

On the use of Multidimensional Data Analysis Techniques for Social Networks ¹

Giuseppe Giordano, Maria Prosperina Vitale ²
Department of Economics and Statistics
e-mail: ggiordan@unisa.it, mvitale@unisa.it

Abstract: Aim of this paper is to describe an unified framework for the treatment of network data with interpretative tools of Multidimensional Data Analysis. Social Network Analysis is an analytical framework concerned with the detection and interpretation of ties patterns among social actors, while Multidimensional Data Analysis is a collection of techniques to explore different kind of statistical data matrices. We highlight the common facets of these two domains while putting emphasis on some peculiarities.

Keywords: Correspondence Analysis, Relational Data, Singular Value Decomposition, Sociomatrix

1. Introduction

Multidimensional Data Analysis (MDA) is a collection of techniques which allow to obtain a synthesis of the information gathered from n statistical units characterized by p qualitative or quantitative variables. The basic data is arranged in the $\mathbf{X}_{(n,p)}$ rectangular matrix. This matrix can be transformed in order to obtain statistical information between pairs of units (similarity/dissimilarity matrix) or between variables (contingency or correlation matrix) according to their data measurement level.

Social Network Analysis (SNA) aims at detecting and interpret patterns of ties among social entities. The typical data structures of SNA (namely sociomatrix) are related either to a set of m nodes, arranged in a squared *adjacency matrix* $\mathbf{A}_{m,m}$ ($actor \times actor$), or to the set of m nodes and the set of q events, arranged in a rectangular *affiliation matrix* $\mathbf{E}_{m,q}$ ($actor \times event$). The computational treatment of relational data arranged in sociomatrixes is useful to derive the associate graph, to compute network indices and to discover groups of equivalent actors in the net (Wasserman, Faust, 1994).

In order to apply the traditional multidimensional techniques to the SNA data structures, it is necessary to provide a statistical insight of the related information. Hence, the statistical treatment of sociomatrixes should be defined according to the final aim of the analysis: describing network structures or explaining individual differences. Indeed, we can define two levels of network analysis, the first one based on actors (as individual, corporate,

¹The present paper is financially supported by grant 2007: 'Applicazione delle tecniche di analisi multivariata al trattamento di dati relazionali' (resp. Maria Prosperina Vitale), and grant 2008: 'Metodi e Modelli per l'analisi statistica delle reti sociali' (resp. Giuseppe Giordano), Department of Economics and Statistics, University of Salerno - Italy

²Address of correspondence: Maria Prosperina Vitale, Department of Economics and Statistics, University of Salerno, Via Ponte Don Melillo, 84084 Fisciano (SA), Italy

or collective social units), the second one based on ties (linkages between actors, e.g. friendship, business, transaction, kinship and so on).

Aim of this paper is to describe an unified framework for the treatment of SNA data with interpretative tools of multidimensional data analysis. We highlight the common facets of the two analytic domains while putting emphasis on some peculiarities. For instance, the use of Multidimensional Scaling (MDS) techniques, as a general scheme able to visualize the similarity of the node-profiles in $\mathbf{A}$ or to jointly display the node-profile and the event-profile in $\mathbf{E}$, could be enriched by statistical reading.

In the next section, we will discuss the ins and outs of using multivariate exploratory techniques for SNA data structures (section 2), then some further developments in the treatment of relational data will be discussed (section 3).

2. Some relations between data matrices of MDA and SNA

The initial steps of MDA involves the definition of different data coding according to the final aim. Several choices have to be made in order to obtain the data structure to be analysed (i.e. disjunctive coding of nominal variables or standardizing quantitative variables). Then, this information is used to build statistical measures (e.g. contingencies, covariances and correlations). Indeed, many multidimensional techniques look for a synthesis of the data structure. The algebraic tool used to decompose the data matrix is the Singular Value Decomposition (SVD) (Eckart, Young, 1936). The facet we wish to highlight is that the algebraic structures and geometrical representations reproduce the statistical meaning which drives to the final interpretation of the results. Hence, the derived results are interpreted according to the statistical information retrieved in the transformed data matrix. For instance, in Principal Component Analysis (PCA) (Jolliffe, 2002) the angle between two vector-variables is a function of their correlation, in Correspondence Analysis (Greenacre, 1984) the proximity between two rows (or columns) profiles is interpreted according to the chi-square distance.

Hence, considering the typical data structures derived from a net, we argue that sociomatrices basically is a coded version of a graph structure from which statistical information have to be derived. For instance, the binary *adjacency matrix* $\mathbf{A}$ just holds the coding of the presence or absence of ties between nodes. Therefore, the analysis of such sociomatrix requires a transformation in order to derive suitable statistical information.

MDA applied to sociomatrices has been used with different aims. One of the main objectives is to represent proximities among actors in low-dimensional sub-spaces using appropriate network measures (e.g. geodesic distance). This goal can be accomplished through Multidimensional Scaling techniques (Cox, Cox, 2001). As a particular case, the use of Correspondence Analysis allows to obtain a joint display of the actors and events of a two mode-network described in the *affiliation matrix* $\mathbf{E}$ (e.g. Roberts, 2000, Faust, 2005). In this case, Correspondence Analysis is used as an optimal scaling method and the statistical interpretation of the association between rows and columns is disregarded. In the same way, Principal Component Analysis is sometimes used to scale network data (Faust, 2002).

