Firm collaboration: exploratory analysis of FVG companies network

M. Franzon T. Rodani

26 Giugno 2020

Abstract

This report briefly explore the network of collaboration between companies in the Friuli-Venezia Giulia region of Italy. The first part introduces the dataset and shows companies distribution in the region with respect to different variables such as company type and location. Then matrices of collaboration between categories of companies are shown, one for each province of the dataset, and then used to choose interesting subsets of the network. The two subgraphs of collaboration chosen show the interaction between manufacturing companies and construction companies in Trieste province, and the interaction between manufacturing companies and retail companies in Udine province. These subgraphs are investigated both in their indirected and directed form in order to find firm communities. Finally, ERG modeling is used to analyze them in order to find good correlations between dataset variables and network topology.

1 Dataset Description

> "Keep in mind that imagination is at the heart of all innovation. Crush or constrain it and the fun will vanish."
>
> -Albert-Laszlo Barabasi

The dataset provided is composed by a sparse square matrix that represent collaborations among firms in the Italian region Friuli-Venezia Giulia in the period 2014-2017. The ij value of the matrix is the number of hours company i has spent in collaboration with company j, likewise sharing co-workers and other activities. As these are collaborations, if ij exists also ji is present, but ij is not equal to ji because every firm can invest a different amount of time in their collaboration. The diagonal of the matrix ii, is the amount of hours the company have spent outside collaborations, that is by itself. The total number of firms is 32020, and for each one are present other four variables, that are stored as dummy in individual tables. The first one is ATECO code[1], the national version of the EU NACE code[2], a statistical classifications of economic activities that follows the logic of the international system of economic classification ISIC. The ATECO table is composed by 21 rows and 32000 columns, where every row is an ATECO code, and every column a firm. If the firm j is part of the ATECO category i, the ij value of this table is 1, otherwise 0. Other dummy variables use the same binary format, where a variable X is stored in a table $K * N$ where K is the number of unique values of X and N the number of firms. The second variable is "Sistema Locale del Lavoro" (SLL)[3], the national version of the EU Labour Market Areas (LMAs)[4]. LMAs are sub-regional geographical areas where the bulk of the labour force lives and works, and are usually different than administrative boundaries. Within these SLLs residents can find jobs within a reasonable commuting distance or can change their employment without changing their place of residence and establishments can find the main part of the labour force necessary to occupy the offered jobs. In Friuli-Venezia Giulia there are presente 11 SLLs, with codes that range from 601 to 611. The third information about firms is the province or "Provincia" which represent the institutional bodies of second level in the Italian Republic, below regions and above municipalities. Friuli-Venezia Giulia region is composed of 4 provinces: Gorizia (GO), Pordenone(PN), Trieste(TS) and Udine(UD). Finally, there is the Hub information, which are geographical areas that refers to different venues of "Agenzia regionale per il lavoro", the regional agency that offers labour services. Friuli-Venezia Giulia region is divided in 5 HUBs: "HUB Giuliano", "HUB Isontino", "HUB Pordenonese", "HUB Medio e alto Friuli", "HUB Udine e basso Friuli".

1.1 Dataset cleaning

All variables stored in binary form where converted in R lists. A brief analysis on these lists has shown that there were some firms with more than one value for different variables. Multiple SLLs,provinces and HUBs were considered valid, as a business can have multiple location, while the companies with more than one ATECO code were removed, along with firms with an ATECO code of ".". The cleaned dataset consist of 30396 firms, as 69 companies had an ATECO code of "." and 1555 had multiple values.

2 Descriptive analysis of the whole network

This section contains the distribution of the firms among the different variables, the network topology and statistics, and the matrix of collaboration among firms aggregated on ATECO codes used to select two subgraphs of Trieste and Udine provinces.

2.1 Quantitative analysis of firms

The first analysis performed is quantitative and shows how firms are spread in the region by province, HUB, and SLL. We can see that the majority of the companies(44,13%) are located in the Udine province (Figure 3), which is indeed the one with the greater population, 43.51% of the region (ISTAT,2019). Looking at the Hub bar plot (Figure 2), we can see that 10465 out of these firms are part of the HUB "Udine e bassa friuliana" while only 4095 are part of HUB "Medio e alto Friuli", showing a concentration toward the city of Udine. An analysis on the ATECO code (Figure 1) among business shows that there is an unbalance towards food services (I, 18,18%), retail(G, 18,07%) and manufacturing(C, 14,98%) which make more than half of all firms dataset.

