# The Complex Interactions of the United Nations Sustainable Development Goals

Benjamin Kemp
MSc Data Science, University of Aberdeen
2021/2022

## SUPPLEMENTARY MATERIAL

### 1. Tools

Our main tool for analysis was python. The analysis was carried out in python using many different publicly available packages. ‘Numpy’ for numerical applications[1]; ‘Scipy’ for correlation[2]; ‘Pandas’ for dataframe to import, store and display data, as well as performing correlation[3]; ‘NetworkX’ was used to display networks and perform centrality calculations[4]; ’Matplotlib’ was used for plotting graphs with options for changing legends, axis labels and combining figures[5]; ‘CDLIB’ was used to perform a community detection algorithm on the data[6]; ‘Pickle’ was used for saving and importing data in python friendly formats[7].

### 2. Database

The country data is freely accessible on the UN stats website[8]. There is data on 261 countries or regions for 17 goals and more than 210 indicators. The available data depends on how good the data collection and storage processes are in each country with more developed countries typically having more available data as they have more money to spend on collection methods. The SDGs were set up in 2015 and since then data for more indicators has become available as countries are aware of the indicators and data they need to track relevant to each goal.

### 3. Code

Code relevant to thesis is provided and contains the information as follows:

1) Data notebook – importing excel data for goals, cleaning data and applying signs, performing correlation, grouping indicators names into goals for later analysis and methods to explore individual goals or indicators
2) Eigenvector centrality – Calculate eigenvector centrality for separate synergy and trade-off networks, plotting eigenvector centrality graphs, plotting network graphs related to eigenvector centrality
3) Community Detection – Perform community detection algorithm, community classification, investigate communities, plot UN SDG classified network and plot community detection classified network

4) Delayed Correlation – Perform correlation and shifted correlation on key (selected) node with all connections, determine difference between shifted correlations, calculate number of indicators with an effect on key node and those the key node effects and finally method of visualising weak causality

Other code not used in thesis but was created during research is as follows:

1) Betweenness centrality – calculate betweenness centrality for separate synergy and trade-off networks, plotting betweenness centrality graphs, plotting network graphs related to betweenness centrality (**Was not used as decision was to focus on one form of centrality**)
2) Eigenvector centrality over time – Data split into two periods (2000-2010 and 2010-2020), eigenvector centrality calculated for each period and overall rank compared with rank in each period (**Not used as due to lack of data in first period many indicators are not able to be considered so analysis is extremely limited, many key eigenvector indicators are not included in analysis**)
3) Betweenness centrality over time - Data split into two periods (2000-2010 and 2010-2020), betweenness centrality calculated for each period and overall rank compared with rank in each period (**Not used as due to lack of data in first period many indicators are not able to be considered so analysis is extremely limited, many key betweenness indicators are not included in analysis**)

All analysis carried out on Brazil, India and China is identical other than naming of variables and initial data used, so only one set of code relating to Brazil is provided as this was the initial main focus.

## 4. Betweenness Centrality

There are many ways to evaluate a network graph with eigenvector centrality being just one method. Another possibility is looking at betweenness centrality, which is a method of determining the influence a node has over the exchange of information in the network[9]. It highlights the node that is most often used as a bridge in the shortest path between all pairs of nodes. This would identify key nodes that link indicators and goals but may work better for a directed network that allows you to specify the direction that information flows in the network.

| Rank | SDG Indicator | Description of Indicator | Betweenness Centrality Synergy Value |
|---|---|---|---|
| 1 | 17.4.1 | Debt service as a proportion of exports of goods and services | 0.0626 |
| 2 | 17.3.2 | Volume of remittances (in USD) as a proportion of total GDP (%) | 0.0482 |
| 3 | 8.3.1 | Proportion of informal employment (%) (**for Males not in agriculture**) | 0.0326 |
| 4 | 12.2.2 | Domestic material consumption per capita in tonnes (**for grazed biomass and fodder crops**) | 0.0290 |
| 5 | 17.1.1 | Total government revenue as a proportion of GDP (%) | 0.0245 |

**Table S. 1.** Top 5 ranked Synergy betweenness centrality nodes (Brazil)

For the top 5 ranked nodes in Brazil's synergy network in terms of their betweenness centrality (Table S. 1) has very different nodes from the eigenvector central network. A key SDG that appears in the betweenness centrality results is SDG 17 (partnerships for the goals). This is unsurprising as the goal is about partnerships and it is closely related to lots of other goals so it is clear through this analysis that it acts as a method of information flow between the other SDGs.

Although this does not give the node that has the strongest positive influence on other nodes, it instead provides the node that would likely have the influence on the greatest number of nodes as it is positioned such that it is reached by many nodes in the network.

