SIS | 2022

51st Scientific Meeting
of the Italian Statistical Society

Caserta, 22-24 June

Università degli Studi della Campania *Luigi Vanvitelli*

Società Italiana di Statistica

www.unicampania.it

Book of the Short Papers

Editors: Antonio Balzanella, Matilde Bini, Carlo Cavicchia, Rosanna Verde

Città di Caserta

1222-2022 800 ANNI

UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE STATISTICHE

STATA

sas

UNIVERSITÀ DEGLI STUDI DEL SANNIO Benevento

Pearson

Matilde Bini (Chair of the Program Committee) - *Università Europea di Roma*
Rosanna Verde (Chair of the Local Organizing Committee) - *Università della Campania "Luigi Vanvitelli"*

PROGRAM COMMITTEE

Matilde Bini (Chair), Giovanna Boccuzzo, Antonio Canale, Maurizio Carpita, Carlo Cavicchia, Claudio Conversano, Fabio Crescenzi, Domenico De Stefano, Lara Fontanella, Ornella Giambalvo, Gabriella Grassia - Università degli Studi di Napoli Federico II, Tiziana Laureti, Caterina Liberati, Lucio Masserini, Cira Perna, Pier Francesco Perri, Elena Pirani, Gennaro Punzo, Emanuele Raffinetti, Matteo Ruggiero, Salvatore Strozza, Rosanna Verde, Donatella Vicari.

LOCAL ORGANIZING COMMITTEE

Rosanna Verde (Chair), Antonio Balzanella, Ida Camminatiello, Lelio Campanile, Stefania Capecchi, Andrea Diana, Michele Gallo, Giuseppe Giordano, Ferdinando Grillo, Mauro Iacono, Antonio Irpino, Rosaria Lombardo, Michele Mastroianni, Fabrizio Maturo, Fiammetta Marulli, Paolo Mazzocchi, Marco Menale, Giuseppe Pandolfi, Antonella Rocca, Elvira Romano, Biagio Simonetti.

ORGANIZERS OF SPECIALIZED, SOLICITED, AND GUEST SESSIONS

Arianna Agosto, Raffaele Argiento, Massimo Aria, Rossella Berni, Rosalia Castellano, Marta Catalano, Paola Cerchiello, Francesco Maria Chelli, Enrico Ciavolino, Pier Luigi Conti, Lisa Crosato, Marusca De Castris, Giovanni De Luca, Enrico Di Bella, Daniele Durante, Maria Rosaria Ferrante, Francesca Fortuna, Giuseppe Gabrielli, Stefania Galimberti, Francesca Giambona, Francesca Greselin, Elena Grimaccia, Raffaele Guetto, Rosalba Ignaccolo, Giovanna Jona Lasinio, Eugenio Lippiello, Rosaria Lombardo, Marica Manisera, Daniela Marella, Michelangelo Misuraca, Alessia Naccarato, Alessio Pollice, Giancarlo Ragozini, Giuseppe Luca Romagnoli, Alessandra Righi, Cecilia Tomassini, Arjuna Tuzzi, Simone Vantini, Agnese Vitali, Giorgia Zaccaria.

ADDITIONAL COLLABORATORS TO THE REVIEWING ACTIVITIES

Ilaria Lucrezia Amerise, Ilaria Benedetti, Andrea Bucci, Annalisa Busetta, Francesca Condino, Anthony Cossari, Paolo Carmelo Cozzucoli, Simone Di Zio, Paolo Giudici, Antonio Irpino, Fabrizio Maturo, Elvira Romano, Annalina Sarra, Alessandro Spelta, Manuela Stranges, Pasquale Valentini, Giorgia Zaccaria.

PUBLISHED BY PEARSON

WWW.PEARSON.COM

ISBN 9788891932310

An Explorative analysis of Different Distance Metrics to Compare Unweighted Undirected Networks

Analisi Esplorativa di Differenti Metriche di Distanza per Confrontare Reti Non Orientate e Non Pesate

Anna Simonetto[1], Matteo Ventura[2], Gianni Gilioli[1]

Abstract Networks are mathematical structures that make it possible to represent complex systems by characterising the relationships existing between the various elements of the network. In order to understand the differences between two different systems, tools are needed to quantify these differences. The purpose of this work is an exploratory analysis of the impacts that different distance metrics have on the comparative evaluation of two networks, at increasing levels of network perturbation. Preliminary results show that, depending on the chosen metric, it is possible to amplify or reduce the effect of perturbation. As the disturbance increases, the differences between the distance assessment systems are reduced.

