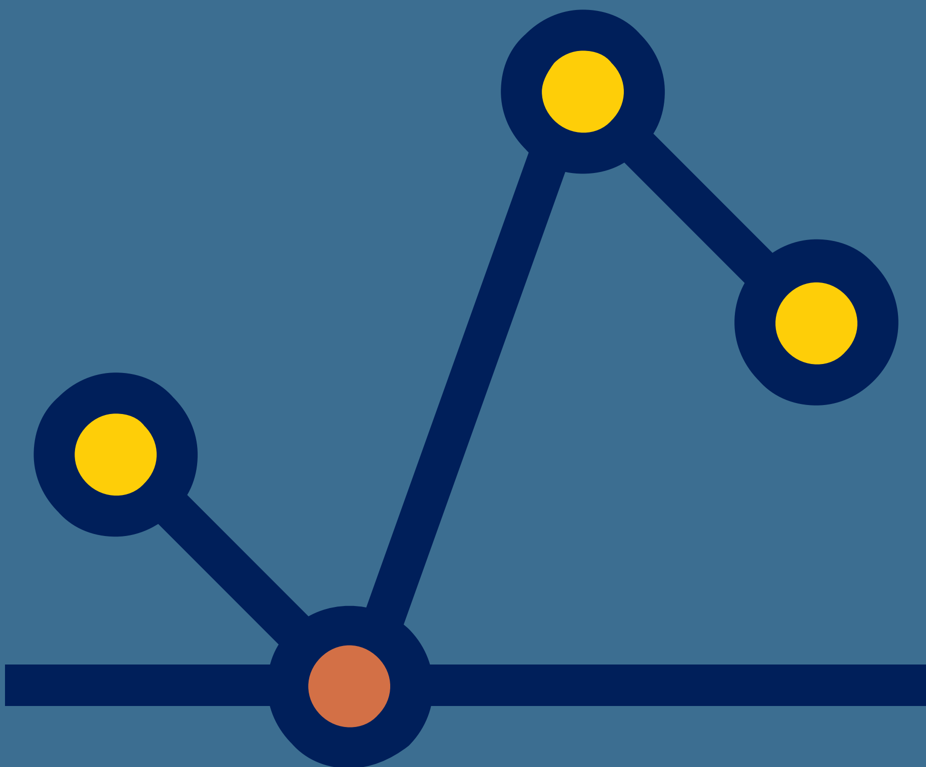
Edited by
Paola Cerchiello · Arianna Agosto
Silvia Osmetti · Alessandro Spelta

Proceedings of the Statistics and Data Science Conference



Pavia University Press

Copertina: Cristina Bernasconi, Milano

Via Salasco, 5 - 20136 Milano
Tel. 02/5836.5751 - Fax 02/5836.5753
egea.edizioni@unibocconi.it - www.egeaeditore.it

Pavia University Press
info@paviauniversitypress.it – www.paviauniversitypress.it

Prima edizione: maggio 2023
ISBN volume 978-88-6952-170-6

Preface

The development of large-scale data analysis and statistical learning methods for data science is gaining more and more interest, not only among statisticians, but also among computer scientists, mathematicians, computational physicists, economists, and, in general, all experts in different fields of knowledge who are interested in extracting insight from data.

Cross-fertilization between the different scientific communities is becoming crucial for progressing and developing new methods and tools in data science.

In this respect, the Statistics & Data Science group of the Italian Statistical Society has organized an international conference held in Pavia on the 27 and 28 of April 2023, attended by over 70 researchers from different scientific fields.

A collection of the presented papers is available in the present Proceedings showing a huge variety of approaches, methods, and data-driven problems, always tackled according to a rigorous and robust scientific paradigm.

