The Regional Dimension

A Bayesian Network Analysis

Marco Scutari
scutari@idsia.ch

Dalle Molle Institute for
Artificial Intelligence (IDSIA)

December 19, 2019

Overview

Bayesian networks:

- Network analysis: graphs and arcs.
- Arcs and correlation.
- Arcs and causality.
- Network analysis and linear regression.
- Model selection.
- Parameter estimation.
- Sensitivity analysis.

To showcase a proof-of-concept model we will use a sample of SDG indicators from a group of African countries and a group of Asian countries.

The General Idea

A network analysis is based on the idea that:

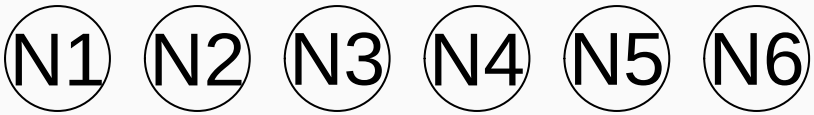
- quantities of interest can be associated to the nodes of a graph; and that
- we can use arcs to represent which variables are correlated with each other.

Hence nodes and variables are referred to interchangeably, as well as arcs and correlations (associations more in general).

The conceptual steps are:

1. identify the variables that are the quantities of interest and draw one node for each of them;
2. collect data on them, gathering a sample of observations;
3. measure whether they are significantly correlated using the data and draw an arc between each such pair.

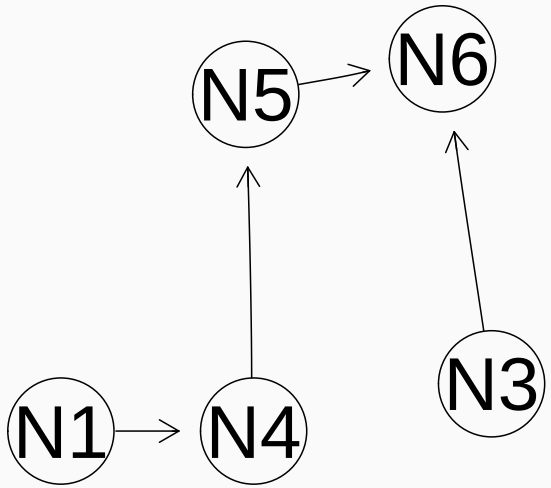
1. identify the variables;



2. collect data;

N1	N2	N3	N4	N5	N6
0.11	0.67	0.37	1.56	-0.77	-0.77
-0.19	-0.52	0.88	-0.44	-0.38	1.28
-1.34	0.19	-1.26	-0.24	-0.31	-0.16
-1.23	-1.01	0.13	-1.15	0.22	0.34
0.85	0.75	0.71	1.17	-0.70	0.29
⋮	⋮	⋮	⋮	⋮	⋮

3. measure correlations and draw the arcs.



Arcs and Correlations

How do we interpret arcs?

- Nodes that are directly connected with each other are directly correlated: changes in one node suggest changes in the nodes that are directly connected with it.
- For nodes that are only indirectly connected, changes in are mediated by the other nodes that are in between.
- If two nodes are not connected at all, that is, there is no sequence of arcs that allows to reach one node from the other, changes in one node will have no effect at all on the other node.

This difference is key to understanding network analysis: any action that affects one node will produce effects that can propagate to directly connected nodes, and from there to nodes that are indirectly connected; but the changes thus produced will become smaller and smaller the farther two nodes are.