Abstract *Le reti sono strutture matematiche che permettono di rappresentare sistemi complessi caratterizzando le relazioni esistenti tra i vari elementi della rete. Al fine di comprendere le differenze tra due sistemi diversi, sono necessari strumenti per quantificare queste differenze. Lo scopo di questo lavoro è un'analisi esplorativa degli impatti che diverse metriche di distanza hanno sulla valutazione comparativa di due reti, a livelli crescenti di perturbazione della rete. I risultati preliminari mostrano che, a seconda della metrica scelta, è possibile amplificare o ridurre l'effetto della perturbazione. All'aumentare della perturbazione, le differenze tra i sistemi di valutazione della distanza si riducono.*

Key words: Undirect graph, Euclidean distance, Spectral method, adjacency matrix, ecological networks.

---

[1] Dip. di Ingegneria Civile, Architettura, Territorio, Ambiente e di Matematica, Università di Brescia; anna.simonetto@unibs.it, gianni.gilioli@unibs.it;

[2] Dip. di Economia e Management, Università di Brescia; m.ventura007@unibs.it.

1 Comparing networks

A network is a data structure describing the interaction pattern existing between a set of elements (e.g., in an ecological perspective the elements could be species, individuals, habitats) by the properties of the links that connect each pair of elements, defined as arcs. In unweighted networks the arc is characterized only by the presence/absence (its value is 1 or 0, respectively). In weighted networks, the weight is proportional to the intensity of the relationship existing between the two elements, usually varying in the range [0, 1] or [-1, 1] in the case of signed networks. The comparison of two networks allows the differences in the two systems to be assessed in terms of the differences in the relationships between the elements.

In research on complex systems, the problem of comparing networks is ubiquitous and not trivial. A tradeoff among interpretability, effectiveness of the results and computational efficiency is needed [8]. In fact, the method applied to measure differences between networks influences the result of network comparison. The choice of the correct method must therefore firstly be based on a clear identification of which differences are to be measured and then on an understanding of how the different methods considered represent these differences.

Some authors have already explored several methods to compare networks. For instance, Tantardini et al. [8] presented a review on methods for comparing networks, both for networks with node correspondence and for networks with unknown node correspondence and tested them on synthetic random networks under some types of perturbations. Wilson and Zhu [9] examined the performance of spectra as graph representation. Specifically, the authors investigated the cospectrality of various matrix representations and compared the Euclidean distance between spectra and the edit distance between graphs.

The aim of the work is to understand how the different used distance metrics describe situations of diversity between networks. Specifically, we explored the performances of three distance measures and three spectral methods applied to compare networks created with the same nodes.

A simulation study was conducted to understand the specific response characteristics of the investigated methods. Specifically, we considered as perturbation the deletion of an edge and with these simulations we aimed at understanding: i) how these metrics react to an increasing number of perturbations, ii) what is the reaction when a node is completely disconnected, and iii) the impact of node degree on the value of the distance.

2 Distance metrics

Mathematically, the network can be represented by a graph. A graph is described by $G = (V, E)$, where $V = \{1, \dots, n\}$ is the set of vertex, i.e. the nodes or elements of the network, and E is the edge set, that is a subset of the set $V \times V$ of ordered pairs of distinct vertices or nodes. The graph is said to be undirected if the edge (i, j) and its

opposite (j, i) are both in E [4]. A graph can be represented by a squared $N \times N$ adjacency matrix $\boldsymbol{A}$ where each row/column corresponds to a node and each cell represents an edge, set to one if the edge exists and zero otherwise. The adjacency matrix of an undirected graph is symmetric [2].

Two classes of methods for comparing networks are exposed hereafter: (1) distances between adjacency matrices, (2) spectral methods.

2.1 Distances between adjacency matrices

Given two $N \times N$ adjacency matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$, four distances are identified: Minkowski, Manhattan, Euclidean and Chebyshev distance.

