JOINT SDG FUND

MEASURING THE INTEGRATION AND NETWORK EFFECT OF THE SDGS

G5055 Practicum Project 2

Columbia University, MA, Quantitative Methods in the Social Sciences

Contents

Core Team Members

Qinyue Hao
qh2231@columbia.edu

Jasmine Hwang
jh4452@columbia.edu

Dan Li
dl3466@columbia.edu

Peishan Li
pl2772@columbia.edu

Rina Shin
rina.shin@columbia.edu

Connie Xu
yx2625@columbia.edu

Hanyu Zhang
hz2697@columbia.edu

Supporting Team Members

Zhiwen Huang
zh2387@columbia.edu

Cara Latinazo
cara.l@columbia.edu

Xingchen Li
xl3005@columbia.edu

Soobin Oh
so2638@columbia.edu

Lizabeth Singh
ls3816@columbia.edu

Mengying Xu
mx2238@columbia.edu

Tianqing Zhou
tz2372@columbia.edu

Executive Summary

Context

The 17 Sustainable Development Goals (SDG's) are currently designed as a network wherein links among goals exist through targets and indicators which refer to multiple goals. The different domains are interconnected and cannot be effectively resolved without being considered as interdependent (Hoff 2011). These linkages should be evidence-based and measurable, but research conducted on the quantitative measurement of SDG linkages/network structure has thus far been sparse.

Potential Users

- Partnered leaders of UN agencies in charge of the 17 goals.
- Government officials of respective departments in each country hoping for improved institutionalization and policy integration in global governance.
- High-level political leaders looking to understand **trade-offs** and **secondary benefits** of SDG-oriented policy enactment (e.g., High-Level Political Forum).

Purpose

1. To explore the **linkages** between UN SDG Indicators.
2. To test for the **predictability** of indicator description similarity on the actual linkages between indicators.
3. To test for the **generalizability** of the indicator network model across countries.

Project Scope & Model Overview

Countries	- **For UN SDG Indicators: Indonesia** (Asia) & **Guatemala** (Central America) - The two countries are selected due to similarity in population density, political stability, etc. as well as (relative) data availability.**
Time Period	- **For UN SDG Indicators:** 2012-2020; - **For UN SDG Indicator Metadata:** Not applicable
Model Summary	- Build exploratory textual model using word embedding methods on SDG **indicator descriptions** accessed through SDG metadata - Build the a social network using correlations between non-disaggregated SDG measurements (composite at at **the indicator level**)

**Note that while there is relatively high non-missing data proportion of the countries of interest, countries of particular interest to the UN often are in the experiencing growth and development phase and have slightly more missing data on average. Additionally, many measures for indicators exhibit missing data for all countries across the selected time period.

Network Model Process Flow

Coefficient-based Network (SDG Indicator Measures)

Download data
from UN SDG Indicator

Clean and Preprocess Data

- Remove disaggregated measurements (Appendix A)
- Pivot on year
- Separate indicators

Text-Based Network (Metadata)

Visualize Text-Based Network for SDG Indicators
Using **tf-idf** similarity scores (Appendix D) on metadata descriptions to construct edges, set threshold for ties between indicators'.

Model Comparison
Between Text-and Coefficient- Networks with QAP Network

Compile Network Visualizations
Indicators = Nodes,
Coefficients = Edges
Color = Goal
Size = Centrality

***Composite Method for Indicator Measurement** (Adopted)*

Impute Missing Data
for PCA Generation
(Appendix C)

- Keep data with >1 year of nonmissing data

Perform PCA
The end result is one measurement per indicator; Recombine the dataset

Build the network

- Calculate the correlation among indicators
- Calculate degree centrality

For the Representative Method not adopted, see Appendix B.

Text Model: Indicators' Definition by Text

Text Model uses 246 indicators' definitions as data source and measures the similarity score between indicators

Index	definition
1.1.1	The indicator "proportion of the population below the international poverty line" is defined as the percentage of the population living on less than $1.90 a day at 2011 international prices.
1.2.1	The national poverty rate is the percentage of the total population living below the national poverty line. The rural poverty rate is the percentage of the rural population living below the national poverty li
1.2.2	The following four series are used to monitor the SDG 1.2.2. 1) Official multidimensional poverty headcount, by sex, and age (% of population) - The percentage of people who are multidimensionally poor 2
1.3.1	The indicator reflects the proportion of persons effectively covered by a social protection system, including social protection floors. It also reflects the main components of social protection: child and Last
1.4.1	The following key concepts were defined to support the indicator in the context of poverty eradication.Basic Services refer to public service provision systems that meet human basic needs including drinki
1.4.2	Indicator 1.4.2 measures the relevant part of Target 1.4 (ensure men and women have equal rights to economic resources, as well as access to ..., ownership of and control over land and other forms of Last r
1.5.1	This indicator measures the number of people who died, went missing or were directly affected by disasters per 100,000 population.
1.5.2	This indicator measures the ratio of direct economic loss attributed to disasters in relation to GDP.
1.5.3	An open-ended intergovernmental expert working group on indicators and terminology relating todisaster risk reduction established by the General Assembly (resolution 69/284) is developing a set ofindic
1.5.4	The Sendai Framework for Disaster Risk Reduction 2015-2030 was adopted by UN Member States inMarch 2015 as a global policy of disaster risk reduction. One of the targets is: "Substantially increase ther
1.a.1	Total official development assistance (ODA) grants from all donors that focus on poverty reduction as a share of the recipient country's gross national income. The OECD/Development Assistance Committee
1.a.2	Total general (local, regional and central) government expenditure on education (current, capital, and transfers), expressed as a percentage of total general government expenditure on all sectors (including
1.b.1	Proportion of government spending towards health and education and direct transfers which benefit directly the monetary poor. Government spending measures public expenditures on health and educat
2.1.1	The prevalence of undernourishment (PoU) (French: pourcentage de sous-alimentation; Spanish: porcentaje de sub-alimentación; Italian: prevalenza di sotto-alimentazione) is an estimate of the proportion
2.1.2	The indicator measures the percentage of individuals in the population who have experienced food insecurity at moderate or severe levels during the reference period. The severity of food insecurity, define
2.2.1	Prevalence of stunting (height-for-age <-2 standard deviation from the median of the World HealthOrganization (WHO) Child Growth Standards) among children under 5 years of age.

Definitions are extracted from the UN metadata PDF files.

Text Model: Methodology

Word Embedding Models convert text into vectors based on word frequency, document frequency, context of words, and etc. **TF-IDF, Word2vec, Doc2vec** as three most used models were chosen initially

Cosine Similarity is used to measure the cosine of the angle between two definitions' vectors, as a proxy for **similarity score** between definitions

EXAMPLE

Indicator 1.1.1

[input]: *"Indicator proportion population international poverty line defined…"*

Word embedding models

[output]: array([[0.03160266, 0.07794142, 0.0407558, …., 0.01125147])

Indicator 1.2.1

[input]: *"national poverty rate percentage total population living national…"*

Word embedding models

[output]: array([[0.14669026, 0.27332506, 0.04717977, …., 0.07041576]])

Calculate **Cosine Similarity** as **Similarity Score**

Range 0-1, 1 as perfect similarity

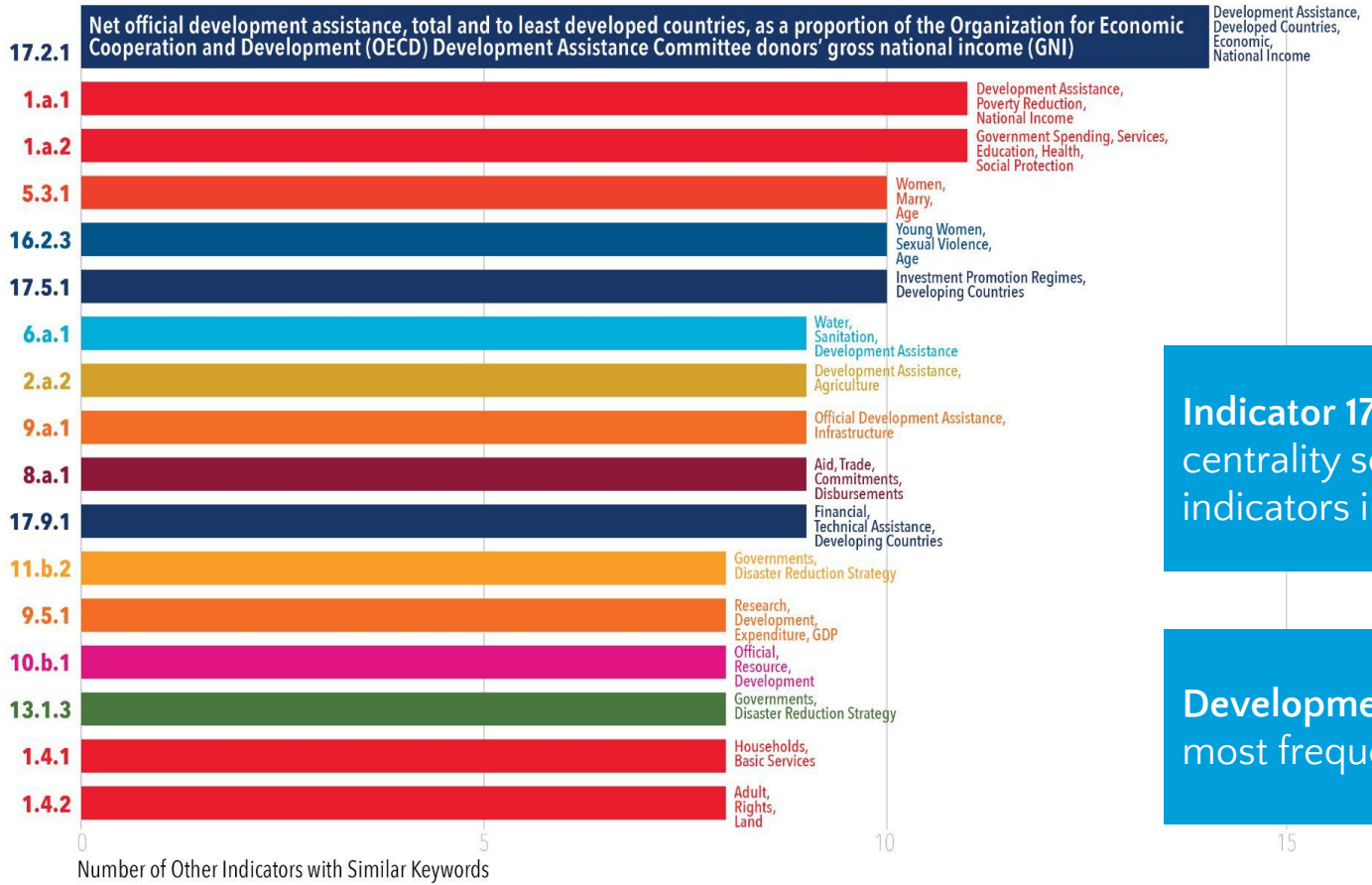
Word Embedding Model details in Appendix D.

