Social Networks as Symbolic Data

Giuseppe Giordano and Paula Brito

Abstract Starting from the main idea of Symbolic Data Analysis to extend Statistics and Data Mining methods from first-order to second-order objects, we focus on network data—as defined in the framework of Social Network Analysis—to define a graph structure and the underlying network in the context of complex data objects. A Network Symbolic description is defined according to the statistical characterization of the network topological properties. We use suitable network measures, which are represented by means of symbolic variables. Their study through multidimensional data analysis, allows for the synthetic representation of a network as a point onto a metric space. The proposed approach is discussed on the basis of a simulation study considering three classical network growth processes.

Keywords Histogram-valued data • Social network analysis • Symbolic data

1 Introduction

In recent years the network paradigm has affirmed as one of the most attractive and valuable cognitive models to describe and represent the complexity of relationships among a wide variety of actors. In a broader sense, the concept of network could be applied to any kind of actors able to establish relationships. However, the concept of network assumes special importance when the actors are individuals and relationships are related to specific state of properties attached to each pair

G. Giordano (✉)
Department of Economics and Statistics, Università di Salerno, Salerno, Italy
e-mail: ggiordan@unisa.it

P. Brito
Fac. Economia & LIAAD INESC TEC, Univ. Porto, Porto, Portugal
e-mail: mpbrito@fep.up.pt

D. Vicari et al. (eds.), *Analysis and Modeling of Complex Data in Behavioral and Social Sciences*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-06692-9_15,

of subjects (personal relations such as trustee, acquaintance, collaborations, and so on). These kinds of networks take into account human beings and the study of their birth, growth, shape and topology is the scope of Social Network Analysis. Nevertheless, the concept of network is so immediate and easy to be generalized that the underlying paradigm has been successfully applied in different fields, ranging from Communication and Transports to Economics as well as Medicine, Physics, Linguistics, Computer Science and many others. As a consequence, different disciplines contributed in different and not exclusive ways to the definition and interpretation of several network measurements. Such metrics have been found as specific features of classes of networks. The definition of network data by complex data objects is based on the different structural information that can be of interest to retrieve. The specific choice of suitable network indices expressed through statistical distributions will be addressed one by one according to the different network features that one wishes to highlight.

The idea of this work is to aggregate information attached to each node in terms of centrality and role (bridge, isolate, transmitter, etc.) in the network and express it as symbolic data by means of histogram-valued variables—see, e.g., Bock and Diday (2000), Noirhomme-Fraiture and Brito (2011)—so that the whole network can be expressed by a "vector" of high-order data (e.g. histograms). In this work, we consider the distributions of some typical network indices, measured at node-level, represented as histogram-valued variables. A symbolic data table will be defined, where each row pertains to one network and the columns hold the network indices. Symbolic data analysis of such data may be applied for the sake of comparisons among networks emerged at different occasions in time, computing similarities among networks, or representing networks as "points" on a reduced metric space, to cite just some possibilities.

The paper is organized as follows. Section 2 gives formal definitions and introduces basic statistical indices commonly used to describe networks. Section 3 recalls the general Symbolic Data framework and more specifically the concept of histogram-valued variables that will be used in the definition of the network symbolic descriptions. These descriptions and the resulting data table are defined in Sect. 4. A Simulation Study based on three network growth models (*Random Graph, Preferential Attachment and Small World*) is carried out in Sect. 5 aimed at describing the procedure and analyze its capability to discriminate among different network structures. Section 6 concludes the paper, pointing out directions for further research.