Minkowski Distance Also known as L_p-metric. It is identified by the following expression [6]:

$$d_{MK}(\boldsymbol{A}_1, \boldsymbol{A}_2) = \left[\sum_{i,j=1}^{N} |a_{1,ij} - a_{2,ij}|^p\right]^{\frac{1}{p}} \qquad p \geq 1. \tag{1}$$

Manhattan Distance Also known as L_1-metric. Its definition is based on Minkowski distance, using $p = 1$ [7]. It corresponds to the sum of the absolute differences, i.e., all the matrices' elements are equally weighted.

Euclidean Distance Also known as L_2-metric. It is a specific case of Minkowski distance, using $p = 2$ [3]. With this distance, greater weight is attached to longer distances.

Chebyshev Distance It is also called $L_\infty-$metric and it is a special case of Minkowski distance where p goes to infinity [1], i.e., only the highest difference influences the metric value:

$$d_{CH}(\boldsymbol{A}_1, \boldsymbol{A}_2) = \lim_{p\to\infty}\left[\sum_{i,j=1}^{N} |a_{1,ij} - a_{2,ij}|^p\right]^{\frac{1}{p}} = \max_{i,j} |a_{1,ij} - a_{2,ij}|. \tag{2}$$

2.2 Spectral methods

These approaches are based on spectral theory, that allows to describe networks' structural properties through eigenvalues and eigenvector of a matrix.

The spectrum $\boldsymbol{\Lambda}$ is the sorted sequence of matrix eigenvalues λ and the spectral distance between two graphs is the Euclidean distance between the two correspondent matrices' eigenvalues [9]:

$$d_\lambda(\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2) = \sqrt{\sum_{i=1}^{N} \left(\lambda_{1,i} - \lambda_{2,i}\right)^2}. \quad (3)$$

In addition to the adjacency matrix, there are two other possible matrix representations of a graph: the Laplacian and Normalised Laplacian matrix.

The Laplacian matrix $\boldsymbol{L}$ is defined as $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$, where is the adjacency matrix of the graph and $\boldsymbol{D}$ is the diagonal degree matrix [9].

The Normalised Laplacian matrix, instead, is defined as:

$$\mathcal{L}(i,j) = \begin{cases} 1 & if\ i = j \text{ and } d_i \neq 0 \\ -\dfrac{1}{d_i d_j} & if\ i \neq j \text{ and } i \text{ is adiacent to } j, \\ 0 & \text{otherwise} \end{cases}$$

where d_i and d_j are respectively the degree of the node i and the degree of the node j. The Normalised Laplacian can be also written as $\boldsymbol{\mathcal{L}} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$.

3 Simulation study

In this section we test the responses of the above proposed metrics through a simulation experiment. All the analyses have been implemented in R [5].

We started from a full connected, unweighted network composed by twelve nodes. At each iteration the distance between the fully connected network and a disturbed network will be calculated. At step 0 the disturbed network is equal to the fully connected network and itself. At each subsequent iteration the comparison network is perturbed by removing a connection arc. After each perturbation the distance of the disturbed network from the full connected one was measured. This is aimed at checking the distances response with increasing number of removed edges, and thus, with decreasing graph density.

To investigate the distances response to the decrease of the node degree and to the complete disconnection of a node, an ordered edge removal criterion was applied. Considering the upper triangular half of the completely connected graph's adjacency matrix, the ordered removal consists in turn into zero the elements of the first row and proceed with the next row only when the previous is full of zeroes; namely, if we consider a graph, the ordered removal consists in removing all the edges from a node and proceed with another node only when the previous one is completely disconnected from the network. This edge removal criterion leads to the generation of a set of nested networks, i.e., the edges removed from the network with n perturbations do not exist in all the subsequent networks with $n + m$ removed edges.