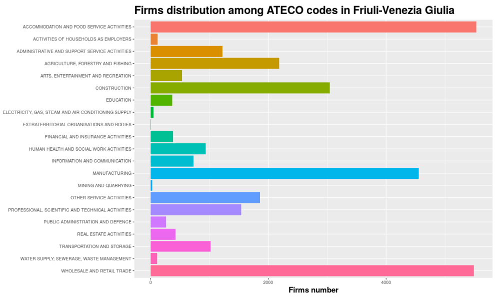

Figure 1:
Barplot of ATECO code distribution

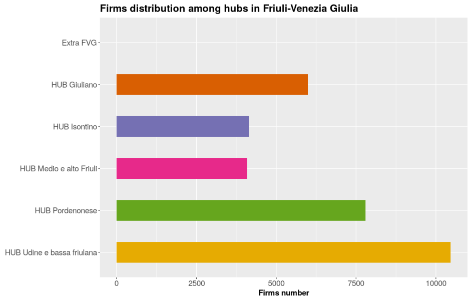

Figure 2:
Barplot of HUB distribution

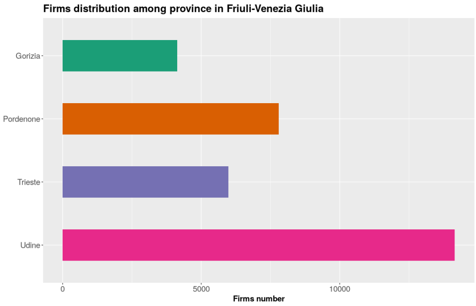

Figure 3:
Barplot of Provinces distribution

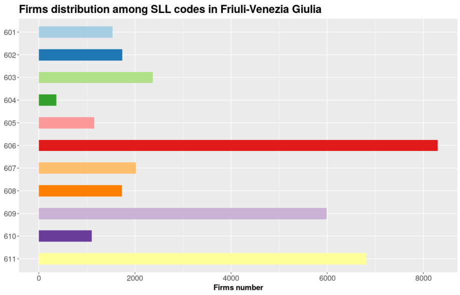

Figure 4:
Barplot of Sll distribution

This is true also in the hours plot where these categories have spent more hours on collaboration than others. A closer look on hours information (Figure 5) showed that a small subset of firms had 0, 1 and other small number of hours incompatible for further analysis. An appropriate threshold was chosen in order to avoid problems later on, so 410 companies with less than 1760 hours were excluded from further analysis. This number of hours is an estimate of the amount of working hours of one employee at full time for one year.

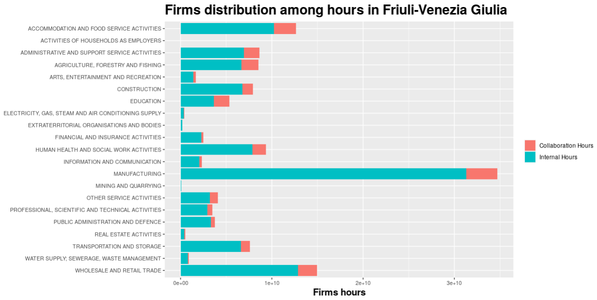

Figure 5:
Barplot external/internal hours

2.2 Network analysis

Once the network was created from the given adjacency matrix (see Section 1) and the variables were loaded as vertex attributes, the first step was to remove the nodes with incorrect ATECO codes (see Section 1.1) and the ones with too little hours (see Section 2.1). After performing this operation the isolates, nodes that were connected only to the ones just removed and thus now without any arc, were also removed. The check of the remaining nodes showed an unexpected number as only 19981 companies remained. The cause of this result is the fact that out of 32020 firms present in the collaboration dataset, an outstanding percentage (29,36%) were not cooperating at all with anyone. The result of this process plus the removal of loops is described in Table 1.

