Clustering Student Mobility Data in 3-way Networks

Vincenzo Giuseppe Genova, Giuseppe Giordano, Giancarlo Ragozini, and Maria Prosperina Vitale

Abstract The present contribution aims at introducing a network data reduction method for the analysis of 3-way networks in which classes of nodes of different types are linked. The proposed approach enables simplifying a 3-way network into a weighted two-mode network by considering the statistical concept of joint dependence in a multiway contingency table. Starting from a real application on student mobility data in Italian universities, a 3-way network is defined, where provinces of residence, universities and educational programmes are considered as the three sets of nodes, and occurrences of student exchanges represent the set of links between them. The Infomap community detection algorithm is then chosen for partitioning two-mode networks of students' cohorts to discover different network patterns.

Keywords: 3-way network, complex network, community detection, mobility data, tertiary education

Vincenzo Giuseppe Genova
Department of Economics, Business, and Statistics, University of Palermo, Italy,
e-mail: vincenzogiuseppe.genova@unipa.it

Giuseppe Giordano
Department of Political and Social Studies, University of Salerno, Italy,
e-mail: ggiordano@unisa.it

Giancarlo Ragozini
Department of Political Science, Federico II University of Naples, Italy,
e-mail: giragoz@unina.it

Maria Prosperina Vitale (✉)
Department of Political and Social Studies, University of Salerno, Italy,
e-mail: mvitale@unisa.it

P. Brito et al. (eds.), *Classification and Data Science in the Digital Age*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-031-09034-9_17

1 Introduction

Many complex relational data structures can be described as multimode or multiway networks in which nodes belonging to different modes are linked. The most common multimode network in social networks is represented by the affiliation network, where two-mode data, actors and events, form a bipartite graph divided into two groups [6]. In the case of tripartite networks, we deal with three types of nodes, and different graph structures can be defined.

Although only a few papers deal with methods for these networks, in recent years, a growing number of works have appeared –especially in bipartite and tripartite cases– to disentangle the inherent complexity of such kinds of data structures. Looking at clustering and community detection algorithms proposed to partition a network into groups, we can identify some strands, all deriving from generalizations of methods suited for one-mode [19] and two-mode networks [2]. A classical approach consists of applying the usual community detection algorithms on a unique supra-adjacency matrix defined by combining all the possible two-mode networks in a block matrix [11, 15]. Alternative methods rely on projecting each two-mode networks and on applying separately the usual community detection algorithms on these matrices [10]. In addition, there are methods adopting both an optimization procedure for 3-way networks [16, 17, 14] by extending the idea of bipartite modularity [2], and an indirect blockmodeling approach by deriving a dissimilarity measure based on structural equivalence concept [3].

In our opinion, approaches based on the analysis of the k-modes examined considering the collection of the $k(k-1)/2$ two-mode networks [10] cannot take into account statistical associations among all modes at same time. Hence, the aim of the contribution is to present a network data reduction method based on the concept of joint dependence in a multiway contingency table [1].

Starting from real applications on the Italian student mobility phenomenon in higher education [12, 21, 7, 8, 13, 22], a 3-way network is defined, where provinces of residence, universities and educational programmes are considered as the three modes. Student mobility flows, measured in terms of occurrences, represent the set of links between them. Assuming that the statistical dependency between the set of nodes provinces of residence and the other two sets of nodes can be captured by the joined pair of nodes (universities and educational programmes), the tripartite network is transformed into a bipartite network, where the two modes are given by Italian provinces of residence (first mode) and the set of nodes given by all possible pairs of universities and educational programmes (second mode). Thus, taking advantage of this approach of network simplification, network indexes and clustering techniques for bipartite networks are available. Hence, the Infomap community detection algorithm is adopted [9, 4] to partition the derived network.

The remainder of the paper is organized as follows. Section 2 presents the details of the proposed strategy of analysis, and the main results are reported from the analysis of student mobility data of Italian universities. Section 3 provides final remarks.

