Estimating Networks of Sustainable Development Goals

Luis Ospina-Forero[1], Gonzalo Castañeda[2], and Omar A. Guerrero[3,4]

[1]Alliance Manchester Business School, The University of Manchester, Manchester

[2]Centro de Investigación y Docencia Económica (CIDE), Mexico City

[3]Department of Economics, University College London, London

[4]The Alan Turing Institute, London

Abstract

An increasing number of researchers and practitioners advocate for a systemic understanding of the Sustainable Development Goals (SDGs) through network analysis. Ironically, the burgeoning network-estimation literature seems neglected by this community. We provide an introduction to more suitable estimation methods for SDG networks. Building a dataset with 87 development indicators in four countries over 20 years, we perform a method-comparison study. We find important differences in the estimated network topologies, as well as in synergies and trade-offs between SDGs. Finally, we provide some guidelines on the potentials and limitations of estimating SDG networks for policy advice.

1 Introduction

In recent years, the concepts of inclusion and sustainability have become central to socioeconomic development. Perhaps, the best example is the Sustainable Development Goals (SDGs), the leading international agenda for national and regional development strategies. As the 2030 Agenda has progressed, it has being recognized that, in order to truly achieve sustainable development, it is necessary to understand how its multiple dimensions interact with each other (Nilsson et al., 2016). In other words, a multidimensional view of development requires well-defined procedures to quantify and operationalize networks of interdependencies between different goals (or their indicators). Following this idea, several studies have attempted to measure such networks through different methods, for example, subjective criteria from expert advice; text mining applied to official documents and names of development indicators; and proximity measures between indicators that are relevant for a given country.

Overall, these efforts are commendable as they shift the discussion on socioeconomic development from topical silos to a systemic view (OECD, 2018). However, scholars from development studies have not provided a –much needed– thorough reflection and analysis on the virtues and shortcomings of the different methods available for the empirical estimation of SDG networks. Furthermore, the number of alternatives considered in these studies has been rather small. This stands in stark contrast with the growing literature on network estimation and causal inference models produced by data and network scientists.[1]

This paper analyses different network-estimation methods with the aim of identifying the most suitable procedures to build networks of development indicators. For this reason, we examine their underlying assumptions, strengths and limitations. Broadly speaking, we classify all available methods into five families: correlation thresholding, Granger causality, chordal information filtering, statistical structure learning and physics-inspired methods. In-

[1]The literature on Quantitative Sustainability Assessments has already made a call for the need to embrace causal methods in SDG networks (Cucurachi and Suh, 2017).

terestingly, only a handful of methods can be employed with empirical development-indicator data. Thus, we perform a comparative study using the methods that show potential to work with this kind of observational data.

Before introducing the reader to network-estimation methods and developing the empirical analysis, we provide a critical account of the current attempts to estimate SDG networks, as well as a brief but necessary discussion on the differences between association, structural-dependence and causation. The paper is structured in the following way: Section 2 reviews the methods that have been used to estimate SDG networks and discusses their contributions and drawbacks. Section 3 provides a general overview of five methodological families that are frequently employed by network and data scientists. Section 4 explains, in more detail, the four methods from these families that are best-suited to construct SDG networks. Section 5 presents a comparative study applying the four selected methods on development-indicator data. Finally, the paper closes with the conclusions in Section 6, where we synthesize empirical results, discuss some of the limitations of causal inference on SDG networks, and propose some guidelines.

2 On networks of sustainable development goals

As discussions on socioeconomic development have shifted toward a multidimensional view (General Assembly, 2015), international organizations and development analysts have come to the conclusion that a network framework is indispensable. At the same time, it has become evident that complexity arises when trying to understand how these multiple dimensions interact with each other. For example, reductions to infant malnutrition may cause children to perform better at the PISA tests. Therefore, a policy originally aimed at improving the dimension of public health may also have implications in the dimension of education. In this example, causality may need to be considered both ways, as improvements in education may also create consciousness about the risks of unhealthy diets. Either way, it is clear that

development does not take place within isolated topics, but it is rather a co-evolutionary process across different dimensions.

Due to the complexity arising from interdependencies between development indicators, it has been argued that policy interventions can only be properly designed when the synergies and trade-offs between different goals have been identified. This is especially important when social, economic and sustainability indicators are jointly considered in a government's development strategy. In order to frame development goals in a setting of complex networks, it is important to take into account the following issues: (1) the way of interpreting an interdependency network (does a link represent causation, structural-dependence or correlation?), and (2) the assumptions underlying the estimation methods (which methods are more suitable given the nature of the available data?). In this section, we discuss issue 1, while leaving 2 for the coming sections. Before entering into these discussions, however, it is important to provide a critical account of the studies that have previously attempted to estimate SDG networks.

2.1 The first generation of SDG networks

The first studies that tried to understand SDGs through the lens of networks appeared in the development literature a decade ago. According to their estimation procedures, we can classify them in two groups: (1) subjective studies that rely on qualitative information (*e.g.*, the conceptual description of the variables), and (2) statistical ones that make use of panel data (countries through time). In group 1, subjectivity comes from determining links based on opinions. Thus, in this group, we can find studies that take either a brainstorming (*e.g.*, expertise and stakeholders knowledge) or a heuristic approach (*e.g.*, informal text mining). In contrast, group 2 applies quantitative techniques to development-indicator data. In the networks estimated in group 2, each node corresponds to a specific indicator, while a link or edge denotes the strength of a relationship between nodes. Usually, these graphs have weighted edges and, in fewer cases, these edges have directions.

Among the subjective studies, we can find Pedercini and Barney (2010); Blanc (2015); Collste et al. (2017); Weitz et al. (2018); Allen et al. (2018). Examples of statistical studies are Czyżewska and Mroczek (2014); Ceriani and Gigliarano (2016); Castañeda et al. (2017); Cinicioglu et al. (2017); Pradhan et al. (2017); El-Maghrabi et al. (2018). It is important to clarify that the first four statistical studies do not make a direct reference to SDGs; nonetheless, we consider them relevant since they are early contributions that use development-indicator data. Finally, the work of Zhou and Moinuddin (2016) seems to be the only one standing in both camps.

For clarity, and to align our exposition with the literature semantics, we use the terms *policy issues*, *development indicators* and *development goals* (or just goals) interchangeably to refer to a specific topic where a policy can be directed (*e.g.*, education, public health, poverty, etc.). Then, in order to refer to the development aspirations of a government or society, we use the term *objectives*. Likewise, an SDG is a development pillar that encompasses a set of policy issues which, *ex-ante*, are thought to be closely associated. Thus, we can think of policy issues as the topics of interest and of the objectives as the final states where the policymaker wants to arrive (*i.e.*, specific values for a set of development indicators belonging to different SDGs).

In most of the first-generation studies the aim has been identifying policy issues with synergistic effects to other goals (positive spillovers) in order to promote them. Similarly, estimating goals that face trade-offs (negative spillovers) is also important to disencourage policies that improve certain indicator but obstruct other objectives. This type of analysis prevails in subjective studies, in which first- and second-order spillovers are inferred (*e.g.*, Blanc (2015); Pradhan et al. (2017); Weitz et al. (2018)). In other studies, different centrality measures are calculated with the purpose of identifying influential policy issues (*e.g.*, **?**Allen et al. (2018).

There are several ways in which subjective studies build SDG networks. First, edges can be constructed through information derived from previous studies. Second, knowledge from

stakeholders and experts in specific fields can be leveraged to propose networks. Third, the text of official documents (*e.g.*, from the United Nations) can be mined to build networks of development goals connected by commonly mentioned policy issues. Regarding statistical studies, we identify three dominant approaches: inference through plain correlations (Zhou and Moinuddin, 2016; Pradhan et al., 2017; Castañeda et al., 2017), Bayesian techniques (Czyżewska and Mroczek, 2014; Ceriani and Gigliarano, 2016; Cinicioglu et al., 2017) and a co-occurrence methodology (El-Maghrabi et al., 2018).[2] The co-occurrence methodology establishes, first, if the performance of a country in a particular indicator is above the average country from nations with similar per capita income. Then, the method infers the likely co-occurrence between two indicators (*i.e.*, if they have proximate mechanisms for delivering an above-average performance).

2.2 Shortcomings of the first generation

In the context of SDGs, where the ultimate goal is to facilitate the prescription and evaluation of public policies (besides inferring the interdependencies of the development process), it is important to specify broad desirable qualities of network-estimation methods. In our view, such attributes are scalability, replicability, specificity, directionality and validity. Next, we elaborate on each of these concepts, and suggest that these desirable properties are rarely fulfilled jointly in the first generation of SDG networks.

2.2.1 Scalability

A network-estimation procedure for SDGs should be easily scalable in order to incorporate as many socioeconomic indicators as possible (appealing to the multi-dimensional nature of development). This has become self-evident as we passed from the 48 indicators considered by the Millennium Development Project to the 232 (and growing) of the SDGs. Such high

[2]Note that, although Czyżewska and Mroczek (2014) briefly discuss the challenge of inferring causality, none of these studies formally attempt to establish causal relationships, yet their results are usually interpreted as if this were the case.

dimensionality calls for representing the policy space as a complex network of goals. This, however, is not trivial when the number of observations for each indicator (usually less than 10) is lower than the number of dimensions typically observed in development-indicator data. Another aspect of scalability is the capacity to estimate the network with relative ease. This is hardly achieved by methods relying on expert advice. For instance, a government may be unable to gather enough experts in so many different issues due to budgetary and time constraints, or simply because such experts are unavailable (something common among poor countries and, at the sub-national level, even in middle and high income nations). Furthermore, as the number of indicators increases, the amount of experts needed to understand all the possible direct and indirect relations becomes impossible to fulfill.[3]

2.2.2 Replicability

Methods that facilitate the replication of empirical studies are desirable in all scientific endeavours. Hence, the techniques through which information is obtained and processed should be accessible to third parties who wish to verify or refine an estimated network. In this view, studies based on expert advise are rarely replicable because human capital is highly scarce in numerous developing countries. Furthermore, even if such experts exist, it does not mean that they are accessible. Moreover, the process through which they arrived to their original estimations might have changed because it may not be systematic or transparent, but rather the result of their individual experiences and perceptions. Finally, brainstorming methods are often biased through 'echo chamber' effects, or the current conventional wisdom (*e.g.*, the relationship between economic openness and job creation). Hence, dealing with such biases requires additional protocols on how to collect expert advise.

[3]This does not mean, however, that this kind of data is not useful to estimate SDGs. In fact, in section 5.2 we argue that expert advise and anecdotal experience can be used to construct complementary *ex post* validation tests.

2.2.3 Specificity

Extensive evidence from real-world development policies has shown that context matters (Rodrik, 2009) and, hence, policies have to be adjusted to the social, economic and political environment where they are supposed to be applied. For this reason, networks that are estimated by pooling data from different countries may fail to be specific to the country of interest (as it is done in Czyżewska and Mroczek (2014); Ceriani and Gigliarano (2016); Cinicioglu et al. (2017); El-Maghrabi et al. (2018)). Similarly, tweaking a 'master' network –previously built for analyzing other countries– also runs the risk of neglecting a country's context (as in Pedercini and Barney (2010)). When time series are too short, a second-best alternative, consists of pooling data from a reduced group of countries with structural similarity (demonstrated through statistical cluster analysis, for example).

2.2.4 Directionality

Since the purpose of building SDG networks is to conduct evidence-based policymaking, one should be cautious when interpreting the outcomes of these estimates. For instance, it is not enough to identify the high centrality of an indicator to argue that the associated policy issue should be prioritized (*e.g.*, when recognizing that the country's electrification is critical for its development because it connects to many other topics). Among other things, it is crucial to identify the direction of edges in these networks. Hence, estimations built on standard text mining, co-occurrence methods or plain contemporary correlations fail to meet the property of directionality.

In principle, 'common sense' can be used to discern between the potential direction of an edge, as there may be indicators that cannot precede others. However, as we move to larger datasets, more complex topologies and higher topical specificity, establishing directions in this fashion takes us back to the scalability problems found when using expert advice. Therefore, detecting directionality in a data-driven way is highly desirable.

2.2.5 Validity

As in any estimation procedure, network inference requires some form of validation to be considered scientifically reliable. In the context of SDGs, external validation through out-of-sample prediction is unfeasible due to the short span of SDG time series. A popular alternative among network scientists is to create data-generating computational models and test how well the estimation method can recover the network specified in such models. This, however, is not possible with SDGs because there is no comprehensive understanding of how development indicators emerge from the interactions of socioeconomic agents at a much more dissagregate level.[4] Therefore, an admissible alternative is appealing to validated methods which were developed in other disciplines, but which are designed for data with a similar structure. Accordingly, none of the methods used in the first generation have passed any formal validation test.

It is important to note that we consider the previous attributes essential to network-estimation methods in the context of SDGs. However, their joint fulfillment represents a major challenge. In addition, it should be pointed out that all of the first-generation studies lack most of these requirements. Although, from our description, it is also clear that some of them have certain strength when evaluating particular aspects.

2.3 What can we infer from first-generation networks?

The first-generation of SDG networks has provided an early approximation to the description of the policy space that is relevant to the 2030 agenda. Among other useful estimates, we find associations between policy issues that share links; clusters of nodes obtained through community detection algorithms; signed edges indicating a positive or negative co-movement between pairs of variables; node centrality (*e.g.*, degree, eigenvector, betweenness or closeness) indicating influence on the network or ability to connect communities; and network

[4]This is also a limitation in the methods reviewed in this paper. In other words, by employing these methods, we are assuming that development-indicator data comes from underlying data-generating processes that follow some common statistical assumptions.

sparsity showing if indicators tend to connect more with issues from their same SDG than with others. Besides these descriptive insights, one of the main virtues of the first-generation networks is that they have helped educating analysts and policymakers about the complex and systemic nature of development.

First-generation networks might also be helpful to improve policy heuristics. For instance, in the co-occurrence approach (El-Maghrabi et al., 2018), it is argued that policy priorities can be determined by means of two criteria: (1) the feasibility of improving an under-performing indicator (with respect to similar countries) given its proximity to other nodes that perform relatively well (density analysis); and (2) their potential for improving other indicators, measured via degree centrality of the intervened policy issue. Nevertheless, one should still be careful when interpreting these criteria. For example, a highly connected node might be the result of incoming links that are not observed when estimating undirected networks and, thus, its promotion could result in poor outcomes.