Arcs and Causality

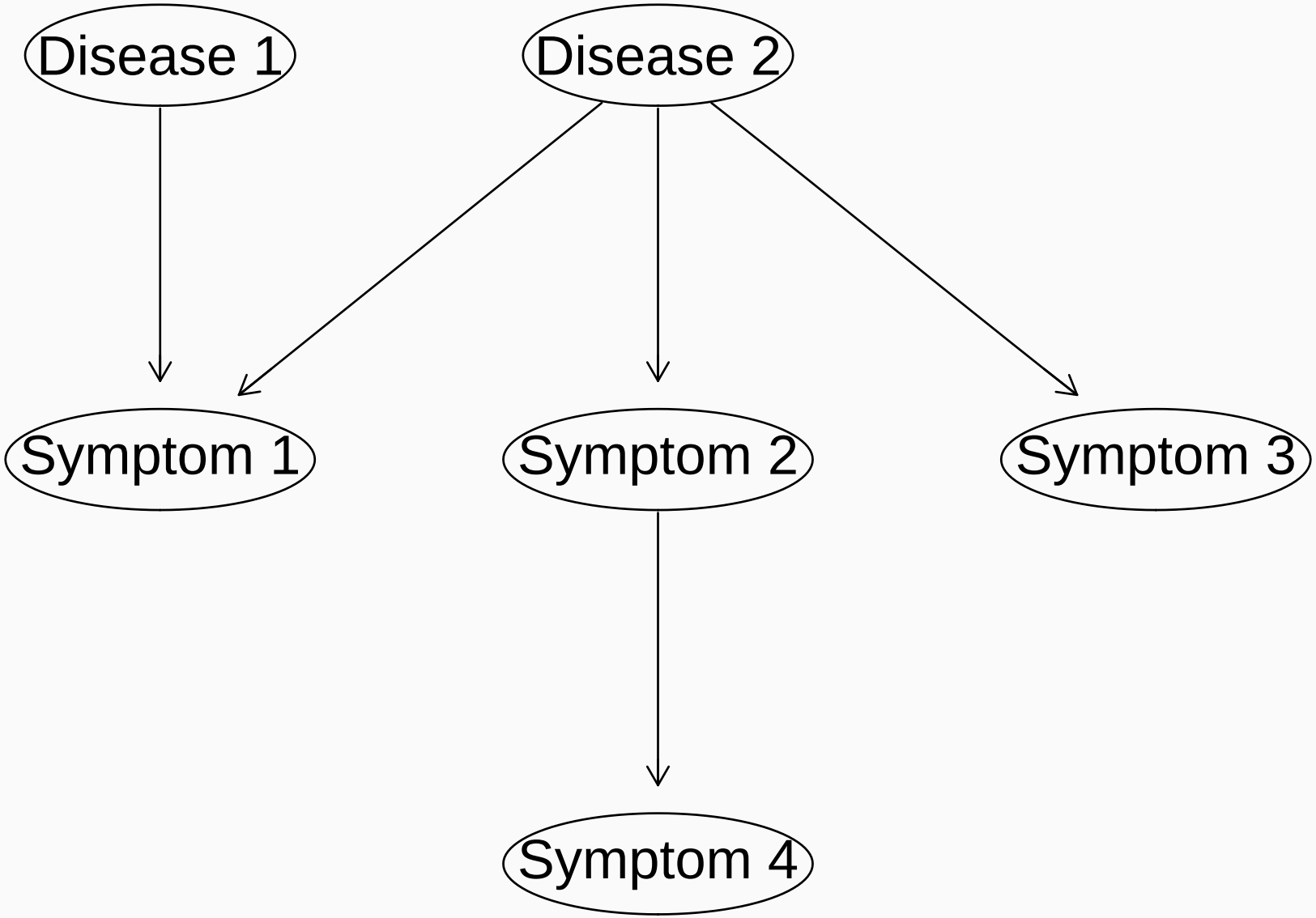
We can give a causal interpretation to arcs, in addition to the statistical interpretation in terms of correlation.

Causation implies correlation: a change in the variable that is identified as the cause may induce a change in the variable that is identified as the effect, and this implies that the two variables are correlated.

The additional information the arc direction provides is that inducing a change in the variable identified as the effect does not imply a change in the variable identified as the cause.

This is not the same as saying that the node identified as the effect provides no information on the node identified as the cause; different effects can be attributed to different causes.

