Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v12n4p774

Student mobility in higher education: Sicilian outflow network and chain migrations
By Genova et al.

Published: 15 December 2019

Electronic Journal of Applied Statistical Analysis
Vol. 12, Issue 04, December 2019, 774-35
DOI: 10.1285/i20705948v12n4p774

Student mobility in higher education: Sicilian outflow network and chain migrations

Genova V. G.[a], Tumminello M.[a], Enea M.[a], Aiello F. *[b], and Attanasio M.[a]

[a] *University of Palermo, Department of Economics, Business, and Statistics*
[b] *University of Enna "Kore", Department of Economics and Law*

Published: 15 December 2019

The most important student mobility (SM) flow in Italy is from the Southern to the Central-Northern regions, a phenomenon that has been magnified by an increasing number of outgoing students from Sicily over the last decade. In this paper, we rely upon micro-data of university enrollment and students' personal records for three cohorts of freshmen, in order to investigate preferential patterns of SM from Sicily toward universities in other regions. Our main goal is to reveal the existence of *chain migrations*, where students from a particular geographical area move towards a particular destination to follow other students that have previously moved. The paper provides aspects that are innovative under the view of the data, of the application, and of the statistical method. The data from each cohort is represented as a tripartite network with three sets of nodes, namely, clusters of Sicilian municipalities, students, and universities. The tripartite network is projected in a bipartite weighted network of clusters and universities, which is, then, filtered, in order to obtain a statistically validated bipartite network (SBVN). The SBVNs of the three cohorts may suggest the existence and evolution of chain migration patterns over time, which are also gender specific.

keywords: student mobility, chain migration, networks.

*Corresponding author: fabio.aiello@unikore.it.

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

1 Introduction

The Italian public universities are subsidized within a competitive framework that economically awards excellence, efficiency, and the capacity of universities to attract students from Italian regions other than its own. However, repeated cuts to public spending over time greatly disadvantaged the Southern universities, also considering the well-known Italian infrastructural and economic North-South divide. Therefore investigating the mechanisms that might explain the presence of a net flow of students from the South to Northern universities appears very important, and "chain migration" might be one of such mechanisms. Over the last decade, student mobility (SM) flows from Southern regions (Attanasio and Enea, 2019; Viesti, 2016) to Northern universities have been growing at an increasing rate, especially from Sicily, in spite of the presence of four universities in the island. Such unidirectional flows imply an increasing loss of human capital for the Southern regions, as most of the students does not come back in the origin region at the end of their studies (Dotti et al., 2013). In the literature, the Italian SM has been studied by following two approaches of analysis, depending on the available data. The first approach uses macro-data to describe flows of students moving from an area to another one and to detect the determinants of mobility. This approach uses information at an aggregate level of macro-area, such as province, region, or macro-region (Giambona et al., 2017; Bruno and Genovese, 2012). The second approach is based on micro-data, which allows to include individual student characteristics (D'Agostino et al., 2019; Lupi and Ordine, 2009). Recently, availability of micro-data at national level allowed to perform longitudinal analyses (Enea, 2018). The previous contributions highlight the reasons for SM, in particular from South and Islands, are based on: *i)* individual student characteristics, such as high school type and final mark, *ii)* the attractiveness of the universities, and *iii)* the regions where these are located in. At the level of university, some determinants of the attractiveness can be represented by the quality of both research and teaching (Ciriaci, 2014), a different training offer, and the availability of facilities and scholarships (Dal Bianco et al., 2010). Instead, at the level of regions, the Central-Northern ones offer higher quality of life and better job opportunities after graduation (Dotti et al., 2014). For these reasons, Central-Northern universities located in towns such as Turin, Milan, Bologna and Rome are often the preferred destination for Southern students who decide to migrate. However, while it is clear which the destination universities are, we believe these determinants are only a part of the reasons that boost the student to migrate. Indeed, students may also choose an out-of-region university by following a chain migration. According to the concept of chain migration (MacDonald and MacDonald, 1964), a student can find initial accommodation and other facilities by means of primary social relationships with previous movers, or migrants in general. In the broadest sense, such a phenomenon has been extensively studied and theorized in economic and sociological contexts (Haug, 2008). Literature contributions on chain migration in higher education mainly focus on international student mobility. Pérez and Mcdonough (2008) investigate on chain migrations of Latinos students in Los Angeles, Daniel (2014) analyzes the migration of Peruvian students enrolled in Brazil, while Ünal (2017) deals with foreign students in Turkey.

Other examples are Brooks and Waters (2010); Parr et al. (2000).

The paper provides new insights from the perspective of both statistical method and application. The statistical method developed in this work is a generalization of the exact test developed by Tumminello et al. (2011) to check for the hypothesis of random co-occurrences in a tripartite network[1]. The application is innovative because we apply a statistically validated network analysis to administrative micro-data of migration flows to investigate the possible existence of a chain migration in SM. The network analysis through the proposed method is to be intended as an exploratory tool and it does not allow to evaluate the simultaneous effects of exogenous factors which could explain the SM. Indeed, our main objective is to test, beyond any other reason of university attractiveness (Ciriaci, 2014), the possible existence of a chain migration (Brooks and Waters, 2010), in which students move to follow other students coming from the same area of origin. Our hypothesis is that flows of Sicilian moving students can significantly differ in the path "area of origin - University of destination", with respect to a null hypothesis of random flows. To define the area of origin, we consider an aggregation level of the Sicilian municipalities, based on some homogeneity criteria, thinner than the 9 Sicilian provinces. The areas we used are 38 clusters of the 390 Sicilian municipalities (Figure 1), aggregated by geographical proximity, and economic and commercial criteria according to D'Agostino and Ruffino (2005). In particular, these clusters were defined starting from 38 main municipalities, detected as gravitational areas, because their larger flows of commuting for work and study, with respect to other municipalities geographically close. Moreover, to enforce the homogeneity of the clusters, the authors hypothesize that the higher the economic and commercial levels, the larger the use of the Italian language instead of the local dialects. The definition criteria of these clusters are consistent with the existence of a social network of communication among people within each cluster. Through the above definition of areas of origin it may be possible to reduce the effect of exogenous factors such as the ones specified at economic and commercial levels. Of course the method does not allow to consider the individual characteristics of the students.

We rely upon micro-data of university first enrollment students and students' personal records for three cohorts of freshmen, the 2008/09, the 2011/12 and the 2014/15.

The data from each cohort is represented as a tripartite network with three sets of nodes, namely, clusters of Sicilian municipalities, students, and universities. The tripartite network is projected in a bipartite weighted network of clusters and universities, which is, then, filtered, in order to obtain a statistically validated bipartite network (SBVN), which represents a generalization of the method introduced by Tumminello et al. (2011), to deal with tripartite systems. Specifically, a directed edge from a Sicilian cluster to a university is set if the flow is statistically significant with respect to a null hypothesis of random flow of students, which takes into account the heterogeneity of both universities (total inflow) and clusters of municipalities (total outflow).

[1] The proposed method is equivalent to test that *integer* weights in a projected bipartite network occur by chance.

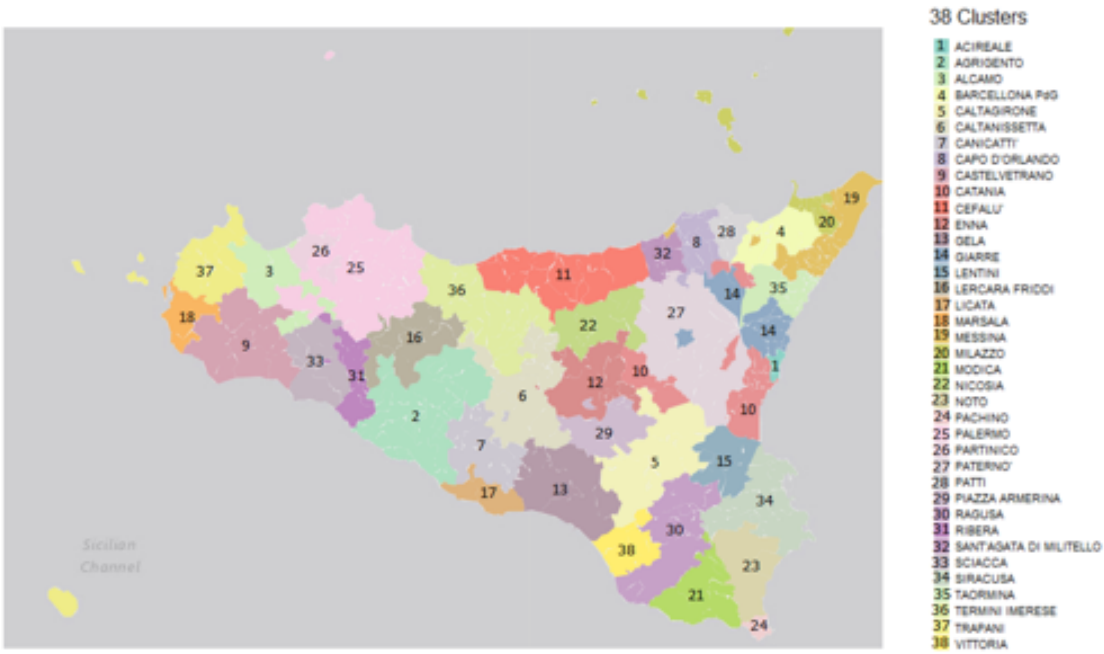

Figure 1: In figure the 38 clusters of Sicilian municipalities.

The structure of the paper is the follow. Section 2 includes a detailed description of the micro-data used throughout the paper; section 3 describes the network representation of the system and the statistical method used to analyse and prune the projected bipartite network of universities and clusters; section 4 includes empirical results from the network analysis of the system; finally, in section 5, we draw our conclusions.

