Identifying patterns and evolution of roles in CO2 emissions

Loredana Salvatore, Digital Humanities and Digital Knowledge, ID number 0001056967

Github repository: https://github.com/Salvadana/NetworkAnalysisProject

1. Introduction

In the context of environmental sciences, the focus is more and more pertaining to the study of various factors and phenomena regarding the analysis of greenhouse gases and their roles in the Earth's ecosystem. Their concentration levels in the atmosphere represent a significant contribution to global warming and to the climate change crisis and have been on the rise due to human activities such as the burning of fossil fuels, deforestation, and industrial processes. This project focuses on the analysis of greenhouse gas emissions, involving the comprehensive collection, careful analysis, and meaningful interpretation of data pertaining to CO2 production in particular. The analysis has as a specific application the assessment and evaluation of CO2 emissions originating from particular industrial sectors and countries.

2. Problem and Motivation

The growing urgency of environmental concerns related to global warming has reached critical levels, posing substantial threats to the delicate ecosystems that sustain human life. Central to these concerns are carbon emissions, which have profound implications not only for the environment but also for the global economy. While it's imperative to pinpoint the drivers of carbon emissions, it's equally crucial to comprehend how the various roles played by countries on the world stage impact these emissions. Countries assume diverse roles in the generation of carbon emissions, and these roles can evolve over time. Exploring these roles through a network analysis lens provides valuable insights that can inform strategies to mitigate emissions. Understanding how a country's position within global perspective influences carbon emissions is particularly pertinent for crafting effective emission reduction plans.

The aim of this project is to explore two distinct networks representing two different moments in time. The chosen dataset spans indeed from 1970 to 2022, enabling to conduct comparative analyses across different timeframes for a more holistic understanding. This exploration seeks to illuminate the relationships between countries and industrial sectors while identifying noteworthy patterns or clusters of interest.

3. Datasets

The project will use the IEA-EDGAR CO2 dataset, a component of the EDGAR (Emissions Database for Global Atmospheric Research) Community GHG database version 7.0 (2022).

The dataset is publicly available and is based on data from IEA Greenhouse Gas Emissions from Energy.[1]

These data provide estimates for emissions of the three main greenhouse gases (CO2, CH4, N2O) and fluorinated gases per sector and country [2]. For the purpose of this project, the considered data are the one regarding the CO2 emissions. The data are available in CSV format and have therefore been stored and read as Pandas DataFrames from which information about nodes and edges have been extracted. In particular, when creating a graph through NetworxX, nodes have been instantiated starting from the columns identifying countries and industries.

Two **bipartite undirected networks** have been created, representing countries and industrial sectors at different points in time (in 1970 and in 2019). The two years have been selected in order to investigate a wide enough timespan allowing to highlight changings and the evolution of the situation. The 2019 has been selected as the last year in order to consider the pre-pandemic emission data, thus excluding possible reduction in emissions due to the lockdowns happened. The two graphs have been stored as *.gephx* files.

Bipartite networks are distinctive because they involve two distinct sets of nodes, where connections occur only between nodes from different partitions. In the networks created each country is connected to the industrial sectors in which it produces emissions: the two node sets are therefore composed by countries and industries. Weighted edges containing information about the amount of CO2 emissions connect each country to the different industrial activities registered for that country in a given year.

The results of the measures applications on the two networks have been compared in order to search for relevant differences in the selected time periods in order to highlight the changing of the network features. Additionally, the creation of **other four graphs** in the last phase of the workflow has been performed to provide a final wider overview of the evolving roles of key nodes.

4. Validity and Reliability

A dataset is considered valid when it accurately represents the real-world phenomenon it aims to measure. Therefore, the network models constructed from the data have been conceived to faithfully reflect the actual relationships and interactions between the nodes in the studied systems.