Arcs and Causality: an Example

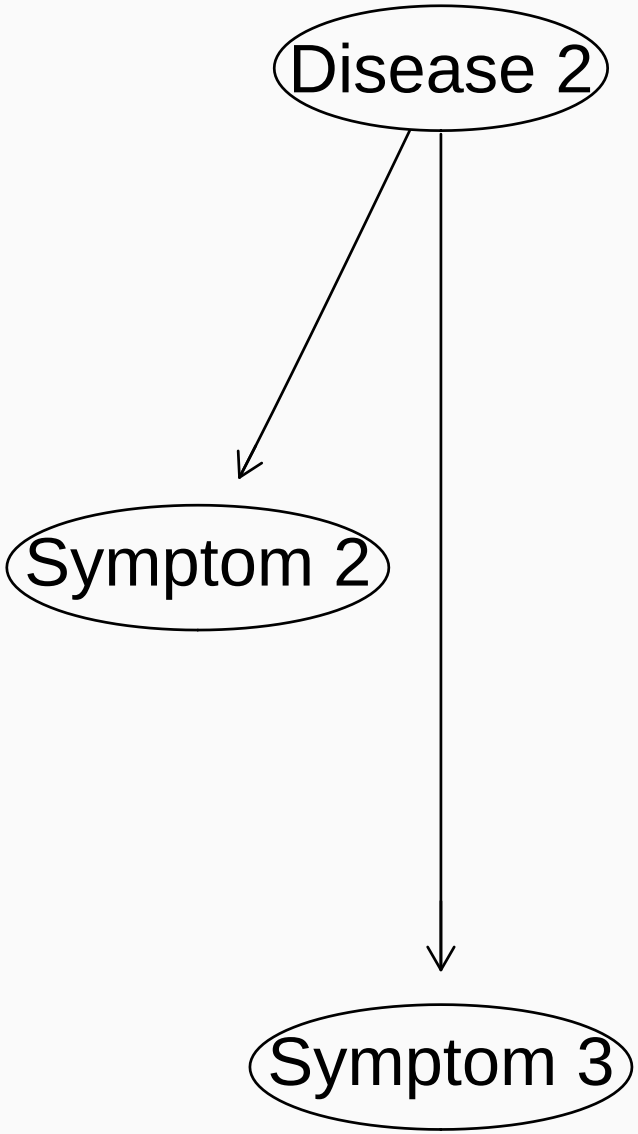


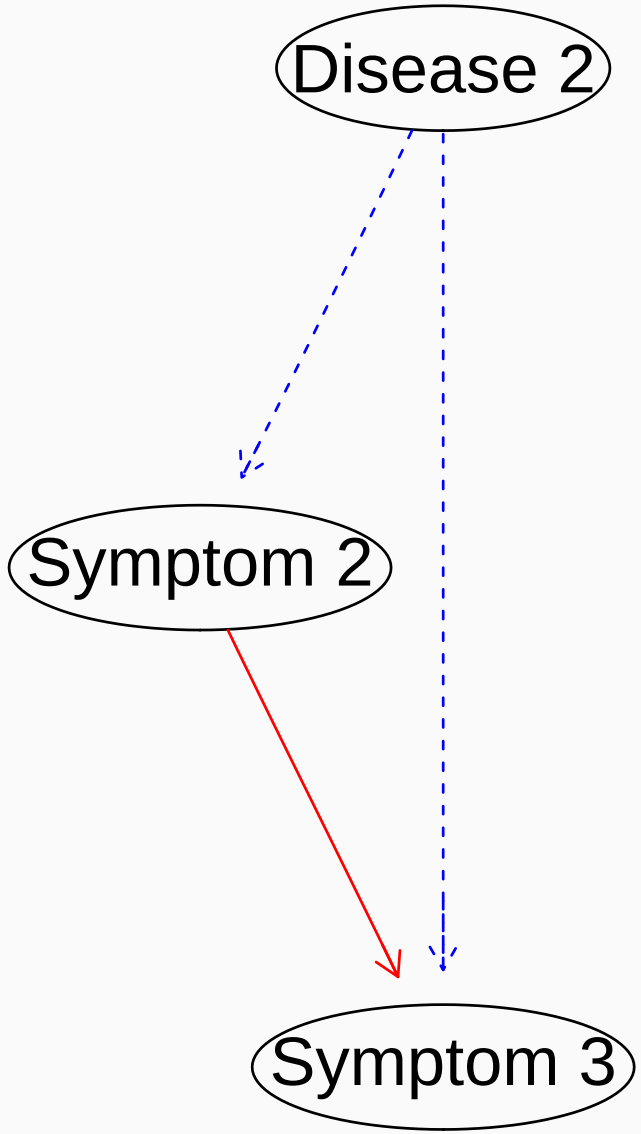
Arcs and Causality: an Example

When a patient requests a visit, the medical doctor will observe various symptoms to diagnose one of several possible diseases.

- Arcs should point from diseases to symptoms since the former cause the latter.
- Diseases and symptoms are correlated, and that is what makes it possible for the medical doctor to decide which disease the patient most likely has; or to present a prognosis with a likely set of symptoms for a given disease.
- Prescribing a therapy that only cures the symptoms will not cure the disease itself; and symptoms will resurface after the therapy stops.
- Directly curing the disease with the appropriate therapy will also eliminate the symptoms.

Difficulties in Getting Reliable Causal Networks





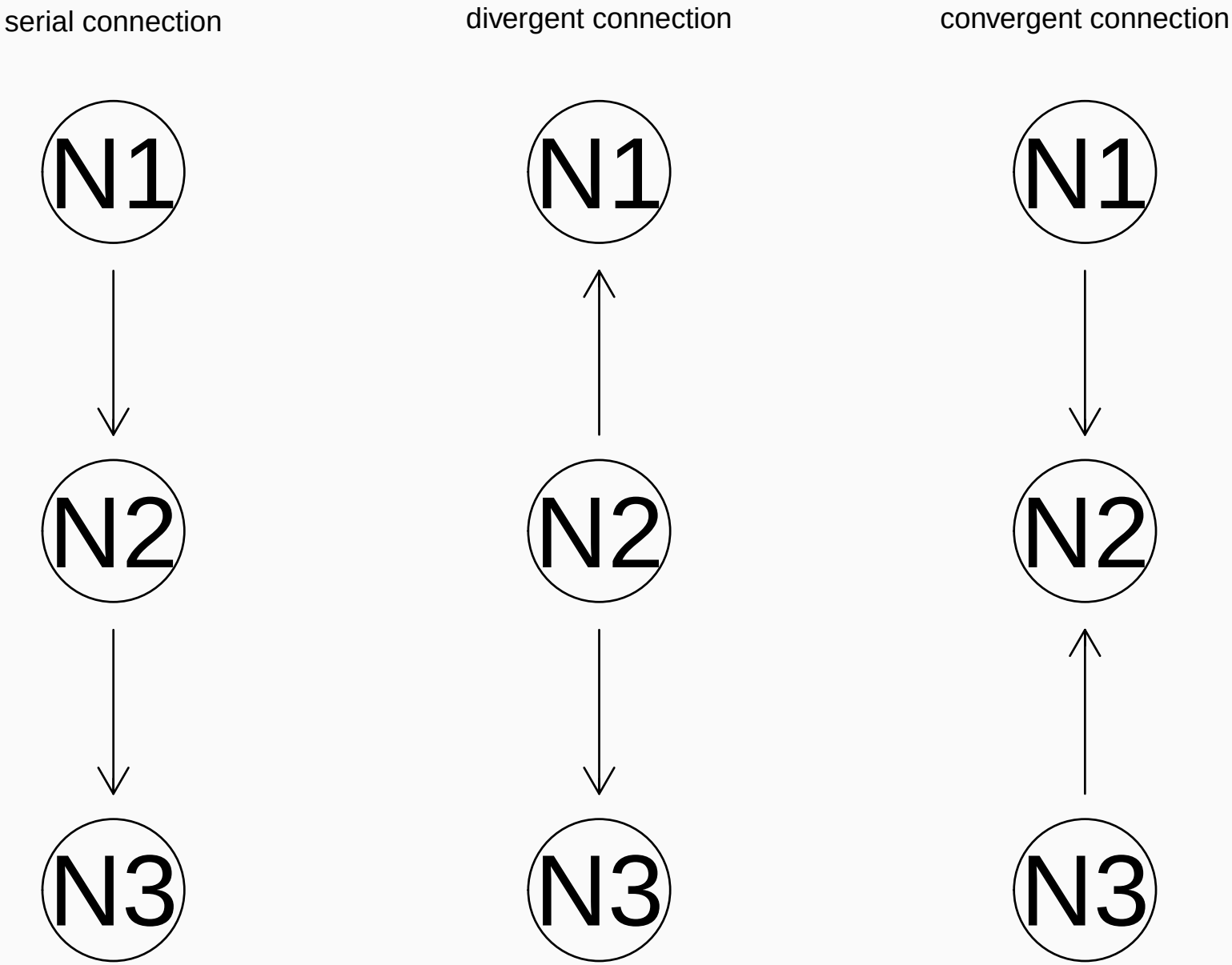
Difficulties in Getting Reliable Causal Networks