Finally, an important way in which first-generation networks can be used is not for direct policy interpretation, but as structural information for more comprehensive models. For example, together with other structural information such as input-output matrices, the tradition of 'system dynamics' incorporates these networks as model components. This approach highlights the fact that policy intervention analysis cannot be solely based on an exercise of synergies and trade-offs of development indicator data alone. It suggests that the causal chain from resource allocation to changes in indicators is complex, limiting our ability to provide policy prescriptions out of network estimates of a partially observed system (i.e. raw sustainable indicator data).

2.4 On causality and policy prescriptions

The recurrent "correlation does not imply causation" phrase should have already come into the reader's mind. Thus, it is pertinent to draw distinctions between the different types of inferences that can be achieved through existing network-estimation methods, as well as

their underlying assumptions and limitations. This is especially important in the context of SDGs, where much of the analysis tends to be used to inform policymakers (unavoidably implying some sort of causation).

First, let us differentiate three types of inference relations: association, structural-dependence and causation. An association is a co-movement between two variables without distinguishing their origin or direction. That is, when X (Y) changes, Y (X) is observed to change too.[5] Structural-dependencies between variables have explicit directions; for example, when X varies, we can also observe changes in Y, but when the latter varies we do not need to observe changes in the former.[6] Finally, causation indicates that there exists 'cause and effect' relationships between variables. That is, X is a cause of another variable Y, when manipulations of the former systematically affects the outcomes of the latter, after controlling for a set of Z variables, possibly related to Y.[7] From this explanation, it becomes evident that, for causal inference to be possible, a directional link between two variables is a necessary but not sufficient condition.

All first-generation SDG networks measure either associations or structural-dependencies, but not causation. However, causality may be the most relevant relationship to be inferred from an SDG network because, ultimately, one would like to provide advice on which nodes to intervene through public policies. Consequently, any policy advice derived solely from association and structural-dependence networks should be taken with a pinch of salt. This makes evident that, besides the recurrent correlation $\neq$ causation, there are other important aspects that are generally overlooked in the estimation of SDG networks, for example, how feasible is it to exercise a policy intervention in an indicator? do interventions take place

[5] An association can be described in different forms: co-occurrence (*i.e.*, a pair of variables that tend to appear in tandem) and contemporaneous correlation (*i.e.*, variables that tend to exhibit a monotonic relationship).

[6] In Bayesian networks, the link's weight is measured through a conditional probability $\Pr(Y|X)$, so that X is the 'parent' node and Y is direct 'child' (X precedes Y). In networks of non-contemporaneous correlations, the direction indicates a time sequence and the weight corresponds to a correlation coefficient (Pearson or Spearman) between a contemporary variable and a lagged one.

[7] In mathematical terms, following the notation of Pearl (2000), causation relates to probabilities of the type $\Pr(Y/\text{do}(X_0))$, where $\text{do}(X_0) = \text{do}(X = X_0)$ means that variable X has been intervened. Thus, classical statistical inference is not the same as causal statistical inference.

at the same level of aggregation as the outcome variables? is it possible to identify all the confounding factors affecting causal relationships and/or the Indicators? etc. While these questions are out of the scope of this paper, we offer some thoughts in Section 6.

It must be noted that the definitions of causation used by the methods reviewed in this paper fall into a particular type: the 'dependence account' (Hall, 2004). Here, causal factors are those whose presence makes possible the existence of one or several outcomes (effects).[8] Because development indicators are aggregate variables (see Section 2.5), by estimating causal SDG networks under the dependence account, one implicitly assumes, among other things, that outcomes and interventions take place at the same level of aggregation; something that may not hold in the context of SDGs. This is the case, for instance, when analyzing the impact of financial development on economic growth. Here, both the 'causal variable' and the outcome are measured at the country level, while policy interventions (*e.g.*, rules of supervision and competition) are implemented at the micro-level (*e.g.*, banks and other financial institutions). Hence, recommendations derived from the dependence account need to be considered carefully.

2.5 Challenges of SDG data

Now that we established the desirable qualities for SDG-network estimations and differentiated the types of inferences, it is important to provide some clarifications on the nature of the data and on the challenges related to building an 'ideal' method. First it should be clear by now that estimating SDG networks through statistical methods requires quantitative data. The relevant data available for this type of analysis consist of development indicators. Second, these indicators have a time-series structure (*i.e.*, today's values tend to depend on

[8] According to Casini and Manzo (2016), acyclic causal graphs make use of a dependence account of causation and a horizontal view of the causal mechanisms. Such view consists of a set of stable relationships in a network of variables. They can be represented as algebraic equations or as network topologies such as cascades ($X \rightarrow Y \rightarrow Z$) and branches ($Y \leftarrow X \rightarrow Z$). An alternative account is, for example, the 'production' one, in which causal factors are those that help to *generate* or bring about specific outcomes. This account, however, requires a different set of techniques that lie beyond the statistical methods here covered.

previous ones). Therefore, an ideal method should take temporal dependence into account. Third, these time series are short (*e.g.*, usually less than 20 annual observations, depending on how many countries and indicators are collected for the study). Thus, the ideal method should be able to handle the statistical problems associated to small samples. Fourth, SDG data is high-dimensional, hence an ideal method should also work in this setting.

The fact that development indicators have temporal dependencies does not imply that time is their only or main determinant. In fact, it is likely that many of the associated policy issues are continuously influenced by policy interventions to the system. In fact, a common concern in time-series analysis is the presence of 'structural breaks' along the time interval under study, especially if the sampling period involves different government administrations. Therefore, while time-series based methods may be useful, there is also something to be learned from intervention-based causal inference. Accordingly, we also consider methodologies that assume that the data come from independent realizations of a distribution.[9]

We should add that an 'ideal' method for estimating SDG networks with development-indicator data should contain directed and weighted edges. Few of the first-generation studies present these two topological features jointly. However, their inclusion is critical in so far as arrows are indispensable to account a correct dependence or causal structure (either of intervention impacts or flows between endogenous variables), and weights are essential to measure the "strenght" or influence of the relationship between goals.

These challenges serve as guidelines to select the methods that are best-suited to estimate SDG networks. Furthermore, in selecting estimation procedures, we privilege those

[9]Another challenge that may be a concern to some readers is the imbalance in the number of indicators between SDGs. The relevance of this problem depends on whether the inferences are made at such an aggregate level. Nevertheless, it is important to mention that this issue is endemic to all studies that try to cover the high-dimensional nature of development. In part, such imbalance is a historical legacy of how different development agendas have developed over the years, giving priority to the measurement of some specific policy issues before others. For example, it is not surprising to see more development indicators about poverty and economic growth than environmental ones. This is so because the Millennium Development Project prioritize the eradication of poverty, while environmental topics are just a recent priority due to the 2030 Agenda.

techniques that are transparent and replicable through the provision of publicly available code.[10]

3 Broad families of network estimation methods

Network estimation methods are abundant in the literature of complex systems, in statistics and in machine learning (*e.g.*, Han and Zhu (2008); Smith et al. (2011); Linderman and Adams (2014); Aragam et al. (2017)). In this paper, we focus on a subset of methodological families that have explicit underlying concepts of causation. Nevertheless, and as a comparative benchmark, we also present results from computing rank correlations (Spearman). In total, we classify the methods into five families (although we don't consider our classification to be definite or exhaustive).

With the aim of making this paper accessible to a broad audience, we try to keep technicalities and jargon at a minimum. However, using some mathematical expressions might be unavoidable. Therefore, before explaining each family, it is useful to introduce some notation. Let us consider the development indicators as variables $\{X_1, X_2, ..., X_p\}$, where p is the total number of indicators. Let G denote a network whose nodes represent individual development indicators. Each link or edge in G is informative about the presence, direction and magnitude of an association, structural-dependence, or causal strength between two nodes (*i.e.*, policy issues or goals). Finally, we denote the weight of edge $i \rightarrow j$ as G_{ij}.

3.1 Correlation thresholding

One of the most commonly used techniques to estimate networks is thresholding correlation matrices (*e.g.*, Boginski et al., 2005; Han and Zhu, 2008; Huang et al., 2009; Dimitrios and

[10]This means that more sophisticated methods that, potentially, could jointly tackle different challenges may not be applied in this study. However, we provide further details about some of these methodologies in the next sections. Finally, we should add that it is particularly important for development studies, a multi-disciplinary field where researchers and consultants are not experts in the science of networks, to have transparent access to novel methods via publicly available code.

Vasileios, 2015). Its popularity owes, to a considerable extent, to its simplicity, straightforward interpretability and lack of restrictive assumptions about the variables' interdependencies (*i.e.*, a particular type of structure that the estimated network should have). Broadly speaking, the approach consists of, first, obtaining a complete correlation matrix. Secondly, given a minimum acceptable correlation magnitude –referred to as threshold– an initially empty network G is populated by edges between nodes i and j. An edge in G exists whenever the magnitude of the correlation between X_i and X_j is larger than the threshold.

The existence of different correlation measures and the possibility of considering lagged values gives place to numerous methodological variations in this family. To mention a few, we can find the Pearson correlation, Spearman, Kendall, etc. In addition there are various ways in which one could select a threshold, for example, by arbitrarily defining it; by choosing a significance level for the p-value of each correlation; or by selecting a threshold under which the resulting network structure best fits a desired property (*e.g.*, number of edges, the distribution of connections, clustering patterns, etc.). Finally, depending on the temporal nature of the correlation, one could generate directed (with lagged time series) or undirected networks (with contemporary correlations).

3.2 Granger causality

The most visible representative of time-series causal inference methods was proposed by Granger (1969). Granger causality (also known as G-causality) can be defined as follows: for a given time point t and time series X and Y, "We say that X_t is causing Y_t if we are better able to predict Y_t using all available information than if the information apart from X_t had been used" (Granger, 1969). This view of causality has given place to diverse methods to estimate networks.

Most methods in this family stem from time series analysis and, more specifically, from the vector auto-regressive family of models (*e.g.*, Castagneto-Gissey et al., 2014; Barigozzi and Brownlees, 2017; Gao et al., 2018). Nevertheless, we can also find G-causality in method-

ologies developed using concepts from physics, for example, mixtures of G-causality and transfer entropy (Hu et al., 2016). A common characteristic among these approaches is the use of hypothesis testing in order to determine the existence of an edge between a pair of nodes($i \rightarrow j$). Such tests evaluate whether the prediction of Y_j at time t –when only using its previous history– improves by including X_i at time t in the set of predictors (*e.g.*, Barigozzi and Brownlees, 2017).

Overall, Granger-causality network-estimation methods generate directed networks and weighted edges as well. However, they have two major drawbacks when it comes to the challenges enlisted in Section 2.5. First, evaluating the presence or absence of an edge is often tested without considering all the other variables in the system. However, in complex systems such as the those driving SDGs, the assumption of separability (*i.e.*, its effect on the response variable does not depend on other variables of the system) is hardly met. Second and most importantly, given that these tests are built on auto-regressive models, they perform poorly when applied to short time series. Strictly speaking, G-causality methods are more in line with structural-dependence than causal inference. This is so because structural-dependence is mainly concerned with discovering the links' directions and describing the links' weights as the intensity of the flow between two nodes (*e.g.* its co-variability in a specific direction), and not in terms of the effect of exogenous interventions.

One of the first studies to explicit address causation in SDG networks was developed by Dörgő et al. (2018), who take a Granger-causality approach. In order to overcome the limitations of G-causality, they pool cross-national data, which, of course, precludes obtaining country-specific networks.

3.3 Chordal information filtering graphs

Filtering graph methods were initially proposed as a mechanistic way to remove edges in highly dense and complex networks as, for example, in a full correlation matrix. Originally, the purpose of this approach was to eliminate spurious links that are irrelevant in terms of

weight and of the structural integrity of the network (Tumminello et al., 2005). Building on these ideas, more recent developments provide network estimation procedures that perform in a fast and scalable fashion (e.g. Massara et al., 2017). Among these proposals, LoGo (Aste and Di Matteo, 2017) stands out as a suitable framework to estimate causation when time series are short.

In broad terms, one of the main characteristics of the filtered graphs family is the way in which they account for the network's structural integrity. For example, methods such as triangulated maximally filtered graphs (TMFG) (Massara et al., 2017) and LoGo (Aste and Di Matteo, 2017) propose to begin the estimation process with a network that satisfies a mathematical property called the 'planarity constraint'. In simple terms, the planarity constraint is a problem in which, given a complex and dense network, one must re-draw it on a surface (a plane) in such way that as many edges as possible are preserved, but that none of them intersect each other. Although there is no clear intuition or causality-driven motivation for using the planarity constraint, it has been shown that the resulting networks can be used in forecasting.[11]

One of the main strengths of this family is its ability for inferring networks from data with more variables than observations. In particular, LoGo can be used to infer the links' weights and to generate a potentially causal interpretation of their direction. Note, however, that chordal filtered networks might be biased toward sub-structures called triangles (3-cliques) and/or other k-cliques (k nodes, all connected to the other $k-1$ nodes). An important drawback of LoGo is that, when a dataset has numerous indicators that are proportionally interrelated (often the case in SDG data), the estimation may result unfeasible. This is so because LoGo performs matrix inversions requiring specific properties (being non-singular) that are unlikely to hold in high-dimensional correlated datasets. Hence, one may need to resort to strategies to ameliorate such correlations, for example, discriminating indicators in

[11] Another common characteristic is that the estimated graphs are 'chordal', a topological property for which there is no clear motivation or interpretation in terms of causality ("every cycle of length > 4 has a chord, an edge not belonging to the cycle that joins two non-adjacent vertices" (Massara et al., 2017)).

somehow arbitrary ways.