Text Model: Methodology & Result

- **TF-IDF** is chosen as the final word embedding model, per the result of manual comparison between the three most frequently used models of TF-IDF, Word2vec, and Doc2vec.
- The similarity score of **0.20** is selected as the **threshold** to retain the indicators that we judged to be substantively similar on average. Indicators with similarity score >= 0.20 are considered to be related.
- A total of **704 links** between **196 indicators** are observed as a result.

More details on Word Embedding Model selection and similarity score threshold in Appendix D.

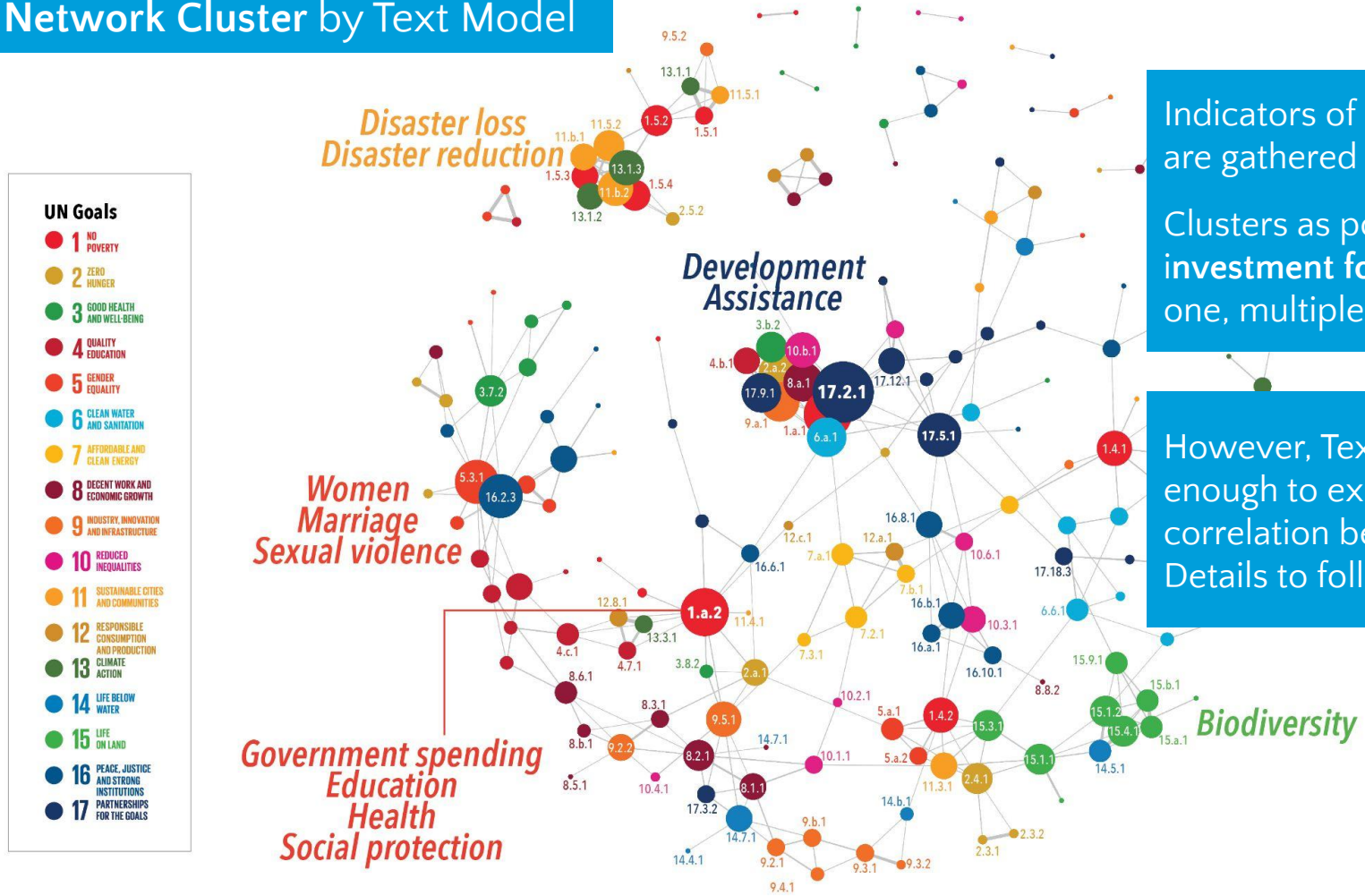
Indicator Centrality by Text Model



Indicator 17.2.1 has the highest centrality score, linking to the most indicators in the network (14 count).

Development and **Assistance** as the most frequently used keywords.

Network Cluster by Text Model



Indicators of similar keywords are gathered as a **clusters**.

Clusters as potentially **investment focus**,investment in one, multiple returns.

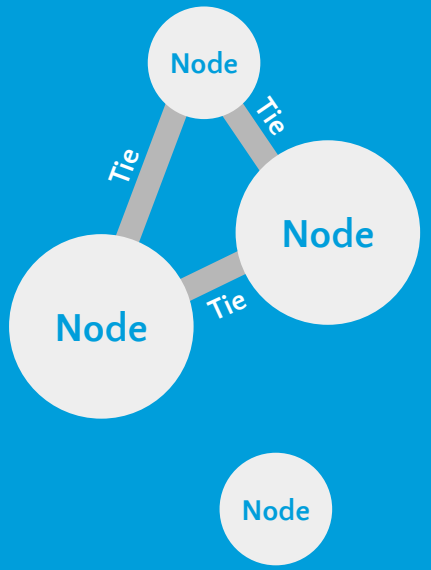
However, Text Model is ***NOT*** enough to explain the correlation between indicators. Details to follow.

Network Construction: Nodes & Ties to Present Linkages

- **Indicators** as nodes, **correlation relationships** between indicators as ties.

  correlation **coefficients** as tie weights.

- A tie is drawn **only** when a pair of indicators are **statistically significant**** related to each other, to ensure a more solid basis for the analysis and modeling.



**Statistically significant when the p-value from regression is < 0.05.

Network Construction: Dimensionality Reduction with Principal Component Analysis

- To construct a network on the indicator level, we need **one measure** for each indicator to do correlations.
- **Principal Component Analysis (PCA)** is a fast and flexible unsupervised method for dimensionality reduction. We use it to **create composite measures** when there are more than one measure under some indicators.
- We projected the original data in a lower-dimensional space, while preserving most of the variance - here we keep **the first principal component** as the representative of each indicator, which accounts for most of the original information.

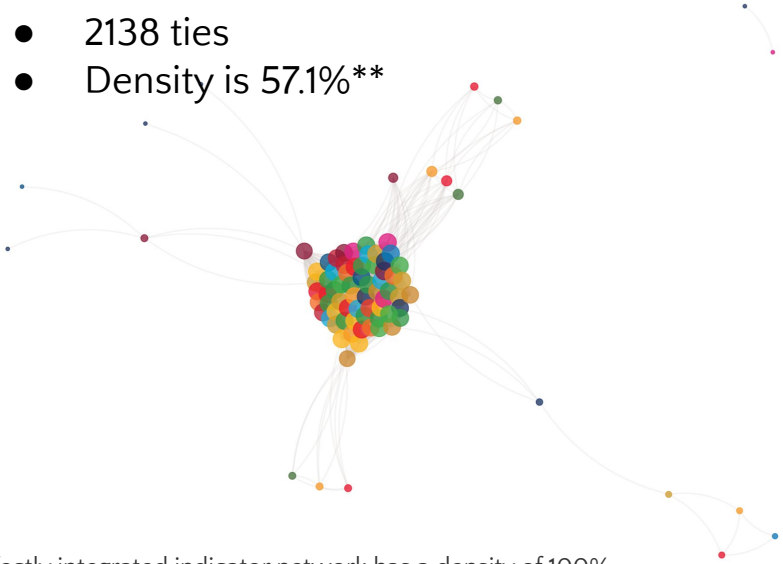
**See the representativeness of composite measures in Appendix B

***PCA does not allow missing data. We used Linear Interpolation to impute and fill in the blanks. See details in Appendix C.

