Third Sector Associations and Cooperation

Diletta Ferri
d.ferri8@studenti.unipi.it
Student ID: 667208

ABSTRACT

The aim of this work is to study Italian associations belonging to the third sector, which consists of private, non-profit entities that promote and carry out activities of general interest, with a particular focus on the Lombardia region.

Firstly, the network is built using the geographical distance between associations and then thoroughly analyzed. Then, the final objective is to identify potential collaborations based on the links between associations and to detect new possible partnerships. [1]

KEYWORDS

Social Network Analysis, Third Sector, Associations, Link Prediction, Cooperation

ACM Reference Format:
Diletta Ferri. 2024. Third Sector Associations and Cooperation. In *Social Network Analysis '24.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Very often, a single association or organization cannot fully address the needs identified through the analysis of the surrounding reality on its own. For this reason, it is increasingly necessary to foster collaboration among different third-sector entities, as well as with public institutions. This approach allows for a shared understanding of needs, the identification of objectives, the allocation of resources, and the planning and implementation of necessary interventions. A collaborative paradigm among different organizations is beneficial for all parties involved. However, collaboration requires effort and it is not always trivial to achieve. [1]

One of the primary objectives of this work is to identify the entities with the highest likelihood of collaboration. Acknowledging that personal acquaintance always facilitates cooperation, the approach chosen is to consider associations located in close geographic proximity as those most likely to collaborate.

2 DATA COLLECTION

To build this network, a series of information is required, particularly geographical data related to ETS ('Enti del Terzo Settore').

Selected Data Sources

The data necessary for this analysis was primarily collected from the portal of the Ministry of Labour and Social Policies dedicated to the RUNTS[2]: Registro Unico Nazionale del Terzo Settore, an electronic register specifically created to ensure transparency and disseminate information about ETS. Additionally, to later confirm the location of the entities in the territory, geographical data from the Lombardia Region was used, available on the portal dedicated to open data released by public italian administrations [3].

Crawling Methodology and Assumptions. The first data source used is the list available directly on the Ministry's website, which includes all registered ETS. In this list, among with other information, each record contains:

- *name of the organization*;
- *municipality*;
- *province*;
- *ETS type*, which will be used as a node attribute in the network.

This list was filtered to include only entities based in a province of the Lombardia Region, while removing information that was not relevant for network analysis. However, the list does not provide the precise address of each organization's headquarters, which is necessary to characterize the network. The missing information was obtained through

[1] **Project Repositories**
Data Collection: https://github.com/sna-unipi/2024_Ferri/tree/main/data_collection
Analytical Tasks: https://github.com/sna-unipi/2024_Ferri/tree/main/network_analysis
Report: https://github.com/sna-unipi/2024_Ferri/tree/main/report

SNA '24, 2023/24, University of Pisa, Italy

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$0.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[2] https://servizi.lavoro.gov.it/runts/it-it/Lista-enti
[3] https://www.dati.gov.it/

web scraping in the 'Entity Search' section[4]. Using the organization's name as the research parameter, each record in the list was enriched with specific geographical details: state, province, municipality, and address. Records missing one or more of these facts, as well as those where the identified province and municipality did not match the original ones, were removed.

Next, latitude and longitude coordinates corresponding to the full address were assigned to each record through the Geocoder library[5]. Using these coordinates, all identified ETS were visualized on a Folium Map, making it easier to spot problematic points that were plotted far outside the considered region's boundaries (Figure 1). To eliminate these errors, a GeoJSON file containing the borders of the Lombardia Region was used[6]. If we visualize the map again (Figure 2) we can see that these points are no longer considered. Finally, any duplicate records in the dataset were removed.

