FACTORIAL CONTIGUITY MAPS TO EXPLORE RELATIONAL DATA PATTERNS[1]

Giuseppe Giordano, Maria Prosperina Vitale

Dipartimento di Scienze Economiche e Statistiche, Università di Salerno
Via Ponte don Melillo, 84084 Fisciano (ITALY)
ggiordano@unisa.it ; mvitale@unisa.it

Abstract

Over the past few decades Social Network Analysis has found increasing application in many social research areas to describe relational ties among social entities. In this paper, we propose to use Multidimensional Data Analysis in the framework of SNA in order to explore the structural properties of a network. In particular, we refer to Contiguity Analysis in order to deal with relational data defined by ties among the actors and described by network centrality and clustering coefficients. The expected results consist of the definition of a network meta-analysis able to synthesise and visualise the pattern of social relationships in a metric space where the related data structure is described. The proposed method is applied to an illustrative example in the context of a virtual learning community.

Keywords: Local Factorial Analysis; E-learning Community; Principal Component Analysis; Social Network Analysis.

1. INTRODUCTION

The analysis of a social network consists of the visualisation, exploration and statistical description of the linkages among interacting units (Wasserman, Faust, 1994). One of the most important facets of Social Network Analysis (SNA) is the opportunity to relate different actors (e.g.: people, companies, countries) according to the degree of relationship (communication, trade exchange, human interaction). Network data can be arranged according to peculiar data matrices and displayed

[1] *This paper has been supported by MIUR grants "ex 60%", (2007): "Applicazione delle tecniche di analisi multivariata al trattamento di dati relazionali" coord. M.P. Vitale and "La valutazione dell'apprendimento on-line attraverso l'analisi delle reti sociali ", coord. G. Giordano.*

through graphs. The two yardstick matrices for social networks are the adjacency – actor per actor – square matrix describing the relationships among actors (one-mode network) and the affiliation – actor per event – rectangular matrix where the rows represent the actors and the columns represent the events to which the actors belong (two-mode network).

Network visualization through a graph makes it possible to explore patterns in relational data and to highlight their structural properties (Freeman, 2005; DeJordy *et al*., 2007). Freeman (2000) describes five historical phases in the use of visual images in SNA: from the earliest graphical representation (*sociogram*) in which the points were placed according to ad-hoc rules proposed by Moreno (1953), to the development of computer systematic procedures for locating points in a metric space. Nowadays, factorial techniques are widely used to represent similarity or dissimilarity measures among the actors on a map. Moreover, different kinds of spring-embedders are used to define the location of points (Kamada, Kawai, 1989) to improve graph readability using general rules, such as the regular spacing of nodes and minimization of edge crossings. While do exist several tools for exploring the structural properties of network data, it is difficult to define the best spatial arrangement of the graph. Indeed, the different ways of locating the nodes and the edges of a network can give emphasis to particular characteristics. It should be argued that the spatial arrangement of network data can influence viewers' perceptions of structural characteristics (McGrath *et al*., 1997). According to Bender-deMoll and McFarland (2006), *"when drawing a graph in* SNA *becomes a methodological tool instead of an illustrative art [...] it is necessary to define a priori criteria for assessing what a good graph layout is"*.

In this paper we aim to define a metric space where the position of the node-points depends on some analytic information about the network structure. We describe a meta-analysis through the factorial decomposition of a data matrix holding a set of network measures (centrality, cohesion, etc.) computed for each node. Hence, starting from the contiguity data analysis described in (Aluja, Lebart, 1984; Benali, Escofier, 1990) we show how these techniques can be profitably used to identify optimal criteria to derive different factorial subspaces. The resulting maps make it possible to give a direct interpretation of the node position consistent with the net structure.

This work is organised as follows: in section 2 a brief review of Contiguity Analysis in the framework of Multidimensional Data Analysis is provided, in section 3 the Network Meta-analysis is introduced. In section 6 we apply the proposed strategy within an illustrative example. Some concluding remarks are provided in section 5.

2. CONTIGUITY FACTORIAL ANALYSIS AND RELATED TECHNIQUES

The notion of contiguity has been undertaken by several authors in order to define multidimensional data techniques able to take into account the dependence structure on a set of statistical units.