2 Data description

The database MOBYSU.IT (2016) has been provided by the Italian Ministry of Education (MIUR) and it includes micro-level longitudinal information on university student careers from 2008 to 2017. The database also includes information on student's high school career, socio-demographics, and from the Bachelor to the Master degree that are not fully available in the publicly available data provided by the National Student Registry Office (ANS). Specifically, MOBYSU.IT database consists of about 200 to 300 variables per record, where each record is a student, and the total number of records is ranging from about 270.000 to about 295.000, depending on the cohort. According to the objective of the present study, information about students coming from Sicilian municipalities has been extracted from MOBYSUD.IT for the time period 2008-2014, including more than 26,000 records per cohort. In this work, we consider the cohorts of enrolled students as coming from a metapopulation, that is, a (numerable infinite) set of subjects, belonging to certain sub-populations, which are independent subsets of the common metapopulation. The subpopulations are renewed every year at the enrollment, according to a process that makes them independent of the previous and the following subpopulations, each one with structure and characteristics common to the metapopulation (representativeness), but, at the same time, carrying peculiar characteristics. Therefore, we assume that they are metasamples, which can be studied from a statistical point of view, by using inferential methods.

Table 1 shows that the percentage of females is grater than the percentage of males at both national and local level.

Table 1: Percentage of freshmen by Gender and Year

	Italy			Sicily		
	Years			Years		
Gender	2008	2011	2014	2008	2011	2014
F	57%	56%	55%	57%	57%	55%
M	43%	43%	44%	42%	42%	44%

Table 2 shows the distribution of students with respect to the high-school they come from. It turns out that there is no apparent difference between Sicilian students and students coming from other regions, in spite of the cohort, although, as expected, we notice relevant differences with respect to the gender of students.

Table 2: Distribution of Sicily freshmen and out of Sicily freshmen students by High-school and Year

		Sicily			All other regions		
		Year			Year		
Gender	Highschool	2008	2011	2014	2008	2011	2014
F	Other	25%	23%	23%	26%	27%	27%
	Classical	24%	28%	28%	18%	18%	18%
	Professional	5%	3%	4%	5%	5%	5%
	Scientific	31%	35%	36%	32%	34%	34%
	Technical	15%	11%	11%	19%	26%	16%
M	Other	5%	5%	5%	9%	9%	9%
	Classical	14%	14%	14%	9%	9%	9%
	Professional	4%	5%	5%	5%	5%	5%
	Scientific	42%	50%	50%	43%	48%	48%
	Technical	35%	26%	27%	35%	29%	29%

Before analyzing the flows from a cluster (source node) to a university (target node), we provide a descriptive statistics of outgoing students (outflow). Specifically, Table 3 shows an increasing number of outgoing students from Sicily over time, a trend that is the opposite of the one observed for the total number of newly enrolled students, i.e., freshmen, which, jointly, amplify the relative magnitude of the phenomenon.

Table 3: Total and outgoing sicilian-freshmen by level degree courses over time, in parenthesis the percentage of outgoing Sicilian students by degree courses and cohort

	Year					
	2008		2011		2014	
Degree	Total	Outgoing	Total	Outgoing	Total	Outgoing
Bachelor	21327	2997 (14.1%)	17210	4115 (23.9%)	16719	4682 (28%)
Master (5/6 years)	5264	675 (12.8%)	5791	1058 (18.3%)	5261	1087 (20.7%)
Total	**26591**	**3672 (13.8%)**	**23001**	**5173 (22.5%)**	**21980**	**5769 (26.2%)**

Furthermore, our analysis aims at distinguishing among 38 Sicilian territorial areas, which are internally homogeneous. This choice is a trade-off between clustering data using 9 provinces (too large and heterogeneous), and just referring to the 300 Sicilian municipalities (too small to analyze migratory effects). In fact, being the selected areas aggregated on the ground of socio-economics characteristics it is possible to evaluate such chain migration effects in spatio-temporal terms by applying a network analysis. According to D'Agostino and Ruffino (2005), we have classified the variable *residence city* into the 38 areas described by D'Agostino and Ruffino. creating a new variable that represents the homogenous residence area of the *i-th* student. In this framework the variables of interest have used in this analysis are the cluster of residence, and university of enrollment.

As it can be seen in Figure 2, a strong increase of the outgoing rate from Sicily is observed from the first cohort, 2008, (panel a) to the last one ,2014, (panel c). Indeed, the number of clusters with an outgoing rate of more than 20% increased a lot through the time window 2008–2014.

Figures 3, 4, and 5 are the heatmaps of Sicilian students flows from the 38 Sicilian clusters (by row) towards the off Sicily universities (by column). Colours in the figures are used to describe the fraction of outgoing students and to distinguish between geographical macro-regions. A straightforward comparison of the figures shows, especially for the last cohort, that such flows are more concentrated (colours orange and yellow) in Central-Northern universities (colours green and grey on the top of the plot). The clustering procedure reported in the left part of the plot, based on student flows, suggests the presence of homogeneous aggregations based on geographical proximity. For example, looking at the 2014 cohort, the cluster formed by Siracusa, Modica, Ragusa, Noto, Caltagirone, and Pachino can be easily interpreted along this line of thinking, since these territories are all located in the Southern-Eastern Sicily.

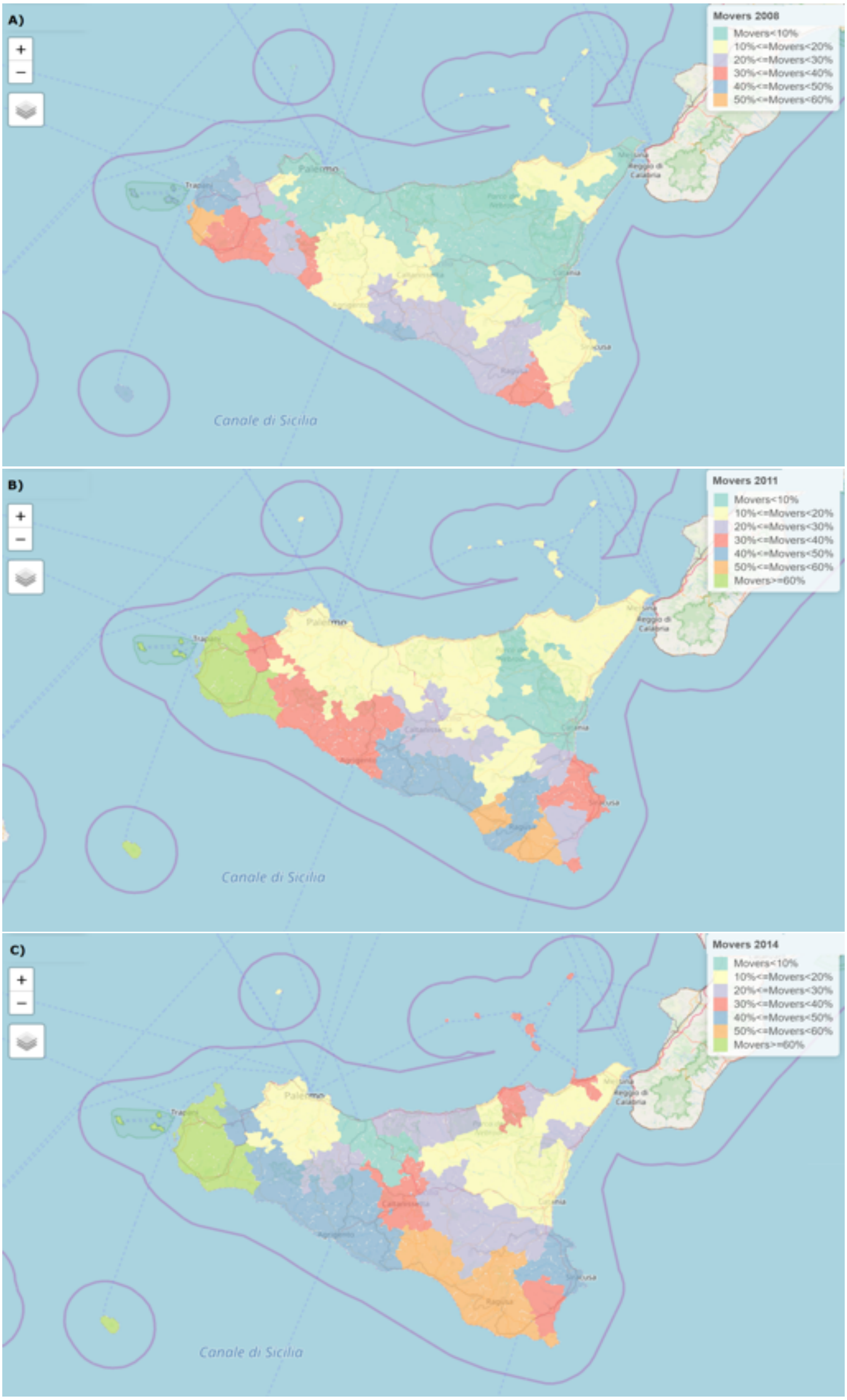

Figure 2: Panels *a*), *b*), and *c*) report the outgoing rates arranged in 7 levels by cohort per cluster for years 2008, 2011, 2014.