3.4 Statistical structure learning

The last three decades have seen numerous studies of the multivariate causal structure of systems (*e.g.*, Lauritzen, 1996; Kalisch and Bühlmann, 2007; Lacerda et al., 2008; Pearl, 2009; Bühlmann, 2013; Hyvärinen and Smith, 2013; Shimizu, 2014; Peters et al., 2017; Ramsey et al., 2017). Currently, there exists a wide variety of statistical methods for this purpose. Many of these techniques were developed from different theoretical standpoints. Therefore, there is a considerable variation in terminology, often referring to similar or even the same ideas. In trying to overcome this lack of clarity, we classify the family of statistical learning methods into causal graphical models and structural causal models. Here, we provide brief description for each one.

3.4.1 Causal graphical models

Graphical models are sometimes referred to as Bayesian networks. They consist of a network of conditional dependencies where the nodes represent the variables and the links their dependencies. Altogether, these networks are modeled as directed acyclic graphs (DAGs)[12]. In statistical terms, a DAG imposes a conditional-dependence structure that induces a probability distribution over the set of random variables (Lauritzen, 1996; Peters et al., 2017). Thus, causal graphical models are designed to estimate the DAG of conditional dependencies.

A typical estimation procedure to construct G consists of performing conditional independence tests across all the observed variables, in order to discard edges from an initial G. This procedure is performed until all the topological properties of a DAG have been met.[13] A problem with this approach, however, is that there may be multiple possible DAGs that can, for example, meet the results of the conditional independence checks, thus making them

[12]A DAG is a network in which no path can return to its starting node.
[13]In addition, likelihood-based approaches can also be employed to estimate G.

all candidate estimates of G.[14] Thus, these methods usually generate multiple estimates of G, which can then be further discriminated by using auxiliary methods. Examples of this family are the PC algorithm (Spirtes et al., 1993) and the IDA framework (Maathuis et al., 2009).

3.4.2 Structural causal models

Structural causal models are also referred to as structural equation models (and should not be confused with econometric structural models (Heckman and Vytlacil, 2007)). In their most general form (Pearl, 2009; Peters et al., 2017), they can be characterized in terms of 'parent-child' relations between variables and noise terms. More specifically, for a given variable of interest X_i (a development indicator), we say that its parents PA_i are a subset of other nodes in the system that are structural determinants of the value of X_i. Together, PA_i and a noise component N_i determine X_i through a function f_i, *i.e.* $X_i := f_i(PA_i, N_i)$. Therefore, depending on the way that one chooses to model the patent-child relationship, it is possible to define a variety of structural equation models as shown in Table 1.

Table 1: Different specifications of structural causal models

Structural causal model	$X_i = f_i(X_{PA_i}, N_i)$
Additive noise model	$X_i = f_i(X_{PA_i}) + N_i$
Causal additive model	$X_i = \sum_{k \in PA_i} f_{ik}(X_k) + N_i$
Linear Gaussian model	$X_i = \sum_{k \in PA_i} \beta_{ik} X_k + N_i,$ and $N_i \sim N(0, \sigma_i)$

Based on Table 7.1 from Peters et al. (2017). Other popular specifications are non-linear Gaussian additive noise models and linear non-Gaussian acyclic models among others.

Clearly, this approach provides ample flexibility to model structural relationships and noise terms. Aragam and Zhou (2015), for example, propose a scalable network estimation

[14] This is known as the "identifiability problem".

method based on a linear Gaussian structural equation model. This approach can handle cases where the sample size is much smaller than the number of variables in the network. Peters et al. (2013), in contrast, provide an alternative approach for multivariate time series (TiMINo) that does not require a linear Gaussianity assumption.

Sometimes, structural causal models are also referred to as graphical models or even Bayesian networks. This confusion in terminology arises because the set of structural equations implies a graphical representation of the dependence structure of the variables $X_1, ..., X_p$. For example, if the parents of X_1 and X_2 are $PA_1 = X_5$ and $PA_2 = X_1$ respectively, then, the functional assignments $\{X_1 = f_1(X_5, N_1), X_2 = f_2(X_1, N_2), X_5 = f_5(N_5)\}$ imply a network $X_5 \rightarrow X_1 \rightarrow X_2$. Note that, in contrast to graphical models, structural equation models provide a data-generating mechanism that can be easily implemented by following the structural assignments. More specifically, following the previous example where, $X_5 \rightarrow X_1 \rightarrow X_2$, values of (X_1, X_2, X_5) can be obtained by sampling a value n_5 from the noise distribution of N_5. Then, we can compute the value $x_5 = f_5(n_5)$, from which $x_1 = f_1(x_5, n_1)$ and, subsequently $x_2 = f_2(x_1, n_2)$ can be obtained as well (where n_1 and n_2 are noise values drawn from the distributions of N_1 and N_2 respectively). On the contrary, in the graphical models of the previous sub-section this data-generating mechanism is not proposed.

In general both causal graphical models and structural equations models introduce some common assumptions that may be considered important limitations. First, the estimated network G is often required to be acyclical. This implies that reinforcing cycles (virtuous and vicious) are not allowed, for example, as with the bidirectional effects between improvements in public health and education. A second important assumption is that the data points are assumed to be independent draws of the distribution that generates the development indicator. This means that the temporal dependence of the data needs to be added to the initial formulation of the models, as done, for example, by Peters et al. (2013).[15]

[15] In this family, there exist more sophisticated methods that are able to deal with the possibility of cycles (e.g. Hyvärinen and Smith, 2013), with non-linearity (e.g. Peters et al., 2013; Hyvärinen and Smith, 2013) and with time dependence.

3.5 Physics-inspired approaches

The methods discussed so far rely on correlation and/or covariance matrices. However, there are techniques with a fundamentally different basis; one inspired on physical and biological phenomena. For instance, Hu et al. (2016); Servadio and Convertino (2018) employ transfer entropy (*i.e.*, the amount of directed information flowing from one variable to another) to infer the association network in multivariate time series. Furthermore, Hempel et al. (2011) developed the inner composition alignment (iota) method to infer directed networks from short time series.[16]

Among the methods inspired in the natural sciences, there is one class that has a strong mathematical backbone: *state space reconstruction methods*. These techniques spawned from the literature on non-linear dynamical systems and, in recent years, they have become relatively popular to estimate causal relationships in physical, biological and ecological systems (*e.g.*, Heskamp et al. (2014); Ye et al. (2015); Tsonis et al. (2015); McBride et al. (2015); Wang et al. (2018)). The basic idea behind state space reconstruction is that any of the variables that are part of a system can be used individually, through its time series, to recover the attractor to which the system tends to evolve (the so called Takens' theorem).[17]

Inspired in this theory, Sugihara et al. (2012) developed the method of convergent cross mapping (ccm). This technique allows estimating directed networks by testing causation between a pair of variables, even if they are weakly coupled (note that bi-directional causality is possible). The test is built around the idea that X causes Y when the dynamics of the former can be recovered (predicted) by the dynamics of the latter; the opposite logic behind G-causality. In other words, the information concerning X is already reflected in the evolution of Y, and cannot be removed from the universe of all possible causative factors.

[16]It is important to note that, although there are methods belonging to a single family, several others share characteristics with different families. For example, Servadio and Convertino (2018) propose a transfer-entropy approach comparable to the one used in correlation thresholding. Similarly, the chordal-filtering-graphs approach of Aste and Di Matteo (2017) also employs transfer entropy to quantify the edges that carry causal interpretations.

[17]For a review of this literature see Ma et al. (2017)

Accordingly, through Y's attractor manifold (a mathematical object that represents a geometric space in which the variable moves, usually denoted as M_y), it is possible to make local-neighborhood predictions on X's manifold M_x. Since local-neighborhood tests can be data demanding, several methods to deal with short time series have been proposed for noisy and high-dimensional systems (*e.g.*, Ma et al., 2014; Clark et al., 2015; Zhang et al., 2017).

Overall, state space reconstruction methods have an appealing theoretical basis and allow the estimation of directed networks without imposing topological constraints. Nevertheless, they also have drawbacks that should be mentioned. On the theoretical side, these models assume that the system is in a 'steady state' fluctuating around the attractor. Thus, for a system that is transitioning to a different attractor, these methods would provide the wrong causal relations. Furthermore, the assumption of remaining near a specific attractor may be in conflict with the principles of socioeconomic development; a process with technological innovations, emerging social norms and considerable political turbulence that may push nations toward different attractors. In addition, the weights of the inferred networks do not have a clear interpretation other than a score about how well one variable predicts the other.[18]

4 Eligible estimation methods for SDG networks

As we have argued in Section 2.5, SDG data conveys several challenges that turn most network estimation methods ill-suited. For example, most methods in the Granger-causality family are inadequate due to the short length of development-indicator time series.[19]. Therefore, and after a comprehensive review of numerous approaches, we have selected four methods to perform empirical estimations. These are: *intervention calculus via the PC algorithm* (pcalg), *concave penalized estimation via sparse Gaussian Bayesian networks* (sparsebn), *inner composition alignment* (iota), and *convergent cross mapping* (ccm). That is, the first

[18]See Cobey and Baskerville (2016) for more technical criticisms on ccm and the conceptual challenges of state space reconstruction.

[19]This explains the limited results obtained by Dörgő et al. (2018)

two belong to the family of statistical structure learning, while the last two correspond to the physics-inspired family.[20]

Overall, no single method 'ticks all boxes' when it comes to the challenges of SDG networks. However, the chosen ones have the common ability of being able to cope with a small number of observations and a large number of variables. Therefore, while the eligible methods are not a definite statement about the correct way to estimate SDG networks, they do provide an important guide on a variety of ways to exploit SDG data. Finally, and as mentioned above, all of the chosen methods have publicly available software. In this section, we explain further details on each of these four methods in order to perform the empirical analysis in Section 5.

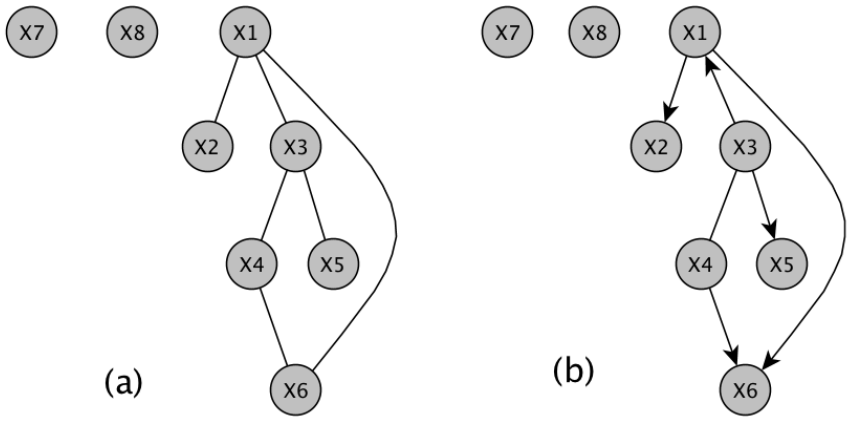
4.1 Intervention calculus via PC algorithm

Intervention calculus when DAG is absent (IDA), proposed by Maathuis et al. (2009), is a causal graphical model designed to estimate causation directly from observational data. The method consists of three steps. First, the basic structure of the dependence network –called the skeleton– is inferred (*i.e.*, a network displaying only undirected edges and no weights). The skeleton is particularly useful as it highlights all the possible causal relations between all variables in the network (see Figure 1-a). In the second stage, edge directions are inferred, so the 'completed partially directed acyclic graph' (CPDAG) is created (see Figure 1-b). The CPDAG is called partial because the inference of some directions is theoretically unfeasible, hence CPDAG displays those edges as undirected. The reason for this unfeasibility lays on the fact that there are multiple (fully) directed acyclic graphs (DAGs) that can be obtained from the CPDAG. These multiple DAGs encode the same distributional and conditional dependence constrains inferred from the data; making them indistinguishable through purely

[20]Although chordal information filtering methods such as LoGo provide accessible code, the required matrix inversions are unfeasible due to the high correlated nature of some development indicators (which could also be caused or heightened by the short length of the time series). Similarly, and in spite its code availability, TiMINo was unable to provide estimates for our data. For this reason, we do not use them in our empirical application. Note that similar problems might also surface in other methods that rely on inverting the variance-covariance matrix.

observational information. Nevertheless, the CPDAG can be thought of as a summary network displaying all the possible directed dependence networks that fit the data constrains. Finally, the third stage estimates the size of the causal effects (edge's weights) for each of the possible relations in the CPDAG, generating a weighted network.

Figure 1: Skeleton and CPDAG of PC algorithm



(a) Example of the skeleton of a graphical model with 8 variables. The undirected edges represent the presence of a dependence relation between nodes. However, the directions are absent. (b) Example of a completed partially directed acyclic graph, providing directions for some of the edges in the skeleton. Some edges, however, remain undirected as the method cannot decide which is the correct direction.

Following Maathuis et al. (2009), the first two stages can be performed via a method called the PC algorithm (Spirtes et al., 1993). The method starts with an undirected graph that includes all possible edges. Then, it iteratively deletes edges that pass (does not reject) a conditional independence test. In the first iteration, a simple independence test is performed between two connected nodes. If the nodes are found to be independent, that edge is deleted from the network. Next, once all pairwise tests are performed, each remaining pair of connected nodes (i, j) is tested for conditional independence given any single node connected to i, or any single node connected to j. Whenever a pair of nodes is found to be conditionally independent, its edge is removed. Then, in a new iteration of conditional independence tests, the size of the conditioning set is increased by one node. This edge deletion process leads to the skeleton of the inferred dependence network.[21]

[21] An alternative method designed to deal with more variables than observations has been developed by

Once with the network skeleton, directions are inferred by considering each triplet of connected nodes $(i - k - j)$. If, in the first step, node k was never part of a 'passed' (does not reject) conditional independence test between i and j, then the triplet takes the form $i \rightarrow k \leftarrow j$; (see (Spirtes et al., 1993) for full details on the PC algorithm). Finally, IDA is used to estimate upper and lower bounds of the weights (of the causal effects). The idea is to estimate the effect that an intervention in variable X_i can have on variable X_j. This effect can be interpreted as the change in the mean of X_j, when X_i is forced to take a given value (see (Maathuis et al., 2009; Lauritzen, 1996) for full details). In this paper, we use the minimum reported bound as the causal effect.[22]

4.2 Concave penalized estimation via sparse Gaussian Bayesian networks

Aragam and Zhou (2015) propose a structural causal model for systems composed of a large number of variables that follow a multivariate Gaussian distribution. They estimate the coefficients of a linear Gaussian structural equation model (SEM) (see Table 1) by solving a convex optimization problem.[23] With the estimated parameters in hand, it is possible to construct a directed weighted network through the system of equations defined in the SEM.