2 Simplification of 3-way Networks

In the present paper, the case of a tripartite network is considered as an example to show how the proposed network data simplification method works. In particular, we consider the real case study of student mobility paths in Italian universities. The MOBYSU.IT dataset[1] enables reconstruction of network data structures considering student mobility flows among territorial units and universities.

More formally, given $\mathcal{V}_P \equiv \{p_1, \ldots, p_i, \ldots, p_I\}$, the set of I provinces of residence; $\mathcal{V}_U \equiv \{u_1, \ldots, u_j, \ldots, u_J\}$, the set of J Italian universities, and $\mathcal{V}_E \equiv \{e_1, \ldots, e_k, \ldots, e_K\}$, the set of K educational programmes, a weighted tripartite 3-uniform hyper-graph $\mathcal{T}$ can be defined, consisting of a triple $(\mathcal{V}, \mathcal{L}, \mathcal{W})$, with $\mathcal{V} = \{\mathcal{V}_P, \mathcal{V}_U, \mathcal{V}_E\}$ the collection of three sets of vertices, one for each mode, and being $\mathcal{L} = \{\mathcal{L}_{PUE}\}$, $\mathcal{L}_{PUE} \subseteq \mathcal{V}_P \times \mathcal{V}_U \times \mathcal{V}_E$, the collection of hyper-edges, with generic term (p_i, u_j, e_k), which is the link joining the i-th province, the j-th university, and the k-th educational programme. Finally, $\mathcal{W}$ is the set of weights, obtained by the function $w : \mathcal{L}_{PUE} \rightarrow \mathbb{N}$, and $w(p_i, u_j, e_k) = w_{ijk}$ is the number of students moving from a province p_i towards a university u_j in an educational programme e_k. Such a network structure can be described as a three-way array $\mathbb{A} = (a_{ijk})$, with $a_{ijk} \equiv w_{ijk}$, and it has been called a 3-way network [3].

To deal with such a complex network structure and aiming at obtaining communities in which three modes are mixed, we wish to simplify the tripartite nature of the graph, without losing any significant information. In statistical terms, the array $\mathbb{A}$ can be interpreted as a 3-way contingency table, and then the statistical techniques to evaluate the association among variables (i.e. the modes) can be exploited [1]. Because a 3-way contingency table is a cross-classification of observations by the levels of three categorical variables, we are defining a network structure where the sets of nodes are the levels of the categorical variables. Specifically, we assume that if two modes are jointly associated –as are, for their own nature, universities and educational programmes– the tripartite network can be logically simplified into a bipartite one. In the student mobility network, we join the pair of nodes in $\mathcal{V}_U$ and in $\mathcal{V}_E$, and then we deal with the relationships between these *dyads* and the nodes in $\mathcal{V}_P$.

Following this assumption, the sets of nodes $\mathcal{V}_U$ and $\mathcal{V}_E$ are put together into a set of joint nodes, namely $\mathcal{V}_{UE}$. The tripartite network $\mathcal{T}$ can now be represented as a bipartite network $\mathcal{B}$ given by the triple $\{\mathcal{V}^*, \mathcal{L}^*, \mathcal{W}^*\}$, with $\mathcal{V}^* = \{\mathcal{V}_P, \mathcal{V}_{UE}\}$. The set of hyper-edges $\mathcal{L}$ is thus simplified into a set of edges $\mathcal{L}^* = \{\mathcal{L}_{P,UE}\}$, $\mathcal{L}_{P,UE} \subseteq \mathcal{V}_P \times \mathcal{V}_{UE}$. The new edges $(p_i, (u_j; e_k))$ connect a province p_i with an educational programme e_k running in a given university u_j. The weights $\mathcal{W}^*$ are the same as in the hyper-graph $\mathcal{T}$, i.e., $w^*_{ij,k} = w_{ijk}$. Note that the weights contained in the 3-way array $\mathbb{A}$ are preserved, but are now organized in a rectangular matrix $\mathbf{A}$ of I rows and $(J \times K)$ columns.

