ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Civil and Environmental Engineering
(XXXIII cycle)

Networks for Change:
Methodologies to Track and Address Global Performances of Innovation and Sustainability

by

Carla Sciarra

Supervisors:

Prof. Francesco	Laio,	Supervisor
Prof. Luca	Ridolfi,	Co-Supervisor
Prof. Guido	Chiarotti,	Co-Supervisor

Doctoral Examination Committee:

Prof. Enrico Bertuzzo, Università Ca' Foscari Venezia
Dr. Tommaso Ciarli, University of Sussex
Prof. Silvana Dalmazzone, Università degli Studi di Torino
Prof. Fabio Fagnani, Politecnico di Torino
Dr. Luz Fernandez Garcia, United Nations Country Office Perú
Dr. Emanuele Pugliese, EU Joint Research Centre Sevilla

Politecnico di Torino

Spring 2021

Declaration

I hereby declare that the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Carla Sciarra
May 2021

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. Degree** in the Graduate School of Politecnico di Torino (ScuDo).

To Roberto, for your intelligence, curiosity and stubbornness.
And to all the bold Women out there, striving and succeeding for things to change.

"The world will say to you, 'There are too many problems.' Do not be afraid to be part of the solutions. Start by discussing the issues. We cannot overcome what we ignore. The more we talk about things, the more we see that the issues are connected because we are connected. [...] Change-making does not belong to one group of people; it belongs to all of us."

– Cleo Wade

Acknowledgment

During these years of PhD, I have accomplished so much, at both personal and professional levels and I had the luck of being accompanied by a lot of people who encouraged me to keep going!

None of this would have been possible if it were not for my supervisors, Francesco, Luca and Guido.

Thank you, Francesco. Despite my initial scepticism, you have believed in me since graduation day and supported me with almost paternal affection. Thanks for your teachings, especially about statistics and (life-applicable) problem solving, and for the laughs during serious contexts. Thank you for having offered me many life opportunities and taught me the "take it or leave it" principle. I am pleased I have changed my mind about pursuing a PhD in Italy.

Thank you, Luca. During these years, you have supported me, pushed me to learn more about any topic, and gave me precious advice about how to become a researcher. Thanks for having filtered, with your equilibrium, the noise of this chaotic team.

Thank you, Guido. Before meeting you, I would have never thought that researchers could be so eclectic. Physics, arts, literature, economics... everything can be harmonized employing passion and willingness, and you are an example to follow.

Thanks to my family. *Grazie Mamma e Papà, perché questo mio traguardo è anche vostro. Grazie perché anche se non sapete cosa sia la teoria delle reti o l'economia complessa, so che ad ogni piccolo successo o sconfitta mi direte sempre che siete fieri di me.* Thanks to my sisters Rita and Amalia and my brother-in-law Rafael for the advice, support, and love. We fight, we argue... still we stand for each other, and we have learnt this has no price. Thanks to Roberto, because you remembered me the beauty of being curious and stubborn, necessary qualities to pursue a career in research, and for all the times you video-called me showing me your lovely big smile while shouting out-loud "zia!".

Thanks to Emanuela and Ilaria for always checking on me. Thanks for their support to Walter, Mariangela, Zappi, Giorgio and in particular Fra, he knew that this PhD was a good choice when I did not yet.

Thanks to Sabrina for being my partner in crime in all the stupid things I have committed in these last years with absolutely no planning.

Thanks to Valentina for being the only one able to understand why stability is not my thing.

Thanks to Marta P., even if I do not remember how we became friends, you have always been there for me in the last three years.

Thanks to my Ciclone: Benedetta, Caterina, Francesca and Marta. Even in the distance and challenging moments, we have supported each other, always trying to laugh about our dramas. Thanks for our talks, trips, runaways, dinners, parades, laughs and beers. Special thanks go to Marta because I could not have imagined a better female role model in academia than you have been. Even if not officially a member, thank you, Centa, for your cooking lessons.

Thanks to all my friend-roommates, Valeria, Angela, Marco, Francesca, Chiara and in particular Giuseppe and Roberta, who supported me with laughs, drinks and food.