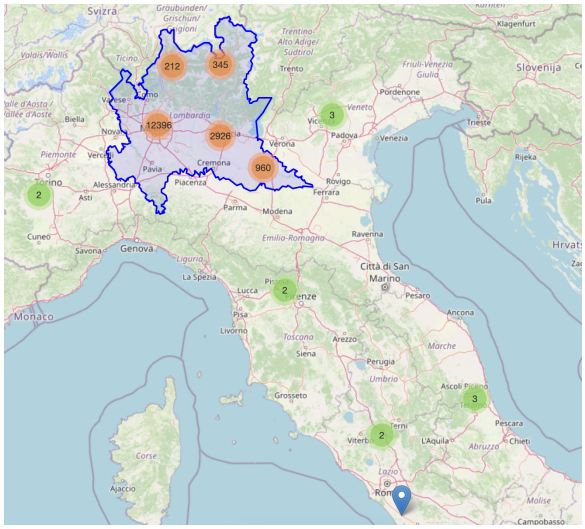

Figure 1: Folium Map plotting all ETS

Based on this cleaned dataframe, a new dataframe of distances between each pair of entities was created, calculated using the Haversine Distance based on the geographical coordinates assigned to each ETS.

3 NETWORK CHARACTERIZATION

This section provides an analysis of the graph created using the dataframes described in the previous chapter.
The nodes in the graph represent the ETS, identified by their

[4]https://servizi.lavoro.gov.it/runts/it-it/Ricerca-enti
[5]https://geocoder.readthedocs.io/
[6]https://www.dati.gov.it/view-dataset/dataset?id=adae2a8f-9b37-4de7-a676-8db1017cb197

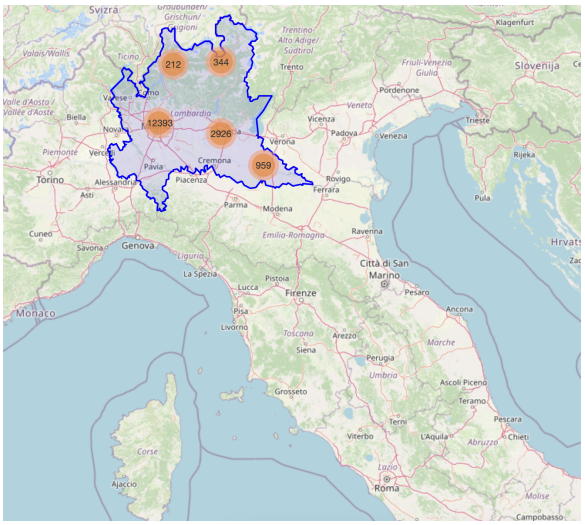

Figure 2: Folium Map after removing points outside Lombardia's boundary

names, with an attribute corresponding to their specific type of third-sector organization.
The edges represent the potential for collaboration between two entities, determined based on their geographic distance: if the distance between two organizations (stored in the distances dataframe) is below a certain threshold, they are considered connected. The underlying assumption is that if two entities are located and operate in close proximity, there is a high probability that the people involved know each other, making cooperation easier. The chosen distance threshold, after plotting the distance distribution (Figure 3) and conducting several tests, is 2 km. This value balances the number of edges and the computational efficiency for subsequent operations with the potential accuracy of the assumption.
The analysis was conducted primarily using the NetworkX library[7].

Degree distribution analysis

The graph consists of 16,828 nodes and 992,029 edges. The first step was to determine the degree of each node and calculate the graph's average node degree, which is 117.9. The degree distribution of the nodes can be observed in the graph shown in Figure 4. As can be observed from the average node degree value and the graph, there are quite a few nodes with a very high number of edges.