The Statistics & Data Science group

Contents

Fractional random weight bootstrap in presence of asymmetric link functions 1
La Rocca Michele, Niglio Marcella, Restaino Marialuisa

Innovation patterns within a regional economy through consensus community detection on labour market network 6
Morea Fabio, De Stefano Domenico

Sparse Inference in functional conditional Gaussian Graphical Models under Partial Separability 12
Fici Rita, Sottile Gianluca , Augugliaro Luigi

A Conformal Approach to Model Explainability 18
Mata Naranjo Juan, Brutti Pierpaolo

A S.A.F.E. approach for Sustainable, Accurate, Fair and Explainable Machine Learning Models 24
Raffinetti Emanuela , Giudici Paolo

Do we really care about data ethics? 30
Ferrara Alfio

Ethical concepts of data ethics between public and private interests 36
Durante Massimo

Being a statistician in the big data era: A controversial role? 42
Manzi Giancarlo

Forecasting relative humidity using LoRaWAN indicators and autoregressive moving average approaches 47
Rojas Guerra Renata, Vizziello Anna, Gamba Paolo

Interpretability of Machine Learning algorithms: how these techniques can correctly guess the physical laws? 53
De Corato Marzio, Ferrara Alfio, Salini Silvia

The role of BERT in Neural Network sentiment scoring for Time Series Forecast 55
Basili Roberto, Croce Danilo, Iezzi Domenica, Monte Roberto

Diagnostics for topic modelling. The dubious joys of making quantitative decisions in a qualitative environment 61
Sciandra Andrea, Trevisani Matilde, Tuzzi Arjuna

Mapping the thematic structure of Data Science literature with an embedding strategy 67
Irpino Antonio, Misuraca Michelangelo, Giordano Giuseppe

Critical Visual Explanations. On the Use of Example-Based Strategies for Explaining Artificial Intelligence to Laypersons 73
Gobbo Beatrice

Visualising unstructured social media data: a chart-based approach 77
Aversa Elena

From teaching Statistics to designers to teaching Statistics through design 85
Mauri Michele, Vantini Simone

Forecasting Spatio-Temporal Data with Bayesian Neural Networks 90
Ravenda Federico, Cesarini Mirko, Peluso Stefano, Mira Antonietta

Oracle-LSTM: a neural network approach to mixed frequency time series prediction 96
Bitetto Alessandro, Cerchiello Paola

Streamlined Variational Inference for Modeling Italian Educational Data 102
Gioia Di Credico, Claudia Di Caterina, Francesco Santelli

The use of magnetic resonance images for the detection and classification of brain cancers with D-CNN 108
Mascolo Davide, Plini Leonardo, Pecchini Alessandro, Antonicelli Margaret

Modeling and clustering of traffic flows time series in a flood prone area . 113
Zuccolotto Paola, De Luca Giovanni, Metulini Rodolfo, Carpita Murizio

Global mobility trends from smartphone app data. The MobMeter dataset 119
Finazzi Francesco

Spatio-temporal statistical analyses for risk evaluation using big data from mobile phone network 124
Perazzini Selene, Metulini Rodolfo, Carpita Maurizio

A Robust Approach to Profile Monitoring 130
Capezza Christian, Centofanti Fabio, Lepore Antonio, Palumbo Biagio

The FDA contribution to Health Data Science 133
Ieva Francesca

A new topological weighted functional regression model to analyse wireless sensor data 139
Romano Elvira, Irpino Antonio, Andrea Diana

Clustering for rotation-valued functional data 145
Stamm Aymeric, Bellanger Lise

Giudici Paolo InstanceSHAP: An instance-based estimation approach for Shapley values 151
Babei Golnoosh, Giudici Paolo

A new paradigm for Artificial Intelligence based on Group Equivariant Non-Expansive Operators (GENEOs) applied to protein pocket detection 152
Bocchi Giovanni, Micheletti Alessandra, Frosini Patrizio, Pedretti Alessandro, Gratteri Carmen, Lunghini Filippo, Beccari Andrea Rosario, Talarico Carmine

Clustering Italian medical texts: a case study on referrals 158
Torri Vittorio, Ercolanoni Michele, Bortolan Francesco, Leoni Olivia, Ieva Francesca

Classification of Recommender systems using Deep Learning based generative models 164
Filali-Zegzouti Sanae, Banouar Oumayma, Benslimane Mohamed

Sparse Inference in Gaussian Graphical Models via Adaptive Non-Convex Penalty Function 170
Cuntrera Daniele, Muggeo Vito M.R., Augugliaro Luigi

Bayesian causal inference from discrete networks 177
Castelletti Federico, Consonni Guido

Sign-Flip tests for Spatial Regression with PDE regularization 182
Cavazzutti Michele, Arnone Eleonora, Ferraccioli Federico, Finos Livio, Sangalli Laura M.