Thanks to all the people of DIATI, all technical and research staff, I had the honour to share this experience with, and in particular, the people of BAD3: Alice, Silvia I, Silvia C, Sofia, Edoardo and Paola. Big, special thanks go to Bert, Roberto, and Davide that, more than anyone else, survived and contributed to my madness during these last months while writing down this manuscript. Most importantly... *Vorrei ringraziare l'Islanda per essere molto lontano ma vicino a Padova.* Thanks to Silvietta for reminding me that the world as we see it has plenty of shining colours and shades, even black and white. Thanks to Paolo and Elena and the random YGA adventure that united us during this PhD.

Gracias a Silvia, Lucía y Mariana, por las risas y los chismes, las discusiones iluminantes y el feminismo, el apoyo y el afecto, siempre están aún con el tiempo y un océano, un continente y una pandemia por medio que nos separan.

Thanks to Tommaso for having pushed me in doing, pretending, more and better from research, work, relationships and food. And thanks to my team-mates at CEST, you all opened my mind about so many life opportunities.

Thanks to the WTF team, aka Water To Food, you showed me a way for science to be accessible to the general public. Also, thanks for all the artichokes and wine we had in Rome.

Thanks to the LOFT gym for being the place where all of my energies and frustrations found burst.

Thank you Clarisa for helping me climb my own mountains.

Thanks to the all anonymous and non-anonymous reviewers for the

instructive discussions about the work presented in this manuscript, in particular Andrea Tacchella and Matthieu Cristelli, and the doctoral committee members Luz Fernández García and Emanuele Pugliese, for their valuable comments about addressing future research works.

Last but not least: "The authors acknowledge the European Research Council funding from the *Coping with Water Scarcity in a Globalized World* project (ERC-2014-CoG, project 647473)."

Thanks to all the people that I did not mention but that, one way or another, they were by my side. Thank you because...

"If you'll believe in me, I'll still believe!"

Abstract

Network science is a mathematical theory that aims at disentangling the complex relationships among the entities in a system. Any system in which connections are present can be interpreted as a *network*: a group of friends, airplane transportation, or river basins. Graphically, the network representation constitutes nodes (i.e., the entities) and links (i.e., the connections). Mathematically, a network is described through a matrix. The clarity of its representation and its versatility have let network science rise in popularity across many fields, ranging from sociology to epidemiology, from economics to engineering, from neuroscience to hydrology. For example, network science has helped us understand the power of 'super-spreaders during the Covid-19 pandemic. But what if its own popularity would compromise the power of such an important tool?

The increasing popularity of network theory also generated several debates about its methodologies and their contexts of applications. Among the most discussed methodologies figure the ones for measuring centrality. Centrality is the feature of being relevant in a network. Many conceptualizations of centrality metrics have been proposed over time. Nevertheless, all the methodologies depend on the particular aims and applications for which they have been idealized. For example, the degree and eigenvector centrality have been introduced to study power and hierarchies in social networks. However, these metrics do not account for the feature of topological distance among the nodes, which other measures as the closeness and betweenness centrality do, instead.

The heterogeneity in the results offered by different measures of centrality is a recurrent problem in the literature. In fact, centrality is often referred to as an instrument to rank the entities in the system and, the differences among the available methodologies unavoidably create differences in the obtained rankings even within the same application. This is the case, for example, of the Economic Complexity theory, a network framework for export data through which explore countries' performances in innovation. Another relevant example of the application of network science frames within the field of sustainable development. The theory has been shown to help unfold the synergies and trade-offs among its sectors. Nevertheless, such a tool's potential has not been fully exploited to rank countries for their status' in achieving sustainable development.

Against this background, in this thesis, we address the need for a common definition in the field of centrality measures by approaching the problem in three manners: firstly, in its general mathematical description; secondly, its application to the area of economic complexity; finally, we introduce a neat centrality framework within the context of sustainable development aimed at measuring countries' status concerning the United Nations' Agenda 2030.

