Clustering and visualizing large cattle-trade networks using relational self-organizing maps

Madalina Olteanu [1,2] & Kevin Pame [1] & Gaël Beaunée [3]
& Caroline Bidot [4] & Elisabeta Vergu [1]

[1] *MaIAGE, INRA, Université Paris Saclay, 78350 Jouy-en-Josas, France*
madalina.olteanu,kevin.pame,elisabeta.vergu@inra.fr
[2] *SAMM, EA 4543, Université Panthéon Sorbonne, 75013 Paris, France*
[3] *Bioepar, INRA, Oniris, La Chantrerie, 44307 Nantes, France*
gael.beaunee@inra.fr
[4] *ISPA, INRA, 33882 Villenave d'Ornon, France*
caroline.bidot@inra.fr

Résumé. Les réseaux d'échanges d'animaux entre fermes, marchés et autres opérateurs, font partie des objets mathématiques les plus étudiés ces dernières années. En effet, de nombreuses recherches issues de champs disciplinaires variés s'attachent à étudier le fonctionnement de ces réseaux complexes, qui évoluent dans le temps. L'un des objectifs est de comprendre ce fonctionnement, extraire de l'information et de la structure, des patterns ou des *clusters* qui puissent mieux expliquer des phénomènes comme les comportements préférentiels, la propagation d'épidémies, ... Nous étudions ici l'apport du clustering relationnel appliqué à trois distances différentes construites sur des chemins temporellement compatibles dans un réseau dynamique. Nous présentons les résultats exploratoires obtenus sur un sous-ensemble représentatif du réseau d'échanges de bovins en France – la Bretagne –, observé avec une fréquence journalière entre 2005 et 2009.

Mots-clés. Réseaux dynamiques, distances temporellement compatibles, clustering relationnel.

Abstract. Farm contact networks or animal trade networks are being one of the most extensively studied mathematical objects these last few years. Indeed, research stemming from various fields – epidemiology, physical statistics, applied mathematics – aims at understanding the behavior of these complex, dynamic networks. One of the goals of these studies is to mindfully explore these massive data, extract meaningful information and structure, and eventually isolate specific patterns and clusters which may contribute to assessing phenomena such as preferential behaviors, epidemics spreading, ... The present contribution addresses the insights that relational clustering trained with various distances computed from temporally reachable paths may bring in the exploratory study of dynamical networks. We illustrate our findings on a representative sub-network of the cattle-trade French network – Brittany –, monitored with a daily frequency between 2005 and 2009.

Keywords. Dynamic networks, temporally reachable paths, relational clustering.

1 Introduction

Time-varying networks have drawn quite a burst of attention in the last years, and two of the related questionings are on the one hand how to understand the underlying structure(s) of the network better, and on the other hand how to visualize simplified version(s) of it. When one aims at bringing into light how groups of entities in a graph are organized and how they interact, clustering and visualization were proven very useful and complementary tools. While the literature on dynamic network clustering and community detection becomes richer and richer, several promising families of methods are emerging. We may cite, for instance, spectral algorithms [3] or model-based algorithms, and particularly dynamical stochastic block models [5], which manage, besides clustering, to capture the temporal evolution of the clusters also, but on relatively small networks.

We focus here on another class of general clustering methods, kernel and relational based, which are not designed specifically for graphs but are able to perform clustering for any similarity/dissimilarity measure between vertices or edges, including mixtures of similarities/dissimilarities. In recent work, there was evidence that kernel or relational self-organizing maps (SOM) are able to uncover structures in static undirected graphs [7], including various sources of information on the vertices and on the edges jointly combined. In this manuscript, we aim at studying how this algorithm behaves on large dynamic networks, which are the resulting partitions and the insights they provide.