[7]https://networkx.org/documentation/stable/reference/index.html

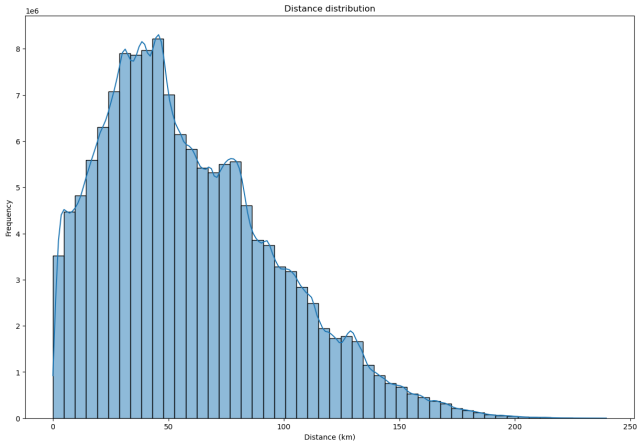

Figure 3: Distribution of the distances between ETS

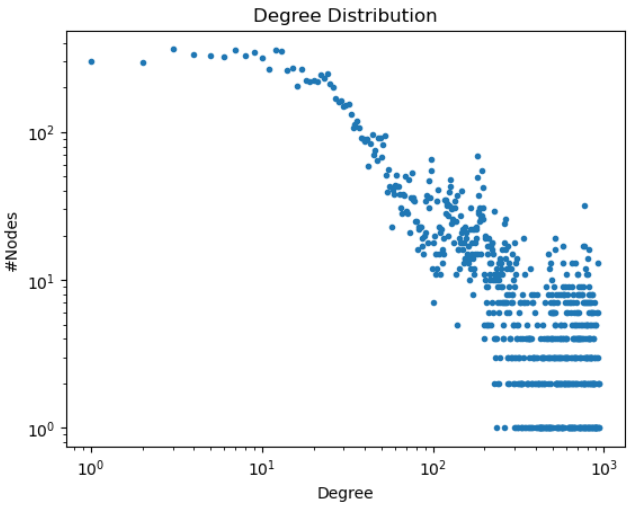

Figure 4: Degree distribution

Connected components and Path analysis

The subsequent analysis focuses on the connectedness of the graph: the graph is not connected. By using the *connected_components()* method, the number of components was determined. A total of 613 components were found, highly variable in size (number of nodes in each component): 173 components consist of isolated nodes, and 432 components contain 5 or fewer nodes.

To better understand the distribution of nodes across components, the size of the 20 largest components was plotted (Figure 5). It is evident that there is a giant component consisting of 11,547 nodes, while all other components are much smaller. For some of the following analyses, particularly in cases where it is necessary to consider connected components, the giant component was used.

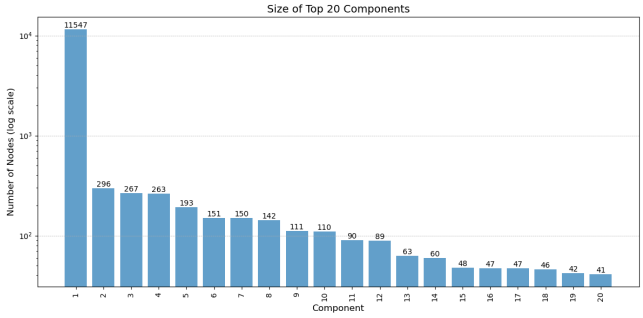

Figure 5: Size of the Top 20 biggest components

To further analyze the distances between nodes, the giant component was used to calculate the diameter and the average shortest path length, which are respectively 113 and 27. The diameter measures the largest distance between any two nodes in the graph, while the average shortest path length represents the mean length of the shortest path between every pair of nodes.

Clustering coefficient and Density analysis

The average clustering coefficient is 0.762, while the network density is 0.007. The average clustering coefficient indicates how well-connected a node's neighbors are to each other. With this high value, it suggests that the nodes have a strong tendency to aggregate, even if the graph can be considered sparse. This tendency is further confirmed by the global clustering coefficient, which considers the closed triangles in the network, providing an overall view of the entire network, and has a value of 0.674.

Centrality analysis

The next step in the analysis is to study centrality, which refers to the importance of nodes within the network.

Degree centrality measures the importance of a node by considering the number of neighbors each node has. For the entire network, the value is 0.007 (the same as the graph density), while for the Giant Component, it is 0.0136. This indicates a quite sparse graph.

