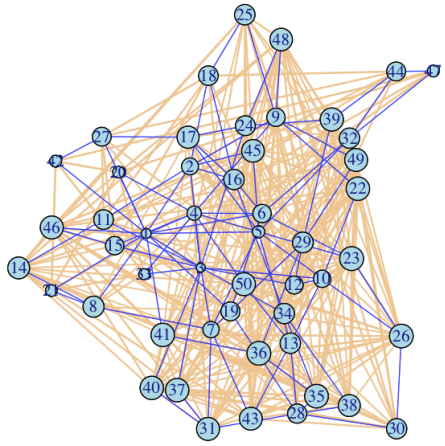
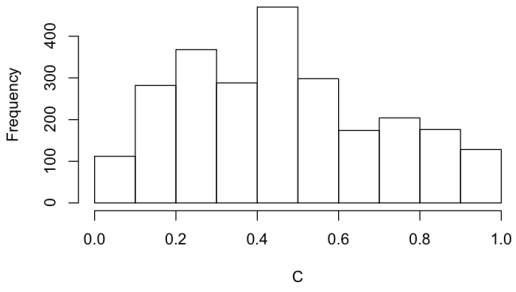


Figure 2: The scatterplot between the node degrees and the average node similarities obtained for the above ER network. A Pearson correlation coefficient of 0.1345 has been obtained, indicating lack of joint variation between these two features regarding this type of uniformly random networks.



(a)



(b)

Figure 3: The considered BA and average networks (a) with $m = 2$ and respective histogram of average node similarity (b), with the size of the nodes reflecting the respective average similarity values.

The considered WS network and its respective histogram of average node similarity is presented in Figure 4. The WS network has $N = 50$ nodes, connections with two neighbors ate each side of a node, and rewiring probability of $p_w = 0.02$. The resulting overall average node similarity is 0.6463 and standard deviation 0.2175, which

is higher than the values obtained for the ER case above. The histogram profile also resulted shifted to the right, as could be expected given that the relatively small rewiring probability has not been enough to substantially change the original topological regularity of this network. It is of particular interest to observe the small average node similarity obtained for the nodes at the extremities of the rewired connections.

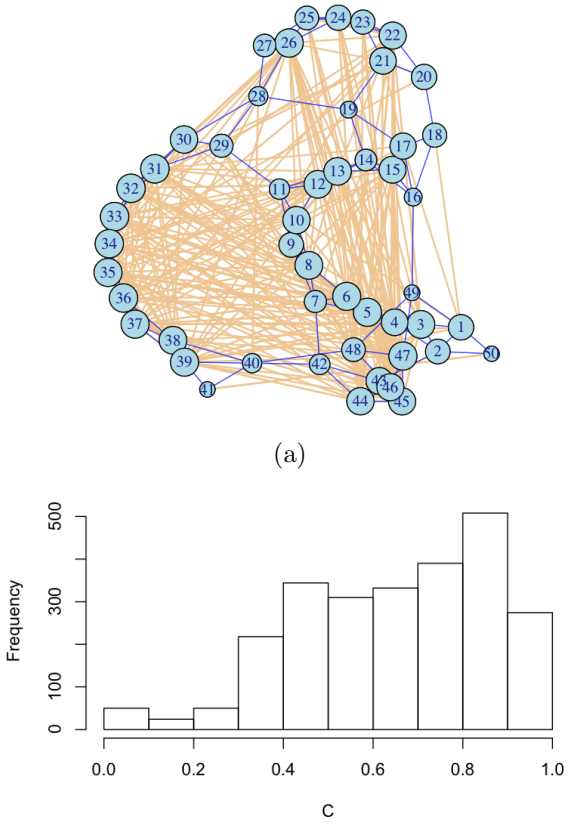
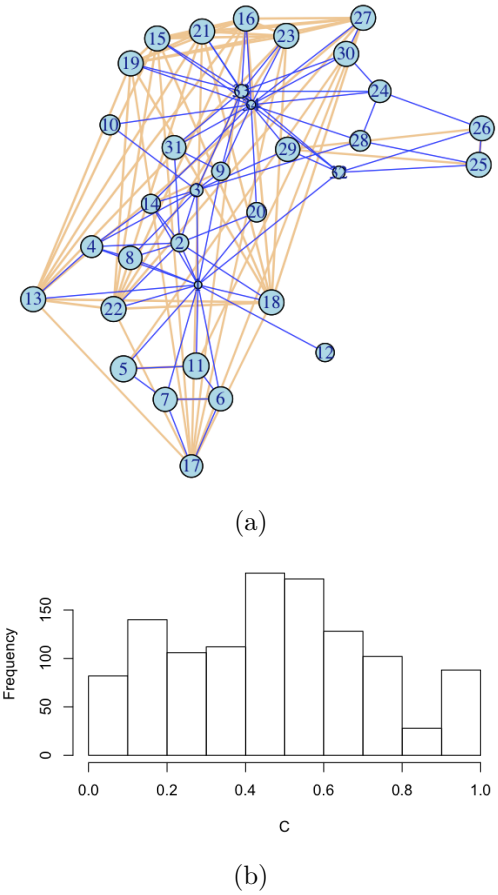