3. Multidimensional Data Analysis for Network Data

Network analysis focus on relational data, i.e. data whose properties cannot be reduced to the attributes of the entities (nodes). However, beyond this relational information, researcher could be interested to auxiliary information (attribute variables) gathered on nodes. Attribute variables can represent both demographic characteristics of nodes (*gender*, *age*, *race*, etc.) and some variables of interest that could be influenced by the network structure (such as *wage*, *performance measurement*, etc.).

The attribute variables on nodes are usually used to enrich the graph interpretation and its visualization. Furthermore, they have been exploited in network modelling to explain the formation of ties (e.g. p^* models, see Robins *et al.*, 1999). However, we notice that exploratory statistical techniques to analyse jointly these sets are still under investigation in SNA.

These two kinds of data can be related and the simultaneous treatment of these data structures can be pursued. In this case, the retrieved information give raise to a structured matrix with two kinds of information, the relational data and the attribute variables. In MDA the presence of different types of information is taken into account by several techniques able to recognize such roles. In particular, non-symmetrical data analysis (Lauro, D'Ambra, 1984) is a collection of exploratory multidimensional data techniques introduced for both quantitative and qualitative data and for two or more sets of variables. While such techniques share the same descriptive goals of unstructured data analysis, they allow highlighting the relations between the two sets of variables.

In the following, we will review some of such data techniques and we will show possible applications in the framework of SNA.

Let $\mathbf{D}$ be the diagonal matrix ($m \times m$) holding the degree of each node in the symmetric and binary adjacency matrix $\mathbf{A}$. Let us consider the data matrix $\mathbf{X}$ containing information on p variables observed on the m nodes. The entry in the adjacency matrix $\mathbf{A}$ can be seen as a particular case of contiguity relation among statistical units. In particular, it produces a sort of partitioning of the units. In this way, it allows a decomposition of the total variance of in $\mathbf{X}$ into two components: the local variance among the adjacent units and the residual variance. The study of these variance components gives raise to two main approaches: the *Smooth Factorial Analysis* and the *Factorial Analysis of the Local Differences* (Benali, Escofier, 1990). The first one allows to depict general patterns in the data by substituting each observation with the average value of adjacent units ($\mathbf{D}^{-1}\mathbf{A}\mathbf{X}$). The second one deals with the analysis of cohesive sub-group variations by replacing each observation with the deviation from the barycenter of its neighbours ($\mathbf{X}-\mathbf{D}^{-1}\mathbf{A}\mathbf{X}$).

When such data structures arise from social network data, the above mentioned techniques allow to visually explore how network structure influences the variables in $\mathbf{X}$. In a recent paper this approach was exploited to analyse the communication patterns among actors in a virtual learning community (Giordano, Vitale, 2007).

When the focus is on the analysis of affiliation matrix $\mathbf{E}$, the classical correspondence analysis approach assumes actors and events to be interdependent. If an association structure between the two node-sets can be hypothesized, we can explore the relationship structure through Non-Symmetrical Correspondence Analysis (NSCA) (Lauro, D'Ambra, 1984). In the framework of SNA, the use of NSCA is a fruitful approach to explore the association pattern between actors and events (De Stefano *et al.*, 2008).

The peculiar features of relational data need specific data pre-treatment and their own nature leads to use ad hoc statistical methods. We warn against the use of such techniques

as merely visualization tools disregarding the statistical meaning.

References

- Aluja B. and Lebart L. (1984) Local and partial principal component analysis and correspondence analysis, in: *COMPSTAT Proceedings*, Havranek T., Sidak Z. and Novak M., eds., Phisyca-Verlag, Vienna, 113–118.
- Benali H. E.B. (1990) Analyse factorielle lisse et analyse factorielle des differences locales, *Revue de Statistique Applique*, 38, 55–76.
- Cox T. and Cox M. (2001) *Multidimensional Scaling*, Chapman and Hall, New York.
- De Stefano D., Ragozini G. and Vitale M. (2008) Non-symmetrical correspondence analysis to explore dependence structure in social networks, in: *Book of Short Abstracts 7th International Conference on Social Science Methodology, RC33 Logic and Methodology in Sociology*, Balbi S., Scepi G., Russolillo G. and Stawinoga A., eds., Casa Editrice Dott. Eugenio Jovene SpA, Napoli, 295.
- Eckart C. and Young G. (1936) The approximation of one matrix by another of lower rank, *Psychometrika*, 1, 211–218.
- Faust K. (2005) Using correspondance analysis for joint displays of affiliation networks, in: *Models and Methods in Social Network Analysis*, Carrington P., Scott J. and Wasserman S., eds., Cambridge University Press, New York, 117–147.
- Faust K., Willert K., Rowlee D. and Skvoretz J. (2002) Scaling and statistical models for affiliation networks: patterns of participation among soviet politicians during the brezhnev era, *Social Networks*, 24, 231–259.
- Giordano G. and Vitale M. (2007) Factorial contiguity maps to explore relational data patterns, *Statistica applicata*, 4.
- Greenacre M. (1984) *Theory and Applications of Correspondence Analysis*, Academic Press, London.
- Jolliffe I. (2002) *Principal Component Analysis*, Springer, New York.
- Lebart L., Morineau A. and Piron M. (2000) *Statistique Exploratoire Multidimensionnelle*, Dunod, Paris.
- Roberts J.J. (2000) Correspondence analysis of two-mode network data, *Social Networks*, 22, 65–72.
- Robins G., Pattison P., Kalish Y. and Lusher D. (1999) An introduction to exponential random graph (p^*) models for social networks, *Social Networks*, 21, 37–66.
- Wasserman S. and Faust K. (1994) *Social Network Analysis: Methods and Applications*, Cambridge University Press, New York.