The other two measures considered are betweenness centrality and closeness centrality, both of which are geometry-based measures, meaning they depend on some function of the node's distance from other nodes in the network.

Betweenness centrality considers the number of shortest paths that pass through a given node. Since it relies on shortest paths, it must be calculated on a connected component,

in this case, the Giant Component. The value of betweenness centrality is 0.0023, indicating that the network is not centered on a few intermediary nodes, but is rather distributed.

Closeness centrality, on the other hand, measures the average of the shortest paths from a node to all others in the network. The closeness centrality value of the network is 0.040, confirming a dispersed structure of the network. As shown in Figure 6, which displays the distribution of closeness centrality values, there are no nodes that can reach all others quickly. Realistically, this is due to the way the edges were defined: since the maximum geographical distance between two connected nodes is 2 km, the length of the path connecting any pair is strictly dependent on their geographical distance.

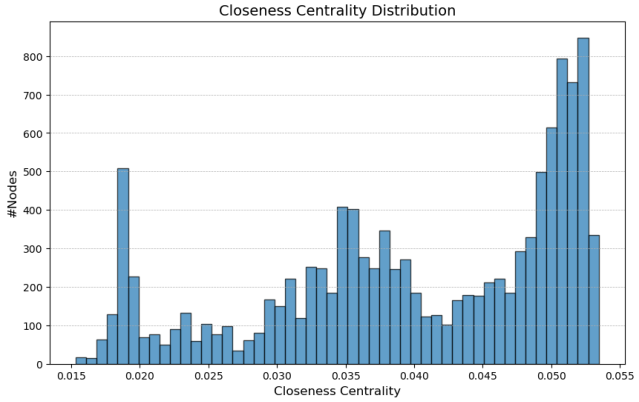

Figure 6: Closeness Centrality Distribution

Comparision with ER and BA graphs

The next step in the network analysis is to compare it with random network models, specifically ER (Erdős–Rényi) and BA (Barabási–Albert), having a similar number of nodes and edges as the analyzed network. The comparison was performed on both the entire network and the Giant Component, with the results presented here referring to the latter.
To construct an ER graph analogous to the network, the probability p (the probability that an edge exists between two nodes) was calculated as:

$$p = \frac{2L}{N(N-1)} = \frac{2 \cdot 906232}{11547 \cdot (11547-1)} = 0.01359 \qquad (1)$$

where L is the number of edges in the real network, and N is the number of nodes.
For the construction of the BA graph, it is necessary to determine the parameter m, which represents the number of links each new node establishes when added to the network.

	Real (GC)	ER	BA
#Nodes	11547	11547	11547
#Edges	906232	906499	894582
Avg Node Degree	156.9640	157.0103	154.9462
Avg Degree Cent.	0.0136	0.0136	0.0134
Avg Closeness Cent.	0.0401	0.4753	0.4744
Avg Betweenness Cent.	0.0023	$9.562e^{-5}$	$9.623e^{-5}$
Avg Clustering Coeff.	0.7620	0.0136	0.0428

Table 1: Comparison of basic measures of the Giant Component with ER and BA graphs

It is defined as:

$$m = \frac{<k>}{2} = \frac{156.9640}{2} = 78.4820 \qquad (2)$$

where $\langle k \rangle$ is the average node degree of the network. The module used to generate the network requires an integer value, so the resulting m was rounded to 78.

In Table 1 some basic measures of the real network are reported and compared with those of the random ER and BA graphs, and in Figure 7 and 8 the comparison of the Degree Distribution of the ER and BA graphs with the real network.