Network Summary

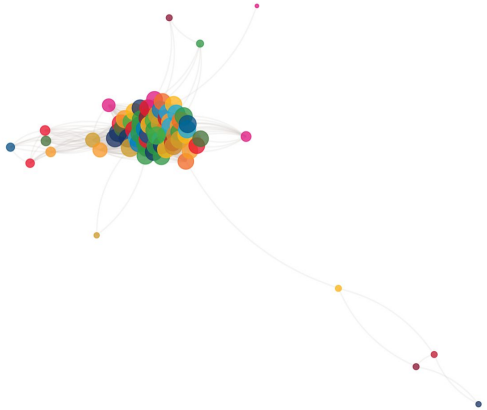
Indonesia
Indicator Network

- 87 nodes
- 2138 ties
- Density is 57.1%**



Guatemala
Indicator Network

- 79 nodes
- 1992 ties
- Density is 64.6%**

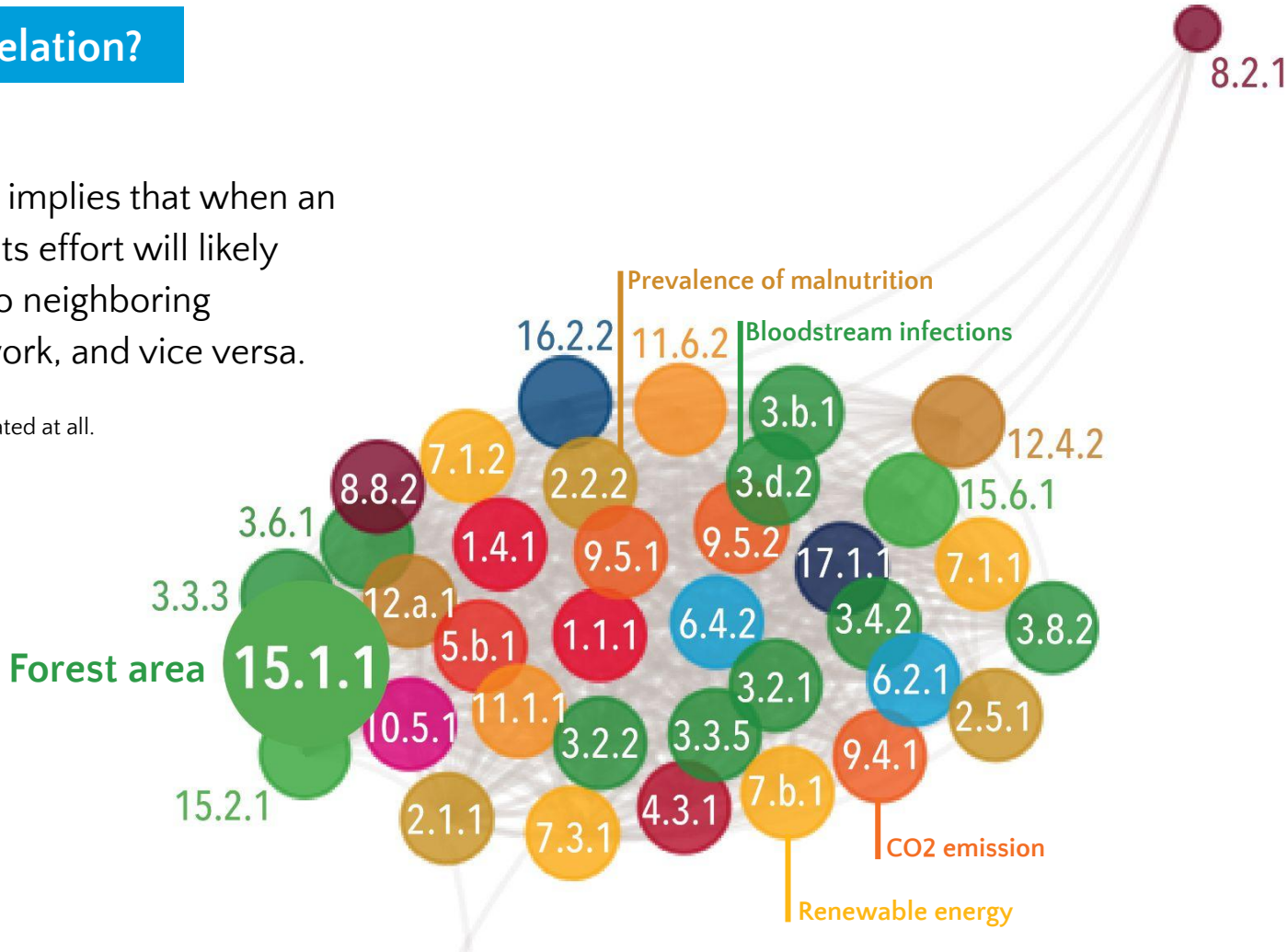


**A perfectly integrated indicator network has a density of 100%.

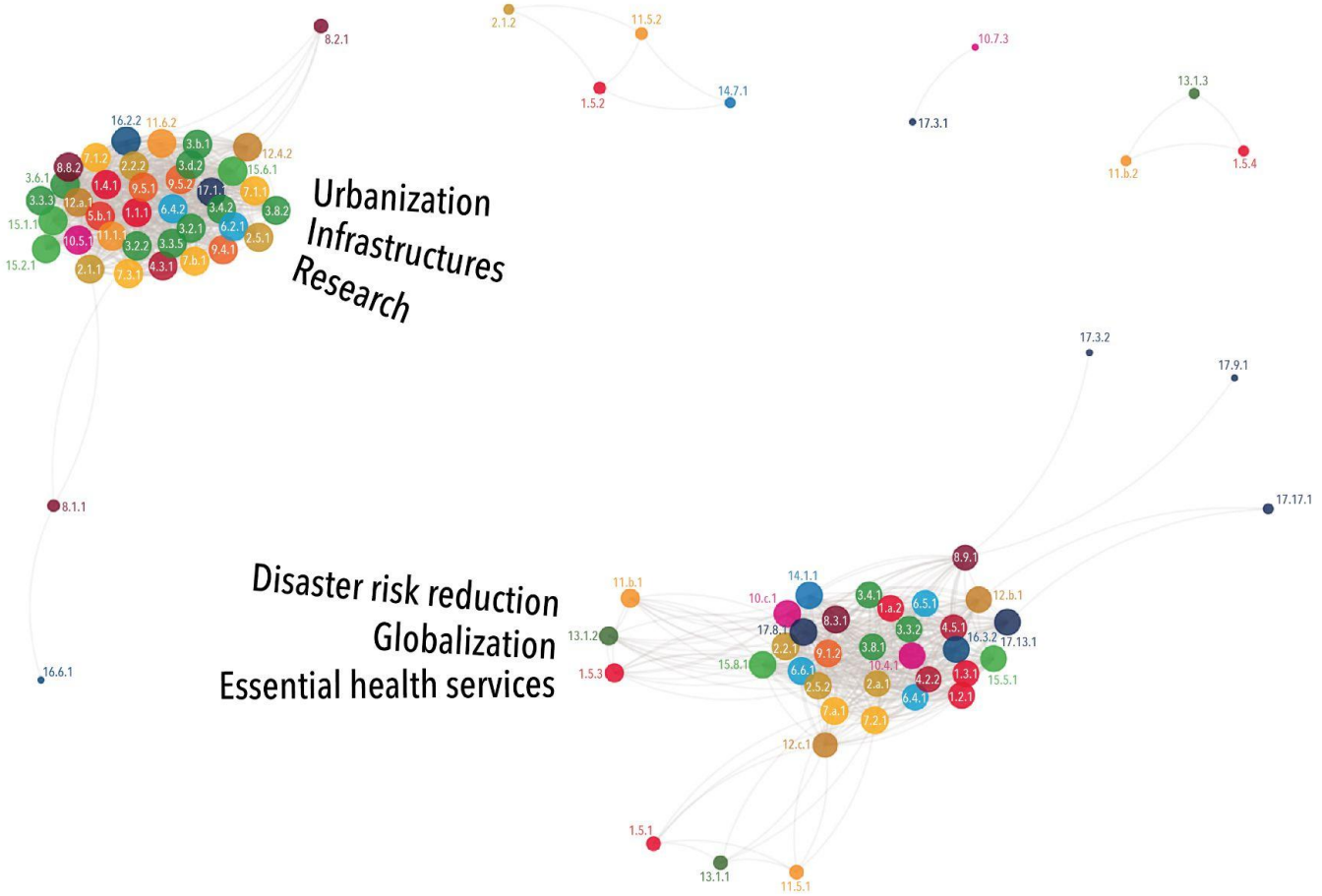
Why Positive Correlation?

A **positive correlation** implies that when an indicator is invested, its effort will likely bring positive effect to neighboring indicators in the network, and vice versa.

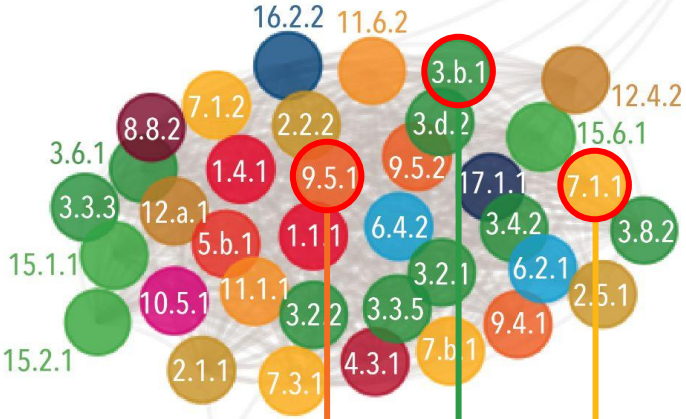
As hard as they may not seem related at all.



Indonesia Network: By Positive Correlation



Indonesia Network: By Positive Correlation



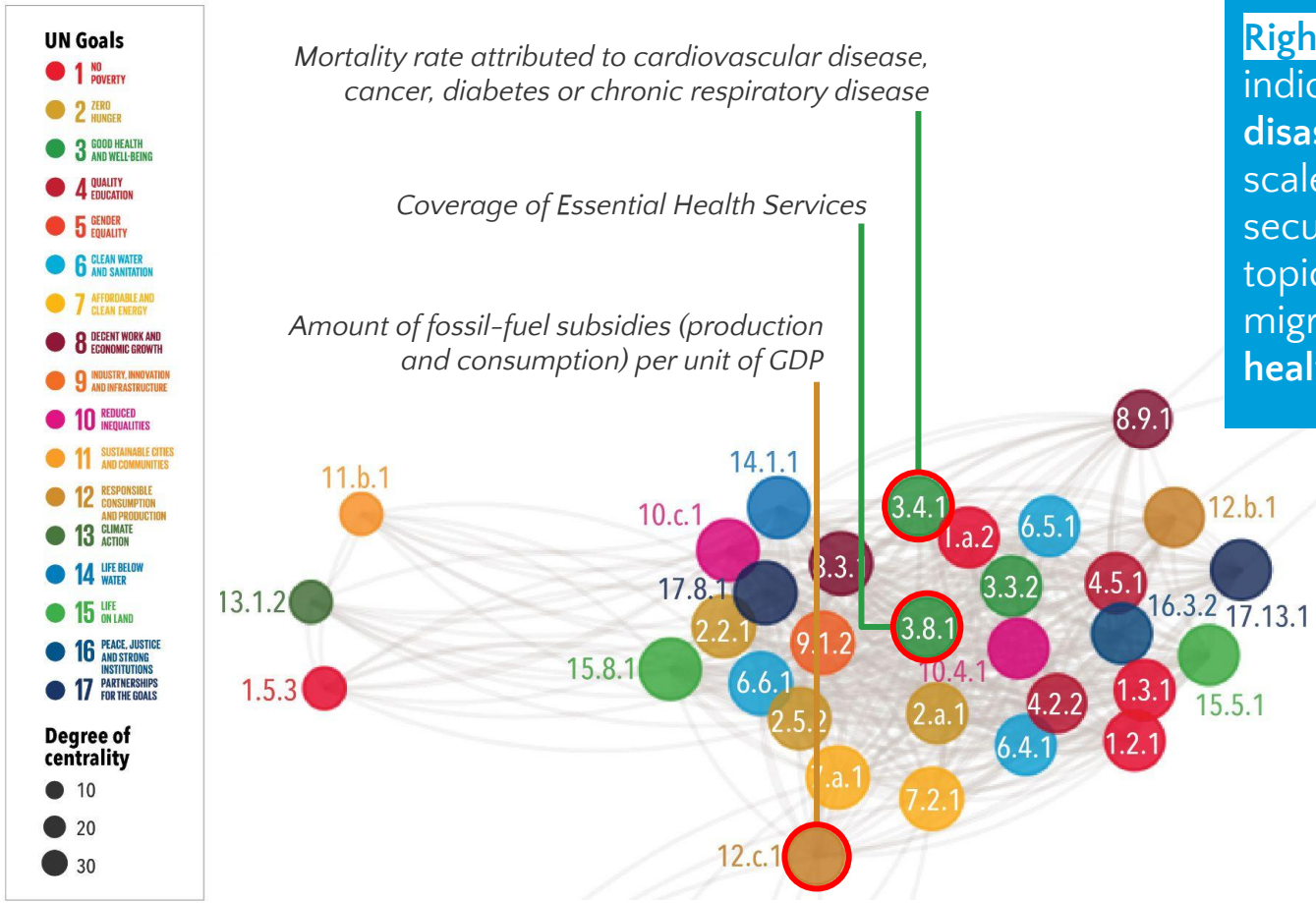
Proportion of population with access to electricity

Proportion of the target population covered by all vaccines included in their national programme

Researchers (in full-time equivalent) per million inhabitants

Left cluster has higher levels of focus on issues related to **urbanization**, **infrastructures, research** and technological development, and agricultural development and issues around human access to **clean food, water**, and **air**.