2 Statistical Description of Network Graph Characteristics

The statistical analysis of a network is basically performed with a descriptive purpose and originated in the framework of Social Network Analysis, see Wasserman and Faust (1994). Formally, network data refer to a set of actors and

their relationships are commonly described and represented in the mathematical framework of *Graph Theory*. A graph data-structure is characterized by two sets: nodes and edges. Let $\mathcal{G}(N, E)$ be the graph represented by the set N of nodes (vertices) with cardinality $n = |N|$ and by the set E of edges with cardinality $m = |E|$. A fundamental concept in the description of a network is the centrality position of each node in the graph; in Social Network Analysis the definition of centrality may vary according to different criteria. The most important node-level centrality measure is the *Degree*. The Degree of a node is defined as the number of edges that connect to it. As a starting point, we consider only undirected simple finite graphs, that is, graphs where edges have no orientation, nodes have no loops, and where no more than one edge exists between any two nodes. There exist many statistical characterizations of a network according to its structural properties; among the important node-level statistics, we consider Closeness, Betweenness and Eigenvector centrality, see Freeman (1979) for definitions and interpretations.

The Degree tells about the number of connections a node has to other nodes and its distribution has been studied for real and theoretical network models, from the simplest Bernoulli random graph (Erdös and Rényi 1960) where it follows a Binomial distribution (limiting to Poisson for large n), to more complex models such as *scale-free networks* whose degree distribution follows approximately a *power law* of the form: $P(d) \sim d^{-\lambda}$, where λ is a constant. An important subclass of scale-free model is the *Preferential Attachment* generation process—also known as *Cumulative Advantage* process—first introduced to study the occurrence of power laws in scientific citation networks, see De Solla Price (1976), and then for explaining the presence of hubs in some parts of the *World Wide Web*, for which the constant λ should vary between *2* and *3*, see Barabási and Albert (1999). Network statistical measures refer either to individual nodes and edges (local measures) or to the network as a whole (global measures). They are defined for binary and weighted variables, for directed and undirected graphs (Kolaczyk 2009). Sometimes we are interested in detecting local measures in specific sub-parts of a graph. The presence of separate components in a graph defines an intermediate level of analysis. In this case the interest is in exploring not all individual nodes/edges but a small part of the graph: a connected subgraph, the giant component (i.e. the connected component with the larger number of nodes). It could also be interesting to study graph partitions induced by hierarchical clustering methods defined by similarity measures among nodes, see Batagelj (1998). Indeed, local network measures may be applied to sub-graphs too.

The definition of network data as a complex object allows considering the different structural information that can be of interest to retrieve, according to some theory-driven structural properties, depending on the field of study.

3 Symbolic Data

In classical statistics and multivariate data analysis the units under analysis are single entities described by numerical and/or categorical variables, each one taking one single value for each variable. Data are organized in a data-array, where each cell (i, j) contains the value of variable j for individual i. However, when analyzing a group rather than a single individual, the within-group variability should be explicitly considered. Consider, for instance, that we are analyzing the staff of some institutions, in terms of age, education level and category. If we just take averages or mode values within each institution, much information is lost. The same issue arises when we are interested in concepts and not in single specimen—whether it is the animal species (and not a specific animal), a model of car, etc. Symbolic Data Analysis, see, e.g., Bock and Diday (2000) and Noirhomme-Fraiture and Brito (2011), provides a framework where the variability intrinsic to a concept as a whole, or resulting from the aggregation of individual observations into groups, is considered in the data representation, and methods developed to take it into account. To describe groups of individuals or concepts, variables assume other forms of realizations; the new variable types, called "symbolic variables", may assume multiple, possibly weighted, values for each entity. Data are gathered in a matrix, now called a "symbolic data table", each cell containing "symbolic data". Each row of the table corresponds to a group, or concept, i.e., the entity of interest. A numerical variable may then be single valued, as in the classical framework, if it takes one single value of an underlying domain per entity, it is multi-valued if its values are finite subsets of the domain and it is an interval variable if its values are intervals. When an empirical distribution over a set of sub-intervals is given, the variable is called a histogram-valued variable. In this study, we shall represent information on networks, expressed by statistical distributions of node-level measures, by histogram-valued variables.