The contiguity structure can be seen as a generalization of partitioned data. It allows a decomposition of the total variance into two components: the local variance among the contiguous units and the residual variance. The study of these variance components gives rise to two main approaches to Contiguity Analysis: the Smooth Factorial Analysis (SFA) and Factorial Analysis of the Local Differences (FALD) (Benali, Escofier, 1990). The former makes it possible to depict general trends in the data by substituting each observation with the average value of contiguous units; the latter deals with the analysis of the local fluctuations by replacing each observation with the deviation from the barycenter of its neighbors. Contiguity is represented through a generic graph structure where the nodes are the statistical units and the edges express the presence of several kinds of relationships. This feature highlights the link with SNA and allows some statistical multidimensional techniques to be expressed in terms of the matrices induced by Graph Theory.

In this paper we use the following notation. Let $\mathbf{C}(m \times m)$ be the adjacency (contiguity) matrix related to the graph G(N,E) where N is the set of m nodes and E is the set of edges, (in this paper **C** is a symmetric binary matrix); let **D** be the diagonal matrix $(m \times m)$ holding the degree of each node. Furthermore, we consider the data matrix **X** containing information on the p variables observed on the m statistical units. The variables in **X** could be any quantitative information recorded on the m nodes. However in the next section we specify them in terms of SNA metric indices. Recalling the definition of the local variance-covariance matrix $\mathbf{V}_l$ for the matrix **X**:

$$\mathbf{V}_l = \frac{1}{2m}\mathbf{X}'\left(\mathbf{D} - \mathbf{C}\right)\mathbf{X} \tag{1}$$

When the variables in **X** depend on the graph structure considered in **C**, the local variance is a biased estimation of the total variance. Let us notice that in the case of partitioned data, the matrix $\mathbf{V}_l$ is a special case of within variance as in Factorial Discriminant Analysis. The diagonalisation of the local variance-covariance matrix will define the Local Principal Component Analysis (LPCA - Aluja, Lebart, 1984). Indeed, expression 1 represents the squared deviations among each node and the average values of the contiguous vertices, the LPCA is close to the local differences variance. These techniques can be seen as special case of PCA in a particular metric. In the same way, if the matrix **X** is substituted by $\mathbf{D}^{-1}\mathbf{CX}$, the PCA of this matrix corresponds to the Smooth Factorial Analysis first proposed to deal with time-dependent data in PCA as stated in Benali and Escofier (1990). It consists of replacing the original values in **X** by the average of contiguous units, so removing local fluctuations. If the matrix **C** represents a complete graph, the FALD traces back to the PCA. In the next section we look at these techniques in the framework of SNA and we show how they can be used to highlight the role of the actors in the network visualisation.

3. THE SOCIAL NETWORK META-ANALYSIS AND THE FACTORIAL CONTIGUITY MAPS

In SNA several indices are used in order to analyse the network structure. For instance, some indices of centrality and prominence have been proposed to describe the position of actors in one-mode and two-mode networks (Faust, 1997), and for egocentric networks (Marsden, 2002). Freeman (1979) defined three concepts of actor centrality based on the number of ties (degree), the bridging role (betweenness), the access to information (closeness) in a network. Moreover, the extent to which the actors are connected in the net can be measured by the clustering coefficient that for each actor stands for the density of its open neighbourhood (Watts, 1999).

Even if we can explore different facets of the network structure through the different proposed structural indices, it should be argued that the whole set of available indices represents a huge quantity of information. On the other hand, when the aim is to deal with complexity in networks (e.g. a large number of nodes with dense connections, longitudinal and multi-relational networks), the graphical representation is hardly able to highlight roles and positions of each actor in the network.

Hence, we propose to exploit the information derived by the different structural indices of a network to provide a factorial synthesis which should be useful both in reducing redundancies and providing a metric reference subspace in which to locate the nodes according to their role in the network. This produces a graphical representation of the nodes as points in a map, where their position and relative distance are based on proximity criteria induced by their index values derived by network analysis. The information drawn from the network indices will be analysed through the methods defined in the context of Multidimensional Data Analysis and in particular by means of contiguity factorial techniques.