Given more time it would have been useful to explore how the betweenness central nodes differ from the eigenvector central nodes and how the key betweenness nodes can be used to improve a country's progress towards the 2030 agenda.

## 5. Eigenvector Centrality over time

Exploring eigenvector centrality[9] rank changes over time shows how the importance of a node can vary between two periods. When comparing rank changes in eigenvector centrality synergy two periods of 10 years inclusive were taken (giving 11 years of data for each). First from 2000-2010 and then from 2010-2020, due to only taking correlation for indicators with 10 data pairs, taking 11 years allowed for the potential of 1 year with missing data to still be included.

For both figures the colours are not linked to SDG they are a part of, and the top 5 nodes are taken from the eigenvector centrality calculated for the whole 655/291 (Brazil/India) indicator network, where there is enough data for both periods.

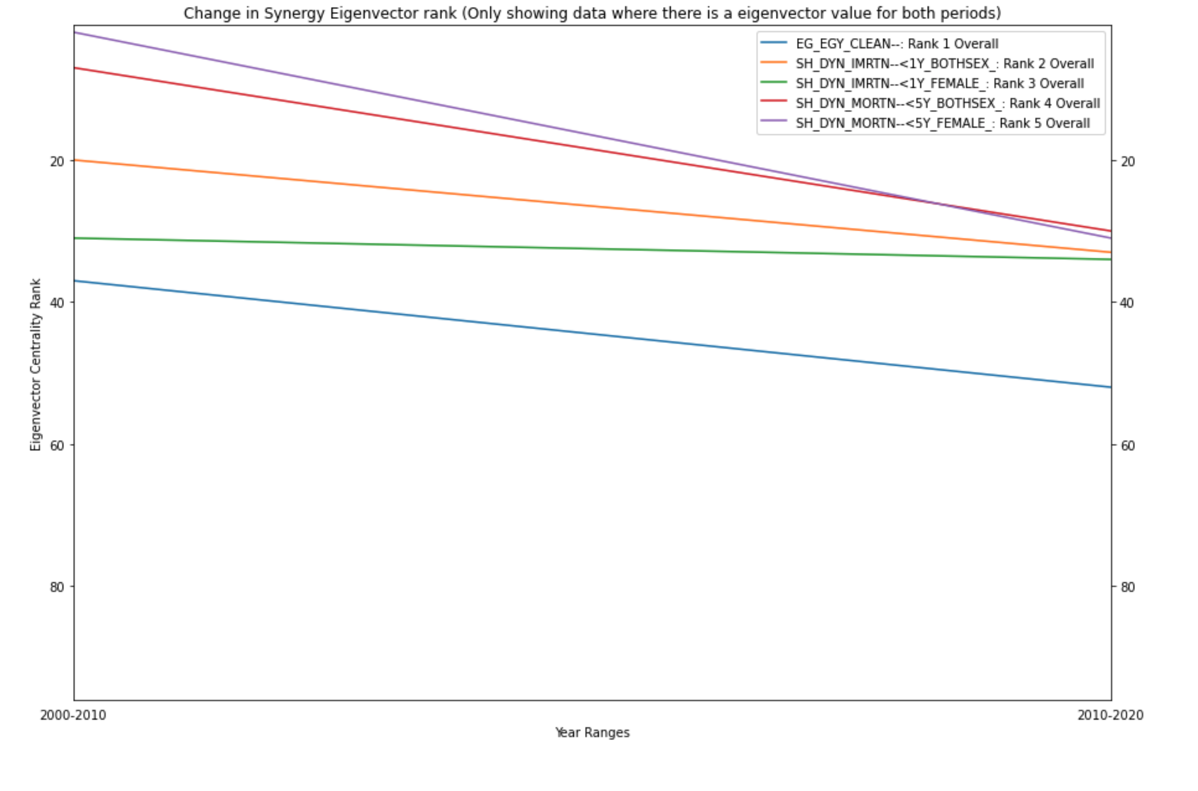

**Figure S. 1.** Changes in eigenvector centrality rank over time (Brazil)

This is again limited by available data as there are only 96 disaggregated indicators (for Brazil) that have enough data points to determine a correlation for both time periods. The top 5 nodes (Fig. S. 1) are displayed in the top right and none of these are the top 5 from the original network. Rank 1 is Primary reliance on clean fuels and technology (SDG 7), rank 2 and 3 are infant deaths (SDG 3) while rank 4 and 5 are Deaths under 5 years (SDG 3).