Figure 4: The considered WS network (a) with rewiring probability $p_w = 0.02$ and 2 connections to each side, and its respective histogram of average node similarity (b).

5 Real-World Networks

In order to complement the study of the application of the average node similarity concept for the characterization of networks, we now consider two real-world structures, namely the Zachary karate club [30, 31] and a network of interconnections between the cortical areas of the *macaque*, [32, 31].

Figure 5 illustrates the Zachary karate club network, which is understood to have two communities, superimposed to the average similarity network (a) and respective histogram of average node similarity. The overall average node similarity was 0.4728 with a standard deviation of 0.2590. Interestingly, this histogram bears similarity with that obtained for the ER case.



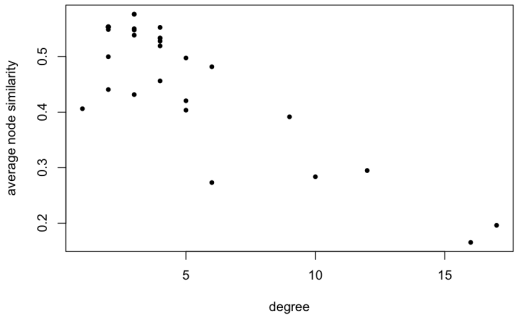
(b)

Figure 5: The Zachary karate club network (a) and its respective histogram of average similarities established by each of the $N = 34$ nodes. A particularly heterogeneous distribution can be observed that resembles that obtained for the BA case in Section 4.

We can observe that the center of the two communities, which are hubs, resulted with substantially small average node similarity in a similar manner to the hubs in the BA case. Interestingly, the nodes mediating the two communities also resulted with small average node similarity as a consequence of their connections to two distinct modules.

Indeed, as it can also be observed respectively to the BA case, the topological links tend to opposite the coincidence links. This tendency can be readily confirmed by the scatterplot between the nodes degrees and the average node similarity presented in Figure 6.

The *macaque* network and respective histogram of the average node similarities are depicted in Figure 7, yielding overall average 0.5647 and standard deviation 0.2065. This originally directed network has been symmetrized

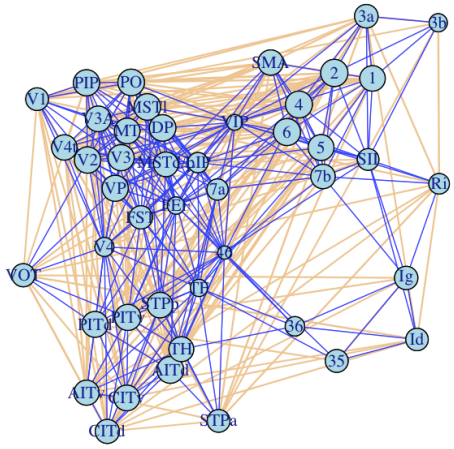


(b)

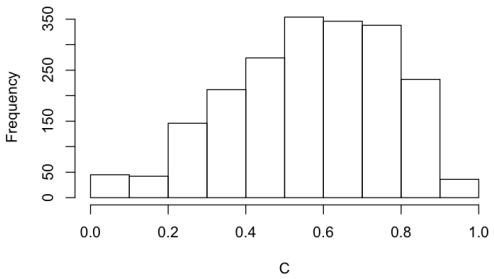
Figure 6: Scatterplot of the average node similarities in terms of the respective node degrees for the Zachary karate club network confirms an opposite tendency of topological and coincidence links for this network. The Pearson correlation coefficient is -0.86.

for the sake of compatibility with the clustering coefficient measurement. An interesting, moderately skewed histogram has been obtained in this case, with median average values, being similar to that obtained for the ER networ.

