

Modal Clustering of Social Networks

Individuazione di gruppi basati sul riconoscimento delle mode nelle reti sociali

Giovanna Menardi and Domenico De Stefano

Abstract The idea underlying modal clustering is to associate groups with the regions around the modes of the probability density function underlying the data. This correspondence between clusters and dense regions in the sample space is here exploited to discuss a possible extension of modal clustering to the analysis of social networks. Such extension, albeit non trivial, seems particularly appealing: conceptually, the notion of high-density cluster fits well the one of cluster in a network, where groups are usually regarded as collections of individuals with dense local ties in their neighborhood. Additionally, modal clustering often resorts to graph theory for the operational detection of clusters, which is another condition that seems particularly appropriate to deal with relational data.

Abstract *I metodi di raggruppamento basati sul riconoscimento delle mode associano i gruppi con le regioni dello spazio campionario attorno alle mode della funzione di densità sottostante i dati. In questo lavoro si discute una possibile estensione di tali metodi all'analisi delle reti sociali. Benché non immediata, tale estensione appare promettente: concettualmente, la nozione di cluster quale regione ad alta densità si adatta naturalmente alle reti sociali, dove i gruppi sono tipicamente definiti come sottoinsiemi di individui fortemente interconnessi. Un altro aspetto che rende i metodi di clustering basati sulle mode particolarmente adatti ad essere applicati alle reti sociali è che, per la concreta individuazione dei gruppi, viene fatto ampio uso di strumenti presi in prestito dalla teoria dei grafi.*

Key words: density, modal clustering, social networks, centrality measures

Giovanna Menardi

Dip. Scienze Statistiche, Università di Padova, e-mail: menardi@stat.unipd.it

Domenico De Stefano

Dip. Economia, Università Ca' Foscari di Venezia e-mail: domenico.destefano@unive.it

1 Introduction

It is often of interest, in the analysis of social networks, to detect the possible existence of clusters of actors in particular regions of a network. Depending on the nature of relationships in the network, this tendency of individuals to interact may be driven by a natural attraction among similar subjects, namely homophily, or because of other mechanisms, such as popularity, ranking or social influence. All these attachment mechanisms are often undisclosed to the researcher, in particular when additional information is unavailable or when he has to deal with large complex network structures (*i.e.*, million of nodes, non-trivial tie formation processes and high overall density). A common approach to reveal – at least in an exploratory way – this tendency to interact is to detect clusters of nodes.

Intuitively, the concept of a cluster in networks is typically associated with a locally dense set of nodes (see, among others, [6]). Similarly to the problem of clustering non-relational data, the lack of a precise definition of cluster in social networks has lead to the proliferation of methods for their partitioning. Among these, some are explicitly designed to handle network data: spectral clustering [3], for instance, hinges on graph theory results and measures network connectivity via eigen-analysis. Blockmodeling [4] is another approach to network clustering, mainly based on finding a suitable permutation of a network matrix so that adjacent subjects can be regarded as equivalent, somehow, according to their social role in the network. Recently, a huge variety of network-based clustering techniques have been developed under the heading community detection, especially in the physics community, based on node centrality measures, flow models, random walks, resistor networks, optimization (see, for instance, [7]). Further clustering methods for social networks are natural adjustments of methods designed for non-relational data. Popular methods as complete or single-linkage clustering, for instance, have found an obvious counterpart in the analysis of social network by simply replacing the concept of distance with that of geodesic distance.

The current work follows the latter route and discusses the extension of the so called *modal* approach for clustering non-relational data to the network framework. The idea underlying modal clustering is to associate clusters with the regions around the modes of the density underlying the observed data, namely clusters correspond to dense regions in the space. This idea of high-density cluster fits well the one of network cluster conceptually. Additionally, modal clustering often resorts to graph theory for the operational detection of clusters.

After a brief review of modal clustering, in the following sections we discuss the extension of such approach to network data and illustrate it on a real data network.

2 Modal clustering

Modal clustering, in principle, delineates a class of methods for grouping non-relational data defined on a continuous, topological space. It builds on the concept

of clusters as “regions of high density separated from other such regions by regions of low density” [5, p. 205]. Formally, the observed data are supposed to be a sample from a multidimensional random variable with unknown probability density function f , and clusters are associated with the maximum connected regions around the modes of f . Any section of the density, at a level k , identifies a level set, namely the region with density above k . When f is unimodal, there is no clustering structure, and the level set is connected for any choice of k . Conversely, when f is multimodal, the identified level set may be connected or not, depending on k . If disconnected, the level set is formed by two or more connected components, corresponding to the regions around the modes of f which are formed by the section at the k level.

Clustering is performed by varying the level which identifies the contour sets and thus giving rise to a hierarchical structure, called cluster tree, where each leaf corresponds to a mode of the density function. Figure 1 illustrates these notions for a simple bivariate example.