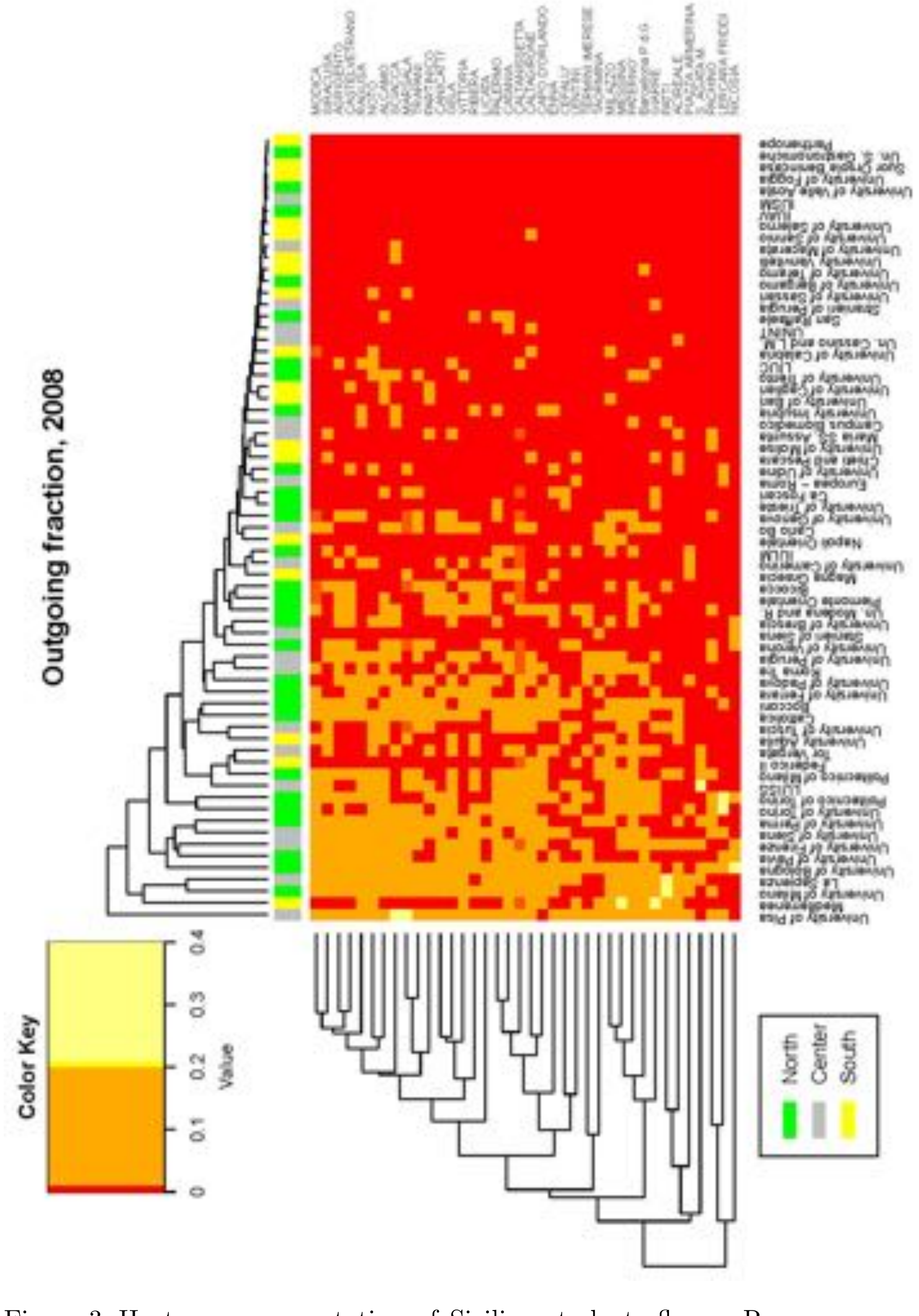

Figure 3: Heatmap representation of Sicilian students flows. Rows correspond to the 38 Sicilian territorial clusters of origin, while columns correspond to off Sicily universities of destination. The order of elements by row and by column is provided by complete linkage cluster analysis. Data correspond to year 2008. Heatmap colours, ranging between red and yellow, describe the fraction of outgoing students, while the other colours distinguish between geographical macro-regions.

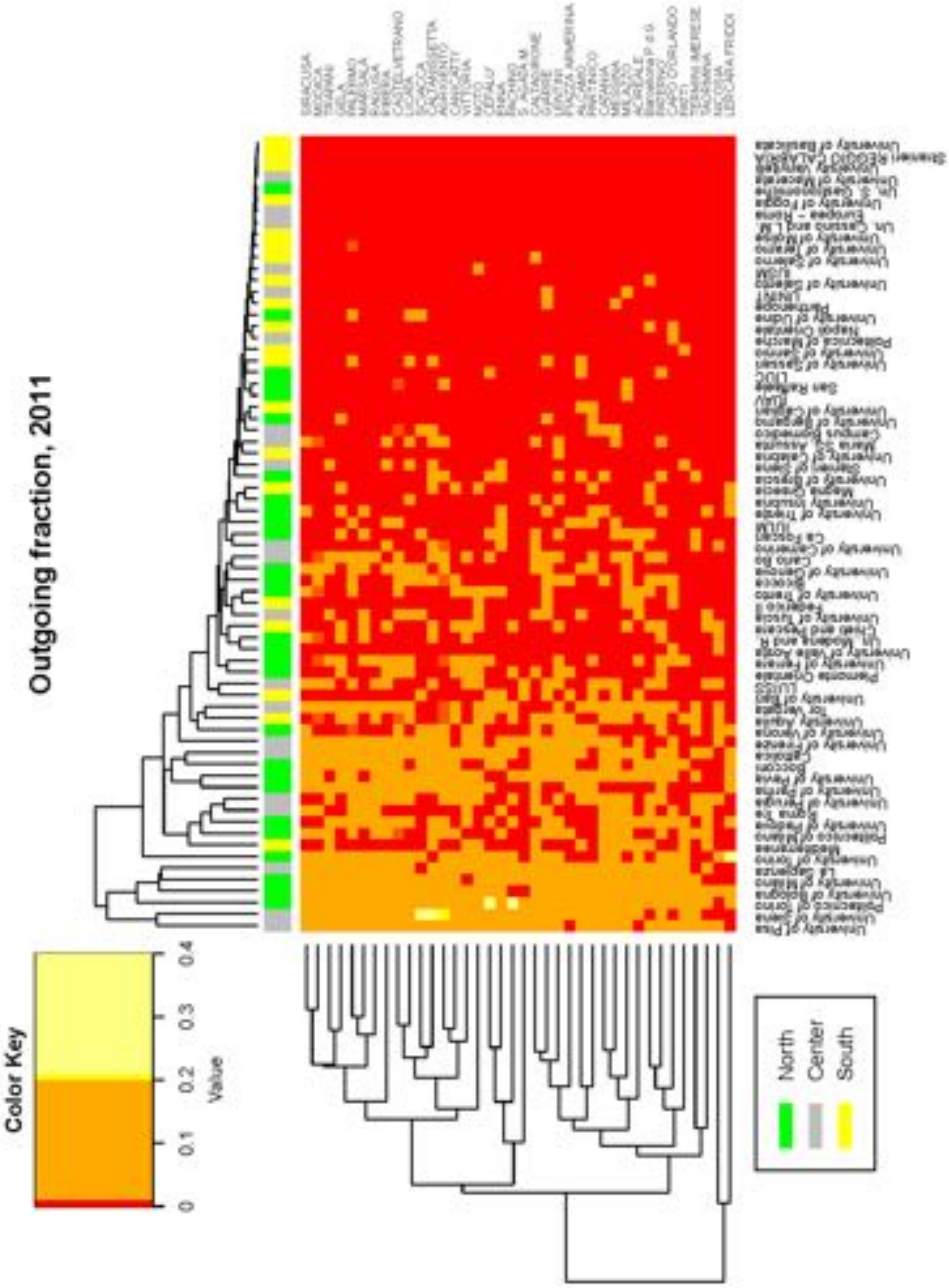

Figure 4: Heatmap representation of Sicilian students flows. Rows correspond to the 38 Sicilian territorial clusters of origin, while columns correspond to off Sicily universities of destination. The order of elements by row and by column is provided by complete linkage cluster analysis. Data correspond to year 2011. Heatmap colours, ranging between red and yellow, describe the fraction of outgoing students, while the other colours distinguish between geographical macro-regions.