A novel sequential testing procedure for selecting the number of changepoints in segmented regression models 187
Priulla Andrea, D’Angelo Nicoletta

On the numerical stability of the efficient frontier 193
Fassino Claudia, Uberti Pierpaolo

Spatial regression with differential regularization over linear networks . . 196
Clemente Aldo, Arnone Eleonora, Mateu Jorge, Sangalli Laura M.

An Estimation Tool for Spatio-Temporal Events over Curved Surfaces . . . 201
Panzeri Simone, Begu Blerta, Arnone Eleonora, Sangalli Laura M.

Gromov-Wasserstein barycenters for optimal portfolio allocation 207
Spelta Alessandro, Pecora Nicolò, Maggi Mario

Online Job Advertisements: toward the quality assessment of classification algorithms for the occupation and the activity sector 214
Catanese Elena, Inglese Francesca, Lucarelli Annalisa, Righi Alessandra, Ruocco Giuseppina

Linear Programming for Wasserstein Barycenters 220
Auricchio Gennaro, Bassetti Federico, Gualandi Stefano, Veneroni Marco

A multi-channel convolution approach for forecast reconciliation 224
Marcocchia Andrea, Arima Serena, Brutti Pierpaolo

Hedging global currency risk with factorial machine learning models 230
Giudici Paolo, Pagnottoni Paolo, Spelta Alessandro

Predicting musical genres from Spotify data by statistical machine learning 236
Biazzo Federica, Farné Matteo

The use of Bradley-Terry comparisons in statistical and machine learning models to predict football results 242
Macri Demartino Roberto, Torelli Nicola, Egidi Leonardo

A new approach for quantum phase estimation based algorithms for machine learning 248
Ouedrhiri Oumayma, Banouar Oumayma, El Hadaj Salah, Raghay Said

A comparison of ensemble algorithms for item-weighted Label Ranking . 254
Albano Alessandro, Sciandra Mariangela, Plaia Antonella

Unsupervised Learning of Option Price in a Controlled Environment: a Neural Network Approach 260
Gatta Federico, Schiano Di Cola Vincenzo, Piccialli Francesco, Cuomo Salvatore

`SEMgraph`: An R Package for Causal Network Inference of High-Throughput Data with Structural Equation Models 266
Grassi Mario, Tarantino Barbara

Dynamic models based on stochastic differential equations for biomarkers and treatment adherence in heart failure patients 271
Gregorio Caterina, Rares Franco Nicola, Ieva Francesca

Detecting anomalies in time series categorical data: a conformal prediction approach 277
Landrò Matteo, Stamm Aymeric, Vantini Simone

The structural behavior of Santa Maria del Fiore Dome: an analysis with machine learning techniques 282
Masini Stefano, Bacci Silvia, Cipollini Fabrizio, Bertaccini Bruno

Statistics and Data Science for Arts and Culture: an Application to the City of Brescia 288
Ricciardi Riccardo, Carpita Maurizio, Perazzini Selene, Zuccolotto Paola, Manisera Marica

Detecting Stance in Online Discussions about Vaccines 294
Francesco Pierri, Pizzo Fabio, Brambilla Marco

Towards the specification of a self-exciting point process for modelling crimes in Valencia 300
Chiodi Marcello, D'Angelo Nicoletta, Adelfio Giada, Mateu Jorge

A Clusterwise regression method for distributional data 306
Balzanella Antonio, Verde Rosanna, de Carvalho Francisco de A.T.