Validity errors such as omission and commission have been avoided through careful data collection and cleaning: for example, when the same industrial sector appeared linked to a specific country two times, due to a distinction regarding the fossil or non fossil source, the total amount of emissions has been summed for each industrial activity of every given country. The total so obtained has been used as a starting point for creating the edges, thus avoiding compromising the weights data. In this way no connection between a country and an industry has been lost, replaced or partially reported, since every edge records the total amount of emissions according to the data.

Regarding the reliability of the study, the use of objective data and measures aims to minimize subjectivity and reduce the potential for bias. The EDGAR made the data publicly available through a Creative Commons Attribution 4.0 International (CC BY 4.0) licence.

[1] Available at www.iea.org/data-and-statistics , as modified by the Joint Research Centre.

[2] https://edgar.jrc.ec.europa.eu/dataset_ghg70#p1

The measures have been applied referring to authoritative sources for the algorithms, based on solid theoretical premises and recognised in the field of Network Analysis. This ensures the obtaining of the same results given the same starting conditions.

5. Measures and Results

Cohesion

- **Density**

The term "cohesion" in network analysis refers to the extent to which nodes within a network are connected to each other. One straightforward way to measure cohesion is by using "density", which is defined as the ratio of the actual number of connections in a network divided by the total number of possible ties.
This measure can be applied to bipartite networks through the “density” method of the bipartite module by NetworkX, helping to understand how closely interconnected the two sets of nodes are. In the case of bipartite networks, the number of potential connections between nodes from different partitions is determined by the product of the sizes of the two partitions. The `nx.bipartite.density` takes this into account, considering the number of possible connections between nodes from different partitions.
Considering the similar dimension of the networks created, both constituted by 234 nodes and about 3000 edges, the comparison through this simple measure can be performed without risking of comparing situations and connection probabilities that are too different one from the other. The obtained results show an increase in density from 0.61 to 0.66, suggesting an increase of industrial activity of the involved countries over time that is aligned with the registered increase of emission due to anthropic activities.

Centrality

Centrality measures help in identifying e the most important or central nodes in a network.
The bipartite module of NetworkX provides also the possibility to compute Degree, Betweenness and Closeness centrality normalized for bipartite graphs. Therefore, these measures have been chosen over the computations on the projected graphs.

- **Degree centrality**

In an undirected network, the degree of a node is the number of edges connected to it.
In unipartite networks, degree centrality values are typically normalized by dividing each node's degree by the maximum possible degree, which is represented by n-1, where n is the total number of nodes within the network.
However, in bipartite networks, this normalization differs slightly. Here, the maximum possible degree of a node within one node set is determined by the number of nodes in the opposite node set. This adjustment accounts for the unique structure of bipartite networks, where nodes in one set interact exclusively with nodes in the other set.

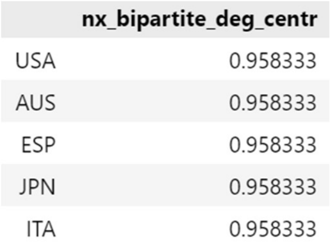
Therefore, the degree centrality for a node v in the bipartite sets U with n nodes and V with m nodes is

$$d_v = \frac{deg(v)}{m}, \text{for} v \in U,$$
$$d_v = \frac{deg(v)}{n}, \text{for} v \in V,$$

where deg(v) is the degree of node v[3].

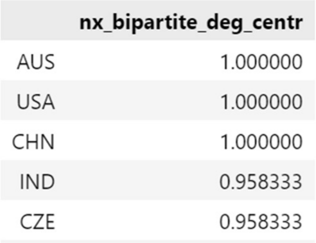
The degree centrality therefore allowed to examine the nodes having the higher number of connections.

The obtained results over the country nodes sets show a slight variation of the most central countries according to the degree: they passed from USA, Italy, Australia, Japan and Spain in 1970 (figure1) to a clear predominance of USA, China and Australia in 2019 (figure 2). This suggests a mostly constant prominence of the USA in the industrial activity, together with increased importance of China in the considered timespan.



	nx_bipartite_deg_centr
USA	0.958333
AUS	0.958333
ESP	0.958333
JPN	0.958333
ITA	0.958333

1.



	nx_bipartite_deg_centr
AUS	1.000000
USA	1.000000
CHN	1.000000
IND	0.958333
CZE	0.958333

2.