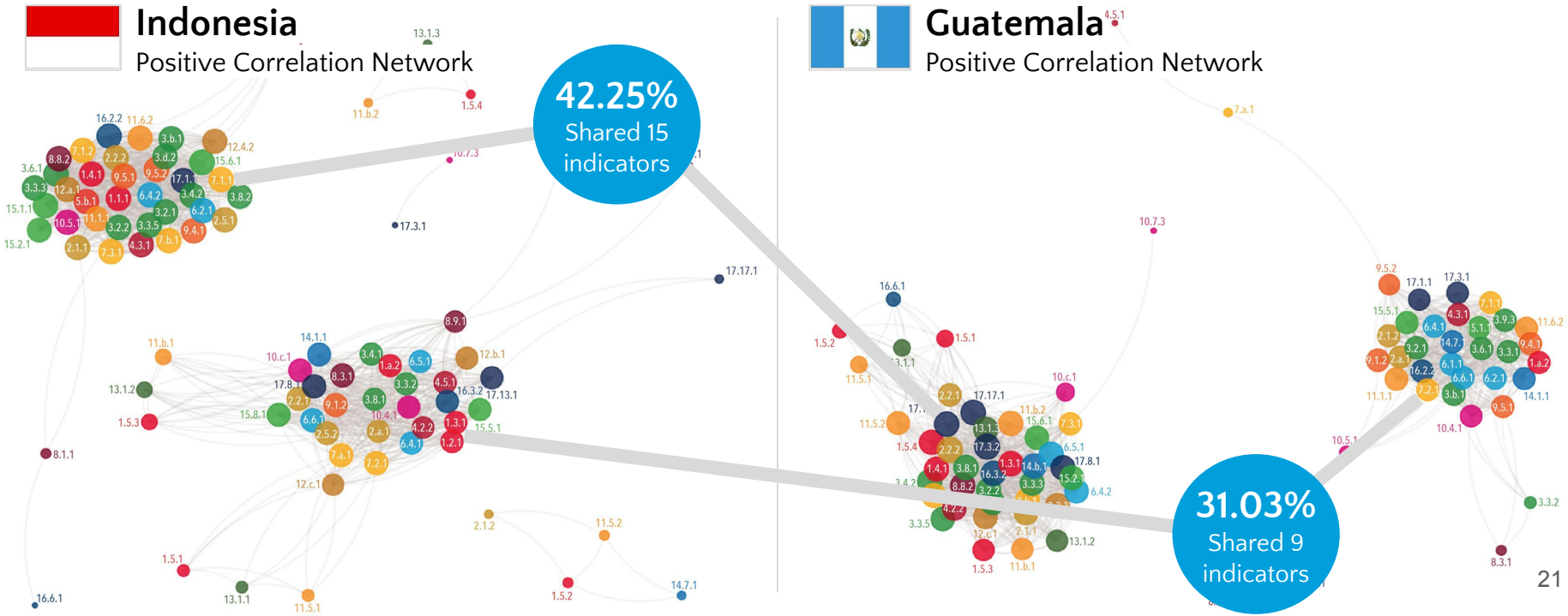
Indonesia Network: By Positive Correlation



Right cluster focuses on indicators pertaining to ***national* disaster risk reduction** (i.e., large scale national level risks, both security and natural) **and** global topics (remittances, aid, and migration), as well as **essential health services** and **disease**.

Correlation Network Between Nations

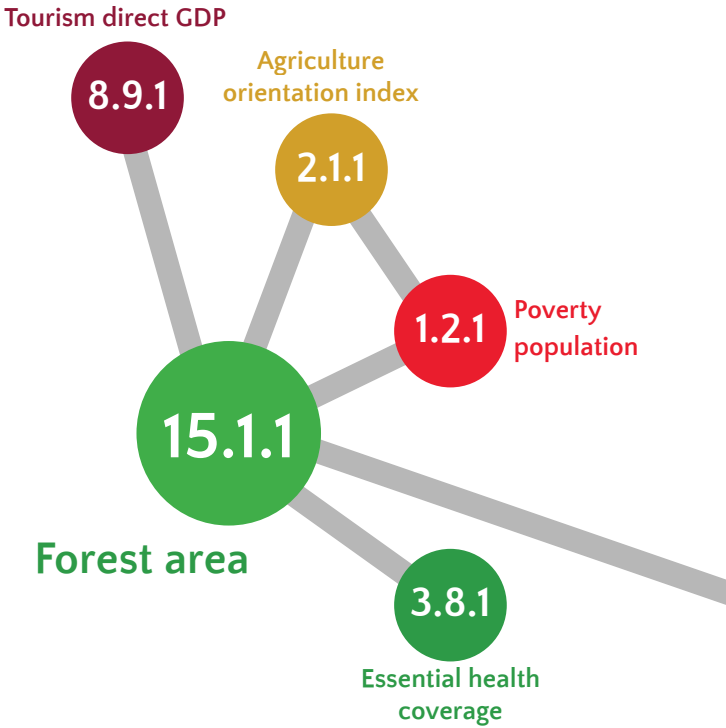
There are only 31.03-42.25% of **shared indicators** between Guatemala and Indonesia's two clusters. It is important to look at network relationships from **country to country** as indicator-level relationships differ from country to country and are not largely consistent.



Why is it Helpful to Also Look at **Negative Correlation**?

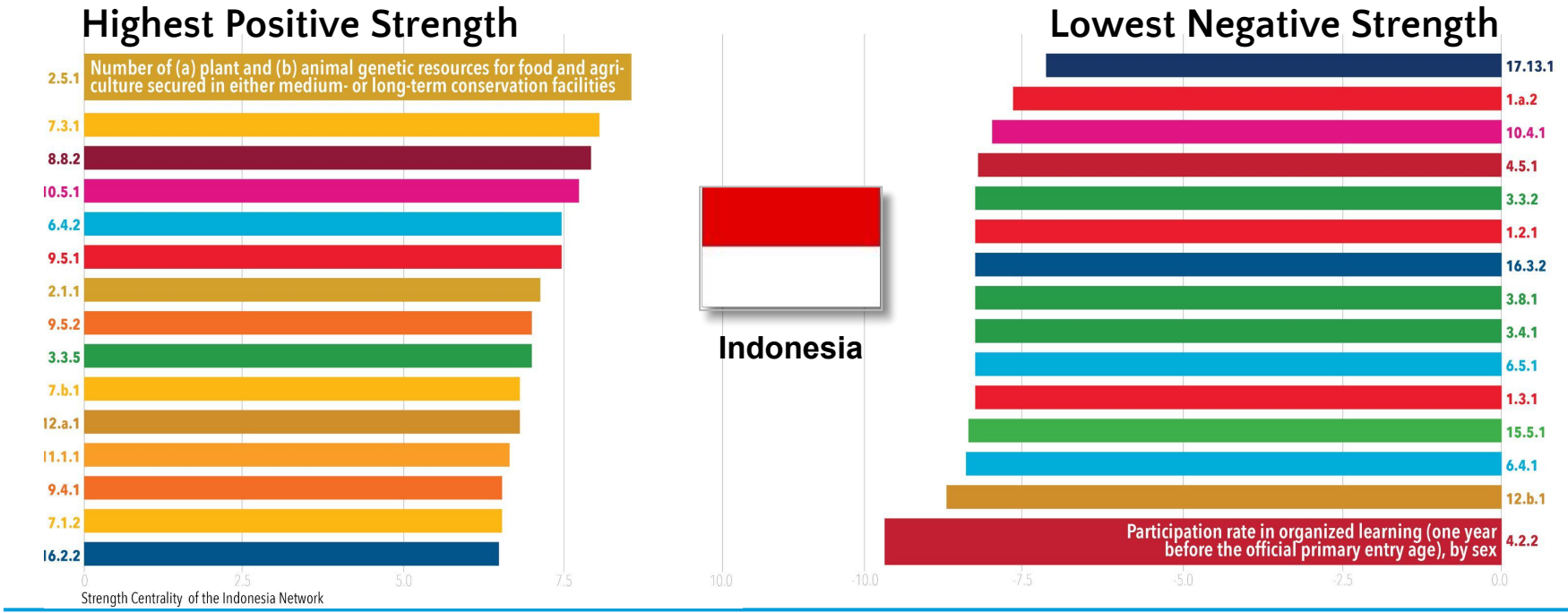
Attentions are often drawn to the positive relationship between indicators that are value-adding to each others.

It is also important to consider the **negative correlation that may reverse the efforts of neighboring indicators**.



Node Strength

Node strength is the sum of weights attached to ties belonging to a node. Below graph presents the sum of coefficients between one indicator and all the indicators it relates to.



Network Model: Additional Findings

- While clusters have dozens of indicators, it is difficult to identify any **particularly central indicators**.
  - Indicators with statistically significant ties to other indicators tend to be relatively consistent in terms of centrality level within their clusters; i.e., **there are even connections in each cluster**.
- It is important to look at network relationships at the **indicator level**.
  - Indicators with same goal did not always show up in the same cluster due to the diversity of sub-topics measured under each goal.