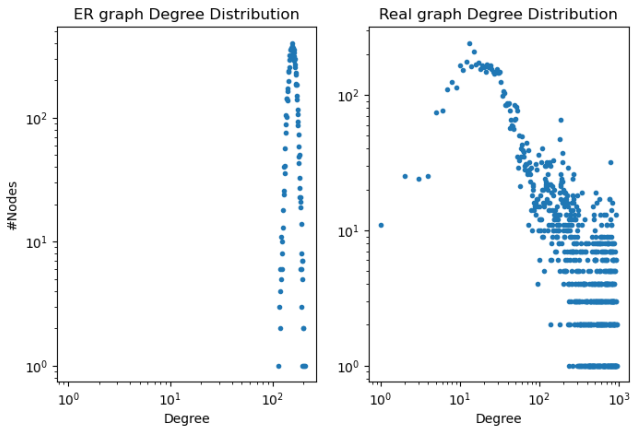

Figure 7: Degree Distribution comparison between ER graph and the GC

The regime of the ER graph is supercritical, as the value of p is greater than the threshold $\frac{\ln N}{N}$ (= 0.0008), meaning that we have a single giant component. It is also evident from the degree distribution (Figure 7) that it does not represent the same edge distribution as the real network, which allows us to rule out the possibility that the edges of the real network connect nodes randomly.

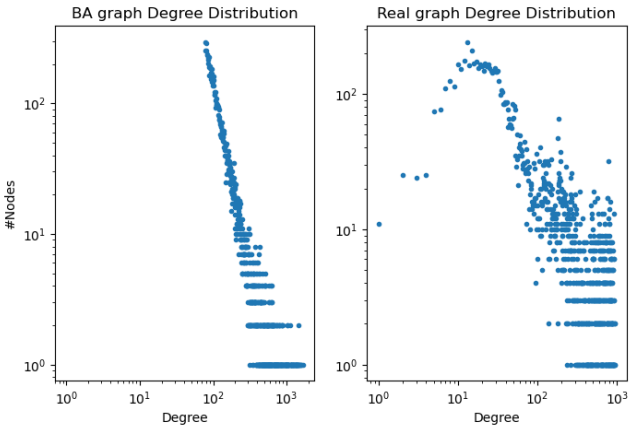

Figure 8: Degree Distribution comparison between BA graph and the GC

The BA graph, on the other hand, is much closer to the real network (Figure 8). It also consists of a single giant component. The fact that this graph is more similar to the giant component of the real network suggests that the way the nodes of the network are connected somewhat resembles the preferential attachment mechanism. However, the clustering coefficient of the BA graph is significantly lower than that of the real network, indicating that the latter has much more tightly connected communities.
Nevertheless, it is still evident, especially from the values of closeness and betweenness centrality, and the clustering coefficient, that neither of the two models can fully capture the structure of the real network.

4 OPEN QUESTION

After analyzing the network of third-sector organizations, this final section shifts the focus to the following questions: how can collaboration among organizations in the area be improved? Which organizations should be encouraged to collaborate?
To answer these questions as realistically as possible, special attention was given to the different types of organizations. First, it is more likely that organizations of the same type share similar areas of interest. Additionally, funding opportunities for activities often target specific categories of organizations. In the considered network, two organizations are connected if they are geographically close to each other, based on the assumption that proximity implies a potential shared network of contacts that could facilitate collaboration. The goal is to leverage these possible relationships to identify new pairs of organizations that could collaborate, effectively translating into new edges in the graph.

A preliminary analysis of the distribution of organizations in the area (Figure 9) highlights a significant presence of 'APS' ('Associazione di Promozione Sociale') and 'OV' ('Organizzazione di Volontariato') types of ETS. To further visualize this distribution, the organizations were plotted on a Folium Map, using different icon colors for each type (a sample portion of the map is shown in Figure 10).