Operation	Order	Size
Init	32020	228352
− Loops	32020	196332
− isolates	22616	196332
− wrong ATECO	21131	123020
− hours < 1760	21017	122720
− isolates	19981	122720

Table 1: Network cleanup process

The obtained network is a full one-mode directed network with mutual weighted arcs. This means that by construction the number of mutual dyads in the network is exactly $\frac{n}{2}$ and the others are null; in the same way the in-degree (popularity) and out-degree (expansiveness) of the nodes are always equal. The density of the network is low, 0.00031, and thus it can be considered a sparse network as $k \approx 6n$, where k is the number of arcs and n the number of nodes. Looking at the centrality scores, we can describe how the different firms are embedded in the labour market and it is possible to identify the role they play within the system. The simplest centrality measures – degree, closeness, betweenness – quantify the position taken by each node (firm) counting the number of connections it has to others (degree centrality), the ability to reach other nodes at shorter path lengths (closeness) and the number of shortest paths passing through it, connecting otherwise disconnected subnetworks (betweenness). These measures are able to identify the so-called authorities and hubs. Authorities represent the ability to find the information needed inside the network, while hubs measure how much "knowledge" is held by a node. In a labour market defined through mobility flows, these actors are crucial for the overall integration and connectivity of the whole system.

	Mean	Median	Min	Max
Degree	12.280	6.000	2.000	704.000
Closeness	$1.191e^{-05}$	$9.661e^{-6}$	$3.294e^{-8}$	$5.005e^{-5}$
Betwenness	$5.205e^{-04}$	0.000	0.000	0.230

Table 2: Centrality measures of the whole network

2.3 Collaboration firms matrices

In order to detect interesting subset of the graph onto which perform further analysis, a matrix of collaboration between firms has been produced for every province in the region. It represent the number of collaboration between ATECO groups in a binary format, so ij is the sum of the number of collaborations between companies of ATECO codes i and j. Analogously, the elements on the diagonal are the number of collaborations between companies of the same category. All values are normalized between 0 and 1 in order to improve readability trough the heatmaps. It is clear that most of the collaborations take place between business in the same category, along the diagonal.

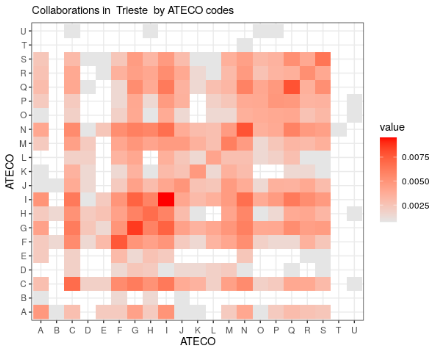

Figure 6:
Matrix of collaboration by Ateco code in Trieste

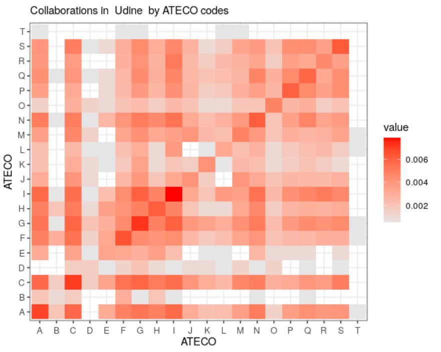

Figure 7:
Matrix of collaboration by Ateco code in Udine

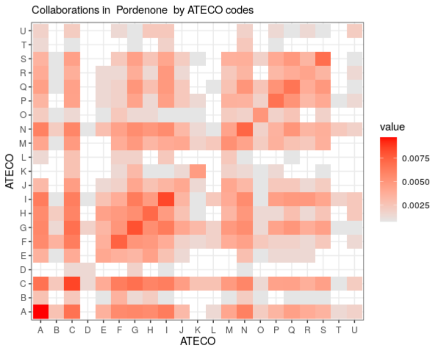

Figure 8:
Matrix of collaboration by Ateco code in Pordenone

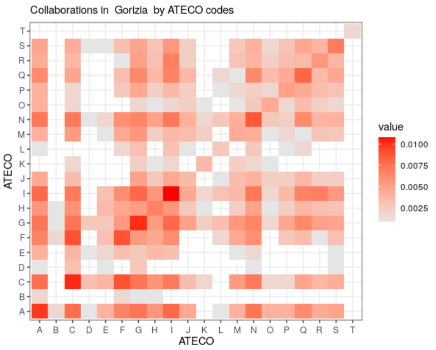

Figure 9:
Matrix of collaboration by Ateco code in Gorizia

3 Descriptive analysis on Trieste and Udine

Using the same metrics of the whole network analysis, was performed a description of two subgraph of Trieste and Udine. The networks chosen are collaborations between construction and manufacturing firms in Trieste and between manufacturing and retail companies in Udine. If collaboration between business of the same category are excluded, construction companies have the higher number of collaborations with manufacturing companies in the Trieste province. At the same time, costruction companies makes part of a goof portion of collaborations for the manufacturing category. This density of collaborations and the relative small size of companies involved were the key points in the choice of the first network. In the Udine network the density of collaboration was also considered but the main factor that led to manufacturing and retail categories was the number of companies. This network order is quite high at 927, making it a good candidate to highlight the differences in the analysis with respect to the relative small network of Trieste, which has order of 152. The tables below shows the cleanup and dimensionality reduction of the two subgraphs, which was performed in order to focus the analysis on the core of the networks, driven by the evaluation of the coreness metric shown in Appendix.