Network Model: QAP Regression Methodology

- Standing problem when making statistical inference with networks models: **non-independence of observations**.
- Problem with standard errors solved by permuting rows and columns in the matrix, while maintaining the underlying relationship -- **QAP** (Quadratic Assignment Procedure).
- Data Preprocessing:
  - Networks recoded into **binary** forms (have tie:1, no tie:0).
  - Select **shared indicators** from different networks.

Similarly Between Text Network & Coefficient Network Structures

Hypothesis: Text similarity is adequately reflecting the actual indicator network structure.

Network Logit Model between the text network and coefficient network

Only the edges with positive scores were kept in the coefficient network to make it conceptually more similar and more comparable to the Text network constructed based on text similarity.

Result: Whether for Indonesia or Guatemala, **the relationship between the coefficient network and text network** is **NOT statistically significant****, therefore **not reliable**.

**Statistically significant when the p-value from regression is < 0.05.

Similarity Between the Network Structure across Different Countries

Hypothesis: The indicator network structure is the same across countries.

Result: Predicting Indonesia indicator network with the Guatemala indicator network has a **72.3% accuracy**, and is highly statistically significant**.

The indicator network shares similar structure across countries, and the similarity is statistically significant.

Interpretation: QAP measures the **generalizability** of the indicator network built on the correlations – if **72.3%** is considered as a high accuracy, Indonesia/Guatemala networks may be used to explain the overall level of integration of SDG Indicators; otherwise, it is suggested to reconstruct network models for other countries, case by case.

**Statistically significant when the p-value from regression is < 0.05.

Similarity Between the Network Structure across Different Countries – Subgraphs

Regression between the **positive/negative subgraphs**:

Results: Predicting **positive ties** in the Indonesia indicator coefficient network with the **positive ties** in the Guatemala coefficient network has a **86% accuracy**, and is highly statistically significant**.
Predicting **negative ties** in the Indonesia indicator coefficient network with the **negative ties** in the Guatemala coefficient network has a **86% accuracy**, and is highly statistically significant**.

The positive/negative subgraphs shares MORE similar structure across countries, and the similarity is statistically significant.

Interpretation: The **generalizability** of the positive/negative subgraphs of the indicator network built on the correlations is better than the whole network. It's better to look into the positive and negative ties separately.

**Statistically significant when the p-value from regression is < 0.05.

Conclusion & Implications

- There are many linkages between UN SDG indicators, some indicators are more integrated and from clusters based on positive linkages.
- Indicators have different overall effect on their related indicators, which should be considered before making investment decisions.
- Indicator description similarity is not reliable for predicting the actual linkages between indicators.
- The indicator network structure are similar to some extent, but not really consistent across countries.

Future Work Directions

1. Look at more countries to test the generalizability of the coefficient network.
2. Consider and discuss spurious correlations.
3. Do time series analysis to exclude serial correlation issues.

Links to Deliverables

- Research Paper
- Blog
- Raw Codes, Data, and Visualizations
- Interactive Visualizations

JOINT SDG FUND

THANK YOU!

APPENDIX

Appendix A

Removal of Disaggregated Measures

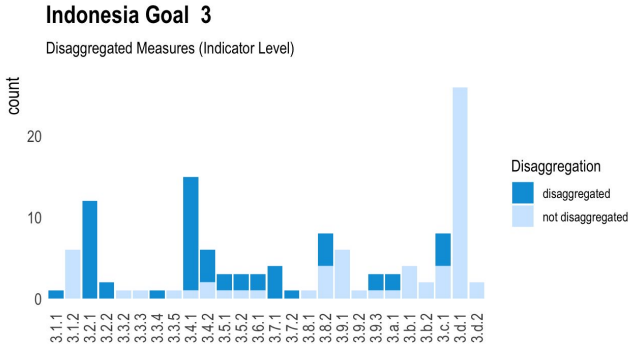
Appendix A: Removal of Disaggregated Measures

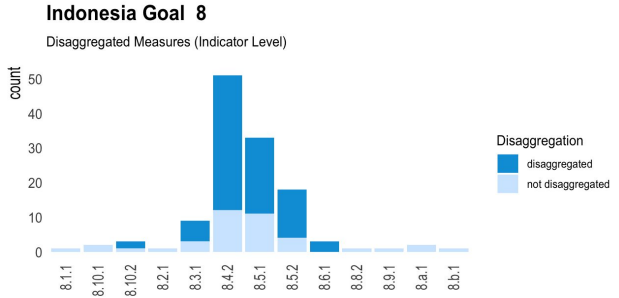
- 'Disaggregated' measures often constitute a substantial proportion (and sometimes all) of our measures for **goals, targets, and indicators**

  * Proportion of disaggregation measures can vary across **goals** and **targets** (see figures at the **right**)

- These measures make our data messy and make it more difficult to have unique observations for measures at any given time

  * (e.g., some measures may have three observations, two of which are disaggregated results by sex).





Some measurements have keywords (e.g., "BOTHSEX", "ALL AREA") indicating coverage of the total population. In such cases, we kept only the natural representative, and removed all the other disaggregated measures.

Appendix B

Representative Method for Indicators

- Initial basis for application
- Summary of method (as applied to UN SDG's)
- Representativeness of composite measures
- Alternative method not adopted

Appendix B: Initial Basis for Application

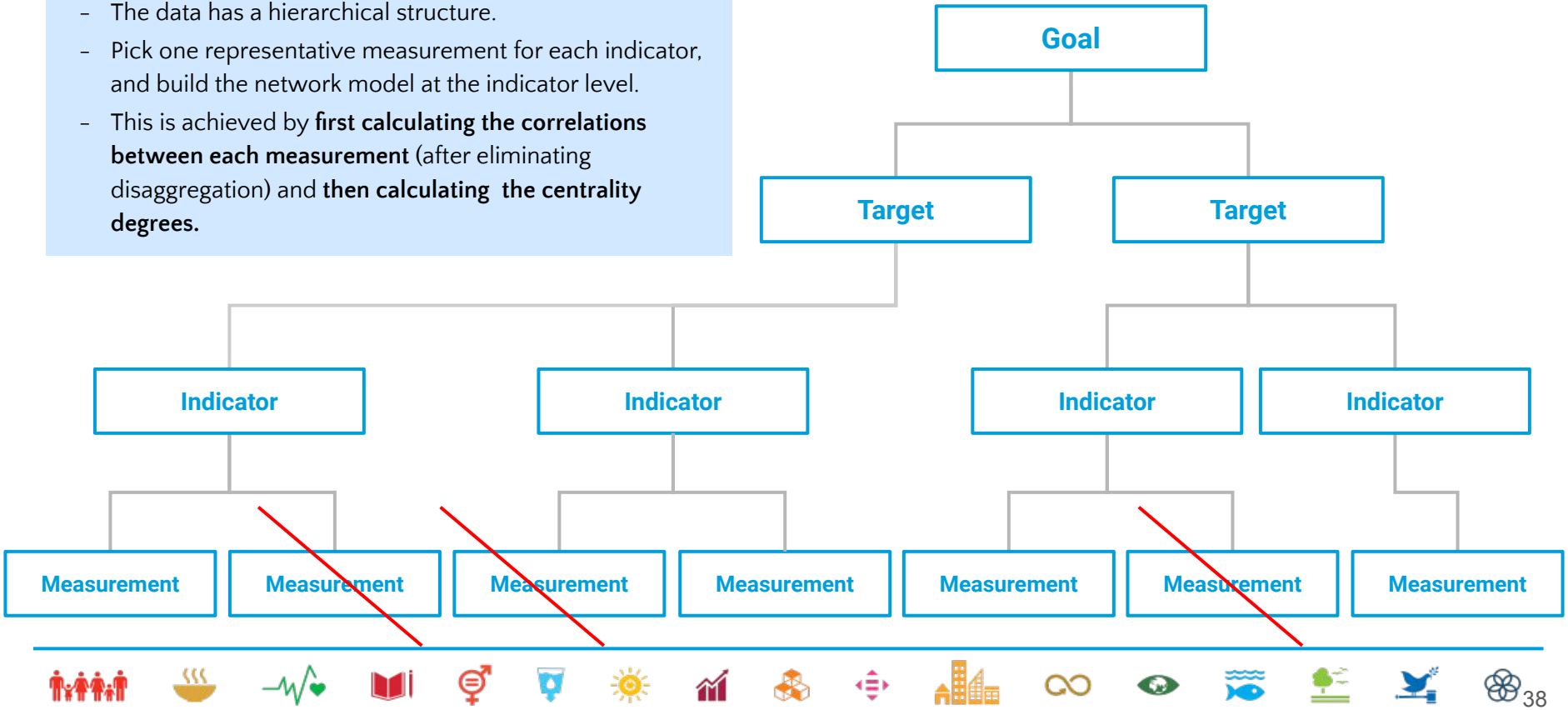
- Similar methods were employed to look at relative representativeness and importance of stocks and movement in the stock market (Hua et al. 2019, Kazemilari et al. 2017)
- Rationale in these papers includes discussion of "complex and heterogeneous [data]... original data sets can be analysed for different purposes and stakeholders." (Hua et al. 2019)
- We felt these characteristics also hold true to SDG measures / indicators. Use of fewer measures / indicators **allows for greater interpretability** (as we can discuss which *measure* is referred to in relation to a target) and **reduces noise and difficulty** in re-weighting measures (in the event that some measures are more important than others).

Blanc, David. (2015). Towards Integration at Last? The Sustainable Development Goals as a Network of Targets. Sustainable Development. 23. 10.1002/sd.1582.