Though important, the degree of a node is not sufficient to evaluate its centrality in weighted networks as the ones created. It is indeed of key importance to consider the amount of emissions linked to specific countries and industries in order to understand their overall impact and role in the CO2 gas production. Therefore, some other centrality measures have been taken in consideration: the weighted degree, together with the Laplacian centrality, have been applied separately to the nodesets of the bipartite graphs allowing to discover a changing of predominance in both industries and countries' roles.

- **Laplacian centrality**

Laplacian centrality goes beyond the immediate neighborhood of a node, revealing deeper structural insights into connectivity and density within a network.

“The Laplacian energy is defined as

$$E_L(G) = \sum_i \lambda_i^2$$

where λ_i's are eigenvalues of the Laplacian matrix of weighted network G. The importance of a vertex v is reflected by the drop of the Laplacian energy of the network to respond to the deactivation (deletion) of the vertex from the network […] The basic idea is that the importance (centrality) of a vertex is related to the ability of the network to respond to the deactivation of the vertex from the network. ”[4].

[3] NetworkX documentation: https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.bipartite.centrality.degree_centrality.html#networkx.algorithms.bipartite.centrality.degree_centrality

[4] Xingqin et al., 2012

Unlike conventional centrality measures like degree, closeness, or betweenness, Laplacian centrality serves as an intermediate measure, bridging the gap between global and local assessments of a vertex's importance. This means that this measure provides a more nuanced understanding of how central a node is within the broader network context, taking into account both local and global aspects of its connectivity.

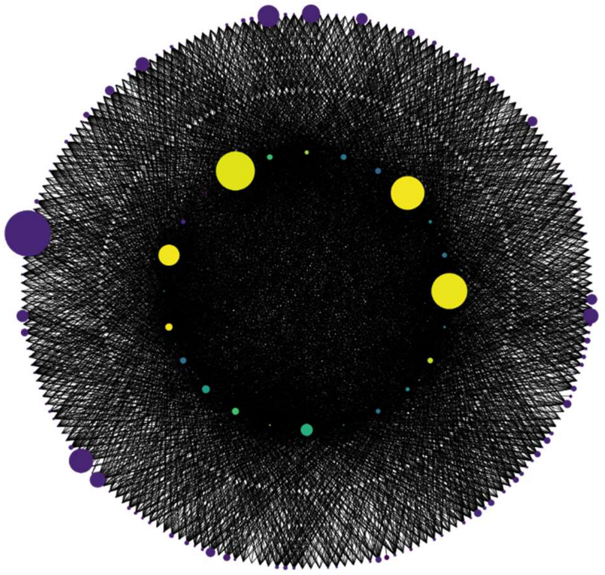


Figure 1: Laplacian centrality plot of the 1970 Graph

By offering insights on how well a node is connected within a network, Laplacian centrality takes into account not only the number of connections a node has but also the quality of those connections.
These measure has allowed to identify the increased importance of India and the overcoming role of China over time, as well as the changing pattern of predominance of industrial emissions per sectors, as it will be discussed in the conclusion..

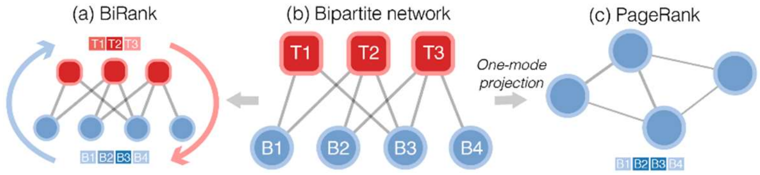
- **Weighted degree**

Through the use of the parameters ‘nbunch’ and ‘weight’ by NetworkX, the weighted degree has been computed, offering the possibility to examine more deeply the node roles in the networks. Its results are mostly aligned with the other weighted measures applied, as it will be shown in fig. 2.