Increasing accuracy in classification models for the identification of plant species based on UAV images 311
Simonetto Anna, Tariku Girma, Gilioli Gianni

Travel time to university as determinant on students' performances 317
Burzacchi Arianna, Rossi Lidia, Agasisti Tommaso, Paganoni Anna Maria, Vantini Simone

The FAITH project: integrated tools and methodologies for digital humanities 323
Ferrara Alfio, Picascia Sergio, Rocchetti Elisabetta, Varese Gaia

Assessing the quality of Automatic Passenger Counter data for the analysis of mobility flows of local public transport systems 328
Urbano Valeria Maria, Burzacchi Arianna, Cherubini Francesco, Arena Marika, Azzone Giovanni, Secchi Piercesare, Vantini Simone

Innovation patterns within a regional economy through consensus community detection on labour market network

Fabio Morea and Domenico De Stefano

Abstract Universities and research centres play a major role in the generation and diffusion of innovation through education, research, spin-offs and technology transfer. This paper examines a further pattern for the spread of innovation within a regional economy, namely the transfer of workers from one employer to another. Our approach is based on the ”labour market” dataset, from which we derive a network by applying an ad-hoc edge weighting strategy. We propose a novel approach to explore the network structure, using a consensus community detection approach that assigns a probability of membership and isolates trivially small communities. Applying the methodology to the Friuli Venezia Giulia region shows that research institutions play a prominent role in innovation patterns, being the leading elements of large communities and often outperforming large industrial groups.

Key words: Unsupervised Clustering Algorithms, Network Analysis, Community Detection, Labour Market data, ISCO-08

1 Introduction

Connections between companies have been studied extensively through the concept of *clusters* using different definitions that include the concepts of spatial proximity, similarity or competition [8]. The use of labour market data to study inter-links between companies is based on the observation that when employees change jobs, they move to another employer geographically close, requiring similar skills and offering better conditions [1] . Increased availability of data and analytical techniques such as

Fabio Morea
Area Science Park, Padriciano 99, Trieste, Italy, e-mail: fabio.morea@areascienepark.it

Domenico De Stefano
Department of Social and Political Sciences, University of Trieste, Piazzale Europa 1, Trieste, Italy
e-mail: ddestefano@units.it

community detection have improved accuracy of these studies. The analysis can be global, such as [7], which uses labour market data from the social network LinkedIn, or regional, such as [4], which use data from Italy's regional labour market observatories. Modularity based methods [5], and specifically the Louvain algorithm [3] are generally used as the community detection algorithm for exploring labour market networks.

2 Data and methodology

Labour market data encodes the information as *events* that can be either the beginning of a new employment contract, or its termination. Each event is associated with a date, an employee, an employer, a professional profile and a location. The full dataset includes 1155342 events involving 74317 local units of companies of all sectors and sizes, as well as universities and research centres, that have either started or terminated an employment contract in the Friuli Venezia Giulia region between 2014 and 2021.

The raw data needs to be cleaned, completed (e.g. adding implicit contract terminations) and processed (e.g. identifying the actual workplace in the case of employment agencies). Moreover, the data is filtered to a subset of interest based on occupations, which for this paper is limited to professional groups ISCO-21 (science and engineering occupations) and ISCO-25 (information and communication technology occupations) as defined by the International Standard Classification of Occupations 2008 [2]. The resulting data set includes about 60164 events, which involve 1890 employers and 16474 employees.

Further analysis is based on a network in which vertices encode employers and edges encode the transition of an employee P from employer A to employer B. Transitions are assigned a weight which represents the relevance of the connection between A and B. The basic option is to assign a weight $W = 1.0$ to each transition; although this leads to valid results, we argue that it does not exploit the potential of the data. In this study, the weights are assigned under the assumption that the experience gained by P while working for A is transferred to B. Our data cannot capture the intrinsic economic value of each transfer, so we have chosen to approximate it with a non-linear parameter W. Let D_P^A be the duration of the contracts of P with A, D_P^B be the duration of the contracts of P with B (both expressed in years), and $maxW$ be a threshold that model the fact that experience gained in previous workplaces is no longer relevant. Our analysis assumes that $W = min(D_P^A, D_P^B, maxW)$ where $maxW = 5.0$.

