Social Networks 38 (2014) 28–40

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

On the use of Multiple Correspondence Analysis to visually explore affiliation networks

Maria Rosaria D'Esposito[a], Domenico De Stefano[b,*], Giancarlo Ragozini[c]

[a] Department of Economics and Statistics, University of Salerno, Via Ponte don Melillo, 84084 Fisciano (SA), Italy
[b] Department of Economics, Ca' Foscari University of Venice, Cannaregio 873, 30121 Venezia, Italy
[c] Department of Political Sciences, University of Naples Federico II, Via Rodinò, 22, 80138 Napoli, Italy

ARTICLE INFO

Keywords:
Affiliation networks
Two-mode structural equivalence
Correspondence Analysis
Doubling
Actor and event covariates

ABSTRACT

In this paper we discuss the use of Multiple Correspondence Analysis to analyze and graphically represent two-mode networks, and we propose to apply it in a Greenacre's doubling perspective. We discuss how Multiple Correspondence Analysis: (i) properly takes into account the nature of relational data and the intrinsic asymmetry of actors/events in two-mode networks; (ii) allows a proper graphical appraisal of the underlying relational structure of actors or events; (iii) makes it possible to add actor and event attributes to the analysis in order to improve results interpretation; and (iv) gives different results with respect to the usual Simple Correspondence Analysis.

1. Introduction

Analysis of the collaboration and affiliation structures arising among different subjects is now more and more of interest in both social and economic sectors. In general, collaboration and affiliation networks are characterized by a set of actors and a set of events in which the actors are involved. These elements give rise to so-called two-mode networks, which contrast with the more widely used and well known one-mode networks.

In the analysis of two-mode networks, Borgatti and Halgin (2011) identify two main approaches: the conversion approach and the direct approach. In the former, a two-mode network is converted into two one-mode networks and the analysis focuses on one mode at time, though this may result in the loss of some information (Borgatti and Everett, 1997; Everett and Borgatti, 2013). The direct approach, which we adopt in this paper, considers instead the two-mode network as it is, and the two modes are jointly analyzed.

Dealing with a two-mode network, one issue is to visualize it and make a graphical appraisal of the underlying relational structure. With this aim in mind, the direct approach mainly makes use of the bipartite graph and spring embedding (Borgatti and Halgin, 2011) or of factorial methods for qualitative data such as Correspondence Analysis (CA) (Benzécri, 1973; Greenacre, 1984). This latter derives low dimensional spaces in which it is possible to represent points corresponding to the actors and the events in order to evaluate similarities of participation/attendance patterns (Wasserman et al., 1989; Faust, 2005; Borgatti and Halgin, 2011).

CA can be applied if we consider the affiliation matrix, corresponding to a two-mode network, either as a two-way contingency table, or as a two-way case-by-variable matrix.[1] The first approach has often been adopted and results in a simple correspondence analysis of the affiliation matrix. This use has given rise to a debate on its pros and cons (Bonacich, 1991; Borgatti and Everett, 1997; Roberts, 2000; Borgatti and Halgin, 2011; Borgatti et al., 2013). In the present paper, we focus on the second approach, i.e. we look at the affiliation matrix as a case-by-variable matrix. Within this framework, in our opinion a proper approach is Multiple Correspondence Analysis (MCA), the extension of CA to the case of many categorical variables (Blasius and Greenacre, 1994, 2006). Given the characteristics of two-mode networks, in this paper we propose to apply MCA by using the complete disjunctive coding in Greenacre's doubling perspective (Greenacre, 1984) and we discuss how this version of MCA can be used to analyze and graphically represent two-mode networks. The interest in this technique is motivated thus: (i) MCA is designed to treat case-by-variable matrices, which are similar in structure to affiliation matrices; (ii) MCA assigns a

* Corresponding author. Tel.: +39 040 558 3041; fax: +39 040 558 7033.
E-mail addresses: mdesposi@unisa.it (M.R. D'Esposito), domenico.destefano@unive.it (D. De Stefano), giragoz@unina.it (G. Ragozini).