There are two key innovations in this approach. The first is a parameter transformation that allows linking the optimization problem to the estimation of high-dimensional Gaussian distributions (*i.e.*, allows estimating large networks). The second is a penalty term in the optimization procedure that allows the estimation of sparser networks (*i.e.*, eliminates edges to address over-fitting).[24] Note, however, that the estimated coefficients (*i.e.*, the weights)

Meinshausen et al. (2016). The authors also provide tools to compare network estimation methods of this family.

[22]The R package `pcalg` (Kalisch et al., 2012) provides all the required tools to estimate these networks on development-indicator data. In this paper we use the code provided for Gaussian variables.

[23]This is an approximation of the non-convex optimization problem that arises when directly considering the optimization of the log-likelihood of the data via the SEM (Aragam and Zhou, 2015).

[24]Additionally, Aragam and Zhou (2015) develop another parameter transformation to convert the non-convex optimization problem into a convex one.

cannot be directly interpreted as causal relations without further analysis of interventional data. Instead the method "focus on finding the most parsimonious representation of the true distribution [of the variables] as a set of structural equations" (Aragam and Zhou, 2015).[25]

4.3 Inner composition alignment

Inner composition alignment (iota) (Hempel et al., 2011) is a method inspired in physical dynamics where a, constantly changing, system can be analyzed by describing its different states and possible ways to transition between them. For example, in a biological system such as gene regulatory networks, one way to infer when genes are coupled is through pairwise tests between time series. More specifically, iota identifies the coupling of two time series when the ranking of the values of one of them is similar to the ranking of the other. In other words, it checks if a non-decreasing ordering applied to time series X_i can also be applied (approximately) to X_j, in order to achieve a non-decreasing ordering of X_j.

Hempel et al. (2011) build on these ideas to relate two time series by how monotonic (increasing or decreasing) X_j is when it is sorted based on the ordering of the other time series (X_i). More precisely, they propose a statistic that captures the degree of monotonicity of the reordered variable X_j in order to test the presence and direction of the edge $i \to j$. This statistic can be understood as an index that depends on the number of intersections between the re-ordered time series X_j and the imaginary horizontal lines fixed at the values of X_j, when the ordering is given by X_i (see Figure 2). Through pairwise computations and a permutation test of the aforementioned statistic, iota aims at estimating a dependence network G. The method works for very short time series and allows inferring the direction of associations between variables [26]. Furthermore, given that it checks for monotonicity on re-ordered variables, this method can also detect non-linear relations between the time series.

[25]The R package `sparsebn` (Aragam et al., 2017) provides all the required tools to estimate these networks on development-indicator data. This tool assumes the data is an i.i.d. sample from a multivariate Gaussian distribution.

[26]Note that, although iota is inspired in time series observations, it does not explicitly consider the temporal dependence of variables.

Figure 2: Intuition behind the iota statistic

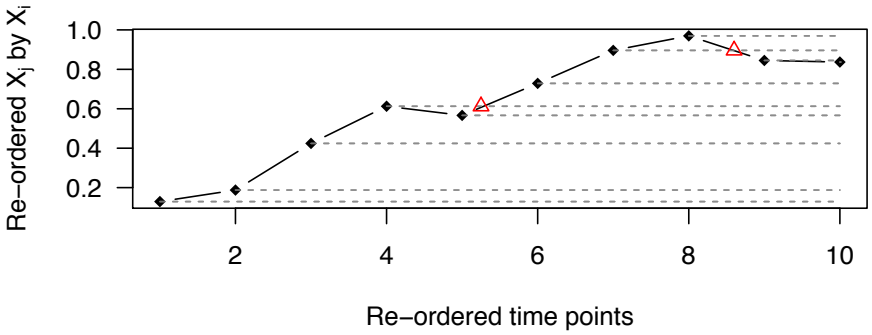


Illustration of the number of crossings (marked as red triangles) between the re-ordered time series X_j and the imaginary horizontal dashed lines fixed at the values of X_j, according to the ordering obtained from X_i.

Note that iota has been constructed for very short time series and does not impose assumptions about the structure of the dependence network. However, rather than being based on a generative theory of the data, it is based on an observed pattern. Therefore, it is not clear if it can actually infer causal or structural-dependence networks.[27]

4.4 Cross convergence mapping

The method known as cross convergence mapping (ccm) was proposed by (Sugihara et al., 2012), based on the idea that complex systems exhibit non-linearities and, thus, they tend to evolve towards strange attractors. Therefore, when two variables (X, Y) are coupled, their corresponding manifolds (M_X , M_y) –obtained through the coordinates of their lagged values– should approximately describe the attractor. Hence, if the information embedded in M_Y can predict the information embedded in M_X, it is argued that X causes Y. Reconstructing the state space, however, requires a large amount of observations when a system is high-dimensional.

In its supplemental material, (Sugihara et al., 2012) claim that ccm works well when studying data of fisheries with approximately 70 species and only 35-40 observations. Moreover, recent extensions attempt to improve the resolution of the manifolds for small samples.

[27]The R package `IOTA` (tocsy.pik-potsdam.de/iota.php) provides all the required tools to estimate these networks on development-indicator data.

This can be done, for instance, by increasing the number of observations through replications of the same unit of analysis but in a different physical space (Clark et al., 2015). An alternative is to use a neural network to estimate the cross map between the manifolds (Ma et al., 2014).

According to the ccm method, causation between two nodes exists when there is convergence in fitting; that is, when the correlation coefficient between the predictor and the observed variable increases rapidly with the series length. In the seminal paper of Sugihara and co-authors, a nearest-neighbor algorithm is used to establish the weights to be used when calculating the predicted values.

Despite the time series of our data having less than 20 observations, we use the original ccm methodology. We do not apply the multi-spatial ccm (Clark et al., 2015) because we only have one observation per country/year. However, since the software associated to this procedure is publicly available, we use it for our estimations (with only one replica)[28]. The ccm approach does not impose topological requirements on G. This means that more nodes imply higher dimensionality of the underlying attractor. Finally, the method takes into account the time-dependence structure of the indicators.

5 Data and results

5.1 Data

In spite of worldwide efforts to build comprehensive datasets tracking the 17 SDGs of the UN's development agenda (General Assembly, 2015), it is still challenging to assemble samples of indicators with a large coverage of countries, indicators and number of observations at the same time. Since, in this paper, we perform a comparative analysis between methods, the estimation of SDG networks demands the following data requirements: (i) all countries

[28]The R package `multispatialCCM` provides all the required tools to estimate these networks on development-indicator data.

should contain the same indicators; (*ii*) all observations must be contemporaneous (they should cover the same sampling period); (*iii*) no indicator should be a linear re-normalization of another (*i.e.*, no two indicators should have a perfect linear correlation); (*iv*) all indicators should exhibit temporal variation (we assume that a constant indicator is independent from all others).[29]

Taking all previous considerations into account, we build a dataset where we reach a compromise between maximizing the length of the sampling period, the number of indicators, the SDGs covered and the number of countries in the sample. In particular, we prioritize the number of indicators and years since our motivation for estimating SDGs is to map a complex web of structural-dependencies between numerous policy issues. Our sample consists four countries from different continents: Egypt (EGY), Indonesia (IDN), Mexico (MEX) and Turkey (TUR). We study the relationships between 87 development indicators (covering 16 of the 17 SDGs) during the 1995–2014 period. Each indicator has been normalized between 0 and 1.[30] The indicators have been adjusted so that larger values denote better outcomes. The main source of the dataset is the United Nations Global SDG Database (United Nations Statistics Division, 2016). In addition, we complement it with the World Bank Sustainable Development Goals Database (The World Bank Group, 2016) and the poverty indices from the World Bank Poverty and Equity Data (The World Bank Group, 2010). Figure 3 presents summary statistics aggregated by SDG and country.[31]

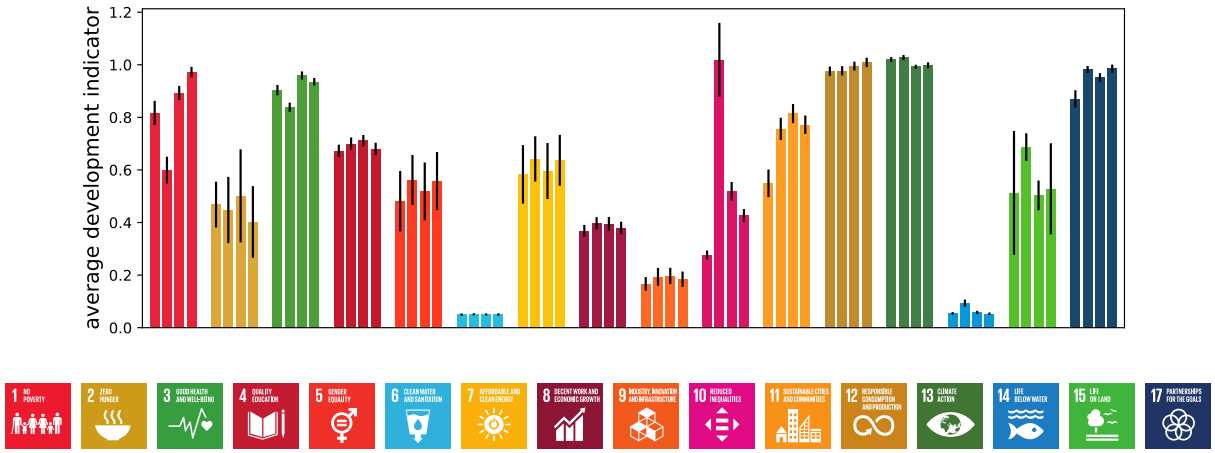
Figure 3 shows differences between the four countries. For example, Egypt is substantially less developed in SDG 10 (reduced inequalities), while Indonesia is in SDG 1 (no poverty). It is also interesting to observe that all the four countries are poorly developed in SDG 6: *clean water and sanitation* and SDG 14: *life below water* (from the four indicators that were

[29]Hence, we focus on non-duplicated time-varying indicators.

[30]For indicator i of country c in period t, we normalize according to the standard procedure $[I_{i,t,c} - \min(I_{i,\cdot,\cdot})]/[\max(I_{i,\cdot,\cdot}) - \min(I_{i,\cdot,\cdot})]$, where I is a matrix with development indicators of all countries, not only those in the sample. The reason for using all countries in the normalization is that cross-national data provide an adequate benchmark to normalize indicators (*i.e.*, zero-values should be assigned to those indicators that equal the lowest levels across all countries and years).

[31]See Table 5 in Appendix A for a complete list of the 87 indicators and their respective SDGs, and Table 6 for the numeric values of the summary statistics.

Figure 3: Summary statistics of development indicators by country and SDG



Each color corresponds to an SDG. All indicators in each SDG have been aggregated by averaging, first, their values through time and, then, through all the indicators in the SDG. Within each color, each bar corresponds to a country in the following order: EGY, IDN, MEX and TUR. The vertical black lines denote the standard error. All indicators in SDG 6 have nearly-zero values. Therefore, in order to better illustrate their comparison against all others, all indicators in this plot have been added the amount of 0.05.

available for the four nations).

With this general picture of the data, we proceed to present the estimated networks. We do this in a twofold fashion. First, we show aggregate results at the level of the SDGs. That is, we study the total number of incoming and outgoing edges (and their weights) in each SDG. This allows gaining insights into the structure of synergies and trade-offs between SDGs at the level of each country. Second, we introduce and compute formal metrics to compare the topologies of the estimated SDG networks, with a special emphasis on comparing the networks produced for the same country via different methods. This provides a more rigorous understanding of the implications of using different frameworks.

5.2 SDG networks

As discussed in Section 4, we present results on four specific methods, *inner composition alignment* (iota), *convergent cross mapping* (ccm), *intervention calculus via the PC algorithm* (pcalg), and *concave penalized estimation via sparse Gaussian Bayesian networks* (sparsebn).

In addition, we provide calculations obtained from correlation networks (lcorr). The latter networks are built by computing pairwise Spearman correlations where the explanatory variable is lagged by one period. In this way, the resulting networks are not necessarily symmetrical. We only keep those edges where the correlation has a p-value lower than 10%.[32]

Table 2 shows the summary statistics about the connectivity in the estimated SDG networks. When one talks about number of links, connectivity is usually referred to as degree. When edges have weights, however, the correct term is strength (the sum of the weights connected to a node). We have split each network into edges with positive and with negative weights. In other words, we analyze synergies and trade-offs separately. Notice that the topological structures of the estimated networks vary considerably, as reflected by their synergies and trade-offs across each country's estimates. However, the estimated weights have slightly different interpretations since each method has a different underlying concept of causality (if any at all). In the light of these variations, one method can be preferred over another depending on the type of the analysis to be pursued, or on the consistency of the estimates with prior information (see below).

In particular, with respect to synergistic effects, the highest average values are observed for the lcorr method and then for the ccm method, irrespectively of the country. While the lowest average values correspond to pcalg, sparsebn and iota methods in ascending order. The same ranking prevails for the estimates of trade-off effects. Notice also that the five methods analyzed here do not identify the same country as the one with maximum synergistic strength. Indonesia is identified twice (lcorr, pcalg), Mexico twice (ccm, sparsebn) and Turkey once (iota). A similar assessment emerges when comparing the identity of the nodes with maximum strength for each country across estimation methods. With the exception of Turkey, where ccm and pcal present the same identifier [11]. While in the case of trade-offs,

[32]In general, if a method performs hypothesis testing to decide the permanence of an edge, we use a 10% p-value threshold. For this study, we rather not use smaller p-values and multiple testing correction as in this work we aim to show the topological differences to the common use of correlation networks in previous studies.