We start by considering the Weighted Principal Component Analysis (WPCA), where each statistical unit (a node) has a different weight according to its own degree. The factorial synthesis is obtained as the solution of the following eigen-equation:

$$\left[\mathbf{X'DX}\right]\mathbf{u}_{\alpha} = \Lambda\mathbf{u}_{\alpha} \tag{2}$$

where $\mathbf{u}_{\alpha}$ are the eigenvectors and Λ is the diagonal matrix holding the eigenvalues. The WPCA makes it possible to enhance the visualisation of the actors with higher values of peculiar structural indices that will be far away from the origin of the factorial plan and will lie opposite to the node-points with lower values. The latter consideration stems from the assumption that most of the centrality indices are positively correlated. Some simulation results on Pearson's correlation coefficient between Centrality Degree and Closeness are reported. Further investigations should be provided for other centrality and clustering measures.

This analysis is useful for an immediate understanding of the actors' roles. The nodes will cluster together in the resulting factorial map according to the size of the network indices arranged in the matrix **X**. However, different weights can be used in the analysis to provide a general method to define the actors' distance.

Next, we consider the two local factorial techniques discussed above. SFA and FALD can be used to take the contiguity information among the units directly into account. In particular, Smooth Factorial Analysis aims to explain the variance in **X** by the contiguity structure in **C**. The SFA eigenequation is:

$$\left[\mathbf{D}^{-1}\mathbf{CX}\right]\left[\mathbf{D}^{-1}\mathbf{CX}\right]'\mathbf{v}_{\alpha} = \Delta\mathbf{v}_{a} \tag{3}$$

where $\mathbf{v}_{\alpha}$ are the eigenvectors and is the diagonal matrix holding the eigenvalues.

In this case the factorial axes will discriminate between contiguous nodes, enhancing the role of those which share similar values (neighbouring effect).

Factorial Analysis of Local Difference will highlight the actors with a prominent role in the network provided that the neighbouring effect has been eliminated (within or residual variance). The FALD eigen-equation to be solved is:

$$\left[\mathbf{X}-\mathbf{D}^{-1}\mathbf{CX}\right]\left[\mathbf{X}-\mathbf{D}^{-1}\mathbf{CX}\right]'\mathbf{w}_{\alpha} = \Omega\mathbf{w}_{a} \tag{4}$$

where $\mathbf{w}_{\alpha}$ are the eigenvectors and Ω is the diagonal matrix holding the eigenvalues. The node-points will lie together on the resulting map if their own role within contiguous units is similar. They are not asked to share the same group but just the same role in a group.

By applying the three factorial techniques on the data matrix **X**, we define the network meta-analysis that summarises the main characteristics of a network. We propose to use the centrality indices (i.e. Centrality degree, Closeness, Betweenness, Eigenvector) and the clustering coefficients (CC1, CC2) because they return actor-based network measures.

As main result, the network meta-analysis produces a configuration of points in an orthogonal factorial subspace. The definition of such a factorial map makes it possible to visually explore the effect of contiguity on the actors' role in the network. Let us notice that the position of the units on the map is now related to the weighted, local and residual variance, respectively for the three analyses (WPCA, SFA, FALD). In the common representation of a network by a graph, the information just concerns the presence or the absence of ties between the set of nodes and metric distances between nodes are not considered.

In the following section, through an illustrative example, we will discuss the main results of this approach and highlight some general interpretative rules.

4. THE ACQUAINTANCE NETWORK IN A VIRTUAL LEARNING COMMUNITY

The factorial techniques proposed above can be applied in different knowledge fields in order to visually explore the actors' position in a network structure. Our approach is used here to analyse the communication patterns among units in a virtual learning community. Several studies exist in the specialised literature to evaluate the interactions in learning environments among students through the indices proposed in the framework of SNA (Haythornthwaite, 1998; Aviv *et al.*, 2003).

In the following, we explore the network structure related to undergraduate students involved in an online course at the University of Salerno (Italy) in the academic year 2006/'07. We focus on the analysis of 36 nodes (the students) with non-zero degree, the ties represent the acquaintance among the students.

The network graph obtained by the KK spring-embedder algorithm implemented in the Pajek software (De Nooy *et al.*, 2005), is shown in Figure 1. This quite simple network shows the presence of one isolated dyad (stud1, stud9), the nodes with relevant betweenness role (stud5, stud17, stud29) and the nodes with high centrality degree (stud4, stud5, stud29, stud36). These characteristics are quite evident and relatively easy to discover at first sight for a small size network. In the following we try to recover this kind of information by means of the proposed approach.