Therefore, this thesis contributes to multidisciplinary literature in network science, economics, innovation, and sustainable development. We introduce a statistical perspective that uniquely defines centrality metrics and compares the information they provide about the network topology. Within this statistical framework, we also introduce multi-dimensional centrality metrics to account for the network's many structural features. Thanks to this multi-dimensional setting on centrality metrics, we reconcile the most notorious economic complexity metrics, i.e., the *Method of Reflections* and the *Fitness and Complexity* algorithm. Therefore, we present the *GENeralized Economic comPlexity* – GENEPY – index: a unique, data-driven, multi-dimensional index of innovation framed within economic complexity. The GENEPY index can track the trajectories of growth of countries without the need to add exogenous information about countries' economies (as, e.g., e Gross Domestic Product). Finally, we introduce the network representation of the Agenda 2030 of sustainable development and, in particular, the countries –Sustainable Development Goals system. Thanks to this representation, we also introduce the use of centrality metrics, especially the GENEPY ones, to shed new light on countries' efficacy in sustainable development.

Hence, our results contribute to defining methodologies to track and address global performances of innovation and sustainability.

Contenuto

La teoria delle reti è una branca della matematica che ha come obiettivo quello di indagare e spiegare le relazioni complesse tra gli enti in un sistema. Qualsiasi sistema in cui gli enti sono collegati tra di loro può essere rappresentato e interpretato sotto forma di *rete*: le relazioni in un gruppo di amici, le rotte che formano il trasporto aereo o l'idrografia di un bacino fluviale. Dal punto di vista grafico, una rete è descritta da nodi (gli enti), collegati tramite linee che ne descrivono il tipo di connessione tra di essi. Dal punto di vista matematico, una rete è descritta per mezzo di una matrice. La chiarezza della rappresentazione e il suo essere adattabile a diversi contesti spiegano perchè tale teoria sia divenuta uno strumento matematico sempre più presente nella letteratura scientifica, con applicazioni che vanno dalle scienze umane all'epidemiologia, dall'ingegneria alla biologia. Ad esempio, grazie all'utilizzo della teoria delle reti è stato definito il ruolo dei 'super-contaminatori' nella diffusione del virus che ha provocato la pandemia Covid-19.

La crescente popolarità di questa branca della matematica ha generato però anche diversi dibattiti scientifici per l'utilizzo delle sue metodologie e dei suoi contesti di applicazione. Tra le più dibattute metodologie in teoria delle reti vi sono quelle di misura della centralità. La centralità è la caratteristica che definisce l'importanza di un ente all'interno della rete e la sua definizione non è univoca, in quanto dipendente dagli scopi per i quali ciascuna specifica misura di centralità è stata ideata. Per esempio, le centralità secondo il grado o secondo l'autovettore sono state introdotte in letteratura per studiare le gerarchie e le scale di potere nei sistemi sociali. Queste due metriche però non tengono conto della distanza topologica tra i nodi, di cui invece tengono conto le misure di centralità dette *closeness* e *betweenness*.

L'eterogeneità dei risultati generati da diverse misure di centralità è un problema ricorrente nella letteratura scientifica. Le misure di centralità vengono spesso utilizzate per stilare delle classifiche di importanza degli enti nel sistema, e l'uso di diverse metriche di centralità genera differenze a livello di risultati anche nella stessa applicazione. Questo è il caso dell'Economia Complessa, un campo in cui la teoria delle reti è usata per rappresentare i dati riguardanti il commercio internazionale e ricavare informazioni circa la capacità di innovazione delle nazioni.

Un altro esempio rilevante di applicazione di tale teoria è quello dello sviluppo sostenibile, campo in cui si è rivelata utile per svelare e comprendere le sinergie e i compromessi tra i vari settori che lo compongono. La teoria però non è mai stata sfruttata per stilare una classifica dei paesi per il loro status in sviluppo sostenibile.

Alla luce di tutto ciò, questa tesi ha l'obiettivo di contribuire alla definizione matematica delle metodologie di centralità in teoria delle reti, soprattutto per quanto riguarda le sue applicazioni nel campo dell'economia complessa e dello sviluppo sostenibile, più specificatamente nell'ambito dell'Agenda 2030 delle Nazioni Unite.