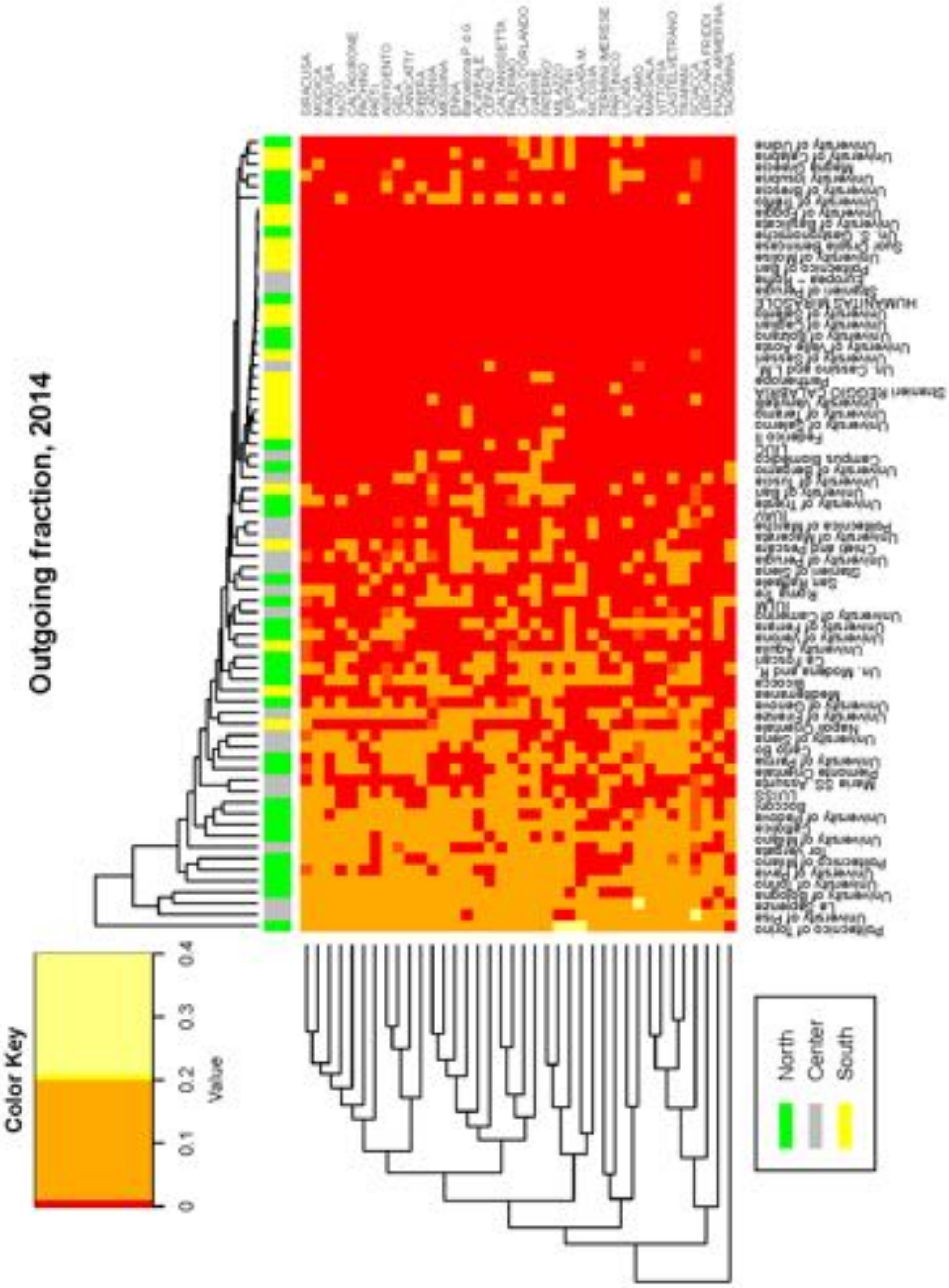

Figure 5: Heatmap representation of Sicilian students flows. Rows correspond to the 38 Sicilian territorial clusters of origin, while columns correspond to off Sicily universities of destination. The order of elements by row and by column is provided by complete linkage cluster analysis. Data correspond to year 2014. Heatmap colours, ranging between red and yellow, describe the fraction of outgoing students, while the other colours distinguish between geographical macro-regions.

3 Methods

Students' mobility system displays several typical features of complex systems, such as, interconnectedness of elements (which is at the basis of the chain migration hypothesis), heterogeneity of elements, that is, heterogeneity of the clusters of municipalities according to number of outgoing students, of the universities, according to the number of enrolled students, heteroscedasticity, modular structure of elements (e.g., of universities), non stationarity, and tipping points (reflecting, for instance, an economic downturn) (Bar-Yam, 1997). Such features make the system very difficult to analyze by using traditional means based on system's decomposition and reductionism (Simon, 1996; Dekker, 2011), and taking a weak-holistic approach (Simon, 1996), such as the one provided by network theory (Newman, 2011), seems to be more appropriate. The most natural network representation of the mobility system is obtained by considering a tripartite network associated with each cohort, where three sets of nodes, namely, clusters of Sicilian municipalities, students, and universities, are distinguished. In this network, a link between any two elements of the same set cannot occur. Specifically, in the tripartite mobility network, a link is set between a student and a university if the student is enrolled in that university, and between a student and a cluster of municipalities if the student comes from a municipality that belongs to that cluster.[2]

The objective of the present paper is to understand whether or not student mobility might be driven, in part, by a chain migration. Along this line, we are interested in eliciting from data the presence of an *excess* of flow from one cluster of municipalities to a university, with respect to a null hypothesis of random flow that takes into account the heterogeneity of both clusters and universities. Such an analysis can be done by considering a straightforward generalization of the Statistically Validated Network method to tripartite networks (Tumminello et al., 2011). Specifically, we adapt this technique to construct a Statistically Validated Bipartite Network (SBVN) of clusters and universities, as detailed in the next section.

3.1 Statistically validated bipartite networks

Starting from the tripartite network of clusters, students, and universities, a bipartite weighted network is obtained by projecting the set of students onto the other sets– see Figure 6. Specifically, a link between a cluster of municipalities and a university is set if at least one student is linked to both nodes in the original tripartite network. The weight of such link in the bipartite weighted network corresponds to the number of students linked to both nodes in the tripartite network (Newman, 2011; Tumminello et al., 2011).

[2]It is worth noticing that the structure of the network is such that no link can be set between a university and a cluster of municipalities (see Fig. 6). This fact implies that the whole information contained in the tripartite network can actually be mapped in a bipartite network with weighted links, that is, a network with only two sets of nodes, clusters and universities, and weighted links connecting elements of the two sets, the weight being equal to the number of students coming from a cluster and enrolled in a university. However, such a simpler representation, would make it harder to describe the configurational model underlying the null hypothesis of random flows described later in the paper. Therefore, we prefer to use the tripartite representation of the system.

Figure 6 shows the original tripartite network (Phase 1), where intermediate nodes between the 38 Sicilian clusters and the 80 universities are the n_q, $q = 1, 2, 3$, outgoing students[3] in a given cohort under analysis, as well as the corresponding weighted bipartite network between clusters and universities, which is obtained as a projection of the original tripartite network. As discussed above, our objective is to determine the presence in the data of an excess flow between a municipality and a university, with respect to a null hypothesis of random flow in a system with a double heterogeneity, which could be interpreted as a mark of the existence of a chain migration in the mobility system. Such an excess of flow can be revealed by performing the following analysis.

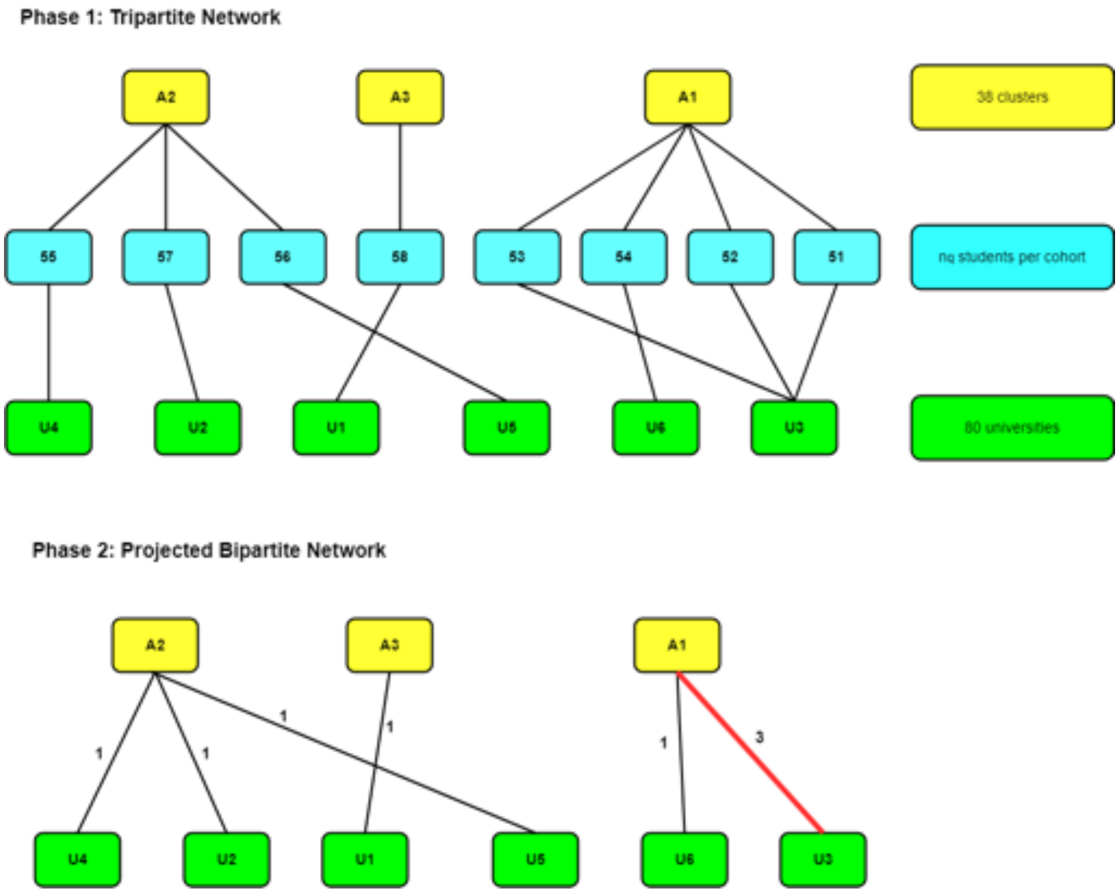

Figure 6: The applied projection of the tripartite students' mobility network to the bipartite one, red line shows a statistically validated link and the thickness the intensity of the flow

[3] $n_1 = 3672$ in 2008, $n_2 = 5173$ in 2011, $n_3 = 5769$ in 2014. See Table 3

3.2 Statistically validated networks: construction

To validate links in the (projected) weighted bipartite network of clusters and universities, the method proposed by Tumminello et al. has been generalized to deal with tripartite systems. In its most general setting, this approach requires the validation of each link of the projected network, against the null hypothesis of random co-occurrence of shared neighbor nodes (students in this case). The null hypothesis of random connections between nodes with assigned degree has been proposed in the seminal paper by Xulvi-Brunet and Sokolov (2004) in terms of a configurational model, which is also able to reproduce assortative mixing. However, in our case, we consider a null hypothesis that does not assume the presence of such an effect. Therefore, the configurational model, where links (of a real network) are iteratively and randomly selected in pairs and swapped is assumed. Accordingly, the co-occurrence of first neighbors in either set can be analytically described through the hypergeometric distribution. In the original formulation of the method, subsets of projecting nodes (students in this case) were considered, in order to deal with the heterogeneity of node degree in the projecting set. However, here, such an heterogeneity is missing, since students only have degree equal to two in the original tripartite network, where one link connects a student to a cluster of municipalities and one link connects her to a university. Therefore, to test the statistical significance of an edge weight in the bipartite weighted network for a given cohort against a null hypothesis (H_0) of random flow, we consider the following distribution of the weight X of a link in the (randomly rewired) projected network that describes hypothesis H_0:

$$H(X|N_A, N_B, N) = \frac{\binom{N_A}{X}\binom{N-N_A}{N_B-X}}{\binom{N}{N_B}}, \tag{1}$$

where:

- N is the total number of students in the tripartite network,
- N_A is the number of students coming from municipality A,
- N_B is the number of students enrolled in university B,
- X is the number of co-occurrences of A and B in the (random) tripartite network associated with H_0, i.e., it is the variable that describes the number of students *flowing* from cluster A to university B.