All nodes see a reduction in their rank between the first time period and the second, meaning that they are more linked and more strongly connected to nodes in the first period. The rank 1 node has a low importance in the first period at rank 37 and even less significance in the second period at rank 52. Due to the lack of data for so many indicators are it not possible to determine whether this decrease in influence is only present in the 96-indicator network or if it also transfers to the full 655 network. The fact all 5 indicators see a reduction in their influence suggests this may be true that certain indicators are becoming less important in describing a country's progress.

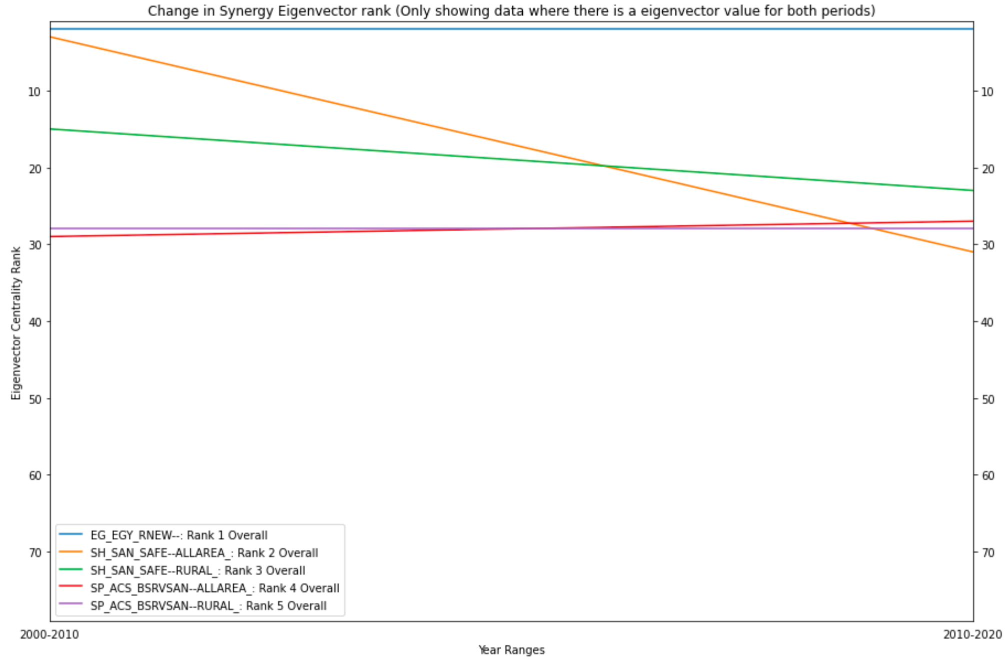

**Figure S. 2.** Changes in eigenvector centrality rank over time (India)

India suffers with the same data issues and has an even smaller 80 disaggregated indicators that can be examined for both time periods. Some of the top nodes (Fig. S. 2) for India are present in the top 5 original network and these are rank 1 which is installed renewable electricity-generating capacity (SDG 7/12), rank 2 and 3 related to population using safe sanitation services (SDG 6) and then 2 new nodes rank 4 and 5 which are about population using basic sanitation services (SDG 1).

For rank 1 this node is the top ranked node in the full 291 indicator network, and it remains the top ranked node for both time periods despite being in a smaller network. The other nodes see either a constant importance or a reduction reinforcing the idea that the key nodes overall may be becoming less influential as time progresses.

Splitting up the data separate smaller time periods allows performance of a single indicator to be tracked overtime to see if it becomes more/ less influential. For the limited data the top indicators begin and remain in the top half of the overall eigenvector centrality measurement, although some see a significant decrease in the rank in the most recent 10 years highlighting the importance of constant evaluation of the SDGs as more data becomes available and changes are made.