Il primo risultato di tale contributo consiste nell'introduzione di una prospettiva statistica sulle misure di centralità dei nodi in una rete. Tale prospettiva consente di confrontare diverse misure a seconda delle informazioni statistiche che queste forniscono sulla rete. Grazie a questa interpretazione statistica, si introducono indicatori multi-dimensionali di centralità come soluzione univoca al problema di definizione della centralità. Tale prospettiva costituisce anche il punto di partenza per l'armonizzazione delle misure di centralità oggi usate nel campo dell'economia complessa. Infatti, grazie alla definizione degli indicatori multi-dimensionali viene introdotto l'indice GENEPY, *GENeralized Economic comPlexity index*, un indicatore che unisce le misure di centralità ottenute dal *Metodo delle Riflessioni* e dall'algoritmo *Fitness and Complexity* e che univocamente definisce lo status di innovazione dei paesi. L'indice GENEPY, misurando l'acquisizione di capacità tramite le similitudini nelle tipologie di prodotti esportati dalle nazioni, è in grado di tracciare le traiettorie dei paesi lungo il loro percorso di crescita economica e di innovazione senza dover ricorrere ad informazioni esogene riguardanti l'economia delle nazioni (p.es., il prodotto interno lordo). L'ultimo contributo offerto da questo lavoro di tesi riguarda l'introduzione delle metriche di centralità nell'ambito dell'Agenda 2030 delle Nazioni Unite. In questa tesi vieni infatti presentata l'interpretazione dell'Agenda 2030 come un sistema complesso di paesi e degli Obiettivi dello Sviluppo Sostenibile (OSS). Questa interpretazione permette di utilizzare le metriche di centralità e del GENEPY al fine di studiare l'efficacia dei paesi nel loro percorso verso la sostenibilità.

Riassumendo, questo lavoro di tesi propone un insieme di metodologie per tracciare e indirizzare le performance globali di innovazione e sostenibilità.

Contents

1

Introduction

1.1 Real Systems & Network Science

"One of the key insights of the systems approach has been the realization that the network is a pattern that is common to all life. Wherever we see life, we see networks." Within his book *"The Hidden Connections: A Science for Sustainable Living"* [1], the physicist Fritjof Capra highlights how all natural systems, in a broader sense *life* he writes, self-organize as networks. In fact, a network is defined when entities in a system connect each other. Plants and pollinators in ecology [2], humans in sociology [3], products in trade [4] and pipes in water supply systems [5]: networking is the key for the evolution and development of systems [6]. The reasons why entities connect in a system usually answer some defined questions: are these plants pollinated by these particular bees? are these people friends? are these products traded by the same country? are these withdrawal points connected by pipelines? Although the answer to these questions may be a simple yes or no, the system ensemble may be complex, with increasing complexity according to increasing number of entities in the system [7]. (Complexity is the feature of systems to be characterized by non-trivial and non-random interactions among many entities [8]). This is why the mathematical network representation schematizing these interactions has found wide acceptance and application to understand the mechanisms of the system under study, helping in monitoring and managing it [6].

In mathematical terms, a network is defined as a graph, the entities are called nodes and the connections among them are the links (or edges) [9]. The modelling of systems as graphs has found wide acceptance and use in literature, starting with Euler's notable solution of the problem of the "Seven Bridges of Königsberg" in the last part of the eighteenth century [7, 9]. By means of network representation, Euler demonstrated that no path exists that crosses the seven bridges of the city of Königsberg without

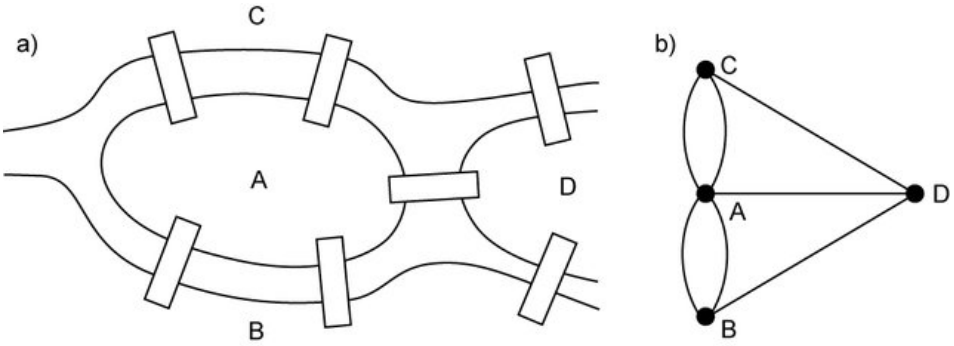

Figure 1.1: The Seven Bridges of Königsberg. Panel a) details the geography of Königsberb, where landmarks are labelled with letters and the seven bridges are represented by rectangles. Panel b) plots the network representation of the problem as proposed by Euler [7, 9]. Figure credits Boguslawski P. [10].