Three properties of relational SOM make it worth of interest in the context of large dynamic networks with additional multivariate information available on the nodes and edges. First, large networks may be clustered by using a bagged and hence parallel-implemented version of the algorithm [4]. In this work, the various parameters of the bagging procedure (number of bags, size of the bags, sampling procedure, ...) were empirically tuned by minimizing the global within-class variance. Second, any temporal dissimilarity computed on the network can be used and one may check whether some are more appropriate than others for revealing clusters, or for obtaining clusters having some desired features. Third, it is possible to adaptively and optimally mix several dissimilarities, including if necessary additional information [6].

2 Mining dynamic networks with relational clustering

For illustrating the above, we explore the French cattle-trade network and more particularly a densely populated region, in terms of both vertices and active edges. We study animal exchanges between farms, commercial operators, farms and commercial operators, in Brittany, during five years (2005-2009) [2]. This yields to a dynamic graph with approximately 30,000 vertices and 2,000,000 instants of activated edges. The graph is directed (from sellers to buyers) and weighted (by the number of exchanged animals).

Three temporal distances are computed and investigated [1]: the instant of the foremost or the earliest arrival, the duration of fastest or the quickest arrival, and the averaged temporal distance as defined in [8].

The resulting clusters are examined and interpreted in terms of homogeneity, sparsity, distribution of exchange distances and other covariates, ... The relevance of the clusters is also assessed by letting a simple susceptible-infected (SI) model diffuse from vertices randomly picked in the network, and then comparing the proportion of infected nodes in the associated clusters and in the whole network.

The earliest arrival distance for instance appears to be very sensitive to epidemic spreading (see Figure 1), and in most cases the cluster where the infection starts is rapidly and largely infected, whereas the rest of the network is only slightly reached. However, the earliest arrival distance is time dependent, a reference initial moment being needed for its computation. If the epidemic is started at this reference moment, then the partitioning is in agreement with the outcome of the infection spreading. However, if there is a delay between the infection starting time and the reference time used for computing the foremost distance, then the vertices in the same cluster as the initial infected one are less or no longer preferentially infected. Since the edge activation process is not stationary in time for the data at study, this effect was expected. Other distances, such as the quickest arrival, are time-independent and they are appealing to use if one wishes to minimize the travel time throughout the entire observed period. On our data, this distance produces homogeneous clusters, of relatively small sizes, and with very low degree of sparsity (most vertices in a cluster are reachable from any vertex in the same cluster).

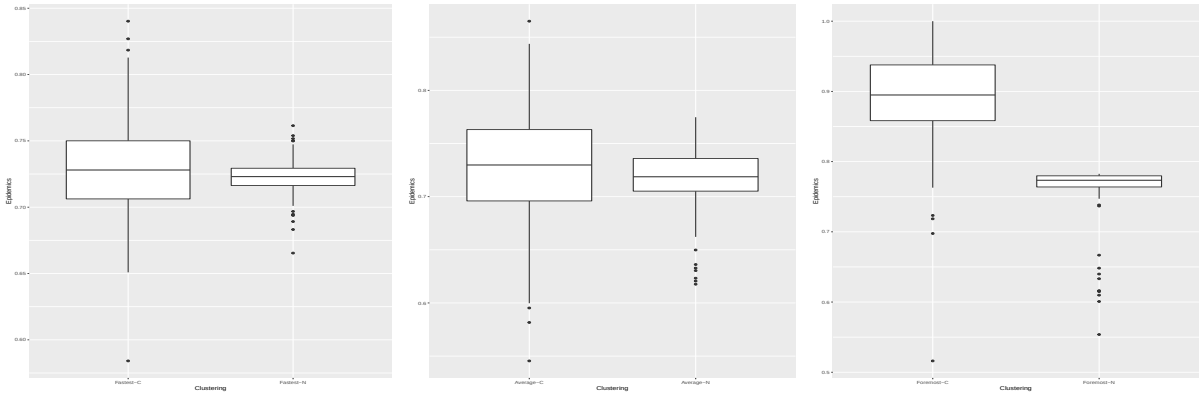

Figure 1: Epidemics propagation in clusters and in the whole network for fastest, average and foremost temporal distances

A complete discussion of the results on the three distances, including reduced network representations with the clusters of vertices projected both on a two-dimensional map and on a reduced directed graph similar to Figure 2, will be provided. We will specifically focus