Kazemilari, M.; Mardani, A.; Streimikiene, D.; Zavadskas, E.K. An overview of renewable energy companiesin stock exchange: Evidence from minimal spanning tree approach.Renew. Energy 2017,102, 107–117

Appendix B: Method Summary

- The data has a hierarchical structure.
- Pick one representative measurement for each indicator, and build the network model at the indicator level.
- This is achieved by **first calculating the correlations between each measurement** (after eliminating disaggregation) and **then calculating the centrality degrees.**



Appendix B: Representativeness of composite measures – Indonesia

Indicator	Representativeness	Indicator	Representativeness	Indicator	Representativeness	Indicator	Representativeness
1.1.1	1	3.4.1	1	8.1.1	1	12.c.1	0.990373888
1.2.1	1	3.4.2	1	8.2.1	1	13.1.1	0.896958719
1.3.1	0.977108462	3.6.1	1	8.3.1	1	13.1.2	1
1.4.1	0.99594841	3.8.1	1	8.8.2	1	13.1.3	0.999999998
1.5.1	0.896958719	3.8.2	0.918019539	8.9.1	1	14.1.1	0.743163202
1.5.2	0.638197434	3.b.1	0.515944086	9.1.2	1	14.7.1	1
1.5.3	1	3.d.2	1	9.4.1	0.704527162	15.1.1	1
1.5.4	0.999999998	4.2.2	1	9.5.1	1	15.2.1	0.862896088
1.a.2	1	4.3.1	1	9.5.2	1	15.5.1	1
2.1.1	0.998490651	4.5.1	0.948131073	10.4.1	1	15.6.1	0.81271137
2.1.2	0.894333428	5.b.1	1	10.5.1	0.632135555	15.8.1	1
2.2.1	0.941614529	6.2.1	0.995175817	10.7.3	1	16.2.2	1
2.2.2	0.985705545	6.4.1	1	10.c.1	1	16.3.2	1
2.5.1	1	6.4.2	1	11.1.1	1	16.6.1	1
2.5.2	1	6.5.1	1	11.5.1	0.896958719	17.1.1	0.913888738
2.a.1	0.902130482	6.6.1	0.7308095	11.5.2	0.638197434	17.13.1	0.533432253
2.c.1	1	7.1.1	1	11.6.2	1	17.17.1	0.682357347
3.2.1	0.999925106	7.1.2	1	11.b.1	1	17.3.1	1
3.2.2	0.999917119	7.2.1	1	11.b.2	0.999999998	17.3.2	1
3.3.2	1	7.3.1	1	12.4.2	1	17.4.1	1
3.3.3	1	7.a.1	1	12.a.1	1	17.8.1	1
3.3.5	1	7.b.1	1	12.b.1	0.689514834	17.9.1	1

Appendix B: Representativeness of composite measures – Guatemala

Indicators	Representativeness	Indicators	Representativeness	Indicators	Representativeness	Indicators	Representativeness
1.3.1	0.78732506	3.6.1	1	8.3.1	1	13.1.3	0.81831407
1.4.1	0.95580359	3.8.1	1	8.8.2	1	14.1.1	0.77179813
1.5.1	0.79727606	3.9.3	1	9.1.2	1	14.7.1	1
1.5.2	0.52195895	3.b.1	0.70829608	9.4.1	0.99217273	14.b.1	1
1.5.3	1	4.2.2	1	9.5.1	1	15.1.1	1
1.5.4	0.81831407	4.3.1	1	9.5.2	1	15.2.1	0.63827793
1.a.2	1	4.5.1	0.85051131	10.4.1	1	15.5.1	1
2.1.1	0.64804687	6.1.1	1	10.5.1	0.63602049	15.6.1	0.87977193
2.1.2	0.99515785	6.2.1	0.91084536	10.7.3	1	16.2.2	0.94298742
2.2.1	1	6.4.1	1	10.c.1	1	16.3.2	1
2.2.2	1	6.4.2	1	11.1.1	1	16.6.1	1
2.a.1	0.93911435	6.5.1	1	11.5.1	0.79727606	17.1.1	0.57907609
2.c.1	1	6.6.1	0.62193496	11.5.2	0.38656667	17.13.1	0.51588222
3.2.1	0.9999533	7.1.1	1	11.6.2	1	17.17.1	0.56914816
3.2.2	0.99958145	7.1.2	1	11.b.1	1	17.3.1	1
3.3.1	1	7.2.1	1	11.b.2	0.81831407	17.3.2	1
3.3.2	1	7.3.1	1	12.4.2	0.64080302	17.4.1	1
3.3.3	1	7.a.1	1	12.a.1	1	17.8.1	1
3.3.5	1	7.b.1	1	12.c.1	0.9987478	17.9.1	1
3.4.1	1	8.1.1	1	13.1.1	0.79727606		
3.4.2	1	8.2.1	1	13.1.2	1		

Appendix B: Central measure as representative (not adopted)

- **Weighted degree centrality scores**: sum of correlation coefficients between a measure and all other measures under the same indicator.
- **Normalized centrality scores**: weighted degree centrality scores divided by the number of other measures under the same indicator (n-1)
- **Cut winner**: If a measure has a normalized centrality score* at least **0.1** higher than the average score, it's defined as the cut winner.

  * For data from Indonesia **(Left)**, **14 out of 40 (35%)** eligible indicators with multiple measurements have a cut winner.

  * For data from Guatemala **(Right)**, **15 out of 42 (35.7%)** eligible indicators with multiple measurements have a cut winner.

Conclusion: we should use composite method.

Indicator	mean_others	top	diff	cut_winner
2.2.2	0.333333333	1	0.66666667	1
2.a.1	0.141831843	0.71544585	0.573614	1
3.b.1	0.094496897	0.4553418	0.3608449	1
15.2.1	0.059244104	0.40402329	0.34477919	1
12.b.1	0.207535522	0.48295547	0.27541995	1
1.5.2	0.182305539	0.44496381	0.26265827	1
6.6.1	0.123471792	0.373905	0.25043321	1
15.1.1	0.25	0.5	0.25	1
1.5.4	0.249999999	0.5	0.25	1
11.b.2	0.249999999	0.5	0.25	1
13.1.3	0.249999999	0.5	0.25	1
11.5.2	0.029168886	0.17798552	0.14881664	1
15.6.1	0.04362281	0.17449124	0.13086843	1
17.13.1	0.071317345	0.19430437	0.12298702	1
10.5.1	0.056070181	0.13464634	0.07857616	0
2.1.2	0.962611105	0.98181972	0.01920862	0
1.5.1	0.982941692	1	0.01705831	0
11.5.1	0.982941692	1	0.01705831	0
13.1.1	0.982941692	1	0.01705831	0
12.c.1	0.990181846	0.99508783	0.00490598	0
3.2.1	0.999840076	0.99990976	6.97E-05	0
1.3.1	0	0	0	0
1.4.1	0.991896821	0.99189682	0	0
14.1.1	0.207592927	0.20759293	0	0
15.8.1	0	0	0	0
17.1.1	0.763923818	0.76392382	0	0
17.17.1	0.99962794	0.99962794	0	0
17.18.3	0	0	0	0
2.1.1	0.996133908	0.99613391	0	0
2.2.1	1	1	0	0
2.5.1	0	0	0	0
3.2.2	0.999763361	0.99976336	0	0
3.4.1	1	1	0	0
3.4.2	0	0	0	0
3.8.2	0.722383885	0.72238388	0	0
3.d.2	1	1	0	0
4.5.1	0.991371216	0.99137122	0	0
6.2.1	0.990351634	0.99035163	0	0
6.4.2	0	0	0	0
9.4.1	0.06940754	0.06940754	0	0

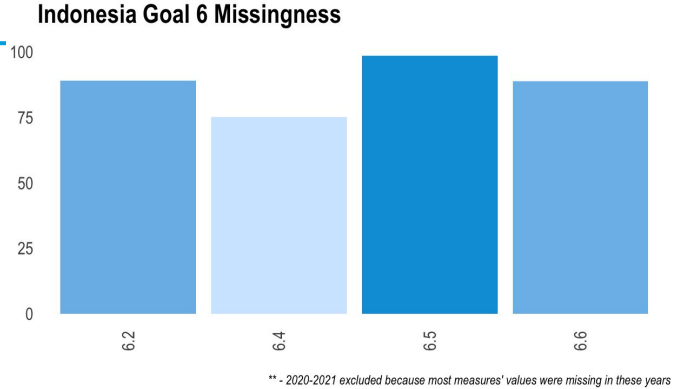
Indicator	mean_others	top	diff	cut_winner
1.5.4	0.39382158	0.78764317	0.39382158	1
11.b.2	0.39382158	0.78764317	0.39382158	1
13.1.3	0.39382158	0.78764317	0.39382158	1
10.5.1	0.29561789	0.64968946	0.35407157	1
15.1.1	0.25	0.5	0.25	1
11.5.2	0.13668096	0.35802967	0.22134871	1
1.3.1	0.08809507	0.28571429	0.19761922	1
1.5.2	0.11068551	0.2959193	0.18523379	1
17.13.1	0.10341806	0.28450858	0.18109052	1
6.6.1	0.11849721	0.2962963	0.17779909	1
15.6.1	0.05799946	0.23199785	0.17399839	1
3.b.1	0.17294101	0.33333333	0.16039232	1
1.5.1	0.74777086	0.85952805	0.11175719	1
11.5.1	0.74777086	0.85952805	0.11175719	1
13.1.1	0.74777086	0.85952805	0.11175719	1
15.2.1	0.01957652	0.11448716	0.09491064	0
12.4.2	0.38064045	0.45153926	0.07089881	0
2.a.1	0.85672529	0.92014852	0.06342323	0
2.1.2	0.96452008	0.9817031	0.01718302	0
12.c.1	0.98732503	0.99458904	0.00726401	0
3.2.1	0.99990206	0.99993879	3.67E-05	0
1.4.1	0.91160719	0.91160719	0	0
12.b.1	0	0	0	0
14.1.1	0.54595408	0.54595408	0	0
15.8.1	0	0	0	0
16.2.2	1	1	0	0
17.1.1	0.1080063	0.1080063	0	0
17.17.1	0.99989777	0.99989777	0	0
17.18.3	0	0	0	0
17.19.2	0	0	0	0
2.1.1	0.13424331	0.13424331	0	0
2.2.1	0	0	0	0
2.2.2	0	0	0	0
2.5.1	0	0	0	0
3.2.2	0.99880477	0.99880477	0	0
3.4.1	1	1	0	0
3.4.2	1	1	0	0
4.5.1	0.81307341	0.81307341	0	0
5.1.1	0	0	0	0
6.2.1	0.77459667	0.77459667	0	0
6.4.2	0	0	0	0
9.4.1	0.97279939	0.97279939	0	0

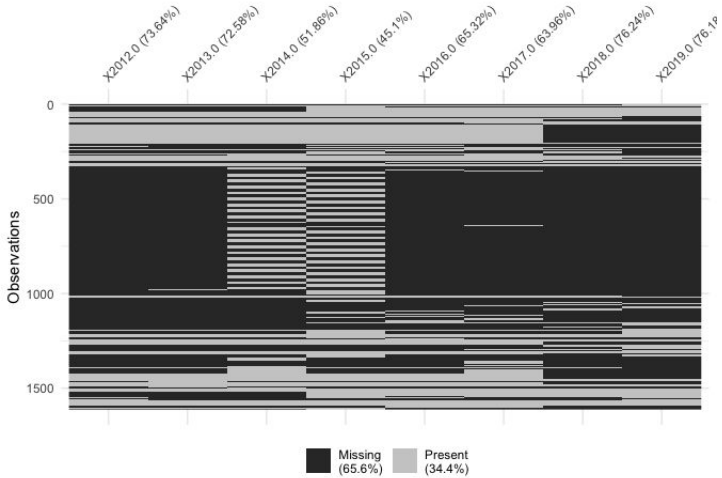
Appendix C

Imputation of Missing Observations in Indicator Measurement Data

Appendix C: SDG Measure Missingness Across Time

- Previous work on missing data showed **a great deal of missing data across time** overall (right), as well as a **number of indicators with only one year of non-null data** for several measures (below).