- **Birankpy centrality measures**

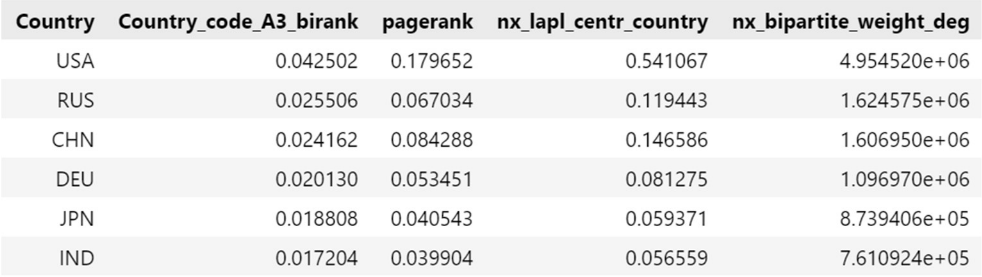
When calculating node centrality measures in bipartite networks, a possible approach is to apply PageRank on the one-mode projection of the network. However, the projection could cause information loss and distort the network topology. Therefore, for better node ranking on bipartite networks, it has been chosen a ranking algorithm that fully accounts for the topology of both modes of the network. The BiRank package (https://pypi.org/project/birankpy/), which implements bipartite ranking algorithms, has been used to perform centrality measures with different normalizers, among which the HITS and the ‘Birank’ one, to apply to the bipartite Networks. The BiRank package contains several ranking algorithms that generalize PageRank to bipartite networks by propagating the probability mass (or importance scores) across two sides of the networks. [5]

[5] Yang et al., 2020



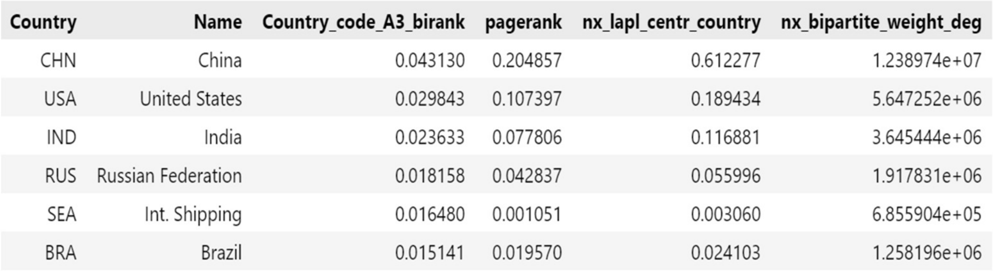
Furthermore, it allowes to apply the pagerank on a unipartite projection and compare the results with the other rankings. This algorithm allows to include the edge weights in the computation too, allowing to preserve the whole information contained in the Graph.

A certain alignment has appeared in the measures results taking into account the node weights, as it can be observed in the following tables reporting the Birank, Pagerank. Laplacian Centrality and Weighted Degree Centrality of the top five nodes in the networks. Slight variations in the ranking regard mostly the Laplacian Centrality, and could be explained with its peculiar nature including the global and local vertex importance.



Country	Country_code_A3_birank	pagerank	nx_lapl_centr_country	nx_bipartite_weight_deg
USA	0.042502	0.179652	0.541067	4.954520e+06
RUS	0.025506	0.067034	0.119443	1.624575e+06
CHN	0.024162	0.084288	0.146586	1.606950e+06
DEU	0.020130	0.053451	0.081275	1.096970e+06
JPN	0.018808	0.040543	0.059371	8.739406e+05
IND	0.017204	0.039904	0.056559	7.610924e+05

a.



Country	Name	Country_code_A3_birank	pagerank	nx_lapl_centr_country	nx_bipartite_weight_deg
CHN	China	0.043130	0.204857	0.612277	1.238974e+07
USA	United States	0.029843	0.107397	0.189434	5.647252e+06
IND	India	0.023633	0.077806	0.116881	3.645444e+06
RUS	Russian Federation	0.018158	0.042837	0.055996	1.917831e+06
SEA	Int. Shipping	0.016480	0.001051	0.003060	6.855904e+05
BRA	Brazil	0.015141	0.019570	0.024103	1.258196e+06

b.