The ability of network analysis to correctly link nodes with arcs requires that all relevant variables are observed and included in the model. If that is not the case, arcs may actually represent indirect correlations (as opposed to direct correlations) or even spurious correlations, and their causal interpretation is more difficult to defend.

This is a difficult assumption to defend in any analysis involving socio-economic data such as SDGs.

Even when a large amount of data are available and all relevant variables are included in the network it is not always possible to identify which node is the cause and which is the effect just using data.

Fundamental Connections



We can identify some arc directions just from the data. The serial and divergent connections describe the same statistical distribution since

$$\underbrace{\mathrm{P}(\mathrm{N1})\,\mathrm{P}(\mathrm{N2} \mid \mathrm{N1})\,\mathrm{P}(\mathrm{N3} \mid \mathrm{N2})}_{\mathrm{N1}\to\mathrm{N2}\to\mathrm{N3}} = \mathrm{P}(\mathrm{N2}, \mathrm{N1})\,\mathrm{P}(\mathrm{N3} \mid \mathrm{N2}) =$$

$$\underbrace{\mathrm{P}(\mathrm{N1} \mid \mathrm{N2})\,\mathrm{P}(\mathrm{N2})\,\mathrm{P}(\mathrm{N3} \mid \mathrm{N2})}_{\mathrm{N1}\leftarrow\mathrm{N2}\to\mathrm{N3}}.$$

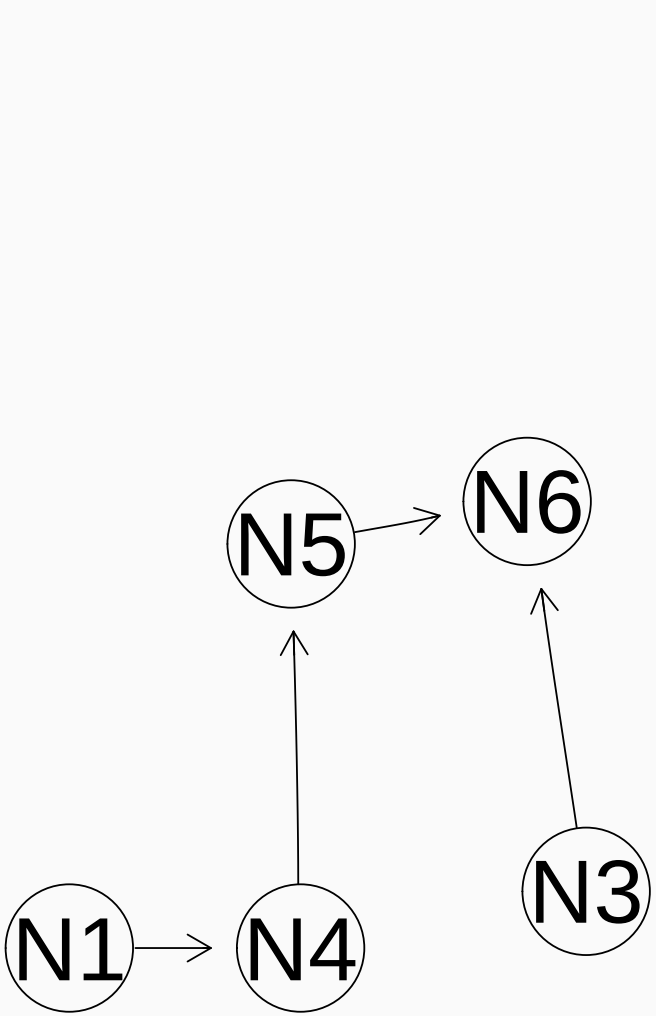
but convergent distribution is not equivalent to the other two. This makes if possible to identify it when it fits the data best, and thus assign directions to most arcs whether we are giving them a causal interpretation or not.

Formally, we call this class of models a Bayesian network (BN). It comprises:

- a directed acyclic graph, that is, a graph in which all arcs have a direction and there are no cycles;
- there is one probability distribution associated with each variable, which in turn is associated with one node in the graph;
- and each distribution is the conditional distribution of the variable given the variables that correspond to the parent nodes in the graph.