	Degree		Closeness		Betwenness	
	UD	TS	UD	TS	UD	TS
Mean	12.940	6.658	$1.011e^{-5}$	$3.581e^{-6}$	0.011	0.049
Median	8.000	6.000	$1.124e^{-5}$	$3.581e^{-6}$	$3.666e^{-4}$	0.013
Min	2.000	2.000	$1.124e^{-5}$	$8.066e^{-7}$	0.000	0.000
Max	134.000	24.000	$1.508e^{-5}$	$5.286e^{-6}$	0.366	0.572

Table 3: Centrality measures of Udine and Trieste subgraphs

Operation	C nodes	G nodes	Order	Size
Init	263	438	701	1208
− isolates	171	351	522	1208
− degree<5	49	109	158	506
− isolates	47	105	152	506

Table 4: Trieste subgraph of C and F ATECO codes

Operation	C nodes	F nodes	Order	Size
Init	1540	1677	3217	9314
− isolates	1240	1261	2501	9314
− degree<5	516	485	1001	6116
− isolates	508	479	987	6116
− degree=2	471	456	927	5996
− isolates	471	456	927	5996

Table 5: Udine subgraph of C and G ATECO codes

Looking at the network dimensions, the number of arcs l correspond to $k \approx 3n$ for Trieste and $k \approx 6n$ for Udine, which can be seen also from the mean degree, where this number is approximately doubled as all arcs are reciprocated. Taking into account also the values of betwenness, it is clear that the smaller network of Trieste is more intra-connected while the larger network of Udine is more sparse. This is clearly represented in the edge density values for the two, which are 0.007 for Udine and 0.022 for Trieste. Closeness centrality, which measure the proximity of node i with respect to the otherd in the network, is reported for each subgraphs while this wasn't possible for the full network as it is not strongly connected.

4 Communities detection

Community detection is key to understanding the structure of complex networks, and ultimately extracting useful information from them. In general, real networks are not random. Weak ties seem to bridge groups of tightly coupled nodes.

In this section the focus is on the possible communities detection of the two subgraphs of Udine and Trieste. First the network between construction and manufacturing firms in Trieste; second, the network between manufacturing and retail firms in Udine.

4.1 Walktrap, a random walks approach

The general idea is that performing random walks on the graph, then the walks are more likely to stay within the same community because there are only a few edges that lead outside a given community. This method runs short random walks of 3-4-5 steps (depending on one of its parameters) and uses the results of these random walks to merge separate communities in a bottom-up manner. Walktrap algorithm could be use also in directed network and its computation time is $O(n^2 log(n))$ in the best case, where n is the number of vertexes.

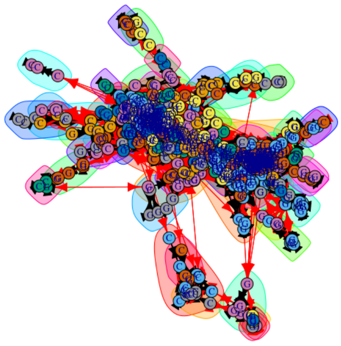

Figure 10:
Community detection with Walktrap algorithm - Udine. Color rappresents the ownership of a specific community

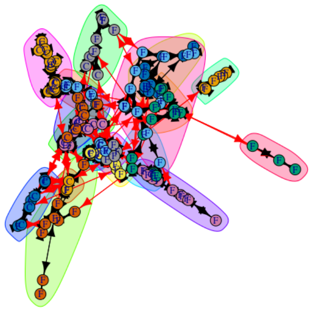

Figure 11:
Community detection with Walktrap algorithm - Trieste. Color rappresents the ownership of a specific community

4.2 Louvain algorithm

Louvain method consists of two phases. First, it looks for ”small” communities by optimizing modularity in a local way. Second, it aggregates nodes of the same community and builds a new network whose nodes are the communities. These steps are repeated iteratively until a maximum of modularity is attained. In order to perform Louvain algorithm, subgraphs were converted into undirected graphs. Also in this case the computation time is $O(n^2 log(n))$ as Walktrap algorithm.