3.1 Histogram-Valued Variables

Let $S = \{s_1, \ldots, s_r\}$ be the set of entities under analysis. For an histogram variable Y (see Bock and Diday 2000) each element $s_i \in S$ is described by a discrete probability or frequency distribution on the set of considered sub-intervals $\{I_{i1}, \ldots, I_{ik_i}\}$ such that $Y(s_i) = (I_{i1}, p_{i1}; \ldots; I_{ik_i}, p_{ik_i})$ with $p_{i\ell}$ the probability or frequency associated to $I_{i\ell} = [\underline{I}_{i\ell}, \overline{I}_{i\ell}[$, $\ell \in \{1, \ldots, k_i\}$, and $p_{i1} + \ldots + p_{ik_i} = 1$. A *Uniform* distribution is assumed within each sub-interval $[\underline{I}_{i\ell}, \overline{I}_{i\ell}[$. For each observation s_i, $Y(s_i)$ can, alternatively, be represented by the cumulative distribution function $F_i(x)$, or by its inverse, the quantile function $q_i(t)$, both piecewise linear functions, given by

Table 1 Distribution of degree (number of friends) for two classes of students

	Degree
Class 1	([0, 4[, 0.2; [4, 8[, 0.5; [8, 12[, 0.2; [12, 16], 0.05; [16, 20], 0.05)
Class 2	([0, 4[, 0.05; [4, 8[, 0.4; [8, 12[, 0.25; [12, 16[, 0.2; [16, 20], 0.1)

$$F_i(x) = \begin{cases} 0\ , & x < \underline{I}_{i1} \\ p_{i1}\frac{x-\underline{I}_{i1}}{\underline{I}_{i2}-\underline{I}_{i1}}, & \underline{I}_{i1} \le x < \underline{I}_{i2} \\ F(\underline{I}_{i2}) + p_{i2}\frac{x-\underline{I}_{i2}}{\underline{I}_{i3}-\underline{I}_{i2}}, & \underline{I}_{i2} \le x < \underline{I}_{i3} \\ \vdots & \\ F(\underline{I}_{i(k_i-1)}) + p_{i(k_i)}\frac{x-\underline{I}_{ik_i}}{\overline{I}_{k_i}-\underline{I}_{ik_i}}, & \underline{I}_{ik_i} \le x < \overline{I}_{ik_i} \\ 1, & \overline{I}_{ik_i} \le x \end{cases}$$

$$q_i(t) = \begin{cases} \underline{I}_{i1} + \frac{t}{w_{i1}}a_{i1}, & 0 \le t < w_{i1} \\ \underline{I}_{i2} + \frac{t-w_{i1}}{w_{i2}-w_{i1}}a_{i2}, & w_{i1} \le t < w_{i2} \\ \vdots & \\ \underline{I}_{ik_i} + \frac{t-w_{ik_i-1}}{1-w_{ik_i-1}}a_{ik_i}, & w_{ik_i-1} \le t \le 1 \end{cases}$$

where $w_{ih} = \sum_{\ell=1}^{h} p_{i\ell}, h = 1, \ldots, k_i; a_{i\ell} = \overline{I}_{i\ell} - \underline{I}_{i\ell}$ for $\ell = \{1, \ldots, k_i\}$.

If this latter representation is chosen, then the observations $Y(s_i)$ should be re-written using the same weight distribution, to allow for the comparison of the corresponding quantile functions, since this procedure leads to functions with the same number of terms corresponding to the same sub-intervals of the unit interval.

Henceforth "distribution" refers to a probability or frequency distribution of a numerical variable represented by a histogram or a quantile function.

Example 1. Consider two classes of students, for which we know the friendship relation among classmates, the friendship networks are described by the respective Degree distributions, as in Table 1. In Class 1 20 % of the students have a number of friends (degree) less than 4, 50 % have degree between 4 and 7, 20 % between 8 and 11, 5 % between 12 and 15, and 5 % between 16 and 20; likewise for Class 2. The units of interest are the classes as a whole and not each individual student. Figure 1a represents the histograms of variable "Network Degree" for Class 1 and Class 2, and Fig. 1b depicts the respective quantile functions.