Table 2: SDG-network strength statistics

Method	EGY		IDN		MEX		TUR	
	Synergy	Trade-off	Synergy	Trade-off	Synergy	Trade-off	Synergy	Trade-off
	30.99	25.1	43.06	24.48	37.82	29.86	41.43	30.62
lcorr	(12.11)	(11.3)	(19.21)	(16.51)	(15.83)	(14.0)	(18.66)	(16.03)
	[31]	[51]	[40]	[43]	[5]	[32]	[45]	[46]
	12.83	11.1	13.01	11.01	11.66	10.44	13.38	11.4
iota	(6.53)	(4.79)	(5.9)	(5.0)	(3.69)	(4.78)	(6.2)	(5.71)
	[5]	[43]	[4]	[7]	[27]	[13]	[36]	[81]
	28.57	21.68	30.83	17.2	33.7	26.41	32.92	20.55
ccm	(14.25)	(10.94)	(17.46)	(12.0)	(15.78)	(11.92)	(18.88)	(12.97)
	[56]	[68]	[12]	[36]	[15]	[36]	[11]	[52]
	1.29	0.64	1.33	0.53	1.26	0.44	1.15	0.55
pcalg	(0.82)	(0.8)	(0.97)	(0.66)	(0.79)	(0.64)	(0.86)	(0.67)
	[74]	[36]	[16]	[35]	[3]	[19]	[11]	[29]
	8.9	7.03	9.63	7.66	10.62	8.0	6.3	3.36
sparsebn	(4.89)	(4.31)	(5.02)	(4.33)	(5.81)	(5.3)	(3.8)	(2.63)
	[85]	[8]	[15]	[42]	[1]	[70]	[3]	[46]

Reported connectivity is the average strength (the sum of the weights) of a node's connections, separated into positive (synergies) and negative (trade-offs) links. Standard deviations are inside round brackets. Nodes with the maximum strength are inside squared brackets.

none of the methods identify the same node for a specific country. These results point out the method-dependent sensitivity of interpretations about the strength of complementary (or conflicting) links between policy issues. Next, we present additional analyses comparing the estimated network.

Since a substantial part of discussions in the sustainable development literature relates to the promotion of complementary policies, let us take a closer look at synergistic networks. Figure 4 shows the structure of synergies as the links that are directed from one indicator to a group of nodes (we provide a similar illustration for trade-offs in Figure 6 of Appendix B). The first feature that stands out in Figure 4 is the striking differences between physics-inspired methods (iota and ccm) and statistical structure learning methods (pcalg and sparsebn).[33]

On the one hand, iota and ccm generate considerably denser SDG networks than pcalg and sparsebn. In addition, there are also important differences between pcalg and sparsebn,

[33]Remember that pcalg and sparsebn assume no temporal dependence in the data.

Figure 4: Estimated synergies between SDGs by country and method

Weights have been normalized so they are comparable in terms of proportion to all other weights in the network. Edges colors indicate the SDG from which they originate. The length of each colored segment of each circle should not be interpreted as they are the result of a visualization algorithm.

with the former estimating considerably less synergies than the latter. These differences vary from country to country, highlighting the importance of not pooling cross-national data to estimate SDG networks. On the other hand, when making within country comparisons, it

is also evident that, depending on the method of choice, the distribution of weights across SDGs is significantly different. For instance, in the case of Egypt, iota suggests that there are many synergies coming in and out from SDG 1 (no poverty). While synergies entering to this node are drastically reduced in ccm and they vanish in the other two methods.

In order to provide more detail on how different can be the inferences derived from these methods, let us rank each SDG in each country and method. We produce this ranking according to how synergistic an SDG is to the others. That is, for a given SDG in a specific network, we take its outgoing edges with destinations in different SDGs. By averaging their weights, we obtain an estimate of how synergistic that SDG is to other goals. Table 3 shows the top five SDGs in each ranking. No country presents the same list of indicators for all methods, and only Egypt and Indonesia exhibit the same list for two methods: lcorr and ccm (although in a different ordering). However, 5 indicators from Egypt are repeated in at least three lists of the top five, 4 indicators from Indonesia, 4 from Mexico, and 3 from Turkey. These results indicate that, given the sensitivity of the different methods, a potential approach could be finding a consensus among networks. This concept will be developed more formally in an appendix below.

Table 3: Top five SDGs by outgoing synergies

Country	lcorr	iota	ccm	pcalg	sparsebn
EGY	15, 14, 2, 11, 3	1, 2, 3, 5, 10	2, 3, 11, 14, 15	3, 7, 10, 14, 15	2, 8, 9, 11, 14
IDN	2, 14, 1, 3, 11	1, 6, 14, 15, 17	1, 2, 3, 11, 14	3, 4, 5, 7, 8	3, 8, 9, 12, 13
MEX	2, 1, 5, 11, 14	3, 4, 6, 11, 15	1, 2, 3, 4, 11	2, 3, 6, 9, 14	7, 8, 9, 12, 14
TUR	2, 3, 1, 5, 17	2, 3, 6, 11, 17	1, 2, 3, 4, 11	3, 7, 12, 13, 14	1, 2, 9, 12, 14

The outgoing synergy of an SDG is computed by averaging the weights of positive outgoing links with destination in nodes (indicators) belonging to a different SDG. In other words we do not count within-SDG synergies. Each set of five SDGs has been sorted from most to least synergistic (left to right).

5.3 Prior-informed tests

As we have explained in Section 2.2, a desirable quality in an SDG network estimation method should be external validity. However, this is hardly achievable in the context of SDGs

since the data-generating process is rather complex and takes place at different aggregation levels and time scales. Therefore, one has to appeal to some form of internal validity, such as a qualitative judgment on the consistency between the properties of the data and the method requirements (*e.g.*, no temporal dependence). Furthermore, there are ways in which prior information may be used to make a judgment on how valid can an estimated SDG network be. For instance, the subjective information employed in some of the first-generation methods can be exploited to –*ex post*– partially validate an estimation. That is, rather than –*ex ante*– relying on subjective data to estimate the network, this information can be leveraged to evaluate whether an estimated set of edges constitutes false positives or whether absent edges translate into false negatives. Here, we show how this approach sheds light on the performance of the estimation methods.

In order to evaluate false positives (FP), we identify 10 pairs of development indicators for which one would not expect any causal or structural-dependence relation in any direction and sign. Therefore a false positive means that a method has estimated a synergy or a trade-off that, *a priory*, is known to have no logic in the context under study (goes against common sense). Thus, there can be up to 40 false positives.[34] For the false negatives (FN), we have identified 10 pairs of indicators that almost certainly have some type of relationship, whether this is causal, an structural-interdependencies, synergistic, a trade-off, or it goes in any direction. A false negative is the complete absence of any type of edge between a pair of nodes in the estimated network when, in reality, that edge exists.[35] Thus, there can be up to 10 false negatives at most. Appendix C provides the two lists of indicators that have been identified for this analysis.

Table 4 presents the number of false positives and false negatives for each estimated network. In relation to the false negatives, pcalg shows the best performance. However,

[34]We can obtain up to 40 false positives for 10 pairs of nodes because, for each pair, there can be 4 possible edges: one positive and one negative from X to Y, and another positive and another negative from Y to X.

[35]For the analysis of false negatives, we have identified indicators whose causal relationship could even be considered tautological. That is, the two indicators in a pair belong to the same SDG and are conceptually closely related.

for the other methods there is no clear indication what method is better irrespective of the country. With respect to false positives (*i.e.*, links that should not be there), lcorr presents the highest number. However, the large number of links established (T) in this method increases the probability of false positives. Thus, if we normalize by T, lcorr and ccm present a relative large number of false positives (> 0.002) for the four country cases. While this threshold is not crossed in two cases for iota (Egypt and Mexico) and for sparsebn (Egypt and Turkey). Not surprisingly, the extremely sparsed networks estimated with pcalg receive the best possible scores. Thus, in our judgement iota, sparsebn and pcalg are potential reliable candidates, although more extensive and rigorous test should be conducted.

Table 4: Prior-informed false positives and false negatives

Method	EGY			IDN			MEX			TUR		
	FP	FN	T	FP	FN	T	FP	FN	T	FP	FN	T
lcorr	12	2	4288	11	2	5130	14	0	5406	18	0	5932
iota	1	0	1195	3	2	1207	1	1	1112	5	2	1279
ccm	8	2	2799	8	2	2825	11	3	3818	10	0	3357
pcalg	0	0	84	0	0	81	0	0	74	0	0	74
sparsebn	0	0	693	4	3	752	2	1	810	0	0	420

FP: false positives. FN: false negatives. T: total number of directed edges estimated in the network.

5.4 Comparing SDG networks

Although the data allow comparisons between countries, we would not want our results to be seen as having potential policy implications because the aim of the paper is not to estimate the definite networks for our country examples. Instead, our purpose is to understand how different methodologies differ in their results. For this reason, we emphasize methodological comparisons rather than country ones.

In order to provide a more rigorous analysis of the structural differences between alternative methods, it is necessary to briefly elaborate on some relevant metrics. A common practice in network science is to measure topological differences between networks by focusing on specific traits. For example, some studies try to quantify discrepancies on how

network connectivity is distributed across the nodes, while others concentrate on how different are the internal communities and clusters. Broadly speaking, we can characterize these different emphases in a spectrum between *connectivity* and *structure*. Different metrics have been developed to compare networks in this spectrum. It should be noted that, as of today, there is no 'gold standard' when it comes to these comparisons. Furthermore, because networks may be directed or undirected, weighted or unweighted, single- or multi-layered, etc., a metric that is popular for certain type of networks might not be well defined for others. For these reasons, we employ three metrics that are well suited for comparing SDG networks and that try cover the spectrum between connectivity and structure. In the three metrics, higher values imply larger dissimilarity between networks.

5.4.1 ψ distance

Xu et al. (2013) propose a comparative measure for the connectivity of two networks called the ψ distance. This measure is general enough to consider weights and directions, and can handle networks with different sets of nodes. In the context of this study, the ψ distance becomes quite simple because we compare networks with exactly the same development indicators (the same sets of nodes). More specifically, this metric is the sum of the magnitudes of the differences between the weights of the same edge in the two different networks. Formally, for two networks G and Q, the ψ distance is defined as

$$d_\psi = \sum_{i,j \in V_{G,Q}} |G_{ij} - Q_{ij}|, \tag{1}$$

where $V_{G,Q}$ is the set of nodes in G and Q, and G_{ij} denotes the weight of edge $i \to j$ in network G.

5.4.2 Hamming distance

Somewhere in the spectrum between connectivity-based and structure-based metrics we can find the popular Hamming distance (Hamming, 1950). Interestingly, Richard Hamming

developed this metric to detect coding errors, but he had a general theory in mind for any object that could be studied through geometry. For instance, the metric has become very popular to study distances between words. In network science, this metric is also known as the edit distance, and it is very intuitive. Given a network G, its Hamming distance from another network Q is the minimum number of edits required to transform it into Q. In our context, since G and Q have the same nodes, edits translate into edge removals and additions. Computing the Hamming distance is usually performed through an algorithm provided in most programming languages.[36] It is important to clarify that this metric does not consider edge weights.

5.4.3 NetEmd

NetEmd (Wegner et al., 2017),[37] is a more recent network-comparison method, initially developed to directly compare the structure of complex systems. Similar to previous structural metrics (Ali et al., 2014), NetEmd measures the frequency of occurrence of different sub-graphs. For example for nodes i, j, k, l, m, the occurrence of triangles is given by the presence of edges $i - j, j - k, k - i$; 4-stars exist if we observe edges $j - i, j - k, j - l, j - m$; square sub-graphs are given by edges $i - j, j - k, k - l, l - i$ and so forth. In total, NetEmd uses 30 types of sub-graphs.[38].

Given G, NetEmd builds the probability distribution of observing a node that is attached to x sub-graphs of a particular configuration (*i.e.*, to one of the 30 configurations).[39] Then, it compares the shape of the distribution to the one obtained from another network Q. The comparison involves some normalization techniques, after which, the final product is a score indicating the average distance between G and Q in terms of their sub-graph structures.[40]

[36]The R package `pcalg` provides the function `shd` to compute the Hamming distance between two networks.

[37]The `NetEmd` package for R can be downloaded from `github.com/alan-turing-institute/network-comparison`

[38]These sub-networks can further be decomposed into 73 configurations when node position is taken into account. By default, NeEmd uses these 73 configurations.

[39]More precisely, to a specific location in one of the 30 sub-graphs.

[40]More specifically, NetEmd uses the minimum Wasserstein distance (Runber et al., 1998) –also known as the earths mover distance– and a few normalization procedures.