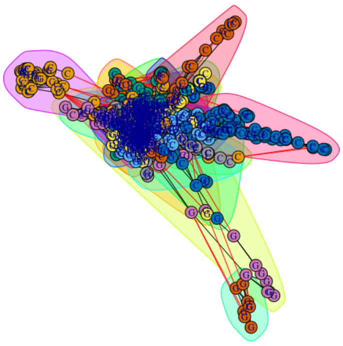

Figure 12:
Community detection with Louvain algorithm - Udine. Color rappresents the ownership of a specific community and each node is labeled with its ATECO code

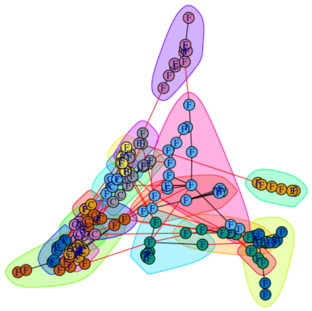

Figure 13:
Community detection with Louvain algorithm - Trieste. Color rappresents the ownership of a specific community and each node is labeled with its ATECO code

4.3 Considerations on community detection

Two different approaches were used to detect community of Udine (C–G) and Trieste (F–C) subgraphs. Walktrap algorithm was used to preserve the direction of the edges, instead with Louvain algorithm was lost the information about direction. Udine subgraph (Figure 10, Figure 12) has higher number of vertex than Trieste, so in both cases the detection is not clearly defined . In Trieste network (Figure 11, Figure 13) the communities are much more defined in both cases and also the number of communities is comparable. In general, could be reasonable assuming that two subnetworks are appoximate as small-world networks. In both it is possible recognize some small-world typcal aspects like hort average path lengths and high clustering. Finally, the metric used to evaluate the efficiency of the two algorithm is the modularity.

The comparison between the two algorithms, considering the modularity and the number of communities respect the total number of vertexes, suggests that the Louvain approach gives a better result in terms of community detection. In both subgraphs it is possible to appreciate that there are a quite large number of small weakly connected communities and few large strong connected communi-

	Walktrap		Louvain	
	UD	TS	UD	TS
Modularity	0.761	0.848	0.799	0.863
# Communities	142	30	35	19

Table 6: Summary of # communities and modularity evaluation of two algorithms.

ties. This is confirmed by the evaluation of the density $\begin{cases} \text{Density}_{UD} = 0.007 \\ \text{Density}_{TS} = 0.022 \end{cases}$. Both subgraphs have low density and considering it as the ratio of the number of edges on the number of possible edges, could be possible assume that the major part of the network is represented by weak relations.

5 ERG models

Exponential random graph models (ERGMs) are a family of statistical models for analyzing data about social and other networks. Also in this case are just considered the two subgraphs of Trieste and Udine, respectively, the network between manufacturing (ID Ateco = C) and construction (ID Ateco = F) and the network between manufacturing (C) and retail (ID Ateco = G). First, the binarization of the adjacency matrix is performed placing a cut-off on the median of collaboration hours, which is respectively 38851 for Udine subgraph and 27365 for Trieste subgraph. Due to the nature of the network, all the edges are reciprocal arcs, so the mutuality in both cases is equal to 1. For this reason the parameters used to fit the ERG models had to be dyad-independent. The metrics used to evaluate the best model was AIC index, according to the rule that *the lower the best*.

5.1 ERG models in Udine (C–G) and Trieste (F–C) subnetwork

The first model was a baseline model in which only edges are used as covariate. Subsequently the homophily on ATECO codes was investigated, which showed poorly results as there is no clear distinction in node parameters from a qualitative firm category perspective. The third simulation involved the extracted memberships from the Louvain clustering method. This choice was made in order to feed the model with an incomplete information on the network topology, because standard methods based on dyads or triads cannot be used as it would led to overfitting. Using extracted membership as covariate, the model should improve at least in the intra-cluster edge placement, which should result in a better score as the result from the Louvain clustering were quite good. This intuition was confirmed by an overall improved score in the simulation. As a proof of this choice the mutual parameter was added in the last simulation, with an expected result of a perfectly reproduced graph. It was predictable, as the mutuality equal to 1 suggests that all vertexes have a mutual relation. Goodness of fit for each simulation along with their respective plot can be found in the Appendix.

```
#Simple simulation with edges as only param
nu.01 <- ergm(nu~edges)

#simulation with edges and homophily on ATECO codes
nu.02 <- ergm(nu~edges+nodematch("Ateco", diff=T))

#simulation with edges and louvain membership
nu.03 <- ergm(nu~edges+nodecov("mem"))

#simulation with edges, louvain membership and mutual ==> perfect fit
nu.04 <- ergm(nu~edges+nodecov("mem")+mutual)
```