Indeed, probability (1) can be obtained by a straightforward calculation of the number of "favorable events" divided by the total number of possible events. Specifically, the result can be obtained by assuming that university B randomly selects N_B students (which allows to keep information about the heterogeneity of universities in the null hypothesis) from the overall N students in the population, N_A of which coming from cluster A (which allows to keep information about the heterogeneity of clusters of municipalities in the null hypothesis) and $N - N_A$ from the other clusters. Therefore, if X represents the number of students picked by university B that come from cluster A, then the number of favorable cases is given by the product of the number of ways in which N_A students can be combined in groups of size X, that is, $\binom{N_A}{X}$, times the number of ways in which the other $N_B - X$ students can be drawn from the set of students coming from the other clusters, that is, $\binom{N-N_A}{N_B-X}$. Finally, the probability (1) is obtained by noticing that all of the possible ways in which N_B students can be drawn from the overall set of students, with size N, is $\binom{N}{N_B}$.

The hypergeometric distribution described in (1) is then used to associate a *p-value* with each pair cluster-municipality, say j and k, respectively, which are connected in the (projected) weighted bipartite network and have a weight (number of outgoing students) $n_{j,k}$:

$$p_{j,k}(n_{j,k}) = \sum_{i=n_{j,k}}^{min(N_A,N_B)} \frac{\binom{N_A}{i}\binom{N-N_A}{N_B-i}}{\binom{N}{N_B}}. \tag{2}$$

Since the structure of the bipartite weighted network allows one to represent it as a two-way contingency table, one may argue about the advantage of using Eq. (2) to associate a *p-value* with each link, instead of the one that could be provided by the standardized residuals of a χ^2 distribution. Actually, the proposed test is exact and, therefore, on the one hand it allows one to better deal with "extreme values" (right tail of the distribution), and, on the other hand, it works even with small samples. The *p-value* calculation reported in equation (2) should be repeated for each link in the weighted bipartite network. Therefore, multiple-test corrections on the threshold of statistical significance of a *p-value* should be considered. The most conservative correction with respect to the family-wise error rate (FWER) is the Bonferroni correction (Miller, 1981), which also applies to the case of dependent tests, whereas, a less restrictive correction is the False Discovery Rate (FDR) (Benjamini and Hochberg, 1995). To say if a connection between two nodes is statistically significant the threshold on the *p-value*, according to Bonferroni, is

$$S_{Bon} = \frac{\alpha}{T}, \tag{3}$$

while, according to the False Discovering Rate, it is

$$S_{FDR} = \frac{\alpha k}{T}, \tag{4}$$

where α is the univariate significance level (e.g., 0.01), T is the number of tests, i.e., the number of links in the weighted bipartite network, and k is the rank of the largest tested *p-value* such that $p_{k_{max}} < k_{max}\alpha/T$. Accordingly, being the Bonferroni correction more conservative than the FDR one, if a link is validated according to Bonferroni, it is also validated according to FDR, whereas, of course, the vice versa does not hold true.

4 Results

Figures 7, 8, and 9 report the bipartite validated networks for Sicilian outgoing students by cohort. There are two types of nodes: the first type corresponds to universities (red circles), with node size proportional to the number of enrolled students from Sicily, while the second one corresponds to clusters of municipalities (cyan rectangles), with size proportional to the number of outgoing students. The solid line is used to represent links that belong to the *Bonferroni* network (and therefore to the FDR too), while the dashed line is used to represent links that only belong to the FDR network. Moreover, the thickness of a link is proportional to the number of outgoing students from a cluster to a given university. The network representation is obtained through the force directed layout method with weight proportional to the binary Pearson's correlation coefficient between a university and a cluster, which implies that the longer the link, the lower the correlation between two nodes is.

According to both Bonferroni and FDR networks, the most attractive universities are the University of Bologna, Milan, Turin, Florence, Pisa, La Sapienza, and Cattolica. Moreover, we observe that the number of (statistically significant) links increases overtime. To better frame this phenomenon from a quantitative a point of view, we include a table with the number of links and nodes for each cohort. Table 4 shows that the number of both nodes and links involved in the statistically validated networks increases between 2008 and 2014. Both the apparent structure of the networks and the increasing number of links over time can be interpreted within the framework of complex systems, by looking at chain migration as reflecting the relevance of private information in the decision process preliminary to the selection of a university by students, and the increasing relevance of such an information with respect to public information. Indeed, in an efficient purely competitive system, where all of the agents share and (rationally) process the same (public) information (Easley and Kleinberg, 2010), the networks reported in Fig.s 7 through 9 should be empty: preferential patterns of mobility shouldn't be observed, since the null hypothesis exactly takes into account the heterogeneity of clusters of municipalities and universities. Such a consideration implies that private information flow acts as a positive feedback mechanism in the system: students from cluster A enroll in university B at time t, then they share their (*positive*) experience, not only concerning the university, but also the city, etc., with their friends at home (information flow from B to A), which, in turn, reinforces mobility from A to B at time $t+1$. Such an information flow can also act in reverse (negative feedback) and break a preferential connection: students from cluster A enroll in university B at time t (preferential pattern from A to B at time t), then they share their (*negative*) experience with their friends at home, discouraging them from enrolling in university B at time $t+1$, which, as a result, tends to cancel out the preferential pattern of mobility from A to B. The fact that the number of validated links in the mobility network increases over time should therefore be interpreted as reflecting the increasing importance of private information to determine the structure of the mobility system, at the expense of public information.

A question that remains yet to be answered concerns the imbalance between the relative influence of the two feedback mechanisms discussed above. If only a positive feedback mechanism was effective to orient the decision of students to enroll in a certain university, then the statistically validated network associated with the cohort of 2014 should properly contain the statistically validated network associated with the cohort of 2008. Therefore, an indicator to look at is the proportion of links from the 2008 statistically validated networks that also belong to the 2014 corresponding networks. Looking at the Bonferroni network, this proportion is 41%, while it is 44% for the FDR network. This result indicates that *old* mobility patterns are destroyed, which suggests that the aforementioned negative feedback mechanism is actually influencing the evolution of the mobility system. On the other hand, given the overall increase of mobility patterns, we can conclude that the role played by the positive feedback mechanism is even bigger, likely driven also by exogenous factors, such as the 2011 sovereign crisis, and the subsequent austerity policy introduced and enforced in 2012 and 2013, which affected differently the South and the North of the country, especially with respect to the labour market, which should have favored the formation of new mobility patterns from the South to the North of the Country.

Finally, to make the discussion about the evolution of mobility system more comprehensive, and to address the question of whether gender-specific patterns occur or not, we look at the similarity between the networks that are separately obtained for male and female students over time. As discussed already, an increasing similarity between networks for subsequent cohorts implies that migration patterns are settling over time, whereas a decreasing trend might suggest a progressively reducing impact of chain migration on students' mobility, with respect to the negative feedback mechanism discussed in the previous paragraphs and other influential factors, such as marketing campaigns, and, broadly speaking, the varying attractiveness of "specific" universities. Here we focus on the difference between gender specific patterns of mobility over time. Similarity between networks can be easily evaluated by associating with each one of the potential links ($N_u \times N_c$, where N_u is the number of universities and N_c the number of Sicilian clusters) a binary variable that takes value 1 if the link is set in the network and 0 otherwise.

Table 4: Number of Links and Nodes per cohort by female, male, and overall.

	Links			Nodes		
Cohorts	Female	Male	Overall	Female	Male	Overall
2008	62	55	95	46	45	52
2011	80	69	110	47	48	54
2014	72	81	123	48	52	62