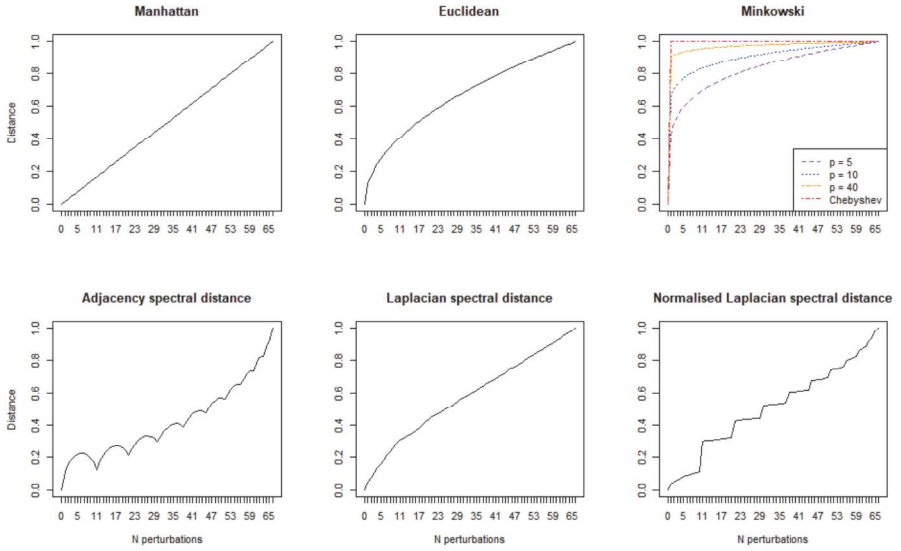
From the results reported in Figure 1, it appears that all the distance metrics have a growing monotonous pattern. We note that, considering the Minkowski distance,

when p increases, the metric distances tend to weight equally the perturbations independently of the number of deleted edges. Namely, the Manhattan distance weight equally all the perturbations, up to the extreme case represented by Chebyshev distance that weights in the same way one perturbation or infinite perturbations.

Instead, spectral methods show some interesting patterns: both the adjacency and the Normalised Laplacian spectral distance show a non-monotonous pattern. We observe level-shifts that seem to be dependent on the position of the removed edges, indeed, spectral methods present a step pattern where the level change happens when, once a node is completely disconnected, the first edge of another node is removed.

This first study shows that, particularly for disturbances of less than 50% of the number of arcs in the fully connected network, the choice of comparison metric influences the assessment of the distance between the compared networks.

Figure 1:Results of the ordered edge removal



4 Conclusions

In this work we compare different distance metrics to estimate the differences between two or more networks.

The simulation study highlights the differences in the distances estimation using four distance metrics and three spectral methods. The networks comparison is performed according to an increasing levels of perturbation. The Manhattan distance estimation resulted not influenced by the disturbance levels, Euclidean distance weighs more heavily on differences at low disturbance levels, Normalised Laplacian spectral

distance weighs more heavily on disturbances that lead to completely disconnecting a network node.
Further developments of the study will be devoted to deep the possible dependance of the distance estimation from the order in which edges are removed.

References

1. Coghetto, R. (2016). Chebyshev Distance. *Formalized Mathematics*, *24*(2), 121–141. https://doi.org/10.1515/forma-2016-0010
2. Coscia, M. (2021). The Atlas for the Aspiring Network Scientist. *ArXiv:2101.00863 [Physics]*. http://arxiv.org/abs/2101.00863
3. Dattorro, J. (2008). *Convex optimization & Euclidean distance geometry* (Version 2008.02.29). Meboo.
4. Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press.
5. R Core Team (2021). R: *A language and environment for statistical computing.* Vienna, Austria. https://www.R-project.org/.
6. Richardson, G. D. (1981). *The appropriateness of using various Minkowskian metrics for representing cognitive configurations*. *13*, 475–485.
7. Sowell, K. O. (1989). Taxicab Geometry—A New Slant. *Mathematics Magazine*, *62*(4), 238–248. https://doi.org/10.1080/0025570X.1989.11977445
8. Tantardini, M., Ieva, F., Tajoli, L., & Piccardi, C. (2019). Comparing methods for comparing networks. *Scientific Reports*, *9*(1), 17557. https://doi.org/10.1038/s41598-019-53708-y
9. Wilson, R. C., & Zhu, P. (2008). A study of graph spectra for comparing graphs and trees. *Pattern Recognition*, *41*(9), 2833–2841. https://doi.org/10.1016/j.patcog.2008.03.011