Figure 2: Overview of the centrality measures in 1970(a) and 2019 (b)

Structural equivalence

Structural equivalence is a count of the number of common neighbours two nodes have.
Similarity measures based on structural equivalence have been applied to provide insights into the degree of similarity between countries in their emissions behavior within shared industrial sectors, in order to identify similar countries and distinguish between the most impactful ones and the least impactful ones. More specifically, the Jaccard's coefficient and the Simrank similarity metric by NetworkX have been chosen.

- **Simrank**

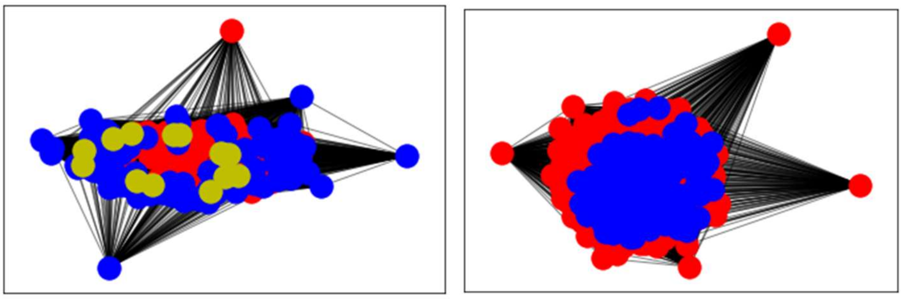
SimRank is a similarity metric according to which two objects are considered to be similar if they are referenced by similar objects.[6]
From these measure's results it can be noticed that the most central countries are more similar to countries that, according to the computed centrality measures, have a lower centrality.

- **Jaccard's coefficient**

For the Jaccard coefficient of two nodes i and j corresponds to the number of common neighbours divided by the total number of distinct neighbours of both nodes.
Also according to the Jaccard's coefficient on the projected Networks, the most central countries tend to result more similar to a certain group of countries which have a different and often lower centrality relevance.

It must be observed that in this computations the similarity depends on the number of common neighbors and does not take into account the weights of the edges connecting countries. Thus, the most similar countries are the ones having the most similar neighbors, e.g. meaning the same industrial activities for countries. It is still interesting, cause it allows to highlight countries wich are mostly tied to certain industrial sectors and contribute to increase their CO2 emissions.

Community detection



Louvain community detection

The Louvain method for community detection is a method to extract non-overlapping communities from large networks.[7] This method optimizes modularity metrics, which measure the extent to which similar nodes are likely to connect to each other. The goal of the algorithm is to find the best way to group nodes in a network to maximize this modularity value.

Since testing all possible node groupings is impractical, heuristic algorithms are employed: the process starts by identifying small communities through local optimization of modularity for all nodes and then aggregates the small communities into single nodes, finally repeating again the optimization process.

This measure has been applied on the weighted projections of the considered graphs performed on the country node sets. Three communities have been detected in the 1970 graph, while two have been found in the 2019 graph, thus pointing to a possible assimilation of countries' industrial activity over time.

[6] Jeh and Wisdom, 2002
[7] Blondel et al., 2008

Figures 3 and 4 show the probability density functions of the countries communities of 1970 and 2019. It can be noticed that the different communities in both timespans split the countries into groups that reach higher PDF values, meaning a most central role in emissions, opposed to other groups that are mostly concentrated over lower values, meaning a lower impact on CO2 emissions.