6 Conclusion

This work explored the distribution of different variables such location, company category and labour market areas among collaborating firms in Friuli-Venezia Giulia from 2014-2017. The dataset contains a lot of information, unfortunately some of it incomplete, and represent an opportunity to understand the relationships between companies of different sectors through their collaboration. The network analysis was focused on community detection on two subsets of the overall graph and on simulation of their topology by ERG modeling. Other interesting research ideas that were excluded due to time constraints are egocentric network analysis of central nodes of some firm sectors, to evaluate the roles of these nodes in the collaboration flows; and blockmodeling on HUB subgraphs in order to detect different density patterns, whose can shown emerging patterns of interaction between business, especially in the purse of innovation through new paths.

References

[1] Istat. ‘ATECO (CLASSIFICATION OF ECONOMIC ACTIVITY) 2007” Istat. 2007. Web. 16 Jan. 2015. https://www.istat.it/en/archivio/17959

[2] Eurostat. ‘Glossary:Statistical classification of economic activities in the European Community (NACE)” Eurostat. 2007. Web. 24 Feb. 2016. https://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Statistical_classification_of_economic_activities_in_the_European_Community_(NACE)

[3] Istat. ‘LABOUR MARKET AREAS” Istat. 2014. Web. 11 Dec. 2019. https://www.istat.it/en/labour-market-areas

[4] Eurostat. ‘Labour Market Areas” Eurostat. 2014. Web. 28 Jun. 2020. https://ec.europa.eu/eurostat/cros/content/labour-market-areas_en

A Appendix

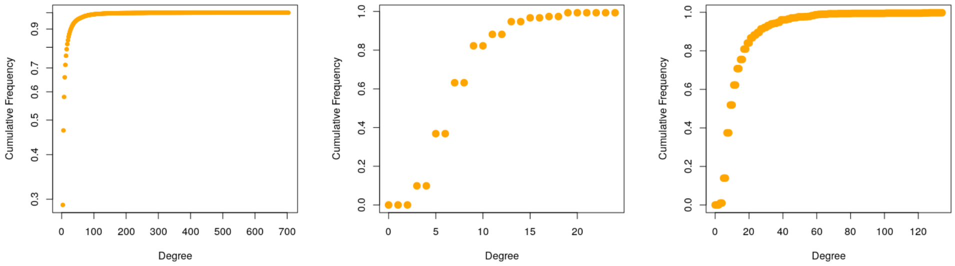

Figure 14:
Degree distribution. From left, the degree distribution of the full network ; in the middle the plot related to Trieste subnetwork; on the right the degree distribution of Udine network

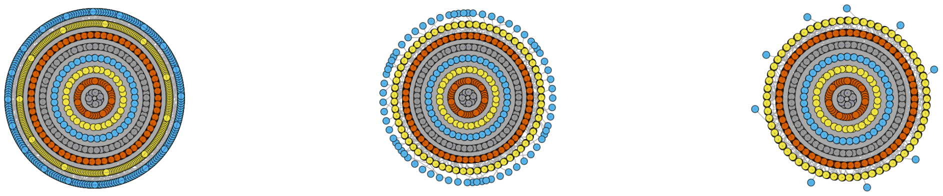

Figure 15:
Coreness plots. From the left, the subgraph of Udine after removing isolates; in the middle after removing nodes with degree <5 and caused isolates; on the right after removing nodes with degree <2 and caused isolates

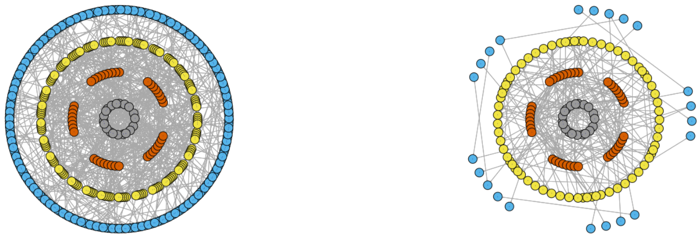

Figure 16:
Coreness plots. From the left, the subgraph of Trieste after removing isolates; on the right the subgraph after removing nodes with degree<2 and caused isolates