- Some of these measures were relatively newly developed (2014-2015) and could be more available (less missing) across countries over time.
- Further, countries of interest to UN SDG are growing and often have relatively higher rates of missing data. (vs. wealthier or 'economically mature' countries not on this list).

Appendix C: Imputation Method

1. Remove measures with only one observation (i.e., difficult to impute remaining observations) - 3% of our columns

2. Use data from 2015-2019 to impute our composite measures.
   a. For imputation, we considered using the column mean but decided against it as column data are temporal (e.g., more trend-based vs stable from row to row).
   b. Instead, we will be using **linear regression to fit a slope for average year-over-year change**

Year <dbl>	SI_COV_POOR <dbl>	SI_COV_CHLD <dbl>	SI_COV_UEMP <dbl>	SI_COV_VULN <dbl>	SI_COV_WKINJRY <dbl>	SI_COV_BENFTS <dbl>	SI_COV_DISAB <dbl>	SI_COV_PENSN <dbl>
2012	NA	NA	NA	NA	NA	NA	NA	NA
2013	NA	NA	NA	NA	NA	NA	NA	NA
2014	NA	NA	NA	NA	NA	NA	NA	NA
2015	NA	NA	NA	NA	15.4	NA	NA	NA
2016	NA	NA	NA	NA	NA	NA	NA	14.0
2017	NA	NA	NA	NA	NA	NA	NA	NA
2018	100	NA	NA	NA	NA	NA	NA	NA
2019	NA	NA	NA	NA	NA	NA	NA	14.1
2020	NA	25.6	0	16.5	22.5	27.8	2.5	14.8
2021	NA	NA	NA	NA	NA	NA	NA	NA

```
Call:
lm(formula = VC_DSR_PDLN ~ Year, data = guat.1.5.1.new, na.rm = TRUE)

Residuals:
ALL 2 residuals are 0: no residual degrees of freedom!

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3436977        NaN     NaN      NaN
Year            1726        NaN     NaN      NaN
```

VC_DSR_PDLN <dbl>	Pdln.new <dbl>
NA	40913
42639	42639
44365	44365
NA	46091
NA	47817

Appendix C: Potential Issues and Considerations

- If **we have two observations with the same value,** we will have the same value throughout the period using this method.

  * For the time being, we are comfortable with this, because certain measures are on a scale with a low range, or are static over time in actuality.

ER_IAS_LEGIS <dbl>
NA
1
NA
NA
NA
1

- Using slope to impute missing data may lead to some illogical/improbable observations being imputed at the beginning or end of the column(s) in question (e.g., negative values in certain years where none should exist; observations beyond the bounds of the indicator's range).

  * These have been looked up by measure unit(s) and data validation steps were added to the regression code after imputation was done.

```
for i in np.logspace(0,10,10):
    if (series_units[series_units['SeriesCode']==str(col)]['[Units]']=='PER_'+str(i)+'_POP') is True:
        missing_rows["inferred_col"] = missing_rows["inferred_col"].clip(upper=i)
if (series_units[series_units['SeriesCode']==str(col)]['[Units]']=='PERCENT') is True:
    missing_rows["inferred_col"] = missing_rows["inferred_col"].clip(upper=100)
elif (series_units[series_units['SeriesCode']==str(col)]['[Units]']=='SCORE') is True:
    missing_rows["inferred_col"] = missing_rows["inferred_col"].clip(upper=10)
elif (series_units[series_units['SeriesCode']==str(col)]['[Units]']=='Ratio') is True:
    missing_rows["inferred_col"] = missing_rows["inferred_col"].clip(upper=1)
```

SI_COV_PENSN <dbl>	Pensn.new <dbl>
NA	0.0
8.3	8.3
26.2	26.2
NA	44.1
NA	62.0

Appendix D

Text Model Methodologies: From Embedding models to Similarity Threshold

- Word Embedding Models Details
- Vote to Select the Final Model
- Similarity Threshold

Appendix D: Word Embedding Models Details

TF-IDF

- Evaluates how **relevant** a word is to a document in a collection of documents.
- This is done by **multiplying two metrics**: how many times a word appears in a document(**TF term frequency)**, and the inverse **document frequency(IDF** which is the log of the total number of documents divided by total number of documents containing the term t(DF))
- So, words that are common in every document, such as **'this', 'what', and 'if', rank low** even though they may appear many times, since they don't mean much to that document in particular. However, if the word '**poverty' appears many times in a document, while not appearing many times in others,** it probably means that it's very relevant.

$$tf - idf\,(t,d) = tf\,(t,d) \cdot idf\,(t)$$

Appendix D: Word Embedding Models Details

Word2vec, Doc2vec

- TF-IDF is essentially bag of words methods, words are taken out of context.
- Word2vec and Doc2vec would consider the context of words/paragraphs. Words in similar context would have similar vectors.
- Doc2vec converts a whole paragraph/document into vectors directly, without averaging the vector of each word.

```
It was really cold yesterday.
It will be really warm today, though.
It'll be really hot tomorrow!
Will it be really cool Tuesday?
```

Appendix D: Cosine Distance

Cosine Distance

- It is the cosine of the angle between two vectors, which gives us the angular distance between the vectors. Formula to calculate cosine similarity between two vectors A and B is.

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Appendix D: Vote to Select the Final Model

Compare the result of 3 models

- The **Top 5 most related indicators list** is different for 3 models

TF- IDF

indicator	related_indicator	similarity_score
1.1.1	1.2.1	0.414994
1.1.1	10.2.1	0.193975
1.1.1	10.7.4	0.158748
1.1.1	16.8.1	0.155508
1.1.1	16.b.1	0.152244

Doc2vec

indicator	related_indicator	similarity_score
1.1.1	1.2.1	0.835552
1.1.1	11.1.1	0.789279
1.1.1	8.5.2	0.702510
1.1.1	11.6.2	0.689911
1.1.1	3.3.1	0.675195

Word2vec

indicator	related_indicator_word2vec	similarity_score_word2vec
1.1.1	1.2.1	0.858479
1.1.1	11.1.1	0.828153
1.1.1	11.6.2	0.700170
1.1.1	3.3.2	0.699983
1.1.1	3.3.1	0.692252

Appendix D: Vote to Select the Final Model

Compare the result of 3 models

- **Sampled 25** out of 246 indicators, and their **Top 5** most related indicators calculated by the models
- **5 members manually voted**, based on which model's result seemed to make most sense judgmentally, to select the final model

	TF-IDF		Doc2Vec		Word2Vec		meta_data		
indicator	related_indicator	similarity_score	related_indicator	similarity_score	related_indicato	similarity_score	linked indicators'	Hanyu	Connie
1.1.1	10.2.1	0.1939752979	11.1.1	0.8712634803	11.1.1	0.82815272	1.2.1, 10.1.1,10.2.1	Tf-Idf	
1.1.1	10.7.4	0.1587481895	11.6.2	0.7619974338	11.6.2	0.7001699939	1.2.1, 10.1.1,10.2.1	Tf-Idf	
1.1.1	16.8.1	0.155508074	17.3.2	0.7530249258	3.3.2	0.69998267	1.2.1, 10.1.1,10.2.1	Doc2Vec	
1.1.1	16.b.1	0.1522444099	3.3.1	0.7486380856	3.3.1	0.6922523048	1.2.1, 10.1.1,10.2.1	Tf-Idf	
1.3.1	17.1.1	0.1591933127	16.b.1	0.6520183871	16.10.1	0.6869434274	3.8.1,3.8.2,1.a.2	All Same	
1.3.1	16.3.2	0.14159404	16.3.1	0.6454389431	10.3.1	0.6845167515	3.8.1,3.8.2,1.a.2	Word2Vec	
1.3.1	16.2.2	0.1329879671	17.1.1	0.6425007771	16.3.1	0.677525626	3.8.1,3.8.2,1.a.2	Tf-Idf	
1.3.1	8.5.2	0.1008440403	10.3.1	0.6343163308	11.7.2	0.6524204451	3.8.1,3.8.2,1.a.2	Word2Vec	
1.3.1	8.7.1	0.09886864258	16.10.1	0.6300875814	16.b.1	0.6468806814	3.8.1,3.8.2,1.a.2	Word2Vec	
2.1.1	7.2.1	0.1172055606	7.2.1	0.9052200558	7.2.1	0.8776728952	2.1.2,2.2.1,2.2.2,2.2.3	All Same	
2.1.1	2.1.2	0.07173966639	7.b.1	0.8326391021	11.5.1	0.7895324392	2.1.2,2.2.1,2.2.2,2.2.3	Tf-Idf	
2.1.1	7.3.1	0.06990966411	12.a.1	0.8273388436	11.4.1	0.7838939546	2.1.2,2.2.1,2.2.2,2.2.3	Tf-Idf	
2.1.1	7.a.1	0.06800952491	7.3.1	0.8243111789	7.b.1	0.7834832389	2.1.2,2.2.1,2.2.2,2.2.3	All Same	
2.1.1	16.9.1	0.06678810757	11.4.1	0.8220610127	13.1.1	0.7827109085	2.1.2,2.2.1,2.2.2,2.2.3	All Same	
3.1.2	5.6.1	0.1279777675	4.2.1	0.891282409	4.2.1	0.8790359515	3.1, 3.2, 3.8	Tf-Idf	

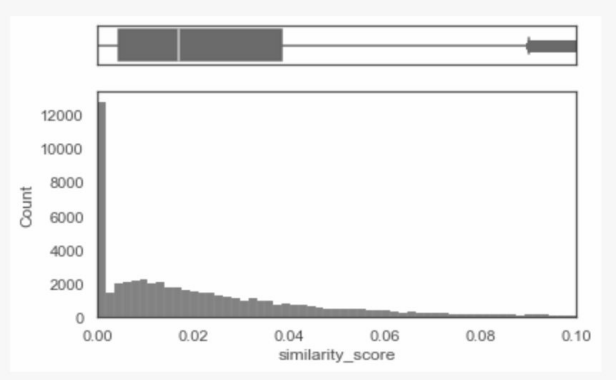
TF-IDF was selected as the final model based on majority voting

Appendix D: Vote to Select the Final Model

Embedding Model	Selected Counts
word2vec	8
Doc2Vec	18
Tf-idf	35
all the same	31

Appendix D: Similarity Threshold

indicator	related_indicator	similarity_score
1.1.1	1.2.1	0.486674568
1.1.1	10.2.1	0.19637961
1.1.1	2.c.1	0.19555602
1.1.1	16.b.1	0.184230509
1.1.1	10.3.1	0.184230509
1.1.1	16.8.1	0.183721566
1.1.1	8.9.1	0.183425276
1.1.1	10.7.4	0.179992993
1.1.1	16.1.2	0.177563457
1.1.1	16.1.1	0.172091635
1.1.1	3.2.2	0.166023278
1.1.1	16.4.2	0.15473371
1.1.1	7.1.1	0.150507241
...		