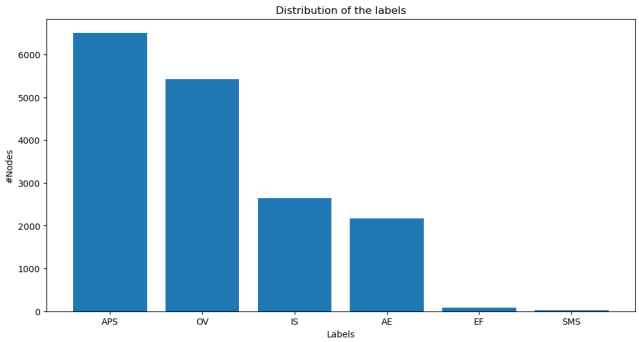

Figure 9: Label distribution

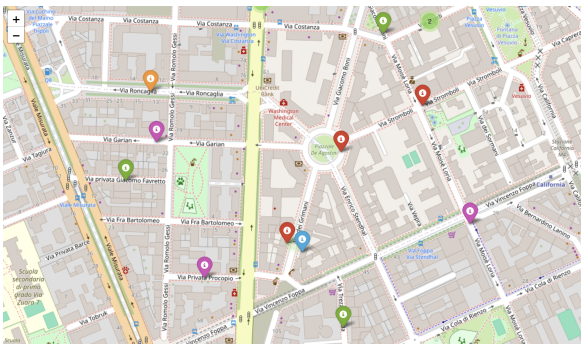

Figure 10: Section of the Folium Map visualizing the different types of organizations

The study for identifying new edges was conducted by restricting the analysis to the giant component, which was divided into a training subgraph (all the nodes and 80% of the edges) and a test subgraph (all the nodes and 20% of the edges).

Various link prediction algorithms are available, each based on different structural properties of the network[2]. To determine which method best suited the network, three different algorithms were tested, each relying on different measures: Common Neighbors and Adamic-Adar (neighborhood-based measures) and Katz (a path-based measure). The performance of the algorithms was evaluated using precision, recall, and F1-score. The results are shown in Table 2.

Considering both the performance and execution time of the different algorithms, the common-neighbor-based approach was chosen to identify new edges. The model was

	Precision	Recall	F1-score
Common Neighbors	0.0817	0.7707	0.1477
Adamic-Adar	0.0817	0.7707	0.1477
Katz	0.0233	0.7709	0.0453

Table 2: Performance of different link prediction algorithms

then trained on the entire subgraph of the giant component to predict new edges. The total number of predicted edges is 1,559,283. Figure 11 presents the heatmap showing the number of possible new edges connecting each pair of organization types.

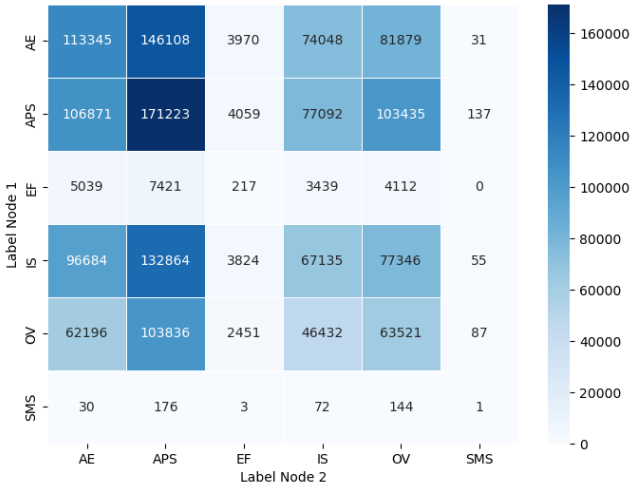

Figure 11: Heatmap of possible new edges for each pair of organization type

The next step was to focus on the newly predicted edges that connect nodes representing organizations of the same type. By narrowing down the edges to this category, the number of new possible edges identified is 415,442 (also visible by summing the numbers on the diagonal of the heatmap in Figure 11). Figure 12 shows the distribution of scores associated with the new links. As can be seen from this distribution, many of the predicted edges (16,929) have an associated score of 1, indicating that the two connected nodes share a single common neighbor, so even though they do have a way of connecting, it's not that probable. Figure 13 presents the number of edges associated with each different type of organization, and we can see that the majority are labeled as 'APS', which is consistent with the fact that the number of nodes with this label is the largest.