The joint visualizations of the original and coincidence similarities obtained for the macaque network, which can be seen in Figure 7(a), has a moderately modular structure. Interestingly, relatively fewer coincidence links can be observed at the core of the denser community, while the hubs are again characterized by small average node similarity. In addition, this network is characterized by a remarkable uniformity of average node similarity values, indicating that it is particularly uniform respectively to the adopted features.



(a)



(b)

Figure 7: The *macaque* network and average similarity networks (a) and respective histogram of the average node similarities obtained for the $N = 45$ nodes of this network.

6 Concluding Remarks

The present work has addressed the still challenging issue of topological characterization of the properties of complex networks in terms of the concept of coincidence similarity, which is capable of particularly selective, sensitive and robust performance. After presenting the concepts and methods, they were respectively applied to three model-theoretic networks, as well as to two real-world structures.

Several interesting concepts and results have been presented and discussed, including: (i) the strongly skewed histogram obtained for the BA case; (ii) the fact that hubs have been found to be invariably characterized by small average node similarity; (iii) a strong negative correlation between the node degrees and the average node similarity in the case of the BA and Zachary networks; as well as (iv) the identification, in the case of the macaque network, that nodes more central to the cores of the modules tend to have small average similarity; (v) the verification that the local topological structure of a network (individual links) bears virtually no relationship with the strengths of the respective links in coincidence networks; and (vi) the understanding of the node(s) with the highest average node similarity as a hub of similarity of the network (or any of its subgraphs or modules), therefore constituting a reasonable respective prototype.

The reported concepts, methods, and results corroborate the potential of the average node similarity for characterizing in a comprehensive and effective manner the topology of virtually every type of network, paving the way to a number of interesting future works. For instance, it would be interesting to extend the analysis to directed networks, as well as to understand the average node similarity as an indication of the regularity or uniformity of networks. Then, we have that all concepts and methods developed with basis on the node degree can be revisited and/or complemented from the perspective of the average node similarity. In addition, by calculating the average node similarity for successively larger subgraphs, relationships can be established with the network self-similarity (fractal). The identification of nodes

with particularly low average similarity also constitutes a possible way to identify outliers. Another particularly promising prospect regards the extension of the similarity concept for quantifying the relationship and matching between two or more networks, including applications to auto and isomorphism.

Acknowledgments.

Luciano da F. Costa thanks CNPq (grant no. 307085/2018-0) and FAPESP (grant 15/22308-2).

Note:

As all other preprints by the author, this work is possibly being considered by a scientific journal. Respective modification, commercial use, or distribution of any of its parts are not possible. This work can be cited by using the respective ResearchGate identification and/or DOI number. Thanks for reading.

References

[1] A.L. Barabási and Pósfai M. *Network Sience*. Cambridge University Press, 2016.

[2] M. Newman. *Networks: An introduction*. Oxford University Press, 2010.

[3] L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242, 2007.

[4] L. da F. Costa, O.N. Oliveira Jr., G. Travieso, F.A. Rodrigues, P.R. Villas Boas, L. Antiqueira, M.P. Viana, and L.E.C. Rocha. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. *Advances in Physics*, 60(3):329–412, 2011.

[5] L. da F. Costa. Further generalizations of the Jaccard index. https://www.researchgate.net/publication/355381945_Further_Generalizations_of_the_Jaccard_Index, 2021. [Online; accessed 21-Aug-2021].

[6] L. da F. Costa. On similarity. https://www.sciencedirect.com/science/article/pii/S037843712200334X, 2022. Physica A: Statistical Mechanics and its Applications, 127456.