---

[1] Database MOBYSU.IT [Mobilità degli Studi Universitari in Italia], research protocol MUR - Universities of Cagliari, Palermo, Siena, Torino, Sassari, Firenze, Cattolica and Napoli Federico II, Scientific Coordinator Massimo Attanasio (UNIPA), Data Source ANS-MUR/CINECA.

Taking advantage of this method, we aim to analyse weighted bipartite graphs adopting clustering methods. Among others, we use the Infomap community detection algorithm [9, 4] to study the flows' patterns in network structures instead of modularity optimization proposed in topological approaches [18, 5]. Indeed, the rationale of this algorithm –*map equation*– takes advantage of the duality between finding communities and minimizing the length –*codelength*– of a random walker's movement on a network. The partition with the shortest path length is the one that best captures the community structure in the bipartite data. Formally, the algorithm defines a module partition **M** of n vertices into m modules such that each vertex is assigned to one and only one module. The Infomap algorithm looks for the best **M** partition that minimizes the expected *codelength*, $L(M)$, of a random walker, given by the following map equation:

$$L(M) = q_{\curvearrowright} H(\mathscr{Q}) + \sum_{i=1}^{m} p^i_{\circlearrowright} H(\mathscr{P}^i) \tag{1}$$

In equation (1), $q_{\curvearrowright} H(\mathscr{Q})$ represents the entropy of the movement between modules weighed for the probability that the random walker switches modules on any given step ($q_{\curvearrowright}$), and $\sum_{i=1}^{m} p^i_{\circlearrowright} H(\mathscr{P}^i)$ is the entropy of movements within modules weighed for the fraction of within-module movements that occur in module i, plus the probability of exiting module i ($p^i_{\circlearrowright}$), such that $\sum_{i=1}^{m} p^i_{\circlearrowright} = 1 + q_{\curvearrowright}$ [9].

In our case, the Infomap algorithm is adopted to discover communities of students characterized by similar mobility patterns. Indeed, to analyse mobility data, where links represent patterns of student movement among territorial units and universities, flow-based approaches are likely to identify the most important features. Finally, in our student mobility network, to focus only on relevant student flows, a filtering procedure is adopted by considering the Empirical Cumulative Density Function (ECDF) of links' weights distribution.

2.1 Main Findings

Students' cohorts enrolled in Italian universities in four academic years (a.y.) 2008–09, 2011–12, 2014–15, and 2017–18 are analysed. The number of nodes for the sets $\mathcal{V}_P$ (107 provinces), $\mathcal{V}_U$ (79-80 universities), and $\mathcal{V}_E$ (45 educational programmes), and the number of students involved in the four cohorts are quite stable over time (Table 1). Furthermore, the percentage of movers (i.e., students enrolled in a university outside of their region of residence) increased, from 16.4% in the a.y. 2008–09 to 20.6% in the a.y. 2017–18, and it is higher for males than females.

Table 1 Percentage of students according to their mobility status by cohort and gender.

Cohort	Gender		Mover status	
			Stayers%	Movers%
2008–09	F	136,381	84.2	15.8
	M	106,950	82.8	17.2
	Total	243,331	83.6	16.4
2011–12	F	126,606	81.7	18.3
	M	102,479	80.9	19.1
	Total	229,085	81.0	19.0
2014–15	F	121,121	80.5	19.5
	M	102,358	80.4	19.6
	Total	223,479	80.5	19.5
2017–18	F	134,315	79.1	20.9
	M	113,496	79.8	20.2
	Total	247,811	79.4	20.6

Following the network simplification approach, the tripartite networks –one for each cohort– are simplified into bipartite networks, and the four ECDFs of links' weights are considered to filter relevant flows. The distributions suggest that more than 50% of links between pairs of nodes have weights equal to 1 (i.e., flows of only one student), and about 95% of flows are characterized by flows not greater than a digit. Thus, networks holding links with a value greater or equal to 10 are further analysed.