In this example, **C** is the (36x36) contiguity matrix, **X** is the (36x6) data matrix holding the four centrality indices and the two clustering coefficients computed on the network data. In order to discuss the main results of the proposed meta-analysis, we report the factorial maps obtained by performing the three techniques (WPCA,

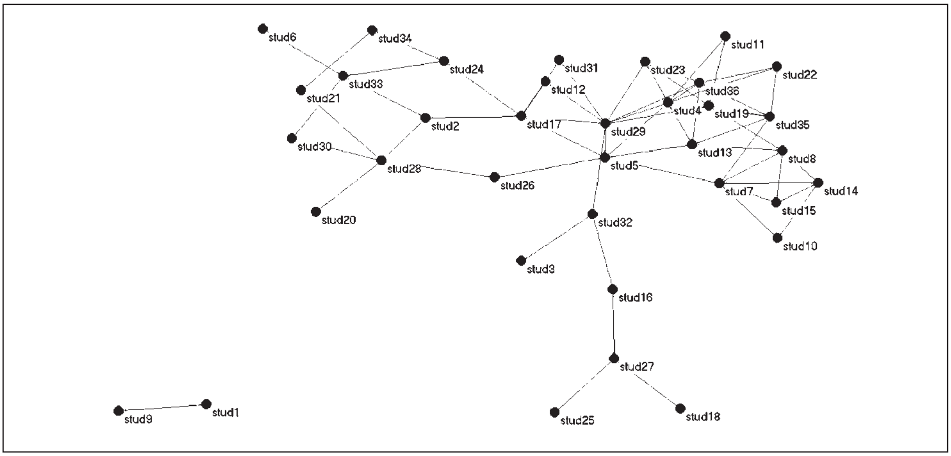

Fig. 1: The acquaintance network of 36 online students.

SFA, FALD) on the data matrices **X** and **C**. The following factorial maps show the scatters of the node-points onto the first two principal axes for each analysis.

In the WPCA we have used the raw degree of each node as weights. In the corresponding factorial map (Figure 2) the location of the 36 nodes is related to the role of the actors in the network: the first axis enhances the correlation between Centrality Degree and Closeness, Eigenvector and CC2. By looking at the scatter of nodes, we observe that the majority of nodes lie on the left-hand side since the network is characterised by just a few nodes with high centrality indices (stud5, stud29 on the top-right, and stud4, stud36 on the bottom-right).

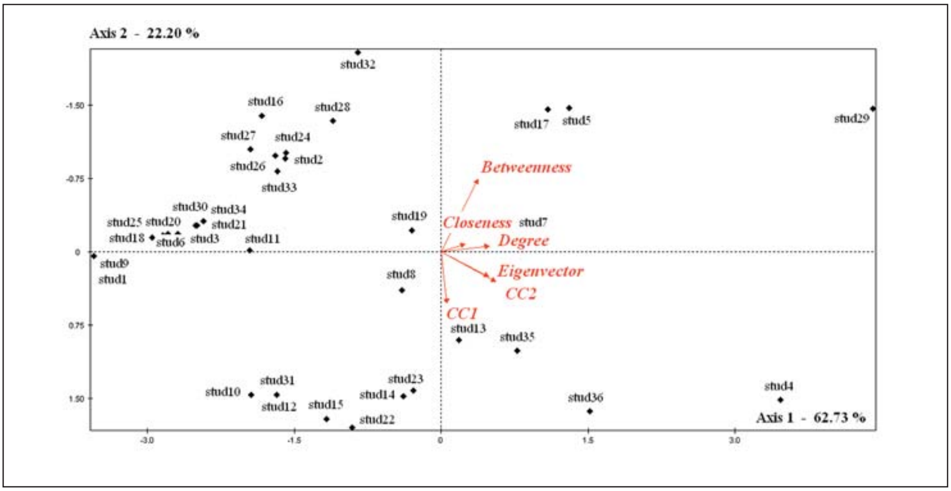

Fig. 2: The WPCA Factorial Contiguity Map.

The second axis is characterised by the opposition between CC1 and Betweenness with straightforward interpretation of the nodes lying along these directions. The factorial plan opposes the students with higher values of the indices (on the right-hand side) versus the students with lower values (on the left-hand side).

In Figure 3, the factorial map obtained by SFA discriminates among nodes belonging to contiguous groups (let us remember that in this technique each value is replaced by the average value of the neighbours). The presence of two main clouds can be clearly seen: the first one, on the left-hand side is characterised by nodes with low values, the second one on the right is related to the central nodes in the network.