In the context of potential collaboration projects, it makes sense to focus on the pairs of organizations that are more likely to collaborate. Therefore, the same type of graph was

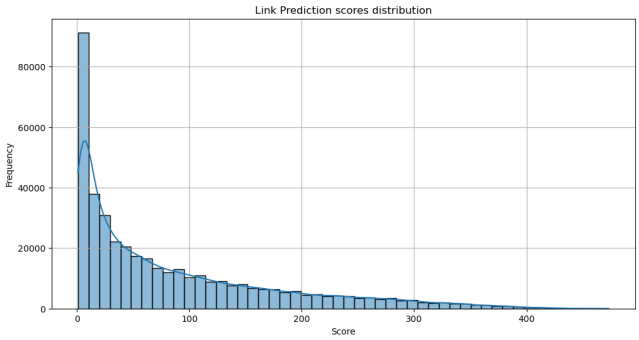

Figure 12: Score distribution edges between same type nodes

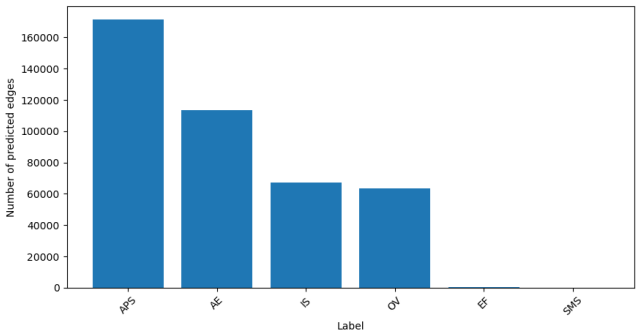

Figure 13: Distribution of labels in all predicted edges between same type nodes

repeated, considering only the Top 10% of predicted edges with the highest scores (Figure 14), resulting in 41,544 edges. We can see that, by considering only these edges, the most numerous category becomes 'AE' ('Altri Enti del Terzo Settore'), which is not one of the largest categories in terms of the number of organizations belonging to it.

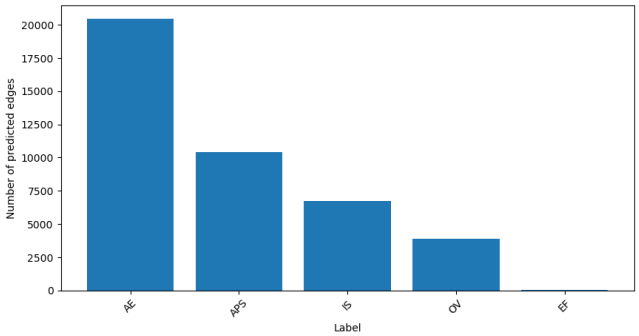

Figure 14: Distribution of labels in the Top 10% of edges between same type nodes

5 DISCUSSION

The aim of this work was to identify new potential connections between third sector organizations in the Lombardia region in order to facilitate new ties within the territory and promote collaboration among those serving the common good. As presented earlier, the data show that collaboration is not always easy, but it is necessary to comprehensively address the needs of the territory.

The decision to connect nodes when the distance between their locations is less than 2 km is a rather restrictive choice, but it was necessary due to the high computing time that would have been required to consider wider distances. Also, we cannot ensure that even if two organizations are located very close to each other (making them connected in this graph) they actually have common acquaintances that facilitate collaboration. Similarly, it is not guaranteed that two organizations of the same type, between which an edge is produced, would be interested in collaborating: they might have different areas of interest and action.

Nevertheless, the links predicted by this algorithm can serve as a good starting point to connect different organizations that might otherwise not know each other.

REFERENCES

[1] G. Marocchi. Pubbliche amministrazioni e terzo settore tra competizione e collaborazione. *Welfare Oggi*, (2 (Focus - verso l'amministrazione collaborativa)), 2018.

[2] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, (58.7), 2007.