The key advantages of BNs are:

- they decompose large models into simpler ones, one for each variable;
- the graph makes it easy to reason qualitatively about the model.



$$
\begin{aligned}
&\mathrm{P}(\mathrm{N1}, \mathrm{N2}, \mathrm{N2}, \mathrm{N3}, \mathrm{N4}, \mathrm{N5}, \mathrm{N6}) \\
&\quad = \mathrm{P}(\mathrm{N1}) \cdot \\
&\qquad \mathrm{P}(\mathrm{N2}) \cdot \\
&\qquad \mathrm{P}(\mathrm{N3}) \cdot \\
&\qquad \mathrm{P}(\mathrm{N4} \mid \mathrm{N1}) \cdot \\
&\qquad \mathrm{P}(\mathrm{N5} \mid \mathrm{N4}) \cdot \\
&\qquad \mathrm{P}(\mathrm{N6} \mid \mathrm{N3}, \mathrm{N5})
\end{aligned}
$$

Each of these distributions is univariate since it only contains one dependent variable.

In the case of continuous variables, we model each node with a linear regression in which:

- the node is the response variable;
- the parents of the nodes are the explanatory variables;
- there is an error term which contains errors that follow a normal distribution with mean zero.

Hence, the parameters of each linear regression are the regression coefficients associated with each parent and the standard error of the residuals.

All parameters are unknown and thus must be estimated from the data. This can be done with a textbook ordinary least squares regression, individually for each node.

The only quantity we need to do that, apart from the data, is the graph, which we need to learn from the data as well since we do not know which arcs it contains.

Before we saw that the BN induces the decomposition

$$\begin{aligned}\mathrm{P}(\mathrm{N1}, \mathrm{N2}, \mathrm{N2}, \mathrm{N3}, \mathrm{N4}, \mathrm{N5}, \mathrm{N6}) = \\ \mathrm{P}(\mathrm{N1})\,\mathrm{P}(\mathrm{N2})\,\mathrm{P}(\mathrm{N3})\,\mathrm{P}(\mathrm{N4} \mid \mathrm{N1})\,\mathrm{P}(\mathrm{N5} \mid \mathrm{N4})\,\mathrm{P}(\mathrm{N6} \mid \mathrm{N3}, \mathrm{N5})\end{aligned}$$

for the nodes; and the distribution for the individual nodes are the linear regressions:

$$\begin{aligned}\mathrm{P}(\mathrm{N1}) : \mathrm{N1} &= \mu_{\mathrm{N1}} + \varepsilon_{\mathrm{N1}} \sim N(0, \sigma^2_{\mathrm{N1}}) \\ \mathrm{P}(\mathrm{N2}) : \mathrm{N2} &= \mu_{\mathrm{N2}} + \varepsilon_{\mathrm{N2}} \sim N(0, \sigma^2_{\mathrm{N2}}) \\ \mathrm{P}(\mathrm{N3}) : \mathrm{N3} &= \mu_{\mathrm{N3}} + \varepsilon_{\mathrm{N3}} \sim N(0, \sigma^2_{\mathrm{N3}}) \\ \mathrm{P}(\mathrm{N4} \mid \mathrm{N1}) : \mathrm{N4} &= \mu_{\mathrm{N4}} + \mathrm{N1}\beta_{\mathrm{N1}} + \varepsilon_{\mathrm{N4}} \sim N(0, \sigma^2_{\mathrm{N4}}) \\ \mathrm{P}(\mathrm{N5} \mid \mathrm{N4}) : \mathrm{N5} &= \mu_{\mathrm{N5}} + \mathrm{N4}\beta_{\mathrm{N4}} + \varepsilon_{\mathrm{N5}} \sim N(0, \sigma^2_{\mathrm{N5}}) \\ \mathrm{P}(\mathrm{N6} \mid \mathrm{N3}, \mathrm{N5}) : \mathrm{N6} &= \mu_{\mathrm{N6}} + \mathrm{N3}\beta_{\mathrm{N3}} + \mathrm{N5}\beta_{\mathrm{N5}} + \varepsilon_{\mathrm{N6}} \sim N(0, \sigma^2_{\mathrm{N6}})\end{aligned}$$

Model Selection and Estimation

The process of estimating such model is called *learning*, and consists in two steps:

1. structure learning: learning which arcs are present in the graph, that is, which nodes are statistically significant regressors for which other nodes;
2. parameter learning: learning the parameters that regulate the effect sizes of those dependencies, that is, the regression coefficients associated with the parents of each node.

The former corresponds to model selection, and the latter to model estimation. Both can be carried out with standard methods from classic literature from linear regression models.

We will showcase how to learn a BN along those lines using, as a proof of concept, as set of indicator for SDGs.