Operationally, clustering involves two main choices: first, a density estimator is required and this is typically selected among the nonparametric methods. Second, for each threshold k it is to establish if the associated level set is connected. Since there is no obvious method to identify connected sets in a multidimensional space, graph theory may come to this aid. Thus, a suitable graph is built on the sample points and the connected components of the subgraphs induced by the selected level sets are then easily detected. A key matter becomes to suitably set the edges of this graph. In [1], for instance, a Delaunay triangulation is adopted; in [8], the minimum spanning tree associated to the nearest neighbor estimate of the density underlying the data is built.

Some of the main strengths of the modal approach to clustering already arise from this short summary. On one hand, there exists a precise notion of clusters, corresponding to a specific characteristic of the probability distribution underlying data; this entails that the problem of clustering, typically regarded as ill-posed just

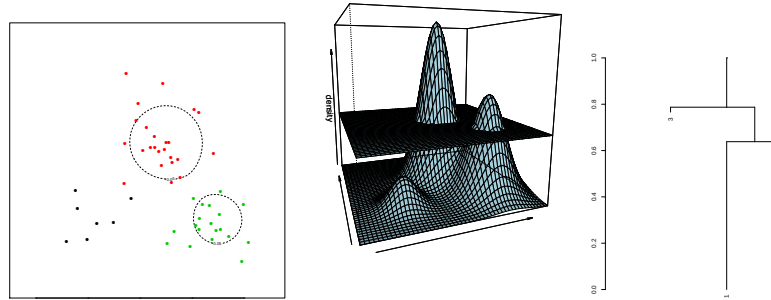


Fig. 1 A sample simulated from three subpopulations and the associated contour set at a level k (left plot). The threshold k defines a section of the trimodal density function (central plot) and identifies two connected regions. On the right, the cluster tree indicates the number of connected components for each k as well as the total number of clusters, corresponding to the leaves.

because of the lack of a formal definition, is somehow circumscribed; it also follows that the number of clusters is an intrinsic property of the clustering structure and its determination is itself an integral part of the estimation procedure. Additionally, this cluster notion is not specifically binded to a particular shape and is then, in principle, close to a “natural grouping of data. The use of flexible distributions or, even, of a nonparametric estimator to model the data, allows to maintain this flexibility.

2.1 Extending modal clustering to network data

Cluster definition as high-density regions suggests a natural counterpart in social network analysis, where clusters are often referred to as sets of actors with dense relationship patterns. Clearly, one has to take into account the inherent limitation of network data, which are not defined in a continuous, topological space as required by modal clustering and whose properties can be established in geodesic terms only. In particular, we have to give up a probabilistic notion of density, and intend it in a less formal way, reflecting some intuitive meaning of cohesiveness.

In the current analysis, we regard to social networks as an undirected graph $\mathcal{G} = \{V, E\}$ consisting of a set $V_{\mathcal{G}} = \{v_1, \dots, v_n\}$ of nodes (the actors of the network) and a set $E_{\mathcal{G}}$ of edges, representing relations between pairs of nodes. In the light of this, we might be tempted to borrow from graph theory the concept of density and akin notions. Density of a subgraph $\mathcal{H} \subseteq \mathcal{G}$ is defined as the proportion of all possible edges of $\mathcal{H}$ which are actually present. In fact, the idea of measuring locally dense nodes by means of network (or subnetwork) density can be misleading. On one hand, this measure is not discriminant enough for clustering purposes, *i.e.* density can be equal in rather different node configurations. On the other hand, density definition as node-wise measure is arbitrary as it cannot be defined in a unique manner on the individual nodes. For instance, one could set $\mathcal{H} = \mathcal{H}(v)$ as the subgraph formed by the nearest neighbors of v or focus on the single node v and reducing to the notion of degree (*i.e.*, the number of relations incident with v).

An alternative set of candidates to quantify local density would be represented by measures of connectivity or measures of centrality, which evaluate, somehow, the position as well as the prominence of each node in a network.

It is worthwhile to observe that the choice of a density measure is not inconsequential in terms of cluster interpretation and different choices might entail a different concept of cluster. For instance, the concept of node degree and, especially, the one of eigenvector centrality – the latter used also in some community detection algorithm [7] – account for the rate of activity and, respectively, the popularity.