The resulting network, after simplification (removal of loops and multiple edges) and pruning of isolated vertices, has a main component of 734 vertices (i.e. employers), two components of size 6 and 4, and 145 other components of size 3 or 2. The subsequent analysis is performed only on the main component. The strength of vertices (i.e. the sum of edge weights of the adjacent edges for each vertex) spans several orders of magnitude, from 0.008 to 309. We assessed the centrality of ver-

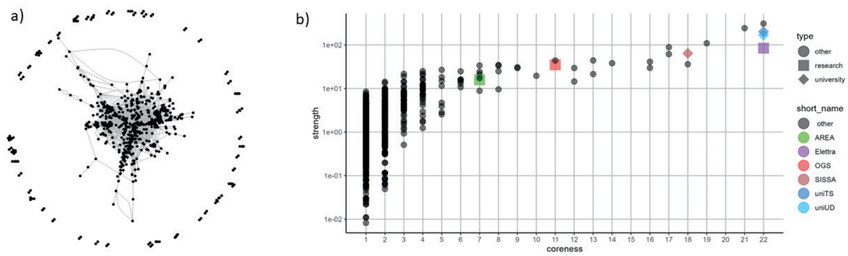


Fig. 1 Universities of Udine and Trieste (blue diamonds), SISSA - *Scuola Internazionale Superiore di Studi Avanzati* (brown diamond) and Elettra Sincrotrone (pink square) are among the top ranking nodes of the network. Other research centers play a mayor role in terms of coreness and strength.

tices by calculating their coreness. The coreness of a vertex is k if it belongs to the k-core but not to the $(k+1)$-core, where the k-core of a graph is a maximal sub-graph in which each vertex has at least k edges. A scatter plot of strength and coreness is shown in Figure 1, providing some insight into the general structure of the network.

We aim to partition our network in a number of communities, in which vertices are strongly connected amongst each other, but weakly connected with vertices belonging to other communities. We require our algorithm to identify only *relevant* communities and to group all sort of *trivially small* communities in a meta-community labeled as community 0. Examples of trivially small communities are those composed of a single vertex or a couple of vertices joined by a single edges; or communities composed by several vertices with extremely weak edges. Finally, we need our algorithm to deliver robust results, that depend as little as possible from random initialisation parameters.

Modularity-based methods are often used for community detection because they meet most of the above requirements. Given a network G partitioned into a number of communities G_i, modularity $Q(G, Gi)$ is a function measuring the extent to which edge density is higher within than between communities [5]. A partition of G that maximises Q results in communities that have strong internal connections and weak connections with other communities. A commonly used method to identify the optimal community structure in labour market networks is the ”Louvain” algorithm, as introduced by [3] and implemented in the iGraph library in the R programming language. It initiates by partitioning the network so that each vertex is assigned to a single community. Then, starting with a random vertex V_i, it computes the potential variation in modularity ΔQ_{ij} that would occur by aggregating V_i to each of its neighbours V_j. If $max(\Delta Q_{ik}) > 0$ then, V_i is removed from its original community and aggregated to the neighbour V_k that maximises the gain. The number of communities is thus reduced, and process is repeated sequentially for all other vertices until $max(\Delta Q_{ik}) \leq 0$. This approach has two known drawbacks. First, the algorithm is greedy and identifies local maxima. Second, the number of communities and the assignment may vary each time the algorithm is run, since the results depend on the random choice of the initial node V_i and the arbitrarily chosen sequence of vertices. A further source of variability is the parameter γ (resolution), which sets an arbitrary trade-off between intra-community edges and inter-community edges, and allows to influence the distribution of community sizes to some extent, as explained by [6]. A typical approach to deal with results depending on random initialization is to run the community detection algorithm for N_i iteration (which leads to N_i different local maxima) and selecting the iteration that produced the highest modularity.