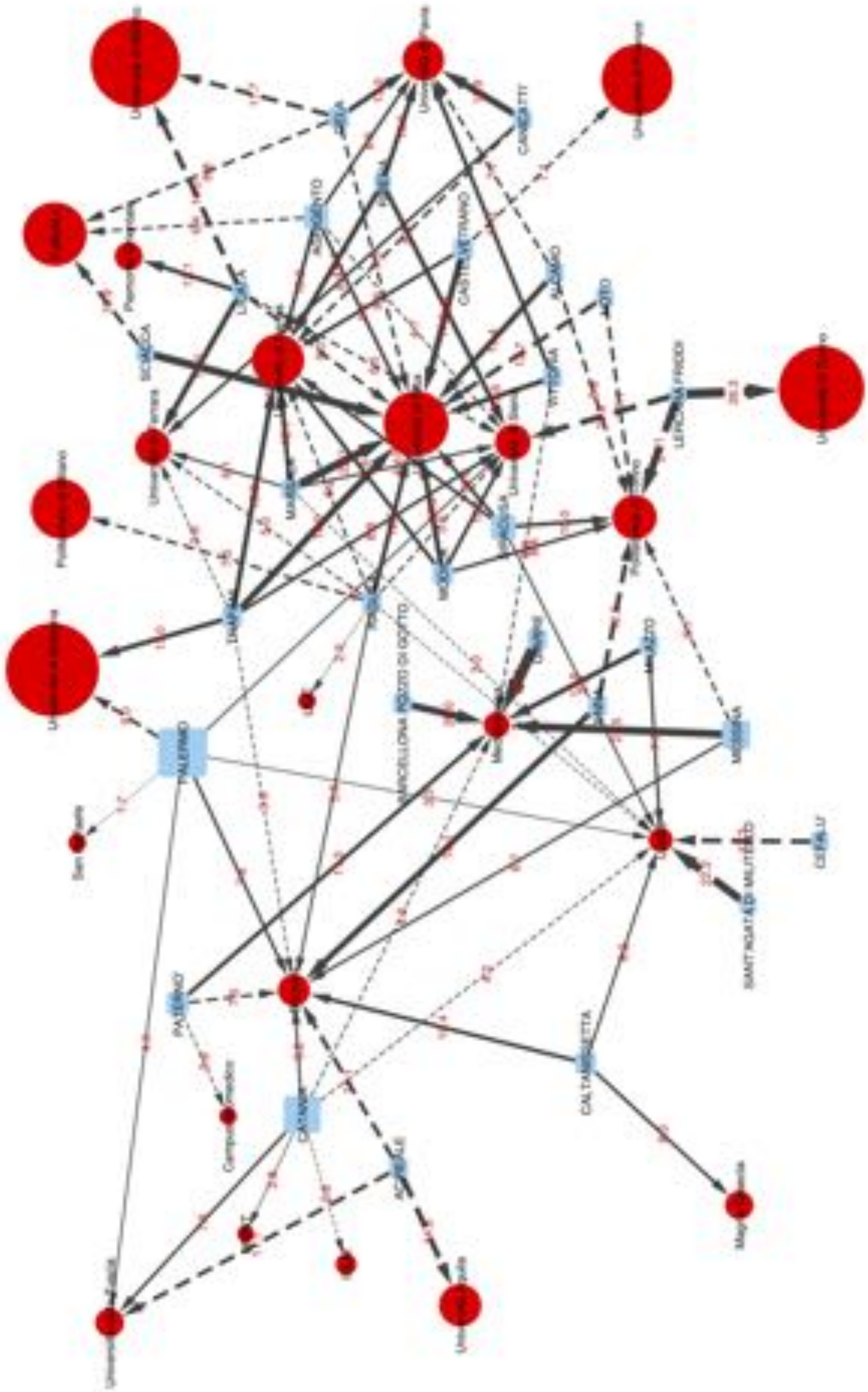

Figure 7: Bipartite validated network for Sicilian outgoing students, year 2008

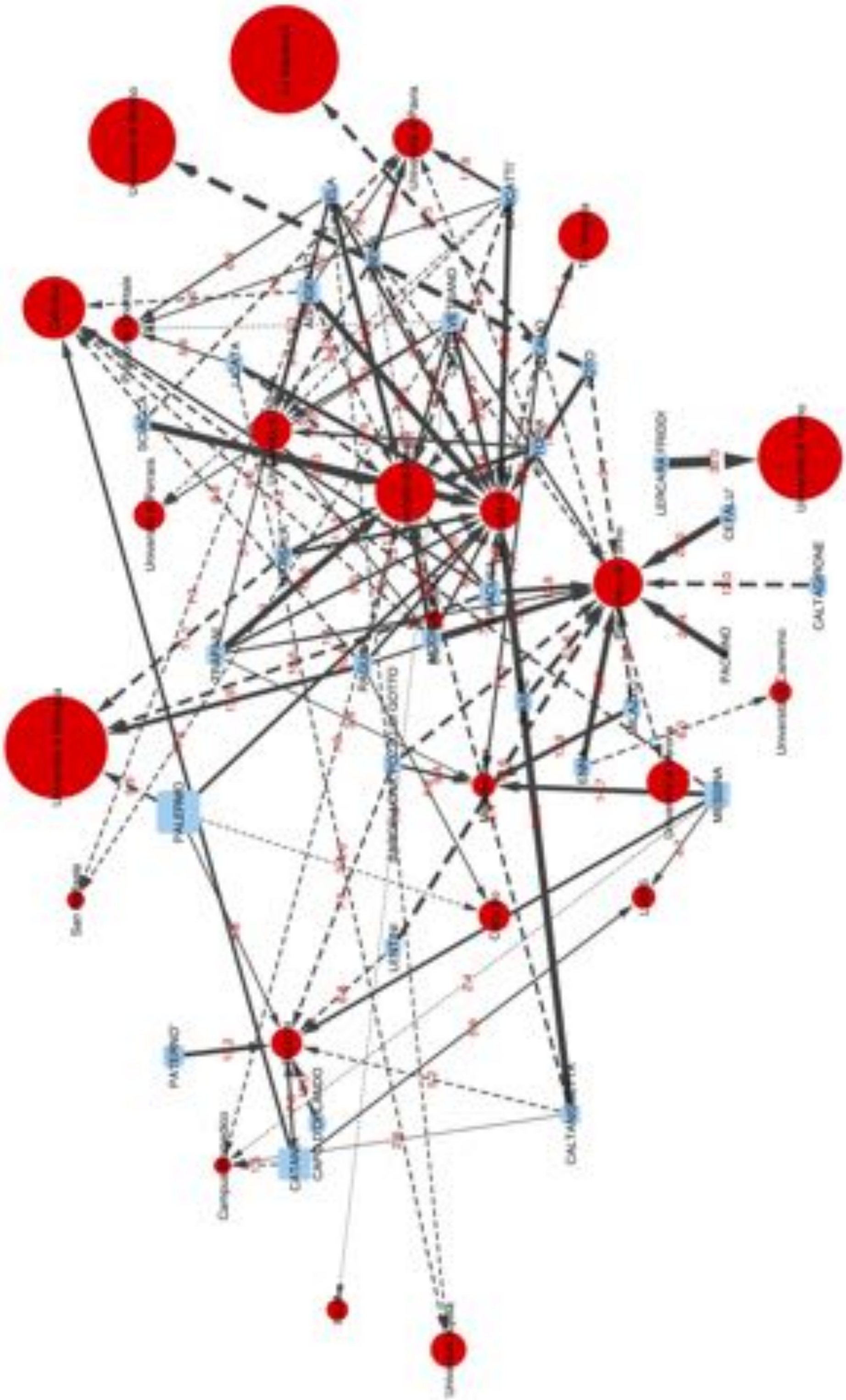

Figure 8: Bipartite validated network for Sicilian outgoing students, year 2011

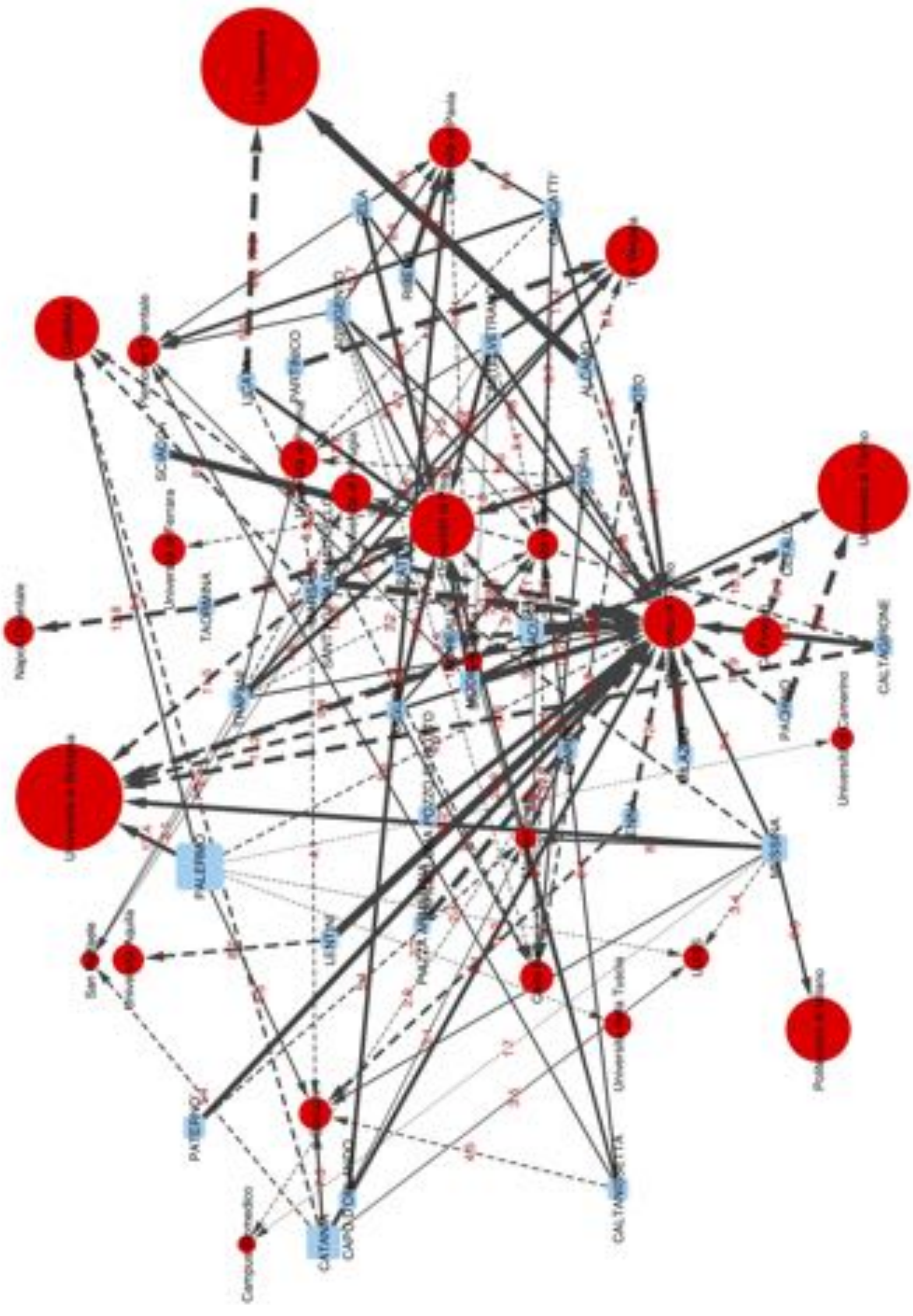

Figure 9: Bipartite validated network for Sicilian outgoing students, year 2014

Here, we use the Matthews correlation coefficient (MCC) (Yingbo et al., 2005) to evaluate the similarity between two networks, since it properly takes into account the heterogeneity of both clusters and universities, and it is symmetric with respect to the exchange of the networks under comparison. The Matthews correlation coefficients for male and female Bonferroni (FDR) networks is on average 0.47 (0.46), and stable over time, which indicates that the relative frequency of gender specific patterns is rather stable throughout the investigated time window.

In summary, these results support the hypothesis that the impact of austerity policies, introduced in 2012 and 2013 in Italy, favoured a change of migration patterns in 2014, by determining the creation of new patterns from Sicily to universities in the North (increasing trend of the number of preferential patterns), but it did not affect the imbalance of gender specific preferences.