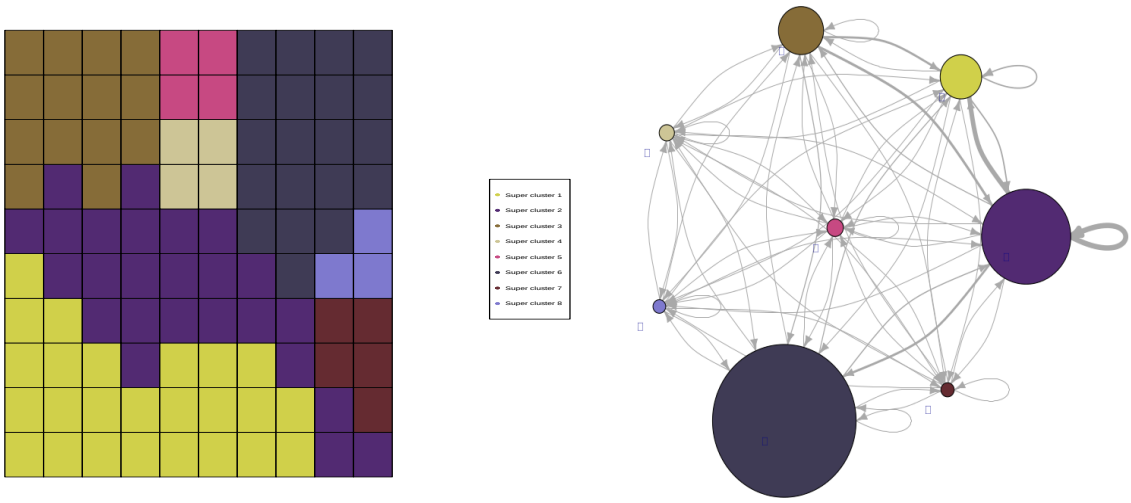

Figure 2: Fastest distance - reduced map and graph to super-clusters

on the pros and cons of using each of the temporal distances in the context of studying a large animal-exchange network. We will also discuss the visualization issue, which is not trivial and should at least contain information about the clusters, the dynamics, and the geographical localization of the vertices.

References

[1] Bui-Xuan, B., Ferreira, A., A., J.: Computing shortest, fastest, and foremost journeys in dynamic networks. International Journal of Foundations of Computer Science **14**(02), 267–285 (2003)

[2] Dutta, B.L., Ezanno, P., Vergu, E.: Characteristics of the spatio-temporal network of cattle movements in france over a 5-year period. Preventive veterinary medicine **117**(1), 79–94 (2014)

[3] Liu, F., Choi, D., Xie, L., Roeder, K.: Global spectral clustering in dynamic networks. Proceedings of the National Academy of Sciences **115**(5), 927–932 (2018)

[4] Mariette, J., Olteanu, M., Boelaert, J., Villa-Vialaneix, N.: Bagged kernel SOM. In: T. Villmann, F. Schleif, M. Kaden, M. Lange (eds.) Advances in Self-Organizing Maps and Learning Vector Quantization (Proceedings of WSOM 2014), *Advances in Intelligent Systems and Computing*, vol. 295, pp. 45–54. Springer Verlag, Berlin, Heidelberg, Mittweida, Germany (2014)

[5] Matias, C., Miele, V.: Statistical clustering of temporal networks through a dynamic stochastic block model. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **79**(4), 1119–1141 (2017)

[6] Olteanu, M., Villa-Vialaneix, N.: On-line relational and multiple relational SOM. Neurocomputing **147**, 15–30 (2015)

[7] Olteanu, M., Villa-Vialaneix, N.: Using SOMbrero for clustering and visualizing graphs. Journal de la Société Française de Statistique **156**(3), 95–119 (2015)

[8] Pan, R.K., Saramäki, J.: Path lengths, correlations, and centrality in temporal networks. Phys. Rev. E **84**, 016,105 (2011)