We suggest a further improvement that exploits the intrinsic variability of Louvain algorithm, using an approach similar to the well known *random forest* algorithm. The Louvain community detection algorithm is repeated N_i times, and at each iteration a randomly chosen fraction α of edges is assigned a weight W_0 (small, but non-zero) and γ is randomly assigned to a range of values around 1.0 . The resulting network is not loosing connectivity (but edges associated with reduced weight are more likely to be assigned to different community at each run) and the size of communities varies at each run.

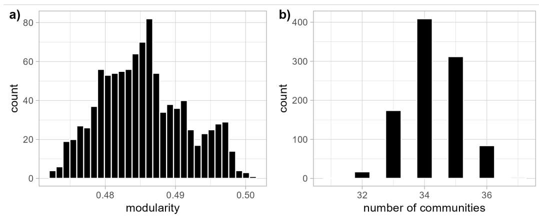


Fig. 2 Variability of results: assignation of a node to a community and total number of communities identified depend on random initialisation and resolution paramter.

For a network G composed of N_v vertices, results are in the form of a matrix A of size $Nv \times Ni$ recording the community assignation for each iteration. The consensus algorithm counts how many times a pair of vertices V_i and V_j are assigned to the same community. The final result of consensus algorithm is a matrix C of size $N_v \times 2$ in which each vertex (employer) is assigned to a community and a proportion of membership $P_{V_i} \in [0,1]$. Vertices that are strongly connected to one another are always assigned to the same community and have $P_{V_i} = 1$; lower values of P_{V_i} indicate that the vertex is not strongly connected to its neighbours, and it may be assigned to two or more communities with some degree of confidence.

Trivially small communities of size $S_{community} < S_{c_{min}}$, and single vertices with $P_{V_i} < 0.5$ are all assigned to a meta-community labelled as "community 0". In presence of more than one component, components of size $S_{component} < S_{k_{min}}$ are also assigned to "community 0".

3 Results and discussion

Communities consist of vertices (i.e. employers) with stronger links to each other than to other communities. In terms of innovation patterns, this can be interpreted as knowledge transfer being more relevant among members of the same community than from one community to another. The fact that research centres are at the heart of their respective communities shows that the transfer of staff is an effective means of transferring knowledge, experience and innovation between academia and industry. Applying the above methodology to in Friuli Venezia Giulia region, we observed that communities are generally characterized by a central vertex (a large company, university or research center), a few prominent elements with a high proportion of membership and a large number of smaller companies. Figure 3 highlights the structure of selected communities in the strength-coreness scatterplot.

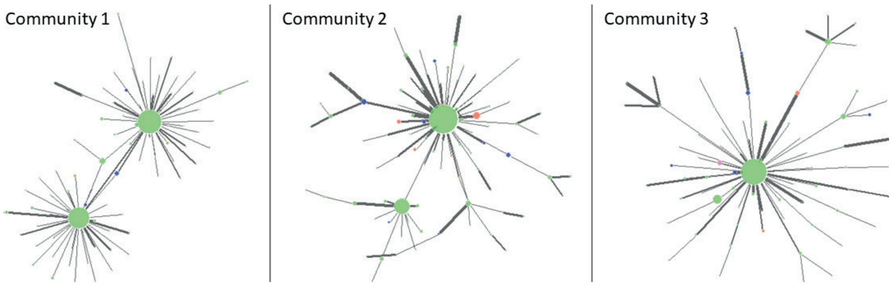


Fig. 3 Some examples of communities. The size of vertices is proportional to their degree, and color scale reflects the proportion of membership (green vertices have a proportion of membership $P_i > 0.9$). Meta-community 0 is composed of several unconnected small communities and individual vertices with $P_i < 0.5$. Most communities have one or two central node of high coreness and strength.