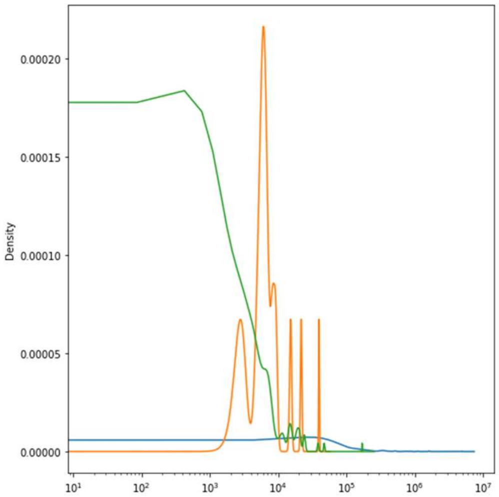


Figure 3: 1970 communities' PDF

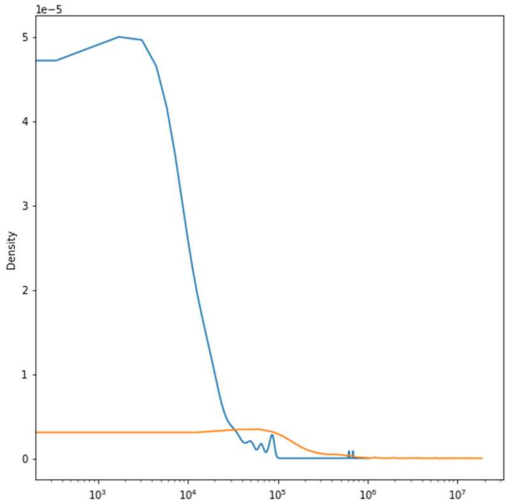
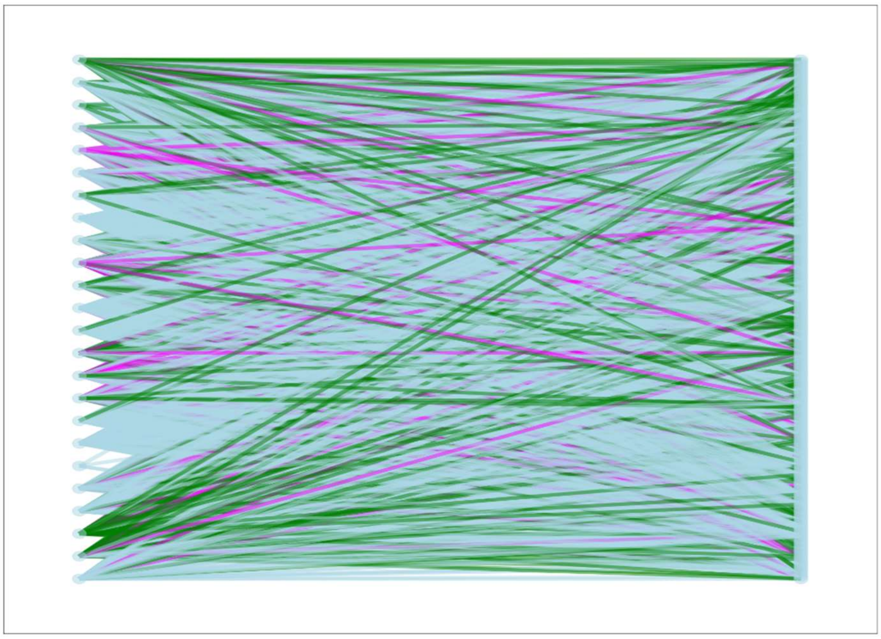


Figure 4: 2019 communities' PDF

Graphs comparisons

Finally, more direct comparisons have been performed on the graphs. The computed difference between the two networks has shown an increase of about 300 units in edges, represented by the green edges in the following graph composition joining together the two networks:

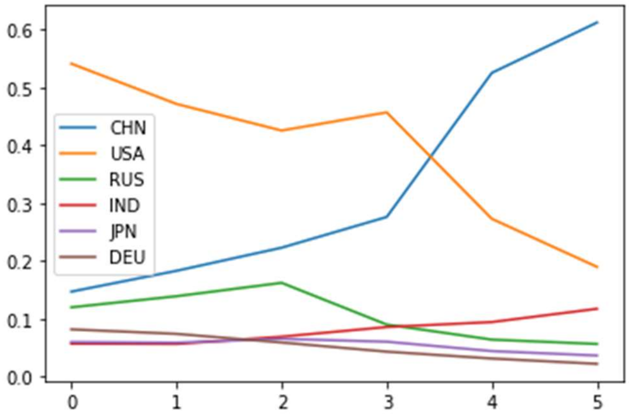


This increase in connections confirms the increase in density initially computed, showing an higher industrial activities of countries over time.