[1] CA has also been widely applied for the analysis of one-mode networks (see e.g. Noma and Randall Smith, 1985; Wasserman et al., 1989; Borgatti and Everett, 1997; Mohr, 1998; Roberts, 2000; Faust and Skvoretz, 2002; Faust et al., 2002; Faust, 2005, 2006).

http://dx.doi.org/10.1016/j.socnet.2014.01.003
0378-8733/

different role to actors and to events, thus allowing distinct features to be highlighted in each mode; (iii) MCA makes it possible to add covariates to the analysis in order to improve results interpretation; and (iv) MCA within the doubling perspective affords visualizations that can be easily interpreted in terms of similarities among actors/events network relations.

The paper is organized as follows. In Section 2 we present and discuss our approach to the use of MCA for affiliation networks, while in Section 3 we highlight the features of MCA useful for two-mode networks. A discussion on how MCA could incorporate external information in the analysis is given in Section 4. In Section 5 we discuss, through an analytical and experimental analysis, how MCA makes it possible to represent the degree of similarity of actor/event profiles and the differences existing with respect to simple CA. Section 6 offers some concluding remarks.

2. Multiple Correspondence Analysis for affiliation networks

Let $\mathcal{G}$ be an affiliation network. It consists of two sets of relationally connected units, actors and events, and can be represented by a triple $\mathcal{G}(V_1, V_2, \mathcal{R})$ composed of two disjoint sets of nodes, V_1 and V_2 of cardinality n and m, and one set of edges or arcs, $\mathcal{R} \subseteq V_1 \times V_2$. By definition $V_1 \cap V_2 = \emptyset$; the set $V_1 = \{a_1, a_2, \ldots, a_n\}$ represents the set of n actors whereas $V_2 = \{e_1, e_2, \ldots, e_m\}$ represents the set of m events. The edge $r_{ij} = (a_i, e_j)$, $r_{ij} \in \mathcal{R}$, is an ordered couple, and indicates whether an actor a_i attends an event e_j. The set $V_1 \times V_2$ can be fully represented by the binary *affiliation matrix* $\mathbf{F} = (f_{ij})$, $i = 1, \ldots, n, j = 1, \ldots, m$, with $f_{ij} = 1$ if $(a_i, e_j) \in \mathcal{R}$ and 0 otherwise. Given $\mathbf{F}$, the row and column marginals $f_{i.} = \sum_{j=1}^{m} f_{ij}$ and $f_{.j} = \sum_{i=1}^{n} f_{ij}$ coincide with the degree d_i of the ith actor and the size s_j of the jth event, respectively, i.e. $f_{i.} = d_i$ and $f_{.j} = s_j$. To the best of our knowledge, the first attempt to apply MCA in a two-mode setting dates back to 1984 (Bourdieu, 1988). Bourdieu used Multiple Correspondence Analysis to analyze the universities, professional affiliations and background variables of Parisian professors. However, he did not treat his data matrix as an affiliation matrix because he considered the affiliations of professors as properties of the individuals and not as relations (De Nooy, 2003). After this unintentional use to analyze affiliation matrices, a particular version of MCA (Carroll et al., 1986) has been used in social network analysis (Wasserman et al., 1989; Faust, 2005). The CA algorithm is here applied to an "edges by actors and events matrix", where the edges $r_{ij} = (a_i, e_j)$ are assumed as units of the analysis and the affiliation matrix $\mathbf{F}(n \times m)$ is transformed into a multiple indicator matrix with $n + m$ columns (one out of n for each actor and one out of m for each event) and as many rows as the edges, i.e. the rows correspond to the 1's in the affiliation matrix and amount to the total number of ties L (Faust, 2005). However, the scaling proposed by Carroll et al. (1986) did not receive a general consensus and has been questioned in the statistical literature (Carroll et al., 1987, 1989; Greenacre, 1989).

In the present paper, we look at actors as observational units and at the participation in events as dichotomous categorical variables. The $\mathbf{F}$ matrix is, therefore, a multivariate case-by-variable data matrix (Gower, 2006), in which a different status is assigned to the rows and columns – in line with the duality perspective (Faust, 2005; Breiger, 1974).

We propose to perform MCA by applying the usual CA algorithm – SVD of the doubled normalized and centered profile matrix – to the *multiple indicator matrix*, or simply *indicator matrix*, $\mathbf{Z}$ derived from $\mathbf{F}$ through a full disjunctive coding.[2]

	e_1	e_2	e_3	e_4
a_1	1	0	0	1
a_2	1	0	1	0
a_3	0	1	1	1

F- Affiliation matrix

	e_1		e_2		e_3		e_4	
	+	−	+	−	+	−	+	−
a_1	1	0	0	1	0	1	1	0
a_2	1	0	0	1	1	0	0	1
a_3	0	1	1	0	1	0	1	0

Z- Full disjunctive coding of F

Fig. 1. A fictitious affiliation matrix $\mathbf{F}$ and the corresponding indicator matrix $\mathbf{Z}$ obtained through full disjunctive coding.