As highlighted in Figure 4 research institutions play a prominent role in the regional labour market, as expressed by the high coreness values and their role within their community. Specifically, the two universities operating in the region (University of Trieste and University of Udine) belong to the largest community (labelled as Community 1, size 89), have comparable values of coreness and largely surpass other large enterprises. Other major research institutions (namely Elettra Sincrotrone Trieste and the National Institute of Oceanography and Applied Geophysics - OGS) belong to the same community as the universities, with comparable strength and significantly lower values of coreness, possibly due to their sectoral specialization. The second largest community (labeled 2, of size 78) is led by two large industrial companies (Danieli Officine Meccaniche and Cimolai), followed by 76 other companies that have remarkably lower values of strength and coreness, thus being much less active in receiving or transmitting knowledge and experience within the regional economy. Similarly, the third community is led by Fincantieri, a major player in shipbuilding, strongly connected by other companies stat are located in the same area, or operate in similar sectors (mechanics, yacht and ship building). Future developments of this research should focus on analysing the temporal evolution of centrality indices and community structure, as well as analysing different groups of professions.

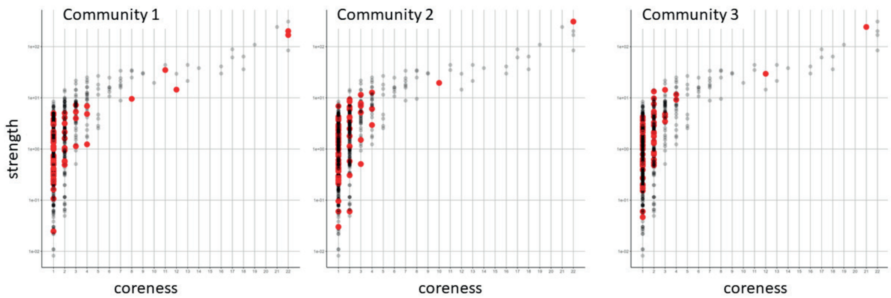


Fig. 4 Leading organizations within selected communities in Friuli Venezia Giulia region. Community 1: University of Trieste, University of Udine, SISSA and OGS. Community 2: Danieli Officine Meccaniche and Cimolai. Community 3: Fincantieri and other companies in the maritime sector.

Code Availability: Code for data analysis associated with the current submission is available at https://doi.org/10.5281/zenodo.7609224. Any updates will also be published on Zenodo.

Acknowledgements This paper was developed as part of the PhD program in Applied Data Science and Artificial Intelligence (www.adsai.it). We would like to thank Dr. Carlos Corvino, head of the Regional Observatory on Policies and the Labour Market of the Friuli Venezia Giulia Region for providing the data used in this study.

References

1. Bjelland, M., Fallick, B., Haltiwanger, J., McEntarfer, E. (2011). Employer-to-employer flows in the united states: estimates using linked employer-employee data. Journal of Business and Economic Statistics, 29(4), 493-505.
2. European Commission (2009), Commission Recommendation of 29 October 2009 on the use of the International Standard Classification of Occupations (ISCO-08), Official Journal of the European Union, 292: 31-47
3. Blondel, V. D., Guillaume, J. L., Lambiotte, R., Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008(10), P10008.
4. Menardi, G., De Stefano, D. (2021). Density-based clustering of social networks. arXiv preprint arXiv:2101.08334.
5. Newman, M. E. (2006). Modularity and community structure in networks. Proceedings of the national academy of sciences, 103(23), 8577-8582.
6. Newman, M. E. (2016). Equivalence between modularity optimization and maximum likelihood methods for community detection. Phys. Rev. E 94, 052315. https://doi.org/10.1103/PhysRevE.94.052315
7. Park, J., Wood, I.B., Jing, E. et al. (2019) Global labor flow network reveals the hierarchical organization and dynamics of geo-industrial clusters. Nat Commun 10, 3449 . https://doi.org/10.1038/s41467-019-11380-
8. Porter, M. E. (2000). Location, competition, and economic development: Local clusters in a global economy. Economic development quarterly, 14(1), 15-34.