The indicators have been recorded on two separate sets of 30 African countries and 26 Asian countries:

- African countries: Angola, Botswana, Burkina Faso, Burundi, Cote d'Ivoire, Cabo Verde, Cameroon, Congo, Eswatini, Ethiopia, Ghana, Kenya, Lesotho, Madagascar, Malawi, Mali, Mauritius, Mozambique, Namibia, Niger, Nigeria, Rwanda, Senegal, Seychelles, Sierra Leone, South Africa, Togo, Uganda, United Republic of Tanzania, Zambia.
- Asian countries: Afghanistan, Armenia, Azerbaijan, Bangladesh, Bhutan, Cambodia, China, Georgia, India, Indonesia, Iran (Islamic Republic of), Kazakhstan, Kyrgyzstan, Lao People's Democratic Republic, Malaysia, Maldives, Mongolia, Myanmar, Nepal, Pakistan, Philippines, Sri Lanka, Thailand, Timor-Leste, Uzbekistan, Viet Nam.

SDG Indicators, 2012--2014

For each group of countries, we consider the following indicators (called *Goal.Target.Indicator*):

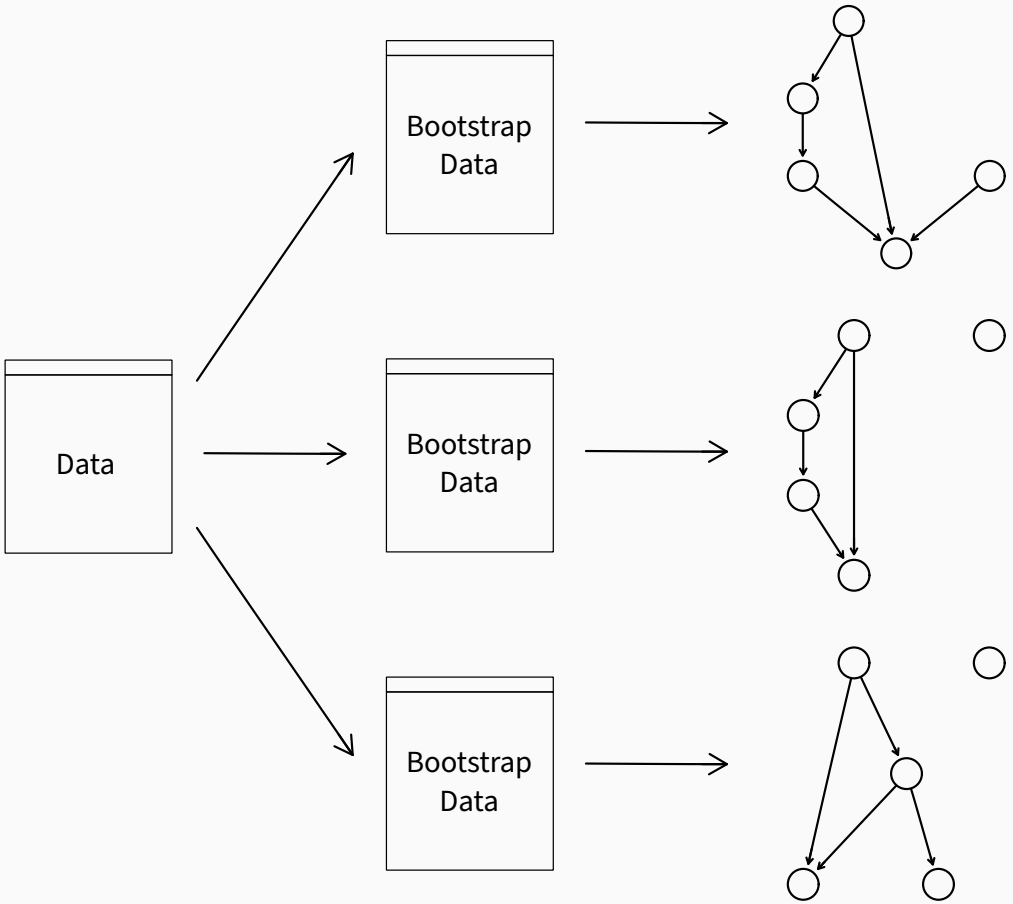
- African countries: 10.b.1, 15.1.2, 15.5.1, 15.a.1, 17.12.1, 17.3.2, 17.8.1, 17.9.1, 2.a.2, 3.2.1, 3.2.2, 3.3.2, 3.b.1, 3.b.2, 4.b.1, 5.5.1, 6.6.1, 6.a.1, 7.2.1, 7.3.1, 8.1.1, 8.4.2, 8.a.1, 9.2.1, 9.a.1
- Asian countries: 15.5.1, 15.a.1, 17.3.2, 17.8.1, 17.9.1, 2.a.2, 3.2.1, 3.2.2, 3.3.2, 3.3.5, 3.b.1, 3.b.2, 4.b.1, 6.6.1, 6.a.1, 7.2.1, 7.3.1, 8.1.1, 8.2.1, 8.4.2, 8.a.1, 9.2.1, 9.a.1

This combination of countries, years and indicators has been chosen to maximise the number of available data in terms of how many indicators are measured simultaneously. Having simultaneous measurements of all variables under investigation across all countries is crucial in measuring how those variables are correlated.