5.4.4 Results

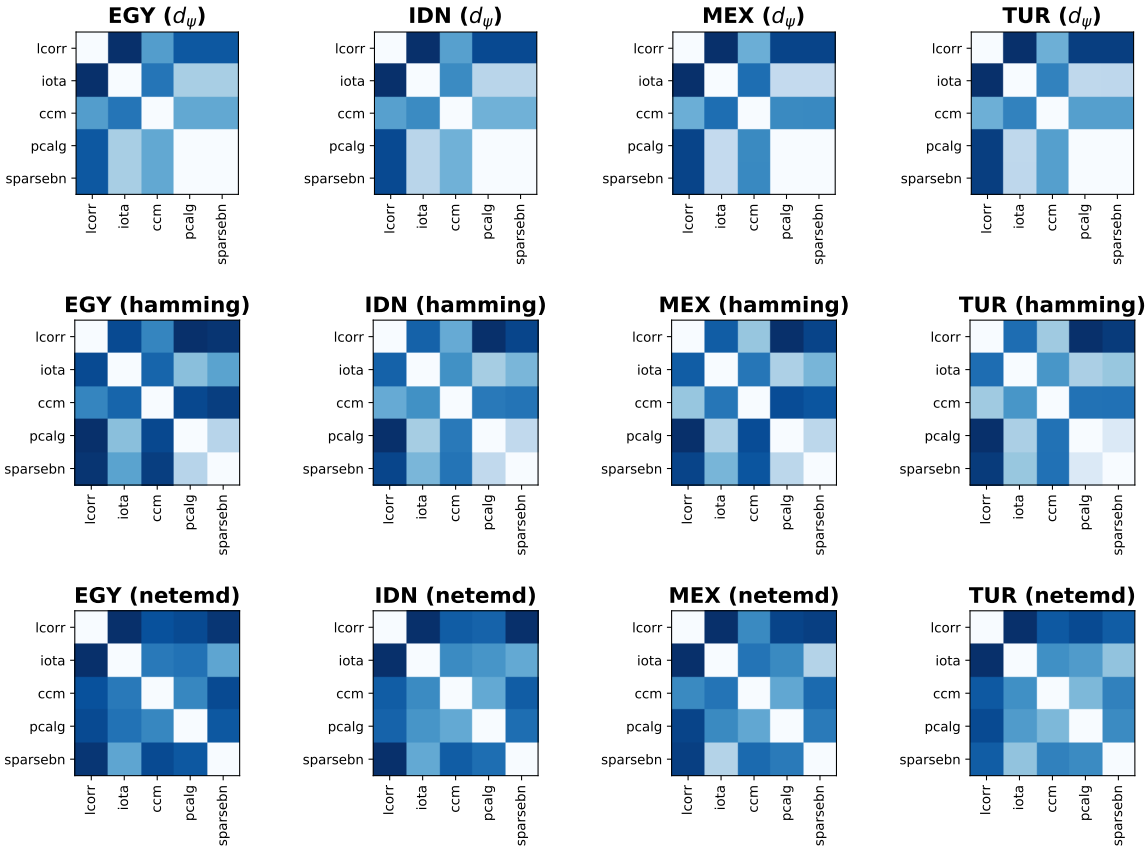
Once the distance metrics have been defined, we can compute them for each estimated SDG network. Figure 5 presents the different comparisons of the five networks per country and per metric. We can see that statistical structure learning methods (pcalg and sparsebn) are very similar in terms of the Hamming and ψ distances, but not under NetEmd. In contrast, the lcorr and iota methods are quite different under the three metrics, while pcalg and sparsebn are also very different in relation to lcorr. We can also say that iota exhibits moderate levels of resemblance with respect to pcalg and sparsebn irrespective of the metric considered. Moreover, the topological discrepancies between ccm and the other methods are salient, yet they vary notoriously depending on the distance metric used. All in all, we conclude that, with the exception of the pair (sparsebn and pclag), the methods here considered produce different topologies. Presumably, the reason for sparsebn and pclag to show closer results is that they belong to the same family and both tend to produce sparse networks.[41]

6 Discussion and conclusions

This paper provides an overview of different network-estimation families suitable for inferring synergies and trade-offs between SDGs. However, warns that policy recommendations directly obtained from these networks should be taken with a pinch of salt, even when using causal inference methods. This is so because network-estimation methods only consider the dependence account of causation, not the production one (Hall, 2004). In the production account, causal factors are those that help generating, or bringing about, specific outcomes. In other words, it calls for modelling the specification of vertical causation which, in the context of SDGs, allows the generation of data from mechanisms that are known to be part of the SDGs system. Among these mechanisms, we would need to include the policymaking process, agent-level interactions, adaptation, among others.

[41] Appendix D presents a consensus-network approach to compare topologies across countries.

Figure 5: Differences between estimated SDG networks by country and metric



Darker tones denote more difference between networks.

Given a lack of observable information, a production account is better-suited for advising development policies because these interventions take place at a much lower level of aggregation than the one provided by the currently available data for SDGs, *i.e.* the development indicators; hence, the need for vertical causal mechanisms becomes self evident. In fact, scenarios of policy interventions from an agent level perspective, and emergent properties at the macro-level is the 'bread and butter' of complex adaptive systems such as markets and ecosystems. Casini and Manzo (2016) argue that agent-based models are a more adequate tool to establish vertical mechanisms and study the causal impact of policy interventions. These models are particularly useful in the absence of empirical information at the multiple levels of aggregation.

A way to conciliate both accounts of causation in the context of the 2030 Agenda is by estimating SDG networks and then using them as inputs for an agent-based model of the policymaking process (with vertical causal mechanisms). Some progress has been made in this vein of research via a political economy game on a network of policy issues (Castañeda et al., 2018). Starting with a set of exogenous objectives, indicators' initial values and an estimated network of structural-dependencies, the model generates endogenous policy priorities that can be used to make causal inference in a variety of topics such as policy resilience (Castañeda and Guerrero, 2018), *ex ante* policy evaluation (Castañeda and Guerrero, 2019), policy coherence (Guerrero and Castañeda, 2019b) and corruption (Guerrero and Castañeda, 2019a). Unfortunately, these studies have only considered socioeconomic indicators with positive interactions; not environmental ones, which are the main source of trade-offs. Therefore, there is ample room to further develop this approach in the context of SDGs. In this setting, the purpose of the network is only to specify how the SDG indicators co-evolve once certain variations are induced by the adaptation of the budgetary allocations.

6.1 Concluding remarks

This paper contributes to the literature on the sustainable development goals by reviewing first-generation network studies and by comparing networks estimated with state-of-the-art methodologies coming from statistics, complexity science and machine learning. In contrast to first-generation studies, estimations based on the newer methods have the potential of meeting more desirable properties for SDG networks (scalability, replicability, specificity, directionality and validity). An endemic property of SDG indicators, however, is that they have a very limited number of observations.

By selecting five specific methods, we estimate networks using data from four countries (Egypt, Indonesia, Mexico and Turkey), covering 20 years and 16 out of the 17 SDGs. The chosen methods for these estimations were: Spearman correlations (icorr), intervention calculus via PC algorithm (pclag), concave penalized estimation via sparse Gaussian

Bayesian networks (sparsebn), cross convergent mapping (ccm) and inner composition alignment (iota). The second and third belong to the statistical structure learning family, while the last two are physics-inspired methods. Although this list is not exhaustive, it offers important insights on the key problems of SGD network estimations.

Our main conclusion is that applied methods –suitable for small sample sizes– produce topologically different networks. Even when these differences are not very sharp, they still produce differentiated implications (*e.g.*, differences in the top-5 synergistic SDGs). Sensitivity to the method of choice and the impossibility of comprehensive validation tests (due to the opaqueness of the data-generating process of SDGs) demands filters for selecting a suitable method. We consider two criteria: discarding those methods whose theoretical assumptions are less consistent with the nature of SDG data and discarding those methods that produce relatively more false positives.

According to our results pcalg is the best method in terms of the number of false negatives, and on the other hand iota and sparsebn are the best methods when analyzing the number of false positives. However, on the theoretical front pcalg, and sparsebn have shortcomings in terms of assuming temporal independence and acyclical graphs. Likewise, the edge weight estimation in iota is not theoretically very sound and, thus, it needs to be combined with other method. Moreover, despite that these five procedures generate very different topologies, their country-level performance can be evaluated in more detail in further studies. In addition, one could consider a consensus networks built from a collection of graphs. Therefore, consensus methods is a topic to be explored in future research.

This paper provides a first view into 'second generation' methodologies for the inference of synergies and trade-offs between SDGs. Naturally, new questions will arise as these methods are disseminated in the sustainability community, so there should be a concerted effort to attack them. Some examples are: How accurate are the different network estimation methodologies? How robust are the methodologies to small samples or to the presence of structural breaks? Are some methods better suited for different types of countries? Which

of the methodologies is better able to capture true causal (or structural-dependence) links? How robust are the methods to violations of their assumptions *e.g.*, temporal dependence, presence of cycles, linearity, Gaussianity?

References

Ali, W., Rito, T., Reinert, G., Sun, F., and Deane, C. (2014). Alignment-Free Protein Interaction Network Comparison. *Bioinformatics*, 30:i430–i437.

Allen, C., Metternicht, G., and Wiedmann, T. (2018). Prioritising SDG Targets: Assessing Baselines, Gaps and Interlinkages. *Sustainability Science*.

Aragam, B., Gu, J., and Zhou, Q. (2017). Learning Large-Scale Bayesian Networks with the Sparsebn Package. *arXiv preprint arXiv:1703.04025*.

Aragam, B. and Zhou, Q. (2015). Concave Penalized Estimation of Sparse Gaussian Bayesian Networks. *Journal of Machine Learning Research*, 16(1):2273–2328.

Aste, T. and Di Matteo, T. (2017). Sparse Causality Network Retrieval from Short Time Series. *Complexity*.

Barigozzi, M. and Brownlees, C. T. (2017). NETS: Network Estimation for Time Series. *SSRN preprint dx.doi.org/10.2139/ssrn.2249909*.

Blanc, D. (2015). Towards Integration at Last? The Sustainable Development Goals as a Network of Targets. *Sustainable Development*, 23(3):176–187.

Boginski, V., Butenko, S., and Pardalos, P. (2005). Statistical Analysis of Financial Networks. *Computational Statistics & Data Analysis*, 48(2):431–443.

Bühlmann, P. (2013). Causal Statistical Inference in High Dimensions. *Mathematical Methods of Operations Research*, 77(3):357–370.

Casini, L. and Manzo, G. (2016). Agent-based Models and Causality: A Methodological Appraisal. *The IAS Working Paper Series*, 2016(7).

Castañeda, G., Chávez-Juárez, F., and Guerrero, O. A. (2018). How Do Governments Determine Policy Priorities? Studying Development Strategies through Networked Spillovers. *Journal of Economic Behavior & Organization*, 154:335–361.

Castañeda, G. and Guerrero, O. (2018). The Resilience of Public Policies in Economic Development. *Complexity*, 2018.

Castañeda, G. and Guerrero, O. (2019). The Importance of Social and Government Learning in Ex Ante Policy Evaluation. *Journal of Policy Modeling*.

Castañeda, G., Íñiguez, G., and Chávez-Juárez, F. (2017). The Complex Network of Public Policies. An Empirical Framework for Identifying their Relevance in Economic Development. Background Paper, Governance and The Law, The World Bank.

Castagneto-Gissey, G., Chavez, M., and Fallani, F. D. V. (2014). Dynamic Granger-Causal Networks of Electricity Spot Prices: A Novel Approach to Market Integration. *Energy Economics*, 44:422–432.

Ceriani, L. and Gigliarano, C. (2016). Multidimensional Well-Being: A Bayesian Networks Approach. Technical Report 399, ECINEQ, Society for the Study of Economic Inequality.

Cinicioglu, E., Ulusoy, G., Ekici, Ş., Ülengin, F., and Ülengin, B. (2017). Exploring the Interaction Between Competitiveness of a Country and Innovation Using Bayesian Networks. *Innovation and Development*, 0(0):1–36.

Clark, A., Ye, H., Isbell, F., Deyle, E., Cowles, J., Tilman, G., and Sugihara, G. (2015). Spatial Convergent Cross Mapping to Detect Causal Relationships from Short Time Series. *Ecology*, 96(5):1174–1181.

Cobey, S. and Baskerville, E. (2016). Limits to Causal Inference with State-Space Reconstruction for Infectious Disease. *PLOS ONE*, 11(12):e0169050.

Collste, D., Pedercini, M., and Cornell, S. (2017). Policy Coherence to Achieve the SDGs: Using Integrated Simulation Models to Assess Effective Policies. *Sustainability Science*, 12(6):921–931.

Cucurachi, S. and Suh, S. (2017). Cause-Effect Analysis for Sustainable Development Policy. *Environmental Reviews*, 25(3):358–379.

Czyżewska, M. and Mroczek, T. (2014). Bayesian Approach to the Process of Identification of the Determinants of Innovativeness. *Finansowy Kwartalnik Internetowy e-Finanse*, 10(2):44–56.

Dimitrios, K. and Vasileios, O. (2015). A Network Analysis of the Greek Stock Market. *Procedia Economics and Finance*, 33:340–349.

Dörgő, G., Sebestyén, V., and Abonyi, J. (2018). Evaluating the Interconnectedness of the Sustainable Development Goals Based on the Causality Analysis of Sustainability Indicators. *Sustainability*, 10(10):3766.

El-Maghrabi, M., Gable, S., Osorio-Rodarte, I., and Verbeek, J. (2018). Sustainable Development Goals Diagnostics: An Application of Network Theory and Complexity Measures to Set Country Priorities. *World Bank Working Paper*, WPS8481:1–22.

Gao, X., Huang, S., Sun, X., Hao, X., and An, F. (2018). Modelling Cointegration and Granger Causality Network to Detect Long-Term Equilibrium and Diffusion Paths in the Financial System. *Royal Society Open Science*, 5(3):172092.

General Assembly (25-Sep-2015). Resolution Adopted by the General Assembly on 25 September 2015. Technical Report A/70/L.1, United Nations.

Granger, C. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37(3):424–438.

Guerrero, O. and Castañeda, G. (2019a). Does Better Governance Guarantee Less Corruption? Evidence of Loss in Effectiveness of the Rule of Law. *arXiv preprint arXiv:1902.00428*.

Guerrero, O. and Castañeda, G. (2019b). Quantifying the Coherence of Development Policy Priorities. *arXiv preprint arXiv:1902.00430*.

Hall, N. (2004). Two Concepts of Causation. In Collins, J., Hall, N., and Paul, L., editors, *Causation and Counterfactuals*, pages 225–276. MIT Press, Cambridge MA.

Hamming, R. (1950). Error detecting and error correcting codes. *The Bell System Technical Journal*, 29(2):147–160.

Han, L. and Zhu, J. (2008). Using Matrix of Thresholding Partial Correlation Coefficients to Infer Regulatory Network. *Biosystems*, 91(1):158–165.

Heckman, J. and Vytlacil, E. (2007). Chapter 70 Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation. In Heckman, J. and Leamer, E., editors, *Handbook of Econometrics*, volume 6, pages 4779–4874. Elsevier.

Hempel, S., Koseska, A., Kurths, J., and Nikoloski, Z. (2011). Inner Composition Alignment for Inferring Directed Networks from Short Time Series. *Physical Review Letters*, 107(5):054101.

Heskamp, L., Abeelen, A., Lagro, J., and Claassen, J. (2014). Convergent Cross Mapping: A Promising Technique for Cerebral Autoregulation Estimation. *International Journal of Clinical Neurosciences and Mental Health*, (1(Suppl. 1)):S20.