Given the network $\mathcal{G} = \{V, E\}$, once that a measure of density $\delta : V \mapsto \mathbb{R}^+ \cup \{0\}$ is selected, clustering of the actors is performed by following any standard method of modal clustering, with the further benefit that the connected components associated to the high-density level sets may be identified as the connected components of the induced subgraphs. The following scheme, for instance, is a suitable adaptation of [1]:

1. Compute the density of the relationships of each actor: $\delta(v_1), \dots, \delta(v_i), \dots, \delta(v_n)$. Clusters will be formed around the modal actors, namely actors with the densest relationship patterns.
2. For $0 < k < \max_i \delta(v_i)$:
 - Determine the level set $V_k = \{v_i \in V : \delta(v_i) \geq k\}$, and define $\mathcal{G}_k \subset \mathcal{G}$ as the subgraph induced by V_k . Subgraph $\mathcal{G}_k$ is formed by removing from $\mathcal{G}$ all the nodes not in V_k and all edges with at least one node among them.
 - Find the connected components of $\mathcal{G}_k$, namely the union of connected pairs that share at least one node (many standard algorithms are available).
3. Build the cluster tree. The cluster tree is a hierarchical structure that associates each level k to the number of connected components of $\mathcal{G}_k$. The total number of leaves corresponds to the number of clusters.

The described way of proceeding gives usually rise to a number of initial clusters, while actors with low-density relationship patterns that do not univocally pertain to a mode are left unallocated. Depending on the aim of clustering and on subjects-matter considerations, part, or, all of these actors may be either left unallocated or be assigned to the initial cluster for which the ratio between the two highest densities, conditional to the group, is maximum.

3 Numerical illustration and discussion

In this section, we show an illustration of the proposed technique on the well-known *Zachary Karate Club* data [9], describing the network of friendships between 34 members of a karate club at a US university in the 1970. We neglect information about the strength of the interactions between actors, which would give rise to a weighted network, and consider the adjacency matrix only. A useful feature of this data set is that, during the period of observation, the club split into two factions, due to a dispute between the administrator and the head teacher. Thus, a true cluster membership of the actors is known and can be used as a benchmark to evaluate clustering methods. Figure 2 shows the *Zachary Karate Club* network and highlights the two factions.

The clustering method described above has been applied on the data by selecting the degree as a measure of local density of the actors. For comparison, two further methods for clustering networks have been applied, namely one approach based on community detection [7] and a k -cores algorithm [2]. Results are displayed on the right side of Figure 2. Performance of the two considered alternative methods is quite unsatisfactory. In both cases four clusters are identified instead of the two. Additionally, cluster composition is homogeneous with the true partition only for the community detection approach. Conversely, modal clustering can identify the true factions well and this result holds also for different measures of local density (not reported for brevity).

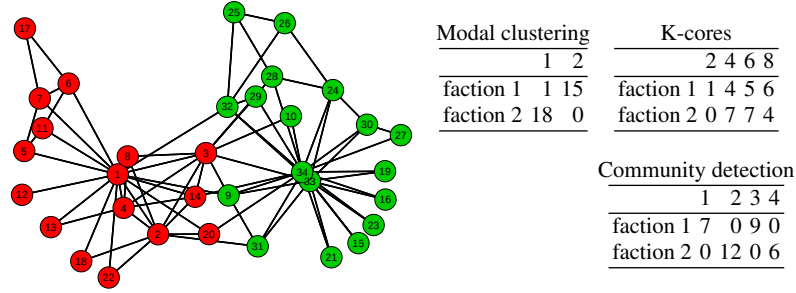


Fig. 2 Zachary Karate Club data: network with the two factions highlighted with different colors (left); number of actors cross-classified by true partitions and partitions obtained by modal clustering and some alternatives (right).

To conclude, the application of modal clustering to social networks is appealing both conceptually, as the notion of high-density cluster fits well the one of cluster in a network, and operationally, because modal clustering often resorts to graph theory tools. An inherent limitation of network data is that they are not defined in a continuous space as required by modal clustering. Thus, a formal notion of density has to be relaxed and intended, for instance, in terms of strenght of relationships between actors. A first exploration of these ideas has shown promising results. A deeper insight is certainly needed, in order to compare different local measures of density as well as taking into account directional networks and the possible presence of covariates for each actor.

References

1. Azzalini, A. and Torelli, N. Clustering via nonparametric density estimation. *Statistics and Computing*, 17:71–80 (2007).
2. Batagelj, V. and Zaversnik, M. Generalized Cores. [arXiv:cs/0202039v1](https://arxiv.org/abs/cs/0202039v1) (2002).
3. Chung, F. *Spectral graph theory*. American Mathematical Society (1997).
4. Doreian, P. and Batagelj, V. and Ferligoj, A. *Generalized Blockmodeling*. Cambridge University Press (2002).
5. Hartigan, J. A. *Clustering Algorithms*. Wiley, New York (1975).
6. Moody, J. Peer influence groups: identifying dense clusters in large networks. *Social Networks*, 23:261–283 (2001).
7. Newman, M. E. Finding community structure using the eigenvectors of matrices. *Physical Review, E*, 74, 036104 (2006).
8. W. Stuetzle, and R. Nugent. A generalized single linkage method for estimating the cluster tree of a density. *Journal of Computational and Graphical Statistics*. 19(2):397–418 (2010).
9. Zachary, W. W. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research* 33:452–473 (1977).