SDG Indicators, 2012--2014

Country	10.b.1	15.5.1	15.a.1	17.3.2	17.8.1	17.9.1	⋯
Afghanistan	1.188	1.000	1.219	0.904	0.891	1.405	⋯
Afghanistan	0.931	1.000	1.068	1.139	0.964	0.799	⋯
Afghanistan	0.880	0.999	0.712	0.955	1.144	0.794	⋯
Armenia	0.731	1.000	0.269	0.972	0.839	0.824	⋯
Armenia	1.402	0.999	0.097	1.062	0.937	0.990	⋯
Armenia	0.866	0.999	2.633	0.965	1.226	1.185	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	
Uzbekistan	0.776	1.000	1.901	1.036	0.824	1.018	⋯
Uzbekistan	1.361	0.999	1.062	1.092	0.935	1.059	⋯
Uzbekistan	0.861	0.999	0.035	0.871	1.239	0.922	⋯
Viet Nam	0.859	1.005	1.158	0.998	0.949	0.967	⋯
Viet Nam	1.255	1.000	0.319	0.998	0.993	1.203	⋯
Viet Nam	0.885	0.994	1.521	1.002	1.057	0.829	⋯

Structure Learning: a Summary



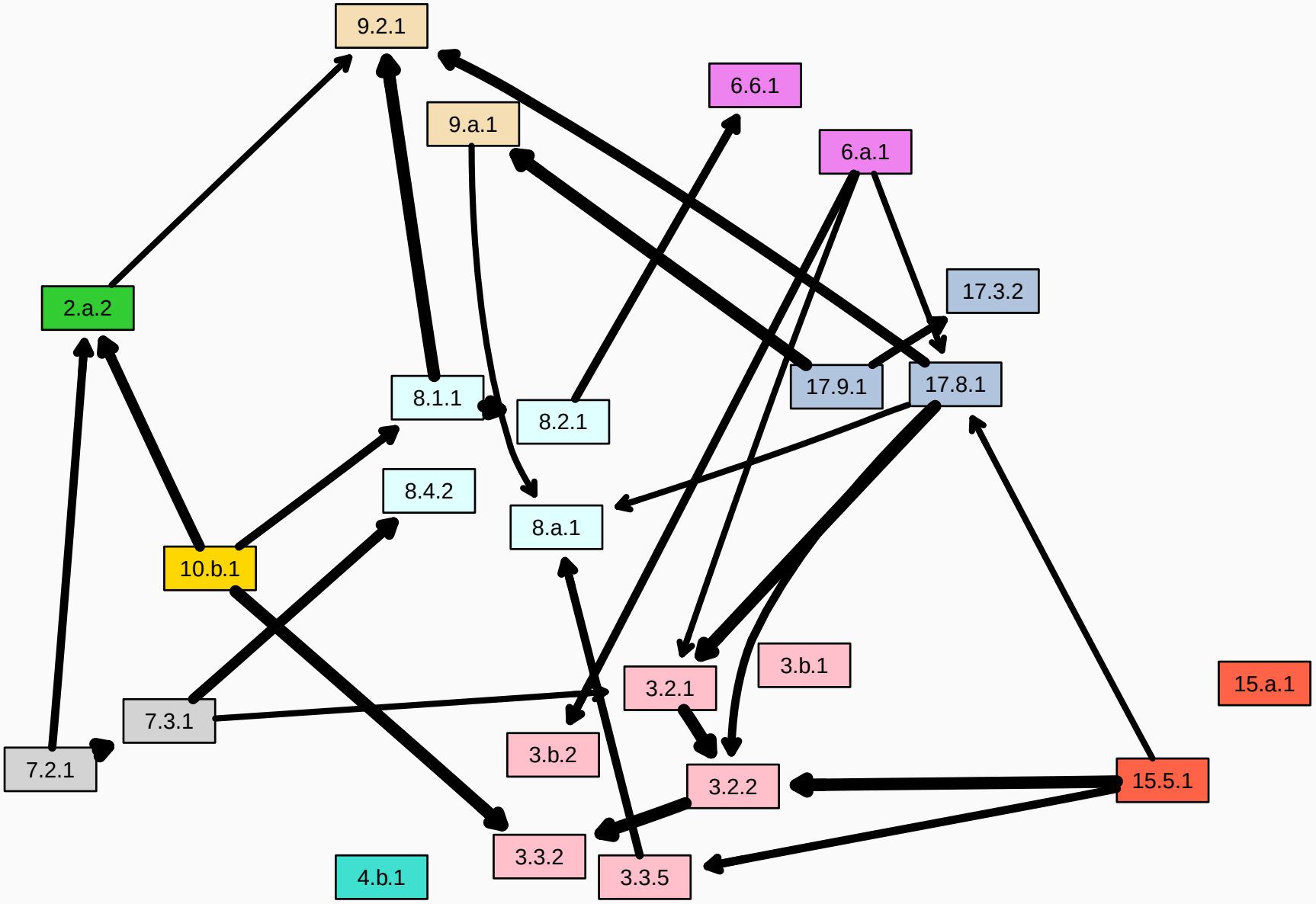
- Perform bootstrap to create artificial samples from the data.
- Learn the structure of the network by finding the network with the best goodness-of-fit from each bootstrap sample.
- Count the frequency of each arc appearing in the learned networks.