Hu, Y., Zhao, H., and Ai, X. (2016). Inferring Weighted Directed Association Network from Multivariate Time Series with a Synthetic Method of Partial Symbolic Transfer Entropy Spectrum and Granger Causality. *PLOS ONE*, 11(11):e0166084.

Huang, W.-Q., Zhuang, X., and Yao, S. (2009). A Network Analysis of the Chinese Stock Market. *Physica A: Statistical Mechanics and its Applications*, 388(14):2956–2964.

Hyvärinen, A. and Smith, S. (2013). Pairwise Likelihood Ratios for Estimation of Non-Gaussian Structural Equation Models. *Journal of Machine Learning Research*, 14(Jan):111–152.

Kalisch, M. and Bühlmann, P. (2007). Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm. *Journal of Machine Learning Research*, 8(Mar):613–636.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M., and Bühlmann, P. (2012). Causal Inference Using Graphical Models with the R Package pcalg. *Journal of Statistical Software*.

Lacerda, G., Spirtes, P., Ramsey, J., and Hoyer, P. (2008). Discovering Cyclic Causal Models by Independent Components Analysis. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, UAI'08, pages 366–374, Arlington, Virginia, United States. AUAI Press.

Lauritzen, S. (1996). *Graphical Models*. Oxford Science Publications. Clarendon Press.

Linderman, S. and Adams, R. (2014). Discovering Latent Network Structure in Point Process Data. In *Proceedings of The 31st International Conference on Machine Learning*, pages 1413–1421.

Ma, H., Aihara, K., and Chen, L. (2014). Detecting Causality from Nonlinear Dynamics with Short-term Time Series. *Scientific Reports*, 4:7464.

Ma, H., Leng, S., Tao, C., Ying, X., Kurths, J., Lai, Y.-C., and Lin, W. (2017). Detection of time delays and directional interactions based on time series from complex dynamical systems. *Physical Review E*, 96(1):012221.

Maathuis, M., Kalisch, M., and Bühlmann, P. (2009). Estimating High-Dimensional Intervention Effects from Observational Data. *The Annals of Statistics*, 37(6A):3133–3164.

Massara, G. P., Matteo, T., and Aste, T. (2017). Network Filtering for Big Data: Triangulated Maximally Filtered Graph. *Journal of Complex Networks*, 5(2):161–178.

McBride, J., Zhao, X., Munro, N., Jicha, G., Schmitt, F., Kryscio, R., Smith, C., and Jiang, Y. (2015). Sugihara Causality Analysis of Scalp EEG for Detection of Early Alzheimer's Disease. *NeuroImage: Clinical*, 7:258–265.

Meinshausen, N., Hauser, A., Mooij, J., Peters, J., Versteeg, P., and Bühlmann, P. (2016). Methods for causal inference from gene perturbation experiments and validation. *Proceedings of the National Academy of Sciences*, 113(27):7361–7368.

Nilsson, M., Griggs, D., and Visbeck, M. (2016). Policy: Map the interactions between Sustainable Development Goals. *Nature News*, 534(7607):320.

OECD (2018). *Policy Coherence for Sustainable Development 2018: Towards Sustainable and Resilient Societies.* OECD, Paris. OCLC: 1040196886.

Pearl, J. (2000). Causal Inference Without Counterfactuals: Comment. *Journal of the American Statistical Association*, 95(450):428–431.

Pearl, J. (2009). Causal Inference in Statistics: An Overview. *Statistics Surveys*, 3:96–146.

Pedercini, M. and Barney, G. (2010). Dynamic Analysis of Interventions Designed to Achieve Millennium Development Goals (MDG): The Case of Ghana. *Socio-Economic Planning Sciences*, 44(2):89–99.

Peters, J., Janzing, D., and Schölkopf, B. (2013). Causal Inference on Time Series using Restricted Structural Equation Models. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 26*, pages 154–162. Curran Associates, Inc.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA.

Pradhan, P., Costa, L., Rybski, D., Lucht, W., and Kropp, J. (2017). A Systematic Study of Sustainable Development Goal (SDG) Interactions. *Earth's Future*, 5(11):1169–1179.

Ramsey, J., Glymour, M., Sanchez-Romero, R., and Glymour, C. (2017). A Million Variables and More: The Fast Greedy Equivalence Search Algorithm for Learning High-Dimensional Graphical Causal Models, with an Application to Functional Magnetic Resonance Images. *International journal of data science and analytics*, 3(2):121–129.

Rodrik, D. (2009). *One Economics, Many Recipes: Globalization, Institutions, and Economic Growth*. Princeton University Press, Princeton.

Runber, Y., Tomasi, C., and Guibas, L. (1998). A Metric for Distributions with Applications to Image Databases. In *In IEEE International Conference Computer Vision*, pages 59–66.

Servadio, J. L. and Convertino, M. (2018). Optimal Information Networks: Application for Data-Driven Integrated Health in Populations. *Science Advances*, 4(2):e1701088.

Shimizu, S. (2014). LiNGAM: Non-Gaussian Methods for Estimating Causal Structures. *Behaviormetrika*, 41(1):65–98.

Smith, S., Miller, K., Salimi-Khorshidi, G., Webster, M., Beckmann, C., Nichols, T., Ramsey, J., and Woolrich, M. (2011). Network Modelling Methods for FMRI. *NeuroImage*, 54(2):875–891.

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search*, volume 81. Springer.

Sugihara, G., May, R., Ye, H., Hsieh, C., Deyle, E., Fogarty, M., and Munch, S. (2012). Detecting Causality in Complex Ecosystems. *Science*, 338(6106):496–500.

The World Bank Group (2010). World Development Indicators (WDI) — Data Catalog. https://datacatalog.worldbank.org/dataset/world-development-indicators.

The World Bank Group (2016). Sustainable Development Goals Database. https://datacatalog.worldbank.org/dataset/sustainable-development-goals.

Tsonis, A., Deyle, E., May, R., Sugihara, G., Swanson, K., Verbeten, J., and Wang, G. (2015). Dynamical Evidence for Causality Between Galactic Cosmic Rays and Interannual Variation in Global Temperature. *Proceedings of the National Academy of Sciences*, 112(11):3253–3256.

Tumminello, M., Aste, T., Matteo, T., and Mantegna, R. (2005). A Tool for Filtering Information in Complex Systems. *Proceedings of the National Academy of Sciences*, 102(30):10421–10426.

United Nations Statistics Division (2016). United Nations Global SDG Database. https://unstats.un.org/sdgs/indicators/database.

Wang, Y., Yang, J., Chen, Y., Maeyer, P., Li, Z., and Duan, W. (2018). Detecting the Causal Effect of Soil Moisture on Precipitation Using Convergent Cross Mapping. *Scientific Reports*, 8(1):12171.

Wegner, A., Ospina-Forero, L., Gaunt, R., Deane, C., and Reinert, G. (2017). Identifying Networks with Common Organizational Principles. *Journal of Complex networks*.

Weitz, N., Carlsen, H., Nilsson, M., and Skånberg, K. (2018). Towards Systemic and Contextual Priority Setting for Implementing the 2030 Agenda. *Sustainability Science*, 13(2):531–548.

Xu, Y., Salapaka, S., and Beck, C. (2013). A Distance Metric Between Directed Weighted Graphs. In *52nd IEEE Conference on Decision and Control*, pages 6359–6364.

Ye, H., Deyle, E., Gilarranz, L., and Sugihara, G. (2015). Distinguishing Time-Delayed Causal Interactions Using Convergent Cross Mapping. *Scientific Reports*, 5:14750.

Zhang, B., Li, W., Shi, Y., Liu, X., and Chen, L. (2017). Detecting Causality from Short Time-Series Data Based on Prediction of Topologically Equivalent Attractors. *BMC Systems Biology*, 11(7):128.

Zhou, X. and Moinuddin, M. (2016). Review of the SDG Index and Dashboards: An Example of Japan's Global Ranking Results. *IGES Working Paper*.

A Data

Table 5: Description of selected indicators for the SDGs

SDG	Indicator	Indicator name
1	1	Proportion of population below international poverty line (%)
1	2	Poverty gap at $1.90 a day (2011 PPP) (%)
1	3	Poverty headcount ratio at $3.20 a day (2011 PPP) (% of population)
1	4	Poverty gap at $3.20 a day (2011 PPP) (%)
1	5	Poverty headcount ratio at $5.50 a day (2011 PPP) (% of population)
1	6	Poverty gap at $5.50 a day (2011 PPP) (%)
2	7	Prevalence of anemia among women of reproductive age (% of women ages 15-49)
2	8	Cereal yield (kg per hectare)
3	9	Maternal mortality ratio
3	10	Proportion of births attended by skilled health personnel (%)
3	11	Births attended by skilled health staff (% of total)
3	12	Maternal mortality ratio (modeled estimate, per 100,000 live births)
3	13	Mortality rate, under-5 (per 1,000 live births)
3	14	Mortality rate, neonatal (per 1,000 live births)
3	15	Adolescent fertility rate (births per 1,000 women ages 15-19)
3	16	Immunization, DPT (% of children ages 12-23 months)
3	17	Immunization, measles (% of children ages 12-23 months)
4	18	Over-age students, primary (% of enrollment)
4	19	Children out of school (% of primary school age)
4	20	Lower secondary completion rate, total (% of relevant age group)
4	21	Adolescents out of school (% of lower secondary school age)
4	22	School enrollment, preprimary (% gross)
4	23	School enrollment, tertiary (% gross)
4	24	School enrollment, primary (gross), gender parity index (GPI)
4	25	School enrollment, primary and secondary (gross), gender parity index (GPI)
4	26	School enrollment, secondary (gross), gender parity index (GPI)
4	27	School enrollment, tertiary (gross), gender parity index (GPI)
4	28	Pupil-teacher ratio, preprimary
4	29	Pupil-teacher ratio, primary
4	30	Pupil-teacher ratio, lower secondary
4	31	Pupil-teacher ratio, secondary
4	32	Pupil-teacher ratio, upper secondary
5	33	Contributing family workers, female (% of female employment) (modeled ILO estimate)
5	34	Contributing family workers, male (% of male employment) (modeled ILO estimate)
5	35	Proportion of seats held by women in national parliaments (%)
6	36	Renewable internal freshwater resources per capita (cubic meters)
7	37	Access to electricity (% of population)
7	38	Renewable electricity output (% of total electricity output)
7	39	Renewable energy consumption (% of total final energy consumption)
7	40	Energy intensity level of primary energy (MJ/$2011 PPP GDP)
8	41	tax revenue (current LCU)
8	42	Tax revenue (% of GDP)
8	43	Exports of goods and services (% of GDP)
8	44	GDP, PPP (constant 2011 international $)
8	45	Foreign direct investment, net inflows (% of GDP)
8	46	Patent applications, nonresidents
8	47	Patent applications, residents
8	48	GDP growth (annual %)
8	49	GDP per capita growth (annual %)
8	50	Employment in agriculture (% of total employment) (modeled ILO estimate)
8	51	Employment in industry (% of total employment) (modeled ILO estimate)
8	52	Employment in services (% of total employment) (modeled ILO estimate)
8	53	Wage and salaried workers, total (% of total employment) (modeled ILO estimate)
8	54	Unemployment, youth total (% of total labor force ages 15-24) (modeled ILO estimate)
8	55	Unemployment, total (% of total labor force) (modeled ILO estimate)
9	56	Individuals using the Internet (% of population)
9	57	Air transport, freight (million ton-km)
9	58	Air transport, passengers carried
9	59	Railways, goods transported (million ton-km)
9	60	Railways, passengers carried (million passenger-km)
9	61	Manufacturing, value added (% of GDP)
9	62	Medium and high-tech industry (% manufacturing value added)
10	63	Income share held by second 20%
11	64	Proportion of urban population living in slums (%)
11	65	Population living in slums (% of urban population)
11	66	Urban population growth (annual %)
11	67	Urban population (% of total)
11	68	PM2.5 air pollution, mean annual exposure (micrograms per cubic meter)
11	69	PM2.5 pollution, population exposed to levels exceeding WHO Interim Target-1 value (% of total)
12	70	Adjusted net savings, excluding particulate emission damage (% of GNI)
12	71	Coal rents (% of GDP)
12	72	Forest rents (% of GDP)
12	73	Mineral rents (% of GDP)
12	74	Natural gas rents (% of GDP)
12	75	Oil rents (% of GDP)
12	76	Total natural resources rents (% of GDP)
13	77	CO2 emissions (metric tons per capita)
14	78	Aquaculture production (metric tons)
14	79	Capture fisheries production (metric tons)
14	80	Total fisheries production (metric tons)
15	81	Forest area (% of land area)
15	82	Red List Index
17	83	Tariff rate, applied, simple mean, manufactured products (%)
17	84	Tariff rate, applied, simple mean, all products (%)

17	85	Tariff rate, applied, simple mean, primary products (%)
17	86	Volume of remittances (in United States dollars) as a proportion of total GDP (%)
17	87	Debt service as a proportion of exports of goods and services (%)