Figure 17: ERG models of Udine subgraph. From the left, the real network; after the simulated one with only edges ; after the one with louvainmembership; the last with mutual

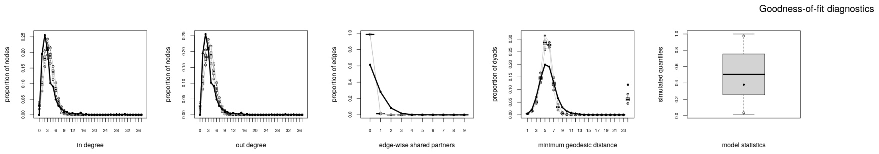

Figure 18:
GOF results from the simulation with only edges - Udine

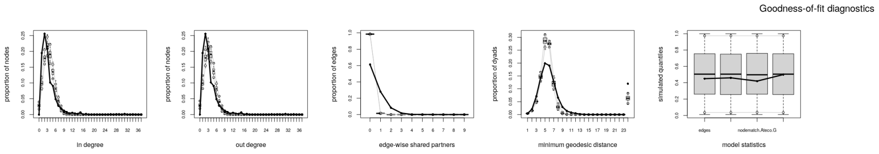

Figure 19:
GOF results from the simulation with edges and louvain membership covariate - Udine

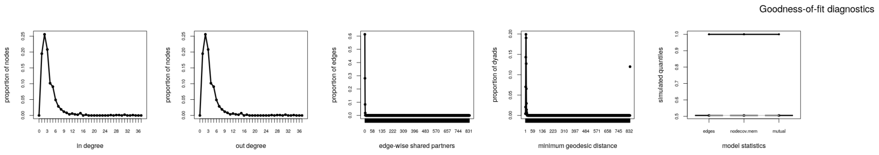

Figure 20:
GOF results from the simulation with edges, louvain membership covariate and mutual term - Udine

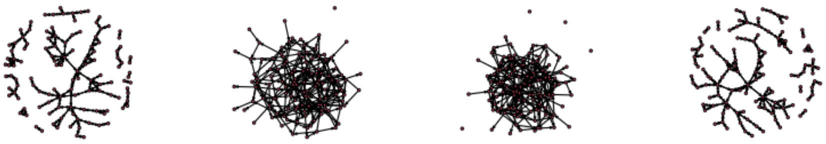

Figure 21: ERG models of Trieste subgraph. From the left, the real network; after the simulated one with only edges ; after the one with louvain membership; the last with mutual

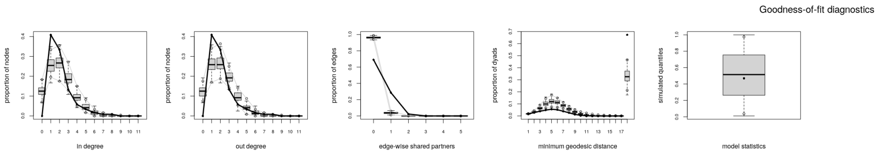

Figure 22:
GOF results from the simulation with only edges - Trieste

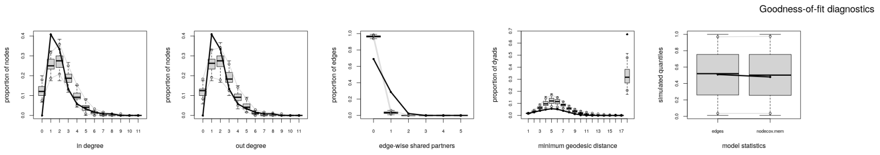

Figure 23:
GOF results from the simulation with edges and louvain membership covariate - Trieste

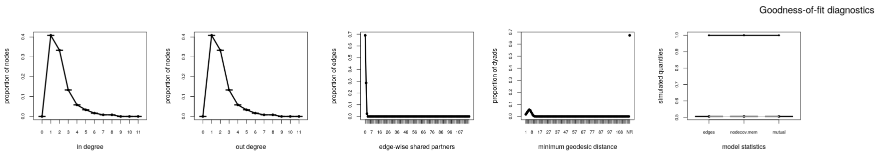

Figure 24:
GOF results from the simulation with edges, louvain membership covariate and mutual term - Trieste