- Each indicator is compared with the rest 245 indicators, so there're 245 similarity scores calculated for each indicator.
- The median of the similarity score is 0.017. However, the team noted that there are no meaningful relationships presented at this level, when judgmentally examining the definitions of indicators.
- So we sampled results with thresholds set at 0.10, 0.15, 0.20 and decided that 0.20 as a filter threshold gives a more reliable result

Appendix D: Visualization for Indicator 17.2.1

What is it?

Zoom in to the most central indicator 17.2.1 and see how it's related to other indicators.

Key takeaway

- 17.2.1 focus on **Development & Assistance,** so it's most related indicators also talk about **aids for developing countries** and **assistance in different areas** such as agriculture,infrastructure scholarships.

Appendix E: Indicators in each Network regression model

1. Indonesia coefficient network ~ Text network
   - Indicators: 87 indicators in the Indonesia Indicator Network
2. Indonesia coefficient network ~ Text network
   - Indicators: 79 indicators in the Guatemala Indicator Network
3. Guatemala coefficient network ~ Indonesia coefficient network
   - Indicators: 74 common indicators in the Indonesia Network and Guatemala Network

Appendix F

Negative Coefficients Network

- Summary
- Network Visualizations
- Most Negative Strength Centrality Scores

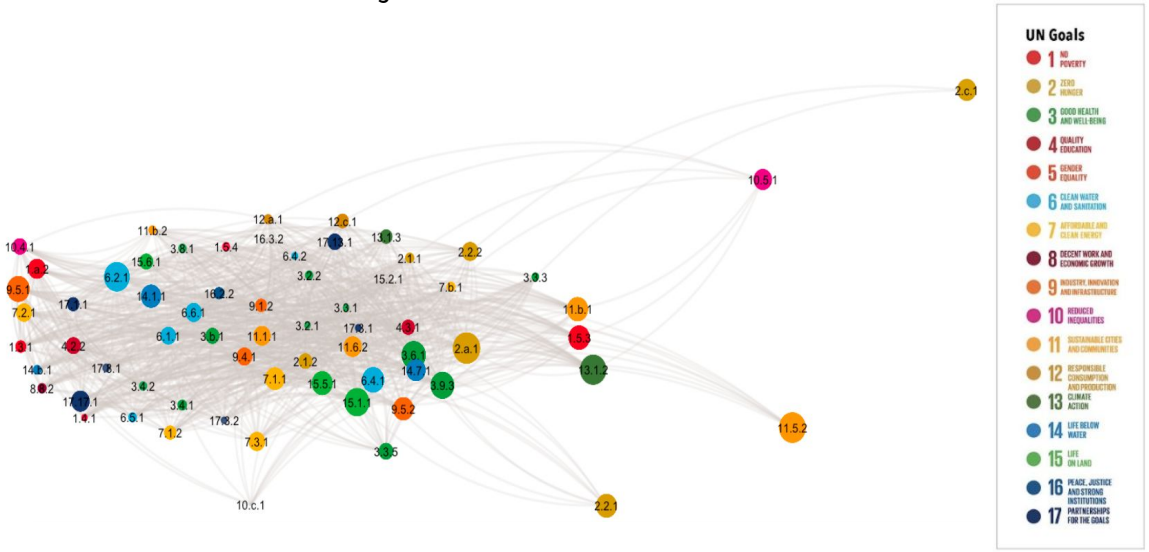
Appendix F: Negative Ties – Summary

- In general, negative ties in network analysis should be used and interpreted with **caution,** as relationships between triads are not always consistent (e.g., while A and B are negatively correlated and A and C are negatively correlated, B and C could be positively correlated, making a full network more difficult to interpret).
- This said, individual negative ties may be interesting to look at, as well as indicators which may exhibit larger degree of negative ties to other indicators (in the aggregate).

Appendix F: Negative Ties (Indonesia)

Coefficient Network

Indonesia network based on negative correlation coefficient



What is it?

This network is analogous to the positive coefficient networks but instead looked at ties with coefficients < 0.

Key takeaway

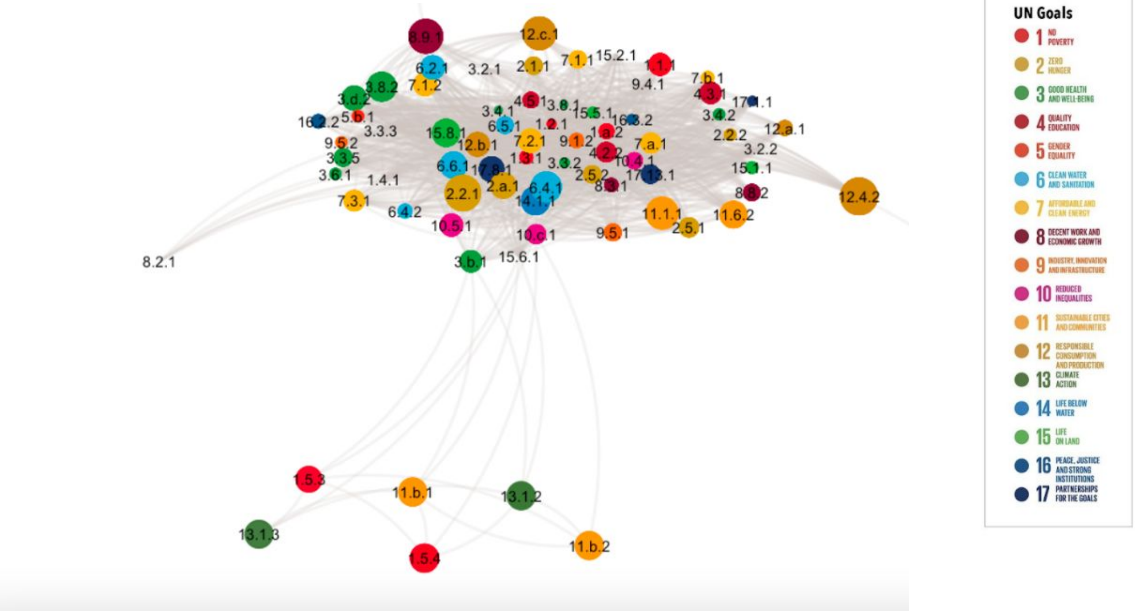
The most connected indicators (i.e., with negative relationships with many other indicators) tend to be those in **Goal 6, Goal 13, Goal 14,** and **Goal 15,** all seemingly focused on ecological themes (clean water, life below water, life on land). This can indicate that **development in other areas** (education healthcare etc.) has not sufficiently included climate action and ecological protection.

Similarly, 11.5.2 appears to be a node with many negative ties, and deals with disasters and loss of life (which may also relate to ecological themes.

Appendix F: Negative Ties (Guatemala)

Coefficient Network

Guatemala network based on negative correlation coefficient



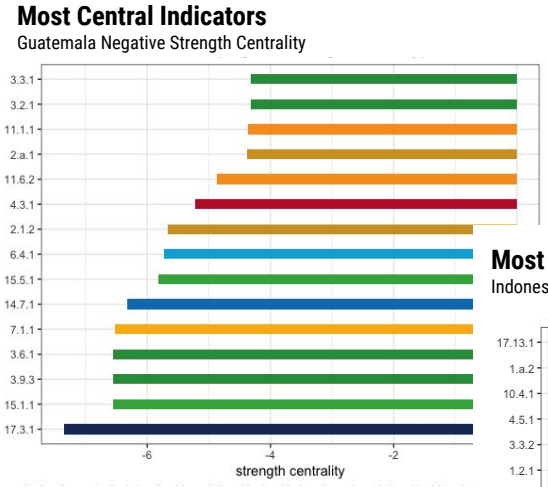
What is it?

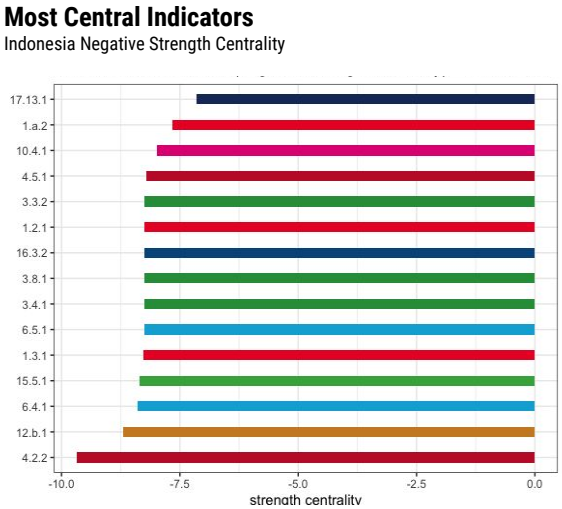
Refer to previous page.

Key takeaway

Similar to Indonesia, **Goal 6, Goal 13, Goal 14,** and **Goal 15,** (clean water, life below water, life on land) have exhibited a number of ties. In this case, **Goal 12** also includes indicators with a large amount of negative ties (e.g., 12.4.2, a measure of *hazardous waste per capita*), as does **Indicator 8.9.1** (tourism revenue proportion of GDP). This also supports the theory that some development, research, innovation, and services seem to correspond to issues with environmental and ecological health / resources.

Appendix F: Negative Ties (Strength Centrality)





- As discussed in the slides above, **strength centrality** measures the sum of statistically significant coefficient ties.

We do not see very much overlap between Indonesia and Guatemala for indicators with largest negative degree centralities. A few indicators from SDG 6 and 15 are consistent across countries; however it does appear that many of the indicators with largest negative strength centrality are from Goal 3 (having to do with health and health risks).