[7] L. da F. Costa. Coincidence complex networks. https://iopscience.iop.org/article/10.1088/2632-072X/ac54c3, 2022. J. Phys.: Complexity, (3): 015012.

[8] M. K. Vijaymeena and K. Kavitha. A survey on similarity measures in text mining. *Machine Learning and Applications*, 3(1):19–28, 2016.

[9] L. da F. Costa. Multiset neurons. https://www.researchgate.net/publication/356042155_Multiset_Neurons, 2021.

[10] R. dos Reis and L. da F. Costa. Enzyme similarity networks. https://www.researchgate.net/publication/360263872_Enzyme_Similarity_Networks, 2022.

[11] S.-H. Cha. Comprehensive survey on distance/similarity measures between probability density functions. *Intl. J. Math. Models and Meths. in Appl. Sciences*, 1(4):300–307, 2007.

[12] C. E. Akbas, A. Bozkurt, M. T. Arslan, H. Aslanoglu, and A. E. Cetin. L1 norm based multiplication-free cosine similiarity measures for big data analysis. In *IEEE Computational Intelligence for Multimedia Understanding (IWCIM)*, France, Nov. 2014.

[13] B. Mirkin. *Mathematical Classification and Clustering*. Kluwer Academic Publisher, Dordrecth, 1996.

[14] L. Hamers, Y. Hemeryck, G. Herweyers, M. Janssen, H. Ketters, H. Rousseau, and A. Vanhoutte. Similarity measures in scientometric research: The jaccard index versus salton's cosine formula. *Information Processing and Management*, 25(3):315–318, 1989.

[15] L. Leydesdorff. On the normalization and visualization of author co-citation data: Salton's cosine versus the jaccard index. *Journal of the American Society for Information Science and Technology*, 59(1):77–85, 2008.

[16] K. Kavitha, B. Sandhya, and B. T. Rao. Evaluation of distance measures for feature based image registration using Alexnet. *International Journal of Advanced Computer Science and Applications*, 9(10), 2018.

[17] M. Brusco, J. D. Cradit, and D. Steinley. A comparison of 71 binary similarity coefficients: The effect of base rates. *PLOS One*, 16(4):e0247751, 2021.

[18] P. Jaccard. Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin de la Société vaudoise des sciences naturelles*, 37:241–272, 1901.

[19] P. Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société vaudoise des sciences naturelles*, 37:547–549, 1901.

[20] B. K. Samanthula and W. Jiang. Secure multiset intersection cardinality and its application to jaccard coefficient. *IEEE Transactions on Dependable and Secure Computing*, 13(5):591–604, 1989.

[21] Wikipedia. Jaccard index. https://en.wikipedia.org/wiki/Jaccard_index. [Online; accessed 10-Oct-2021].

[22] A. Schubert and A. Telcs. A note on the Jaccardized Czekanowski similarity index. *Scientometrics*, 98:1397–1399, 2014.

[23] J. Hein. *Discrete Mathematics*. Jones & Bartlett Pub., 2003.

[24] D. E. Knuth. *The Art of Computing*. Addison Wesley, 1998.

[25] W. D. Blizard. Multiset theory. *Notre Dame Journal of Formal Logic*, 30:36—66, 1989.

[26] W. D. Blizard. The development of multiset theory. *Modern Logic*, 4:319–352, 1991.

[27] P. M. Mahalakshmi and P. Thangavelu. Properties of multisets. *International Journal of Innovative Technology and Exploring Engineering*, 8:1–4, 2019.

[28] D. Singh, M. Ibrahim, T. Yohana, and J. N. Singh. Complementation in multiset theory. *International Mathematical Forum*, 38:1877–1884, 2011.

[29] L. da F. Costa. Supervised and unsupervised pattern recognition and their performance. https://www.researchgate.net/publication/360936159_Supervised_and_Unsupervised_Pattern_Recognition_and_their_Performance, 2022.

[30] W. W. Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, 33(4):452–473, 1977.

[31] G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

[32] L. Négyessy, T. Nepusz, L. Kocsis, and F. Bazsó. Prediction of the main cortical areas and connections involved in the tactile function of the visual cortex by network analysis. *European Journal of Neuroscience*, 23(7):1919–1930, 2006.