Structure Learning: Arc Strength

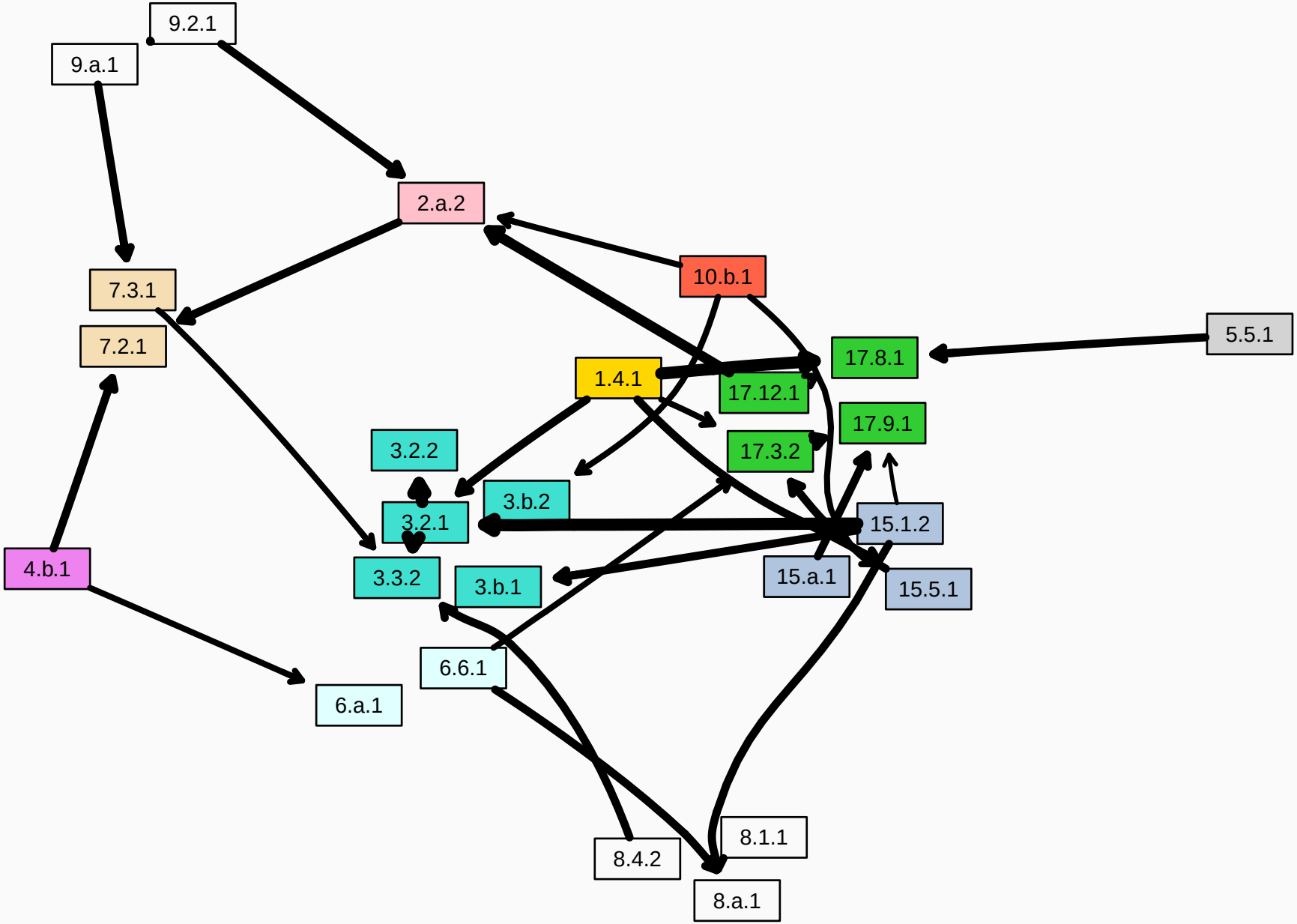
Thanks to this procedure we can quantify our confidence in each arc (its strength) and its direction, and we can create consensus networks with the arcs whose strength is above a threshold.

from	to	strength	direction
1.4.1	15.5.1	0.815	0.887
1.4.1	17.12.1	0.295	0.763
1.4.1	17.3.2	0.700	0.879
1.4.1	17.8.1	0.990	0.874
1.4.1	17.9.1	0.125	0.800
1.4.1	2.a.2	0.055	0.636
1.4.1	3.2.1	0.820	0.912
1.4.1	3.3.2	0.060	0.917
1.4.1	3.b.1	0.600	0.592
1.4.1	3.b.2	0.090	0.611
⋮	⋮	⋮	⋮
17.9.1	15.5.1	0.195	0.513
17.9.1	3.2.2	0.065	0.769
17.9.1	4.b.1	0.125	0.600
17.9.1	6.6.1	0.555	0.518
17.9.1	7.3.1	0.310	0.613

Regional Network for the Asian Countries



Regional Network for the African Countries



Considerations on this Proof of Concept

- Linear regression models are not very well suited to capturing nonlinear relationships, and can be misled by outliers.
- Classic linear models assume observations are independent, but indicators for each country are clearly not, and...
- data should be collected in consecutive years to minimise changes in the surrounding economic conditions that may act as confounders.
- Different ways of normalising and/or de-trending the data lead to different BNs, with no clear winner.
- The number of observations could be artificially improved by increasing the frequency or the geographical granularity with which Indicators are recorded, at the cost of increasing noise.
- The selection of Sustainable Development Goals could be refined to answer specific questions, leading to simpler models that are easier to interpret.

Thanks!

Any questions?