To reveal groups of universities and educational programmes attracting students, the Infomap community detection algorithm is applied. Looking at Table 2, we notice a reduction of the number of communities from the first to the last student cohort, suggesting a sort of stabilization in the trajectories of movers towards brand universities of the center-north with also an increase in the north-north mobility [20], and a relevant dichotomy between scientific and humanistic educational programmes. Network visualizations by groups (Figures 1 and 2) confirm that the more attractive universities are located in the north of Italy, especially for educational programmes in economics and engineering (the Bocconi University, the Polytechnic of Turin and the Cattolica University).

Table 2 Number of communities, codelength, and relative saving codelength per cohort.

Cohort	Communities	Codelength	Relative saving codelength
2008–09	14	0.96	83%
2011–12	17	1.72	70%
2014–15	3	5.23	12%
2017–18	3	1.00	83%

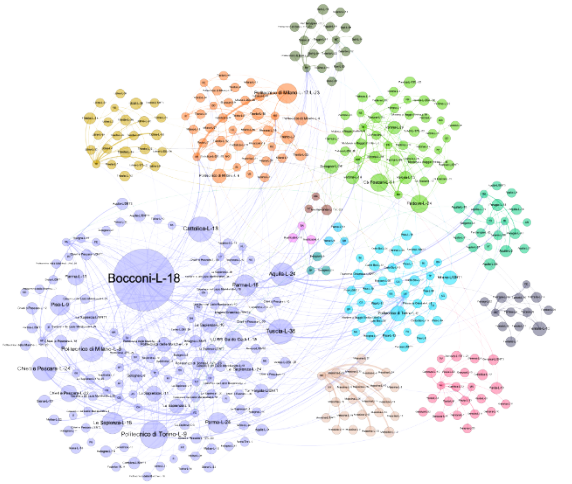

Fig. 1 Network visualization by groups, student cohort a.y. 2008–09.

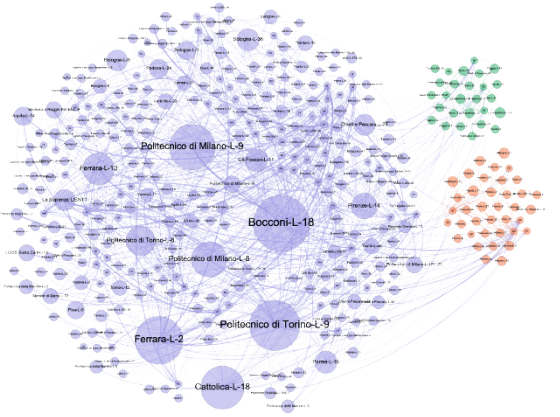

Fig. 2 Network visualization by groups, student cohort a.y. 2017–18.

3 Concluding Remarks

The proposed simplification network strategy on tripartite graphs defined for student mobility data provides interesting insights for the phenomenon under analysis. The main attractive destinations still remain the northern universities for educational programmes, such as engineering and business. Besides the well-known south-to-north route, other interregional routes in the northern area appear. In addition, the reduction in the number of communities suggests a sort of stabilization in terms of mobility roots of movers towards brand universities, highlighting student university destination choices close to the labor market demand.

Hyper-graphs and multipartite networks still remain very active areas for research and challenging tasks for scholars interested in discovering the complexities underlying these kinds of data. Specific tools for such complex network structures should be designed combining network analysis and other statistical techniques. As future lines of research, the comparison of community detection algorithms that better represent the structural constraints of the phenomena under analysis and the assessment of other backbone approaches to filter the significant links will be developed.

Acknowledgements The contribution has been supported from Italian Ministerial grant PRIN 2017 "From high school to job placement: micro-data life course analysis of university student mobility and its impact on the Italian North-South divide", n. 2017 HBTK5P - CUP B78D19000180001.