4 Representation of Networks by Symbolic Variables

To represent a network as symbolic data, we consider the empirical distribution of network measurements referring to each node (Degree, Closeness, etc.) and represent them by histograms. In the resulting symbolic data table, each row pertains

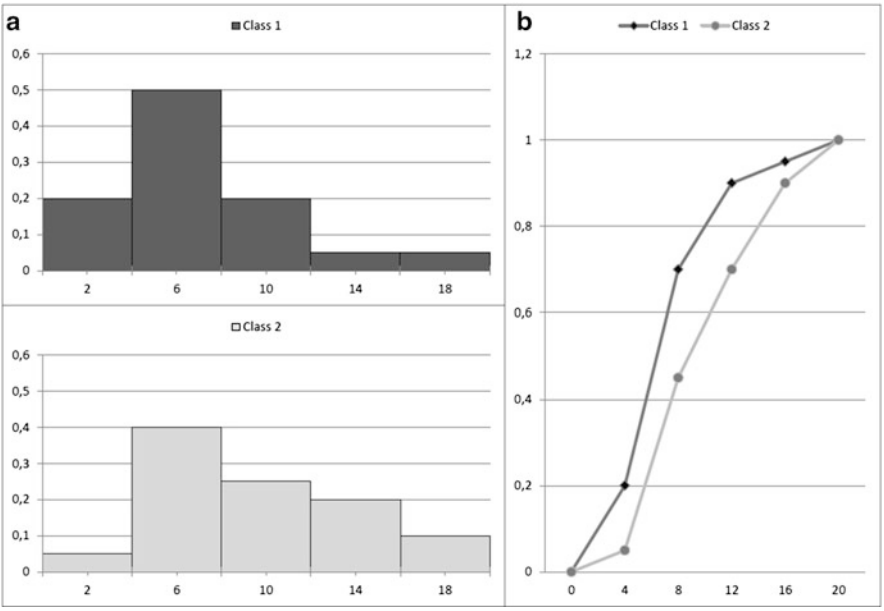

Fig. 1 Representation of the Degree of Class 1 and Class 2 by histograms (**a**) and quantile functions (**b**)

to a different network and each column to a network index, each cell recording the distribution of the corresponding index in the given network. Consider, for example, that the degree of each node of each network under study has been computed. Then, for each given network, the distribution of the degree values may be obtained, and represented in the form of a histogram (or by the corresponding quantile function). The same may be done for all network indices under analysis. Thereby, each row of the data array corresponds to a description of a network, as a "vector" of histograms. Distances between such descriptions may be used to compare several networks and represent them as "points" in a reduced metric space.

5 Simulation Study

A simulation study is carried out to generate several network data structures. The simulation scheme controls for two attributes: Graph order and Generating process. A third attribute, the parameter regulating the process is specific to each process. Each of the three factors has three levels, leading to a total of 27 different network data structures. Each type of network is replicated 100 times. The considered levels are: (1) the order of the graph: $n \in \{100; 300; 500\}$; (2) the generating process: $P \in \{$*Random Graph*; *Preferential Attachment*; *Small-World*$\}$