Within the framework of oriented networks, the degree of a university indicates the "popularity" of the university in terms of number of Sicilian clusters displaying a preferential pattern pointing to it. Accordingly, we decided to calculate the degree ranking of universities as reported in Table 5. The order of universities in the Table is based on the ranking of the median degree rank (reported in parenthesis in the Table) across all of the networks reported in the Table. Universities that showed a degree equal to zero in at least one network are not displayed in the Table.

Looking at Table 5 is clear that University of Pisa, on average, is the big favourite, but decomposing its degree by gender, it's clear that University of Siena is the most chosen university for female students overtime. In light of these results, we focused our case-study network analysis on University of Siena.

Table 5: Degree ranking of universities by cohort in the FDR networks, for female, male, and overall. In parenthesis the conditional ranking (see text for details on the selection of listed universities).

		2008			2011			2014		
Ranking	Universities	Female	Male	All	Female	Male	All	Female	Male	All
1	University of Pisa	9 (2)	9 (1)	13 (1)	13 (2)	7 (3)	14 (3)	9 (2)	13 (2)	13 (2)
2	University of Siena	10 (1)	2 (8)	9 (3)	17 (1)	15 (1)	17 (1)	3 (11)	8 (3)	7 (4)
3	Polytechnic University of Turin	1 (15)	3 (7)	7 (7)	7 (4)	11 (2)	15 (2)	10 (1)	23 (1)	28 (1)
4	Bocconi	3 (8)	5 (3)	9 (5)	6 (5)	4 (5)	8 (5)	5 (3)	4 (5)	6 (7)
5	University of Pavia	4 (7)	4 (5)	6 (8)	4 (8)	4 (6)	5 (8)	5 (4)	5 (4)	7 (5)
6	University of Parma	9 (3)	6 (2)	9 (4)	8 (3)	6 (4)	9 (4)	2 (12)	1 (16)	5 (8)
7	Mediterranea	6 (5)	5 (4)	7 (6)	3 (9)	3 (9)	5 (7)	4 (7)	3 (9)	4 (11)
8	University of Bologna	1 (10)	1 (14)	2 (12)	2 (10)	1 (12)	4 (9)	4 (5)	1 (13)	7 (3)
9	LUISS	7 (4)	2 (12)	9 (2)	1 (18)	2 (10)	2 (15)	1 (20)	3 (10)	3 (14)

Figures from 10 to 12 show the statistically validated subnetworks that include all of the Sicilian clusters pointing to University of Siena. Regardless the gender, in 2008 the Sicilian cluster involved in the flow to Siena are 9, where the clusters of Lercara Friddi, Ribera, and Modica have the greater flow. In 2011, clusters involved in the choice of Siena University increase rapidly–from 9 clusters observed in 2008, they become 17 in

2011 (for all students). Such an effect, however, slowed down in 2014, where the clusters involved for all the students are 7. Looking at the network links according to gender, the number of clusters from where female students have moved to the university of Siena increase from 10 in 2008 to 17 in 2011, and it has been the most favorite destination for female in 2011, but this growth slowed down in 2014 for both gender. As it can be seen in Table 5, Polytechnic University of Turin moves over from the 7^{th} position in 2008 to the 1^{st} position in 2014.

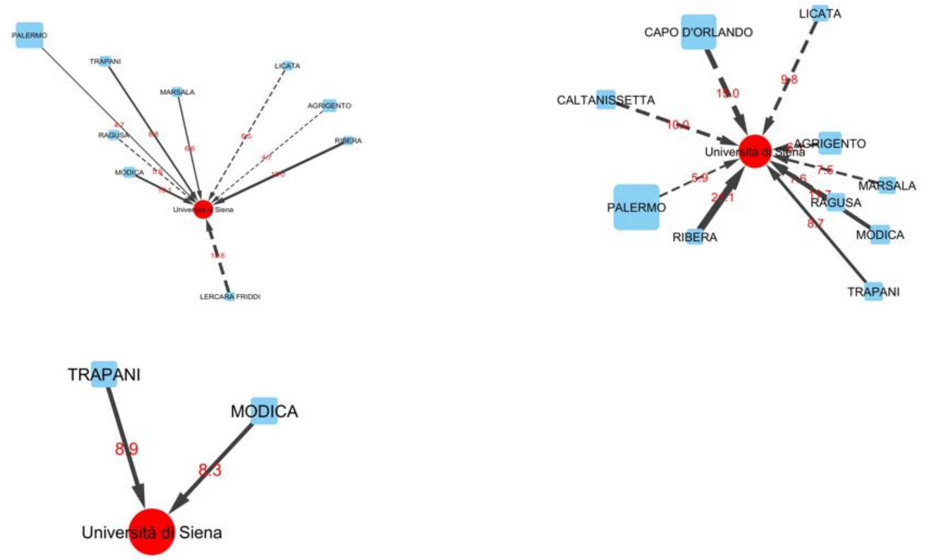

Figure 10: Bipartite validated network for University of Siena, focus on Sicilian outgoing students (left top), female students (top right), and male students (bottom left), year 2008

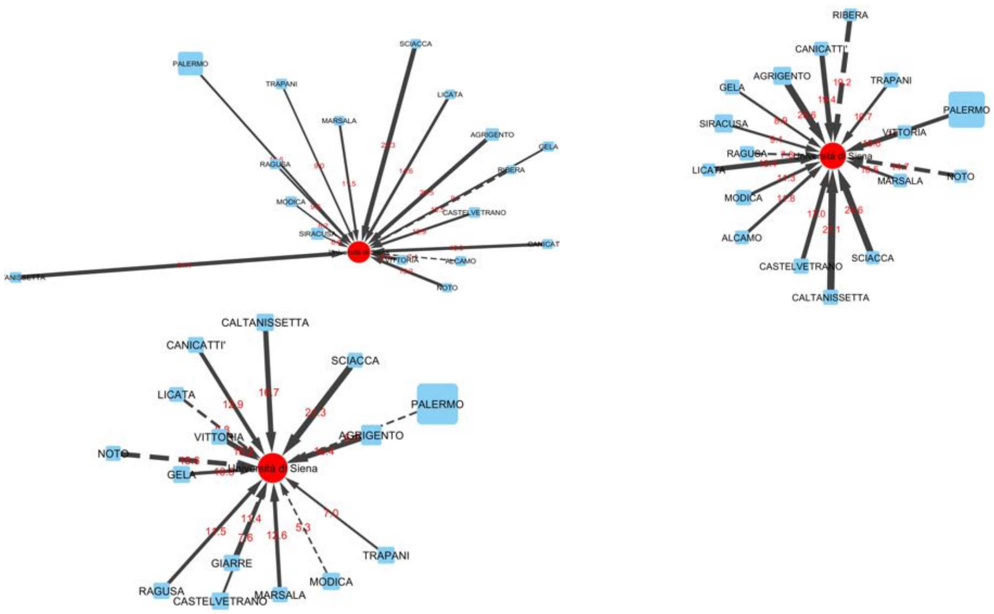

Figure 11: Bipartite validated network for University of Siena, focus on Sicilian outgoing students (left top), female students (top right), and male students (bottom left), year 2011

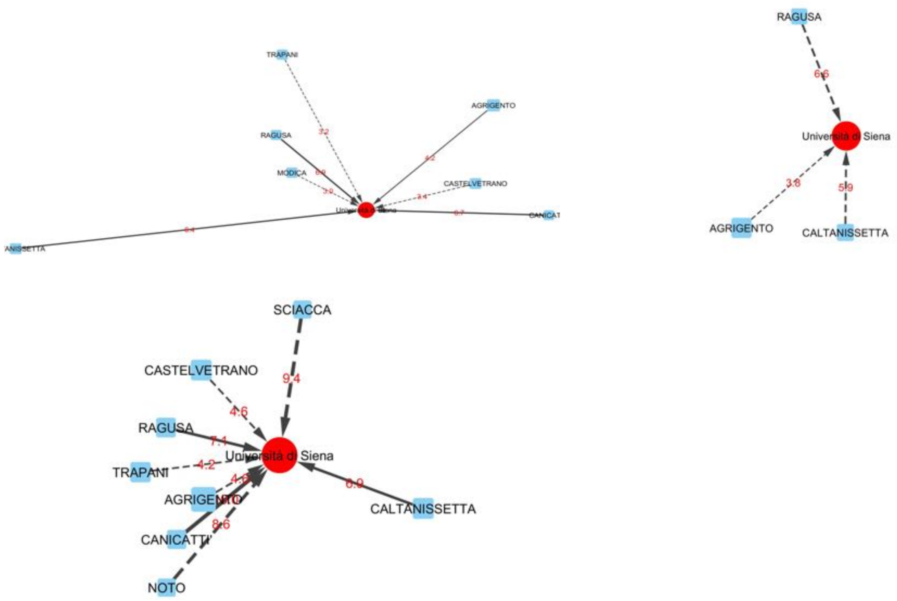

Figure 12: Bipartite validated network for University of Siena, focus on Sicilian outgoing students (left top), female students (top right), and male students (bottom left), year 2014

Finally, it's worth to note that contiguous clusters might display similar (statistically significant) patterns of mobility, likely due to the presence of exogenous factors influencing specific mobility routes. Indeed, the considered null hypothesis, which is modeled through the hypergeometric distribution, allows one to indirectly take into account the presence of factors influencing the nodes–e.g., the labour market of the cluster of origin, and the attractiveness of specific universities–through the parameters N_A (the number of students coming from municipality A) and N_B (the number of students enrolled in university B). However, it is possible that some cluster-university link is favored (or disfavored) by factors affecting the connection itself, instead of the connected nodes, factors such as the transport system. For instance, let's consider the case of the contiguous clusters of Marsala and Trapani. People from these clusters live close to the same airport, namely, Birgi airport. Looking at the connections of these clusters in the Statistically Validated Bipartite Network (SVBN) obtained by using the FDR correction, it turns out Trapani is linked to $n_T = 6$ universities, Marsala to $n_M = 5$, and $n_{TM} = 4$ destinations are in common, in 2008. Considering that, overall, there are 78 universities in the system, one can apply the original method developed by Tumminello et al. (2011) to the (bipartite) SVBN and calculate the probability to observe a value of co-occurrence at least as extreme as the observed one ($n_{TM} = 4$) is 0.00005, according to the hypergeometric distribution. Similar results are obtained, by looking at the cohorts of 2011 and 2014. Specifically $n_T = 5$, $n_M = 4$ and $n_{TM} = 3$, in 2011, with an associated p-value of 0.0005, whereas, $n_T = 7$, $n_M = 6$ and $n_{TM} = 3$, in 2014, determining a p-value of 0.007[4]. This results suggest the presence of common factors influencing the mobility of students from the two clusters. However, we believe that such an analysis deserves more attention and it will be considered for future research.