Let us notice the position of the dyad stud1, stud9 which share the same coordinates on the bottom-left of the map, as well as the units stud31 and stud12 on the top-right. It is interesting to compare them with their role and position shown in Figure 2. Hence the correlation structure of the SFA factorial map has the same shape as the WPCA. Finally, let us highlight the position of the nodes stud26 and stud32 which lie along the direction spanned by the variable Betweenness and let compare their bridging role in Figure 3.

The factorial map obtained by the FALD approach is presented in Figure 4, here we are interested in those points far away from the origin of the plan, since these points represent nodes with idiosyncratic roles with respect to their neighbours.

This technique explains the variance within contiguous units. The proximity among units depends on the role played by each node within its contiguous group.

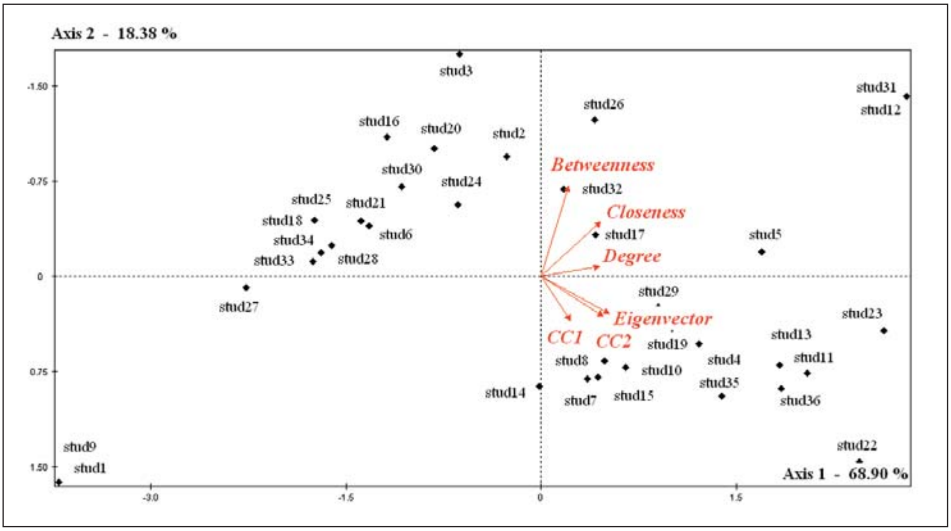

Fig. 3: The SFA Contiguity Map.

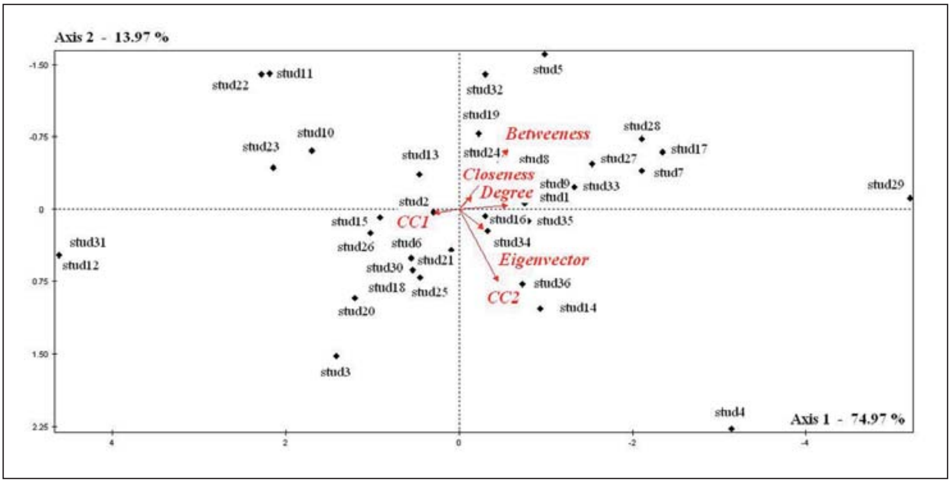

Fig. 4: The FALD Contiguity map.

The kind of role played in the map is explained by the direction of the vector variables. For instance, let us consider the position of stud4 and stud29 in Figure 4. The two node-points, which seem quite similar for the role and position in the graph, have a peculiar role in terms of CC2 (stud4) and Centrality Degree (stud29).