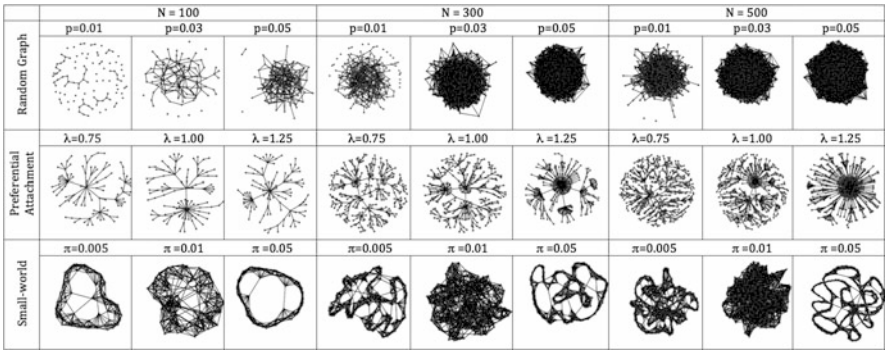

Fig. 2 Examples of the networks generated for each of the 27 configurations

and (3) the process parameter: for each generating process a specific parameter controls, respectively: the density (p) of the Random Graph: $p \in \{0.01; 0.03; 0.05\}$; the power ($\lambda$) of the Preferential Attachment: $\lambda \in \{0.75; 1.00; 1.25\}$; the rewiring probability (π) of the Small-World model: $\pi \in \{0.005; 0.01; 0.05\}$. Figure 2 shows examples of the networks generated for the different network data structures.[1]

For each of the 100×27 networks, we have computed the Betweenness Centrality, the Closeness, the Degree and the Eigenvector Centrality, and the corresponding order statistics; these were then summarized by the corresponding median values for each of the 27 network types. These data were then standardized, given the different network sizes involved. Finally, the network data are represented in a 27×4 symbolic data matrix, containing in each cell the distribution of the index in column for the network type in that row. Different multidimensional symbolic data analysis could be performed on the obtained symbolic data array. In this work, ascending hierarchical clustering has been carried out, using the Mallows' distance, which is adapted to the kind of data at hand—see also Irpino and Verde (2006). The Mallows' distance is computed using the quantile functions of the distributions in the symbolic data table. For each variable Y_j, $j = 1, \ldots, 4$ (each network index in our case), we have $d_M^2(R_{i_1 j}, R_{i_2 j}) = \int_0^1 (q_{i_1 j}(t) - q_{i_2 j}(t))^2 \, dt$ where q_{ij} is the quantile function of the distribution of variable Y_j for network R_i. The global squared distance between network types is then computed additively on the variables, $D_M^2(R_{i_1}, R_{i_2}) = \sum_{j=1}^{4} d_M^2(R_{i_1 j}, R_{i_2 j})$. The use of the same quantile representation for each network makes it possible to directly compare the quantile functions. Moreover, in Irpino and Verde (2006) it is proved that if a Uniform distribution is assumed in each sub-interval of the histograms, the squared Mallows' distance may be re-written using the midpoints and half-ranges of these sub-intervals (in number

[1] Simulations and network statistics are obtained by: *R version 2.15.2 (2012-10-26). Base packages: base, datasets, graphics, grDevices, methods, stats, utils; other: igraph 0.6.5-1, sna 2.2-1.*

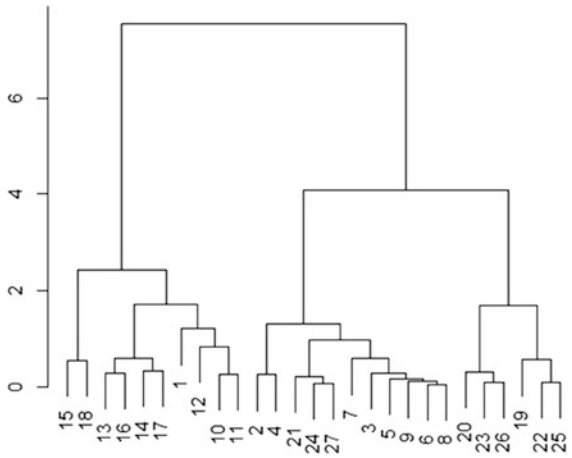

Fig. 3 Dendrogram on the 27 network types, using the Mallows' distance on standardized data, and the Ward aggregation criterion