## 6. China Data

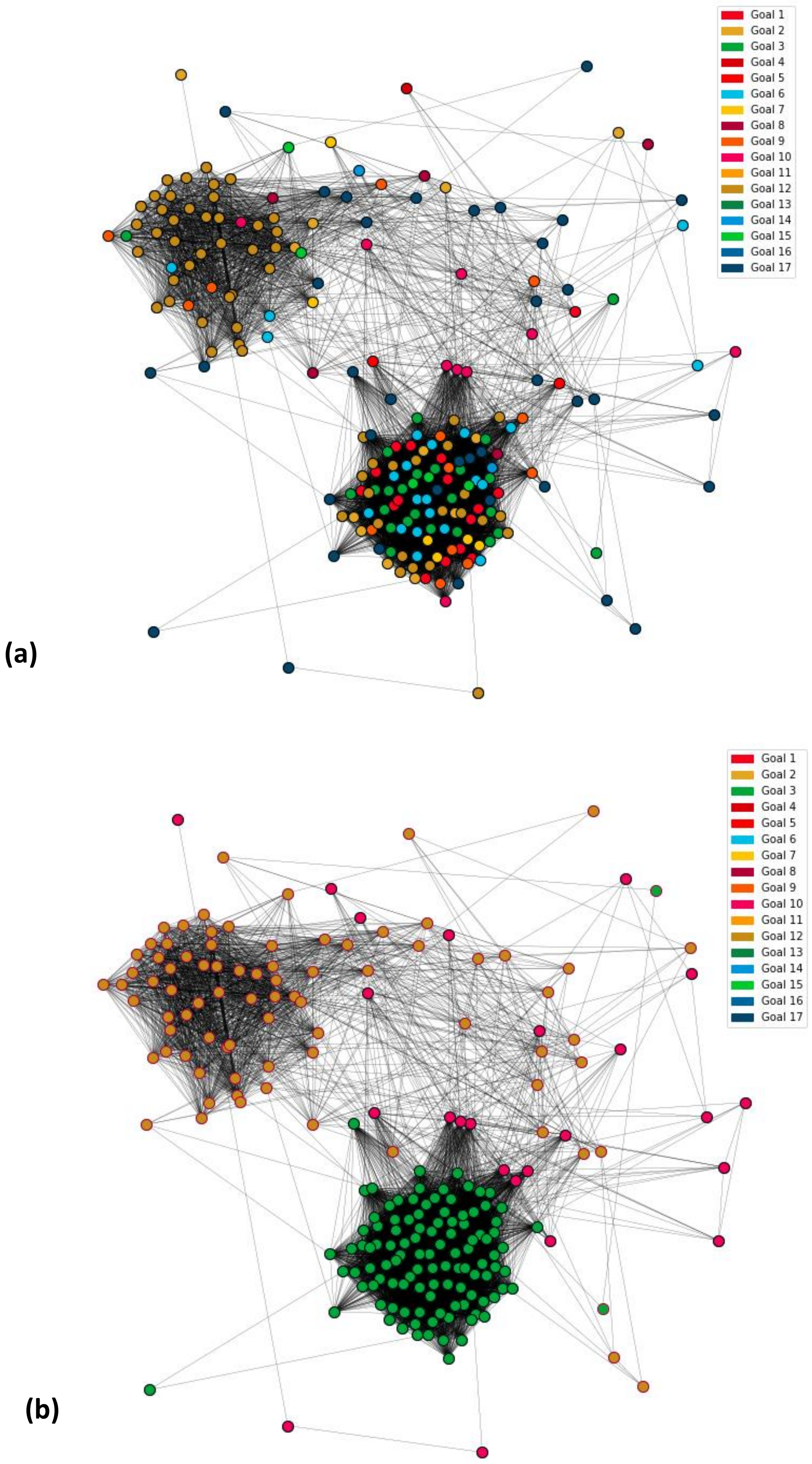

**Figure S. 3.** UN classification of all 17 Goals (a) and community detection (b) of China

Displaying China's network of indicators (Fig. S. 3a) produced the same results as Brazil and India, forming two distinct clusters. When applying the Leiden[10] community detection algorithm, the result (Fig. S. 3b) was extremely similar to the findings for India where the larger dense cluster was classified as SDG 3 and the smaller cluster classified as SDG 8/12.

## 7. Additional Discussion

The network evaluated was undirected since there is no significant causality measure available between the correlation pairs. If there was more data available to enable more concrete causality analysis this would allow the eigenvector centrality process to identify the individual indicators that would provide the largest positive/negative impact on the network. Even applying the weak causality method to all the data pairs in the network would allow an initial understanding of the most important nodes to be made.

## References

1. Numpy documentation https://numpy.org/
2. Scipy documentation - https://scipy.org
3. Pandas documentation - https://pandas.pydata.org
4. NetworkX documentation - https://networkx.org
5. Matplotlib documentation - https://matplotlib.org
6. Rossetti, G., Milli, L & Cazabet, R. CDLIB: a python library to extract, compare and evaluate communities from complex networks. *Appl Netw Sci* **4**, 52 (2019). https://doi.org/10.1007/s41109-019-0165-9
7. Pickle documentation - https://docs.python.org/3/library/pickle.html
8. UN Data - https://unstats.un.org/sdgs/dataportal/database
9. Centrality methods - https://towardsdatascience.com/unveiling-important-nodes-in-a-network-4992a2ea1cca
10. Traag, V.A., Waltman, L. & van Eck, N.J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* **9,** 5233 (2019). https://doi.org/10.1038/s41598-019-41695-z