References

1. Agresti, A.: Categorical Data Analysis (Vol. 482). John Wiley & Sons, New York (2003)
2. Barber, M. J.: Modularity and community detection in bipartite networks. Phys. Rev. E, **76**, 066102 (2007)
3. Batagelj, V., Ferligoj, A., Doreian, P.: Indirect Blockmodeling of 3-Way Networks. In: Brito P., Cucumel G., Bertrand P., de Carvalho F. (eds) Selected Contributions in Data Analysis and Classification. Studies in Classification, Data Analysis, and Knowledge Organization, pp. 151–159. Springer, Berlin, Heidelberg (2007)
4. Blöcker, C., Rosvall, M.: Mapping flows on bipartite networks. Phys. Rev. E, **102**, 052305 (2020)
5. Blondel, V. D., Guillaume, J. L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech.-Theory E, **10**, P10008 (2008)
6. Borgatti, S. P., Everett, M. G.: Regular blockmodels of multiway, multimode matrices. Soc. Networks, **14**, 91–120 (1992)
7. Columbu, S., Porcu, M., Primerano, I., Sulis, I., Vitale, M.P.: Geography of Italian student mobility: A network analysis approach. Socio. Econ. Plan. Sci. **73**, 100918 (2021)
8. Columbu, S., Porcu, M., Primerano, I., Sulis, I., Vitale, M. P.: Analysing the determinants of Italian university student mobility pathways. Genus, **77**, 34 (2021)
9. Edler, D., Bohlin, L., Rosvall, M.: Mapping higher-order network flows in memory and multilayer networks with infomap. Algorithms, **10**, 112 (2017)
10. Everett, M. G., Borgatti, S.: Partitioning multimode networks. In: Doreian, P., Batagelj, V., Ferligoj, A. (eds.) Advances in Network Clustering and Blockmodeling, pp. 251-265, John Wiley & Sons, Hoboken, USA (2020)

11. Fararo, T. J., Doreian, P.: Tripartite structural analysis: Generalizing the Breiger-Wilson formalism. Soc. Networks, **6**, 141–175 (1984)
12. Genova, V. G., Tumminello, M., Aiello, F., Attanasio, M.: Student mobility in higher education: Sicilian outflow network and chain migrations. Electronic Journal of Applied Statistical Analysis, **12**, 774–800 (2019)
13. Genova, V. G., Tumminello, M., Aiello, F., Attanasio, M.: A network analysis of student mobility patterns from high school to master's. Stat. Method. Appl., **30**, 1445–1464 (2021)
14. Ikematsu, K., Murata, T.: A fast method for detecting communities from tripartite networks. In: International Conference on Social Informatics, pp. 192-205. Springer, Cham (2013)
15. Melamed, D., Breiger, R. L., West, A. J.: Community structure in multi-mode networks: Applying an eigenspectrum approach. Connections, **33**, 18–23 (2013)
16. Murata, T.: Detecting communities from tripartite networks. In: Proceedings of the 19th international conference on world wide web, pp. 1159-1160. (2010)
17. Neubauer, N., Obermayer, K.: Tripartite community structure in social bookmarking data. New Rev. Hypermedia M., **17**, 267-294 (2011)
18. Newman, M. E., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E, **69**, 026113 (2004)
19. Newman, M. E.: Modularity and community structure in networks. Proceedings of the National Academy of Sciences, **103**, 8577-8582 (2006)
20. Rizzi, L., Grassetti, L. Attanasio, M.: Moving from North to North: how are the students' university flows? Genus **77**, 1–22 (2021)
21. Santelli, F., Scolorato, C., Ragozini, G.: On the determinants of student mobility in an interregional perspective: A focus on Campania region. Statistica Applicata - Italian Journal of Applied Statistics, **31**, 119–142 (2019)
22. Santelli, F., Ragozini, G., Vitale, M. P.: Assessing the effects of local contexts on the mobility choices of university students in Campania region in Italy. Genus, **78**, 5 (2022)