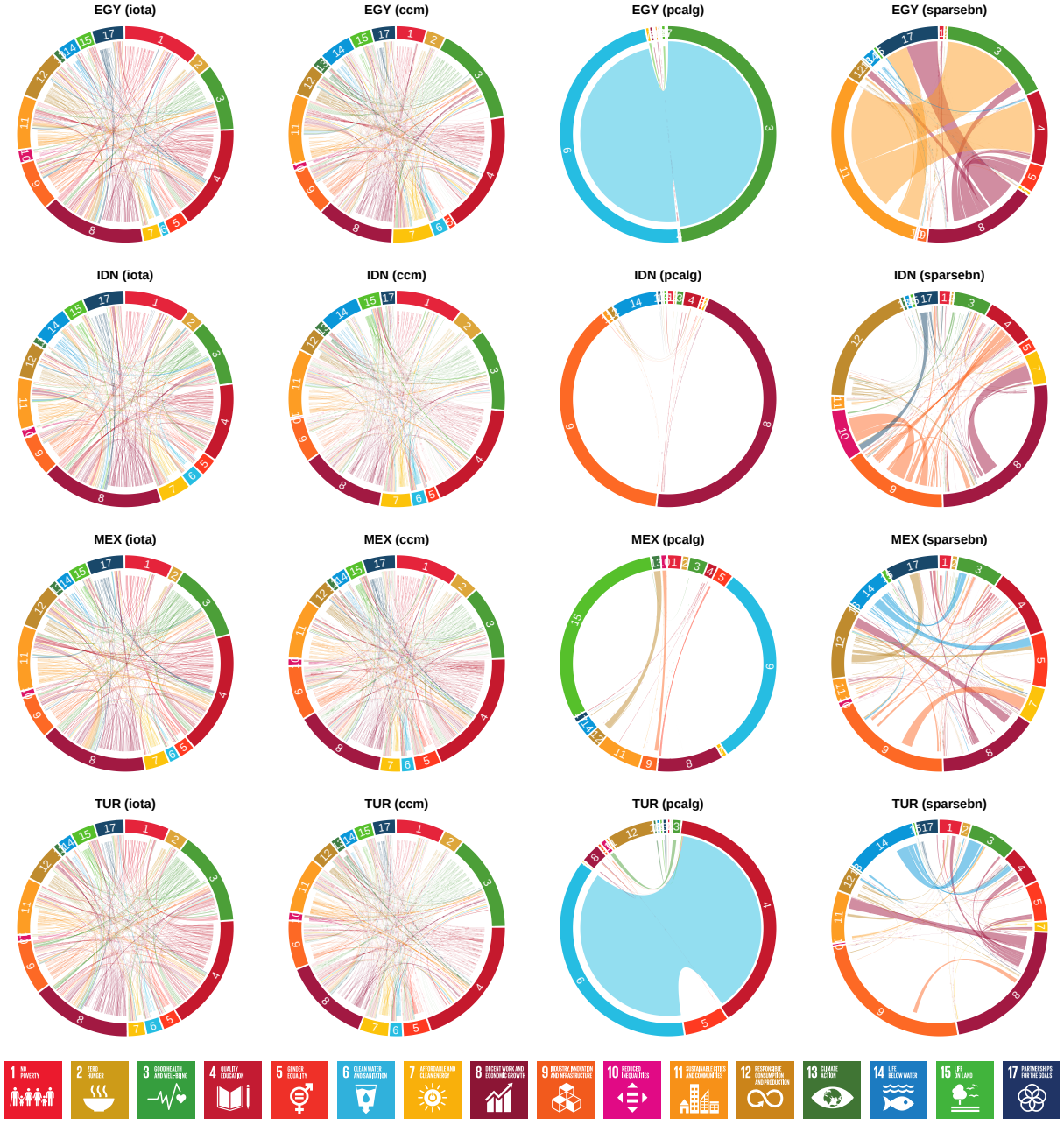
Table 6: Average indicator by SDG and country

SDG	EGY	IDN	MEX	TUR
1 no poverty (N =6)	0.76725	0.54939	0.84259	0.92224
	(0.04127)	(0.04647)	(0.02264)	(0.01556)
2 zero hunger (N =2)	0.41806	0.39757	0.45148	0.35207
	(0.08294)	(0.12176)	(0.17261)	(0.13204)
3 good health and well-being (N =9)	0.85357	0.78856	0.90912	0.8856
	(0.01548)	(0.01318)	(0.01137)	(0.01012)
4 quality education (N =15)	0.62278	0.6497	0.66144	0.62992
	(0.0191)	(0.01961)	(0.01737)	(0.0193)
5 gender equality (N =3)	0.43131	0.51191	0.46852	0.50791
	(0.11062)	(0.09066)	(0.10564)	(0.10615)
6 clean water and sanitation (N =1)	0.0	0.00085	0.00036	0.00032
	(0.0)	(7e-05)	(3e-05)	(3e-05)
7 affordable and clean energy (N =4)	0.53328	0.59188	0.54591	0.58666
	(0.10695)	(0.08192)	(0.10224)	(0.09259)
8 decent work and economic growth (N =15)	0.31846	0.34711	0.34486	0.32953
	(0.01839)	(0.01919)	(0.02196)	(0.01921)
9 industry (N =7)	0.11709	0.14339	0.1467	0.13468
	(0.0214)	(0.02978)	(0.0267)	(0.02471)
10 reduced inequalities (N =1)	0.22666	0.96867	0.46898	0.37651
	(0.0127)	(0.13654)	(0.031)	(0.02065)
11 sustainable cities and communities (N =6)	0.49914	0.70617	0.76453	0.72196
	(0.04837)	(0.0381)	(0.03208)	(0.03084)
12 responsible consumption and production (N =7)	0.92537	0.92716	0.945	0.95924
	(0.01377)	(0.01341)	(0.01271)	(0.01356)
13 climate action (N =1)	0.97044	0.97751	0.94283	0.94803
	(0.0047)	(0.0056)	(0.00306)	(0.00715)
14 life below water (N =3)	0.0046	0.04355	0.00799	0.00315
	(0.00072)	(0.00922)	(0.00197)	(0.0006)
15 life on land (N =2)	0.46285	0.63705	0.45329	0.47814
	(0.23121)	(0.04842)	(0.05232)	(0.16902)
17 partnerships for the goals (N =5)	0.82021	0.93225	0.9015	0.93548
	(0.02858)	(0.00892)	(0.01246)	(0.0111)

Sample mean across years and indicators within the same SDG. Standard deviation in parentheses. N indicates the number of indicators within a single SDG.

B Estimated trade-offs

Figure 6: Estimated trade-offs between SDGs by country and method



Weights have been normalized so they are comparable in terms of proportion to all other weights in the network. Edges colors indicate the SDG from which they originate.

C False positives and false negatives

Table 7: Prior-informed false positives

Goal	Indicator	Indicator name
5	34	Contributing family workers, male (% of male employment) (modeled ILO estimate)
7	37	Access to electricity (% of population)
5	35	Proportion of seats held by women in national parliaments (%)
12	73	Mineral rents (% of GDP)
8	50	Employment in agriculture (% of total employment) (modeled ILO estimate)
6	36	Renewable internal freshwater resources per capita (cubic meters)
3	9	Maternal mortality ratio
4	19	Children out of school (% of primary school age)
11	67	Urban population (% of total)
12	71	Coal rents (% of GDP)
3	11	Births attended by skilled health staff (% of total)
9	61	Manufacturing, value added (% of GDP)
11	66	Urban population growth (annual %)
15	82	Red List Index
3	17	Immunization, measles (% of children ages 12-23 months)
12	74	Natural gas rents (% of GDP)
11	69	PM2.5 pollution, population exposed to levels exceeding WHO Interim Target-1 value (% of total)
4	22	School enrollment, preprimary (% gross)
8	53	Wage and salaried workers, total (% of total employment) (modeled ILO estimate)
17	84	Tariff rate, applied, simple mean, all products (%)

Contiguous rows with the same shade indicate a pair of indicators for which an edge is considered a false positive.

Table 8: Prior-informed false negatives

Goal	Indicator	Indicator name
1	1	Proportion of population below international poverty line (%)
1	3	Poverty headcount ratio at $3.20 a day (2011 PPP) (% of population)
3	11	Births attended by skilled health staff (% of total)
3	12	Maternal mortality ratio (modeled estimate, per 100,000 live births)
4	20	Lower secondary completion rate, total (% of relevant age group)
4	22	School enrollment, preprimary (% gross)
5	34	Contributing family workers, male (% of male employment) (modeled ILO estimate)
5	35	Proportion of seats held by women in national parliaments (%)
7	39	Renewable energy consumption (% of total final energy consumption)
7	40	Energy intensity level of primary energy (MJ/$2011 PPP GDP)
8	42	Tax revenue (% of GDP)
8	43	Exports of goods and services (% of GDP)
8	52	Employment in services (% of total employment) (modeled ILO estimate)
8	53	Wage and salaried workers, total (% of total employment) (modeled ILO estimate)
9	61	Manufacturing, value added (% of GDP)
9	62	Medium and high-tech industry (% manufacturing value added)
11	67	Urban population (% of total)
11	68	PM2.5 air pollution, mean annual exposure (micrograms per cubic meter)
12	71	Coal rents (% of GDP)
12	72	Forest rents (% of GDP)

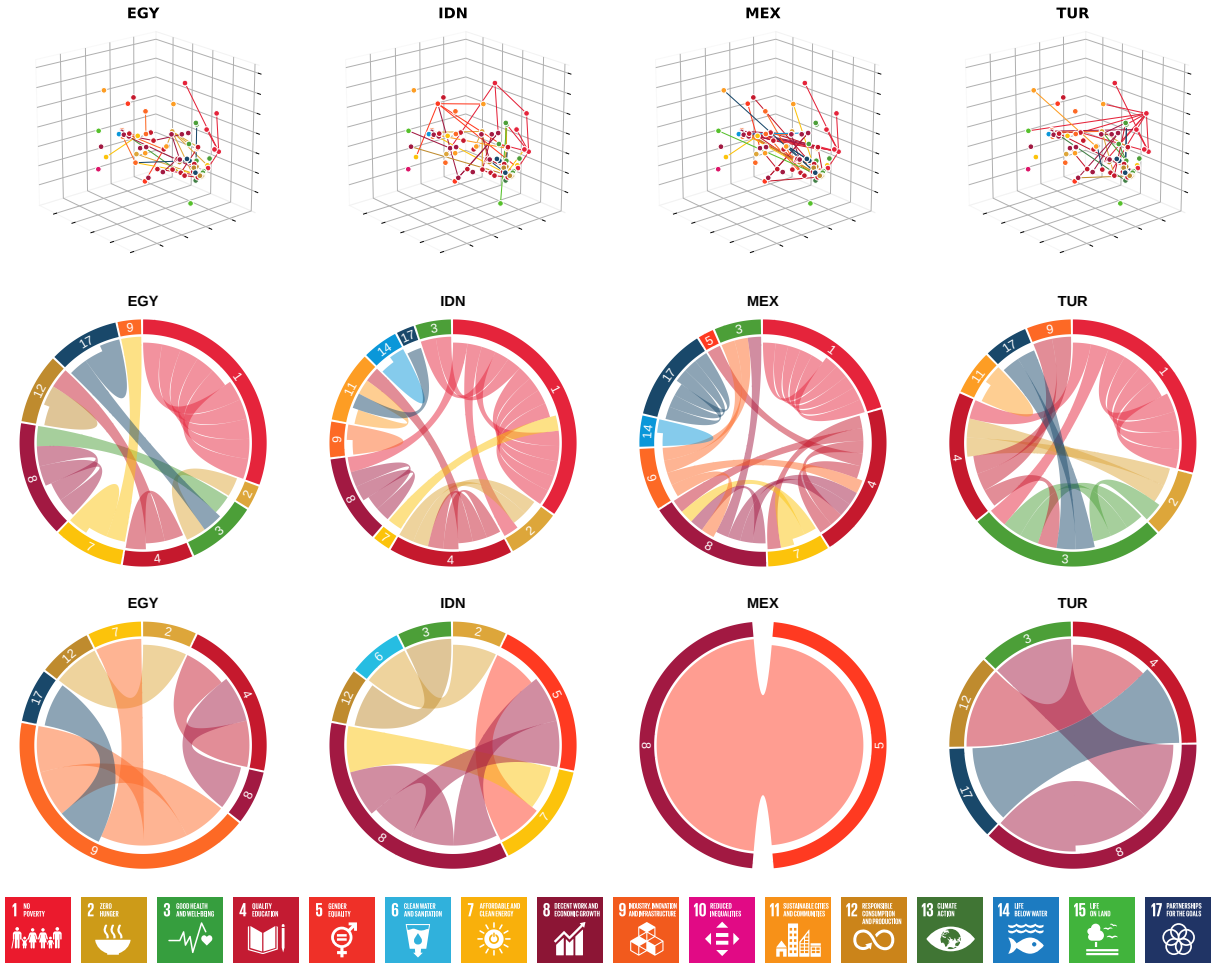
Contiguous rows with the same shade indicate a pair of indicators for which the complete absence of edges is considered a false negative.

D Consensus networks

Given multiple graphs, a *consensus network* consists of all the nodes and edges that these networks have in common (it does not consider weights). Hence, this approach reduces the probability of false positives. Since pcalg produces SDG networks with very few links, we build consensus networks only considering lcorr, iota, ccm and sparsebn. Each consensus network is built at the level of each country, allowing us to compare differences between nations.

Figure 7 shows the consensus networks of the four countries through three different plots. In the top row, we have drawn the nodes in a 3D space through a principal-component analysis that uses all the data points in the sample. In other words, the coordinates of the nodes are the same for every country. Then, we draw the edges corresponding to each country's consensus network. Note that, in the consensus networks, several nodes are completely disconnected and some others are not part of the giant component. Furthermore, the chordal plots of consensus synergies and trade-off allow us to see clear differences in the overall structure between countries. For instance, in the Mexican network the SDG 17 has synergistic effects only within the same SDG. In contrast, this SDG in Egypt and Turkey has important effects on SDG 3, and in Indonesia on SDG 11.

Figure 7: Consensus SDG networks



Top panels: consensus networks overlaid on a 3D principal-component projection of the entire set of development indicators. It contains both synergies and trade-offs. Each coordinate was obtained by pooling the data across countries and years, and reducing it to three dimensions via principal-component analysis. Thus, each dot represents the average level of development of each indicator in the sample, but reduced from four countries and 20 years to 3 dimensions. Middle panels: chordal plots of consensus synergies. Bottom panels: chordal plots of consensus trade-offs.

Finally, the middle and bottom rows of the plot correspond to the synergies and trade-offs of the consensus networks at the level of SDGs. As previously suggested, this type of networks can provide more robust estimates in so far as they are based on a wisdom-of-the-crowd methodology. Note that the estimated synergies and trade-offs are quite different for these four countries (context matters). Nonetheless, a common feature is that there are many complementarities within SDGs. Positive spillovers between SDGs are more frequent in Mexico and Turkey than in Egypt and Indonesia. Negative spillovers are widespread in the Mexican indicators, while they are very focused in Indonesia.