of K_j) : $d_M^2(R_{i_1 j}, R_{i_2 j}) = \sum_{\ell=1}^{K_j} p_\ell \left[(c_{i_1 j \ell} - c_{i_2 j \ell})^2 + \frac{1}{3}(r_{i_1 j \ell} - r_{i_2 j \ell})^2 \right]$, where, for the histogram-valued variable j and network i, $c_{ij\ell} = \frac{\overline{I}_{ij\ell} + \underline{I}_{ij\ell}}{2}$ is the midpoint of the interval $I_{ij\ell}, \ell \in \{1, \ldots, K_j\}$ and $r_{ij\ell} = \frac{\overline{I}_{ij\ell} - \underline{I}_{ij\ell}}{2}$ the corresponding half-range. In our implementation, we have described each distribution by a histogram with 100 sub-intervals, defined by the distributions' percentiles, i.e., $K_j = 100, j = 1, \ldots, 4$ and $p_\ell = 0.01, \ell = 1, \ldots, K_j$. Applying the Mallows' distance $D_M(R_{i_1}, R_{i_2})$ to these data (see Irpino and Verde 2006), we obtained a 27×27 distance matrix, on which hierarchical clustering with the Ward criterion has been performed. Figure 3 represents the obtained dendrogram.

Looking at the obtained hierarchy, we may conclude that the used distance is able to discriminate the group labeled as 10–18 (the Preferential Attachment processes) and the 0–9 group (Random graphs); as regards the group 19–27 (Small World) three out of nine cases have been confused with the Random graph group, this is likely to happen when Small World processes have higher values for the parameter π, and we have the higher rewiring probability ($\pi = 0.05$) in graphs 21, 24 and 27, in this experiment.

The results are therefore promising, allowing discriminating quite well the different processes. However further study should be devoted to establish true sensitivity and robustness of the proposed approach as well as finding suitable network statistics to discriminate among different kind of networks.

6 Conclusions

A Network Symbolic Data Analysis approach has been proposed. Network symbolic descriptions have been defined that represent network indices by histogram-valued variables, leading to a symbolic data matrix allowing for multivariate data analyses. A simulation study has shown that complex information may be dealt with by this approach and a distance matrix among networks can be defined. Application to real data-sets can take advantage of the proposed approach in terms of (1) comparison among different network structures, (2) exploring sub-graphs or components of complex networks, as well as (3) in longitudinal studies where the same network is observed in different occasions. Indeed, the possibility of obtaining a distance matrix among different networks may lead to the application of classical factorial techniques. Transforming a whole network to a point in a metric space is one the major advantages of the proposed approach. Representing networks as points in a factorial subspace, for instance, may help discussing their proximity, their clustering or analysing trajectories of such *network-points* in order to explore their dynamics.

Acknowledgements This work is financed by the ERDF—European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT—Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project "FCOMP-01-0124-FEDER-037281".

References

Barabási, A. -L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439), 509–512.

Batagelj, V. (1988). Generalized Ward and related clustering problems. In H.-H. Bock (Ed.), *Classification and related methods of data analysis* (pp. 67–74). Amsterdam: North-Holland.

Bock, H. -H., & Diday, E. (2000). *Analysis of symbolic data*. Berlin: Springer.

De Solla Price, D. J. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science, 27*(5), 292–306.

Erdös, P., & Rényi, A. (1960). *On the evolution of random graphs. Publication of the Mathematical Institute of the Hungarian Academy of Science, 5*, 17–61

Freeman, L. C. (1979). Centrality in social networks I: Conceptual clarification. *Social Networks, 1*, 215–239.

Irpino, A., & Verde, R. (2006). A new wasserstein based distance for the hierarchical clustering of histogram symbolic data. In V. Batagelj, et al. (Eds.), *Proceedings of IFCS 2006* (pp. 185–192). Heidelberg: Springer.

Kolaczyk, E. D. (2009). *Statistical analysis of network data. Methods and models*. Springer Series in Statistics. New York: Springer.

Noirhomme-Fraiture, M., & Brito, P. (2011). Far beyond the classical data models: Symbolic data analysis. *Statistical Analysis and Data Mining, 4*(2), 157–170.

Wasserman, S., & Faust, K. (1994). *Social networks analysis: Methods and applications*. New York: Cambridge University Press.