5 Conclusions

The descriptive statistics of mobility data reveals an increase of the movers from Sicily towards the central and northern Italian universities over time. In fact, it has been observed an increase of both the rate of outgoing students from Sicily and the number of clusters with an outgoing rate of at least 30%, which, initially (2008's cohort), were mainly located in the western clusters of Sicily, whereas, at the end (2014's cohort), they were uniformly distributed over the whole Center and South of Sicily, from the western clusters to the eastern ones.

The network analysis of preferential patterns of mobility has shown that, among universities of central Italy, Pisa and Siena are (on average) the most attractive ones for students coming from western and southern Sicily, and that Polytechnic University of Turin is the favorite northern university for the last cohort of students (2014).

In addition, the network analysis provides some elements to support the existence of migratory chains, over time, by means of the increase of the number of:

- new links, which become significant in the networks of the most recent cohorts

[4] The total number of considered universities in the system is 80 in 2014.

and were not in the previous ones, and their thickness which also increased in the networks of the last cohort;

- the new nodes (and their size) that appeared in the networks of the last cohort.

Specifically, it has been observed and increase of the number of clusters with movers and the number of off-Sicily universities chosen by the movers (preferential patterns). At the same time, the number of both movers from some clusters and students towards a given off-Sicily universities (the thickness of their link–outgoing flows), as well as the number of movers from Sicily who enrolled in off-Sicily universities increased.

By investigating the similarity between the networks according to the sex of movers (through the MCC), we observe a stability of similarity over time. This result implies that, even if mobility patterns vary over time, such a variation occurs by keeping almost constant the similarity between gender specific patterns. The case of Siena also suggests that gender specific patterns may both vary over time in a non-monotonous way (pick in 2011).The proportion of preferential links that are preserved over time slightly exceeds 40%. This result indicates that some *old* patterns of mobility are destroyed, while new ones appear, especially in the networks of the last cohort as a possible drawback on the financial crisis, which affected differently the South and the North of the country, especially with respect to the labour market, and favoured the formation of new mobility patterns from the South to the North of the Country. Finally, the observed 40% stable links, however, implies that the role played by chain migration in explaining mobility patterns of students is rather important.

Finally, it is worth to include an assessment of the strengths and limitations of the present study, in relation to the statistical methods used to highlight chain-migration effects on students' mobility. The null hypothesis considered in the paper involves a probability distribution, the hypergeometric distribution, which is conditioned to both the total number of students moving from each cluster and the total number of students enrolling in each university. Therefore, implicitly, such a null hypothesis incorporates information about the specific attractiveness of universities, as well as cluster-specific factors that might influence the decision of students to enroll in a university out of the region. So, the considered null hypothesis appears to be appropriate to reveal chain-migration effects, since it allows one to untangle chain migration from factors mainly related to node-specific characteristics. Furthermore, the considered null hypothesis properly takes into account the heterogeneity of clusters, in terms of number of moving students, as well as the one of universities, in terms of number of enrolling students. However, factors associated with a specific cluster-university link, such as an airport that directly connects a cluster to a specific destination, are not incorporated in the null hypothesis, neither directly or indirectly. In the last paragraph of the discussion section, we have provided an example of the impact of such link-specific factors, by showing how they may determine a significant similarity between the patterns of mobility from two contiguous clusters. The example of Trapani and Marsala clusters suggests that, based on the present analysis, we cannot claim that chain-migration is the only factor determining the revealed preferential patterns of mobility, since other link-specific factors (e.g., the transport network) might also come into play. A possibility to deal with such an issue would be to use the transport network to group together clusters with a similar connectivity, and perform the analysis of preferential patterns separately for each group of clusters. Another possibility might be to use the Wallenius non-central hypergeometric distribution, by assigning a specific weight to groups of students, sorted according to the groups of clusters of origin and groups of destinations with similar connectivity in the transport network (Puccio et al., 2019). Such an analysis left for future research.

Acknowledgement

This paper 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. 2017HBTK5P.

References

Attanasio, M. and Enea, M. (2019). La mobilità degli studenti universitari nell'ultimo decennio in italia. In De Santis, G., Pirani, E., and Porcu, M., editors, *Rapporto sulla popolazione. L'istruzione in Italia*, pages 43–58. Bologna, Il Mulino.

Bar-Yam, Y. (1997). *Dynamics of complex systems.* Reading, Massachussetts, Addison-Wesley.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Brooks, R. and Waters, J. (2010). Social networks and educational mobility: the experiences of UK students. *Globalisation, Societies and Education*, 8(1):143–157.

Bruno, G. and Genovese, A. (2012). A spatial interaction model for the representation of the mobility of university students on the italian territory. *Networks and Spatial Economics*, 12(1):41–57.

Ciriaci, D. (2014). Does university quality influence the interregional mobility of students and graduates? the case of italy. *Regional Studies*, 48(10):1592–1608.

D'Agostino, A., Ghellini, G., and Longobardi, S. (2019). Out-migration of university enrolment: the mobility behaviour of italian students. *International Journal of Manpower*, 40(1):56–72.

D'Agostino, M. and Ruffino, G. (2005). *I rilevamenti sociovariazionali – Linee Progettuali. Atlante Linguistico della Sicilia.* Palermo, Centro studi filologici e linguistici siciliani. Dipartimento di Scienze Filologiche e Linguistiche, University of Palermo.

Dal Bianco, A., Spairani, A., and Ricciari, V. (2010). La mobilità degli studenti in italia: un'analisi empirica. *Rivista di Economia e Statistica del Territorio*, 1(1):123–143.

Daniel, C. (2014). Building a south-south connection through higher education: the case of peruvian university students in brazil. *Cahiers de la Recherche sur l'éducation et les savoirs*, (13):119–137.

Dekker, S. (2011). *Drift into failure. From hunting broken to understanding complex systems.* Farnham, Ashgate.

Dotti, N., Fratesi, U., Lenzi, C., and Percoco, M. (2013). Local labour markets and the interregional mobility of italian university students. *Spatial Economic Analysis*, 8(4):443–468.

Dotti, N., Fratesi, U., Lenzi, C., and Percoco, M. (2014). Local labour market conditions and the spatial mobility of science and technology university students: evidence from italy. *Review of Regional Research*, 34(2):119–137.

Easley, D. and Kleinberg, J. (2010). *Networks, Crowds, and Markets.* Cambridge, Cambridge University Press.

Enea, M. (2018). From south to north? mobility of southern italian students at the transition from the first to the second level university degree. In Perna, C., Pratesi, M., and Ruiz-Gazen, A., editors, *Studies in Theoretical and Applied Statistics*, pages 239–249, Cham. Springer International Publishing.

Giambona, F., Porcu, M., and Sulis, I. (2017). Students mobility: assessing the determinants of attractiveness across competing territorial areas. *Social Indicators Research*, 133(3):1105–1132.

Haug, S. (2008). Migration networks and migration decision-making. *Journal of Ethnic and Migration Studies*, 34(4):585–605.

Lupi, C. and Ordine, P. (2009). Family income and students' mobility. *Giornale degli Economisti e Annali di Economia*, 68 (Anno 122)(1):1–23.

MacDonald, J. S. and MacDonald, L. D. (1964). Chain migration ethnic neighborhood formation and social networks. *The Milbank Memorial Fund Quarterly*, 42(1):82–97.

Miller, R. G. (1981). *Simultaneous Statistical Inference.* New York, Springer-Verlag.

MOBYSU.IT (2016). *Database MOBYSU.IT, Mobilità degli studi universitari italiani, Protocollo di ricerca MIUR - Università degli Studi di Cagliari, Palermo, Siena, Torino, Sassari, Firenze e Napoli Federico II, Fonte dei dati ANS-MIUR/CINECA.*

Newman, M. (2011). *Networks. An Introduction.* Oxford, Oxford University Press.

Parr, N., Lucas, D., and Mok, M. (2000). Branch migration and the international dispersal of families. *International Journal of Population Geography*, 6(3):213–227.

Pérez, P. and Mcdonough, P. (2008). Understanding latina and latino college choice: A social capital and chain migration analysis. *Journal of Hispanic Higher Education*, 7(3):249–265.

Puccio, E., Vassallo, P., Piilo, J., and Tumminello, M. (2019). Covariance and correlation estimators in bipartite complex systems with a double heterogeneity. *Journal of Statistical Mechanics: Theory and Experiment*, 5(30 May 2019):e53404.

Simon, H. A. (1996). *The Sciences of the Artificial.* Cambridge, MIT Press.

Tumminello, M., Miccichè, S., Lillo, F., Piilo, J., and Mantegna, R. (2011). Statistically validated networks in bipartite complex systems. *PLoS ONE*, 6(3):e17994.

Ünal, S. (2017). The new actors of international migration: a comparative analysis of foreign students' experiences in a medium-sized city in turkey. *People's Movements* *in the 21st Century: Risks, Challenges and Benefits*, page 231.

Viesti, G. (2016). *Università in declino. Un'indagine sugli atenei italiani da Nord a Sud*. Roma, Donzelli.

Xulvi-Brunet, R. and Sokolov, I. (2004). Reshuffling scale-free networks: From random to assortative. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 70(6 2):066102/1–066102/6.

Yingbo, L., Jiujun, C., Xiao, W., and Fuzhen, C. (2005). Research on the matthews correlation coefficients metrics of personalized recommendation algorithm evaluation. *International Journal of Hybrid Information Technology*, 8(1):163–172.