In MCA for measurement data, each row of $\mathbf{Z}$ represents an observation and each column represents one modality of each categorical variable, with $z_{ij} = 1$ if the jth modality of the qth categorical variable is present in the ith case, 0 otherwise. In the case of Q categorical variables the indicator matrix is $\mathbf{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_Q]$. If the qth variable has J_q categories, Z_q is $I \times J_q$, and $J = \sum_{q=1}^{Q} J_q$ is the total number of categories. In this case the row margins are constant and equal to Q, while the column margins correspond to the category frequencies.

In the case of affiliation matrix $\mathbf{F}$, in order to apply the full disjunctive coding, we think of each event e_j as a dichotomous variable with categories e_j^+ and e_j^-. In the full disjunctive coding, each category is describing by a dummy variable, i.e. e_j^+ is a dummy variable coding the participation in the event, and e_j^- is a dummy variable coding the non-participation. $\mathbf{Z}$ will then contain two orthogonal columns for each e_j (Fig. 1).

The $\mathbf{Z}$ matrix is a $n \times 2m$ matrix of the form: $\mathbf{Z} \equiv [\mathbf{F}^+, \mathbf{F}^-]$, where $\mathbf{F}^+ = (e_j^+) = \mathbf{F}$, and $\mathbf{F}^- = (e_j^-) = \mathbf{1} - \mathbf{F}^+ = \mathbf{1} - \mathbf{F}$, where $\mathbf{1}$ is an $n \times m$ all-ones matrix.

The indicator matrix $\mathbf{Z}$ turns out to be a *doubled matrix*, and the MCA we are proposing is equivalent to the CA of this doubled matrix. This allows us to interpret results in a doubling perspective.

The $\mathbf{Z}$ row marginals $z_{i.}$ are constant and equal to the number of events m, while the column marginals $z_{.j}$ are equal to the event size s_j, when associated to e_j^+, or to $n - s_j$, when associated to e_j^-.

The profile matrix $\mathbf{P}$ and weight matrices $\mathbf{D}_a$ and $\mathbf{D}_e$ proper of the CA algorithm and involved in the normalization are:

$$\mathbf{P} = \frac{\mathbf{Z}}{nm}, \quad \text{with} \quad nm = \sum_{i=1}^{n} \sum_{j=1}^{m} z_{ij},$$

$$\mathbf{D}_a = \mathrm{diag}\left(\frac{1}{n}, \ldots, \frac{1}{n}\right),$$

$$\mathbf{D}_e = \mathrm{diag}\left(\frac{s_1}{nm}, \ldots, \frac{s_{2m}}{nm}\right).$$

The centered and doubled normalized $\mathbf{Z}$ is the matrix $\mathbf{S}$

$$\mathbf{S} = \mathbf{D_a}^{-1/2}\left(\frac{\mathbf{Z}}{nm} - \mathbf{D_a}\mathbf{1}\mathbf{1}^T\mathbf{D_e}\right)\mathbf{D_e}^{-1/2} = \sqrt{n}\left(\frac{\mathbf{Z}}{nm} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{D_e}\right)\mathbf{D_e}^{-1/2},$$

where $(1/n)\mathbf{1}$ is the vector of actor weights and $\mathbf{1}^T\mathbf{D_e}$ is the vector of the event weights.

The SVD of $\mathbf{S}$ gives:

$$\mathbf{S} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$$

where $\boldsymbol{\Sigma}$ is the diagonal matrix of singular values, and $\mathbf{U}$, $\mathbf{V}$ are the matrices of the left and right singular vectors, respectively.

[2] We choose this approach, instead of the mathematically equivalent CA algorithm applied to the squared super-matrix of cross-tables (the so-called Burt matrix), in order to highlight the role of actors and to emphasize the relation of this approach with the doubling perspective. The idea behind doubling is the allocation of two complementary sets of data to a given rating scale, one labeled as the "positive" pole of the scale and the other as the "negative" pole (Greenacre, 1984).