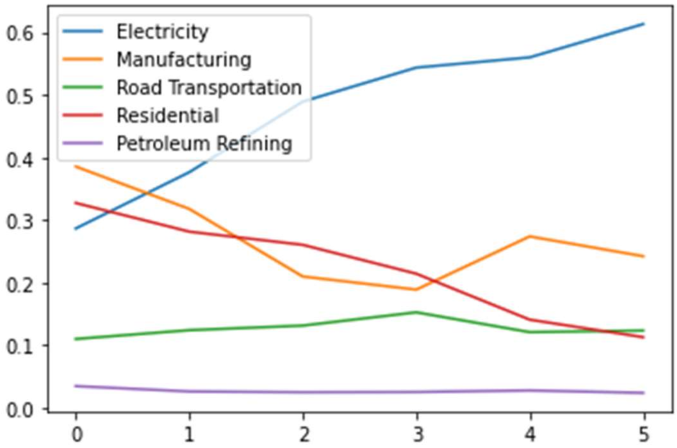
6. Conclusion

To allow a deeper analysis of the roles played by countries and industrial sectors, a final comparison has been performed including additional graphs.

Starting from 1970, a graph has been created following ten-years intervals until 2019(thus, having 9 years in the last interval). The Laplacian centrality has been chosen as a weighted centrality indicator to compute and visualize for each of the previously mentioned graphs over time, thus showing some interesting results: It can be observed that China has overcome USA across time, while other countries like Russia, Germany and Japan have decreased their relevance in the network (in contrast, India acquired relevance overcoming all three of them). These results confirm the tendencies highlighted by the centrality measures computed in the previous phases of the study.



Also in the industrial activities, the Electricity sector has evolved in its centrality shifting above all the other ones, while for example Manufacturing and Residential industries have decreased their relevance in the CO2 production panorama.



These additional results confirm the tendencies individuated applying the described centrality measures. It must be nonetheless observed that a decreasing in centrality does not necessarily coincide with a lower CO2 production, but only with a less important role in the overall emissions.

7. Critique

The conducted research makes a valuable contribution by shedding light on the pivotal roles played by countries and industrial sectors in the realm of Greenhouse Gas Emissions. It not only offers insights into the intricate structure and interconnections within this complex system but also lays the foundation for more in-depth analyses, particularly in terms of network clustering and segmentation.

However, there is room for further investigation. Exploring emission quantities in greater detail could uncover additional nuances, as well as a more comprehensive exploration of clustering and community detection may reveal deeper insights and trends within both countries and industrial communities. This suggests that there are exciting avenues for future research in this field.

Bibliography

Blondel, Vincent D; Guillaume, Jean-Loup; Lambiotte, Renaud; Lefebvre, Etienne (9 October 2008). "Fast unfolding of communities in large networks". *Journal of Statistical Mechanics: Theory and Experiment*. **2008**

G. Jeh and J. Widom. “SimRank: a measure of structural-context similarity”, In KDD’02: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 538–543. ACM Press, 2002.

Xingqin Qi, Eddie Fuller, Qin Wu, Yezhou Wu, Cun-Quan Zhang,
Laplacian centrality: A new centrality measure for weighted networks,
Information Sciences,Volume 194, 2012, Pages 240-253.

Yang KC, Aronson B, Ahn YY. BiRank: Fast and Flexible Ranking on Bipartite Networks with R and Python. J Open Source Softw. 2020.