5. CONCLUDING REMARKS

The definition of a factorial map, based on the contiguity structure of relational data, can profitably support the visualization of the main characteristics of a network. In order to enhance the visual display of the network structure we have defined a meta analysis of the traditional results of SNA through peculiar factorial techniques. The adopted method makes it possible to exploit the contiguity information among nodes and produce different visualisations able to highlight different sources of variability (weighted, local and residual). By comparing the resulting factorial maps, the researcher should be able to recognise and understand several roles of the actors in the network. The strategy we have described can be generalised to comprise different data matrices holding auxiliary information on the actors. Further investigation we need to provide a major insight into large scale networks.

REFERENCES

AVIV, R., ERLICH, Z., RAVID G., GEVA, A. (2003), Network analysis of knowledge construction in asynchronous learning networks, *Journal of Asynchronous Learning Networks*, **7**.

ALUJA BANET, T., LEBART, L. (1984), Local and Partial Principal Component Analysis and Correspondence Analysis, In: *Havranek, T., Sidak, Z., Novak, M. (Eds.) COMPSTAT Proceedings*, pp. 113-118, Phisyca- Verlag, Vienna.

BENALI, H., ESCOFIER, B. (1990), Analyse factorielle lissée et analyse factorielle des différences locales, *Revue de Statistique Appliquée*, **38**, 55–76.

BENDER-DeMOLL, S., McFARLAND, D.A. (2006), The Art and Science of Dynamic Network Visualization, *Journal of Social Structure*, **7**.

DE NOOY, W., MRVAR, A., BATAGELJ V. (2005), *Exploratory Social Network Analysis with Pajek,* CUP, Cambridge University Press.

DeJORDY, R., BORGATTI, S.P., ROUSSIN, C., HALGIN, D.S. (2007), Visualizing Proximity Data, *Field Methods*, **19**, 239–263.

FAUST, K. (1997), Centrality in affiliation networks, *Social Networks*, **19**, 157–191.

FREEMAN, L.C. (1979), Centrality in social networks conceptual clar*Giuseppe Giordano and Maria Prosperina Vitale* ification, *Social Networks,* **1**, 215–239.

FREEMAN, L.C. (2000), Visualizing Social Networks, *Journal of Social Structure*, **1**.

FREEMAN, L.C. (2005), Graphic Tecniques for Exploring Social Network Data, In: *Carrington, P., Scott, J., Wasserman, S. (Eds.) Models and Methods in Social Network Analysis*, pp. 248-269. Cambridge University Press, Cambridge.

HAYTHORNTHWAITE, C. (1998),A social network study of the growth of community among distance learners, *Information Research*, **4**.

KAMADA, T., KAWAI, S. (1989), An Algorithm for Drawing General Undirected Graphs,*Inform. Process. Lett.* **31**, 7-15.

KILKON, K., KYOUNG, J.L., CHISUNG, P. (2007), Rethinking Preferential Attachment Scheme: Degree centrality versus closeness centrality, *Connection*, **27**, 53–59.

MARSDEN, P.V. (2002), Egocentric and sociocentric measures of network centrality, *Social Networks*. **24**, 407–422.

McGRATH, C., J. BLYTHE, KRACKHARDT, D. (1997), The effect of spatial arrangement on judgments and errors in interpreting graphs, *Social Networks*, **19**, 223–242.

MORENO, J.L. (1953), *Who Shall Survive?,* Beacon, N.Y.: Beacon House Inc.

WASSERMAN, S., FAUST, K. (1994), *Social Network Analysis: Methods and Applications*, Cambridge University Press, Cambridge.

WATTS, D. J. (1999), *SmallWorlds*, Princeton University Press: Princeton, New Jersey.

TECNICHE FATTORIALI LOCALI PER ESPLORARE DATI RELAZIONALI

Riassunto

In questo lavoro è proposta una strategia per integrare la tipica rappresentazione grafica di una rete con le tecniche di visualizzazione derivate dall'Analisi Fattoriale Locale. L'obiettivo finale è la definizione di una meta-analisi nell'ambito dello studio delle reti sociali che consenta di sintetizzare e visualizzare la struttura dei legami relazionali in uno spazio metrico. Il metodo proposto è illustrato attraverso un caso studio relativo ai flussi comunicativi in un ambiente di apprendimento online.