crossing one bridge twice. In his scheme, Fig 1.1, the nodes are the city's landmarks and the bridges the connections among them.

Since Euler's proof, network science became an important tool to model real systems in an easy way [6]. In fact, graphs are easily described by matrices, detailing which nodes are connected and how [9, 11]. The power of network science stands here: matrices are easier to handle than other mathematical tools [9]. We owe most of the basic knowledge about networks to sociology, where graphs were first introduced to study power and hierarchies [9, 12, 13], opinions and influences [14, 15]. As of today, network theory applications range from epidemiology [16–19] to economics [20–23], from sociology [24] to engineering [25, 26] and neuro-sciences [27, 28]. The World Wide Web is among the most relevant examples of complex networks nowadays: thousands of web pages are connected each other by hyperlinks, directing a possible surfer from the one page to the other [7]. Most recently, the experience we have lived with the *COVID-19* pandemic has clearly shown the (un)luckiness of being connected: the virus spread thanks to the transportation networks – airplanes, trains and cruising ships – from China to more than 100 countries [29–31]. The global transportation network was the luck of the virus and the bad luck of ours.

Although the theory has more than two centuries of history, network science presents some controversial methodologies as many more applications come by. In fact, following the growing number of problems framed within network science, many methodologies have been introduced in time, each of which has been tailored to the particular application they refer to. However, because of their use to define change-making actions, we argue that some

order in network science applications and tailored methodologies is essential. In particular, in this thesis we try to address the debate concerning with the metrics of network centrality, their theoretical formulations and applications to the field of economics and sustainable development. Therefore, neatness in the applications hence represents the *fil rouge* of this thesis, which we argue is required to empower the use of network science in addressing, as we show, global performances of innovation and sustainability.

Some introduction to basic notions of network science is due to the reader to follow along. Please note that in the common language, the words *graph* and *network* have become interchangeable.

Let G be a graph, with N nodes and E edges. G is mathematically described by the so-called *adjacency matrix* $\mathbf{A}$, whose ij-th element is 1 if node i and node j share an edge, zero otherwise. The matrix $\mathbf{A}$ has dimension $N \times N$, since the system has N entities interacting each other. If the network is *undirected*, i.e., the connections between the nodes are mutual, the adjacency matrix is symmetric with $A_{ij} = A_{ji}$ for $i \neq j$. This is the case of a social network: people know each other. An example of undirected network is given in Fig 1.2, which represents the marriage relations among the most notables Renaissance families in Florence [13]. The network has 15 nodes connected by 20 edges: marriage is, of course, mutual so that the connections among the families have no directions (Fig 1.2) and the adjacency matrix is symmetric, as shown in Table 1.1. Instead, the adjacency matrix is asymmetric if the network is *directed*: the edges are directional, with nodes pointing to others and $A_{ij} \neq A_{ji}$ for $i \neq j$. An example of this network is the World Wide Web: the hyperlinks are not mutual, so that if Google sends us to visit the webpage of Politecnico di Torino if searching for it, the landing web page does not link to Google.

Generally, the self-connections of the nodes (loops) are null, such that the diagonal values of the adjacency matrix, $A_{ii} = A_{jj}$, are null. Some systems, as the protein interaction network of yeast, naturally present self-loops as characteristic feature [7, 9]. Moreover, some systems may require the presence of weighted connections to be described, where the strength (or weight) of the connections among the nodes replaces the binary ij-th element of the adjacency matrix [9].

A different kind of network is obtained if we aim at representing the interactions between two different sets of nodes, instead. In this case, the links only exist between the sets, while no connections exist among the nodes belonging to the same group [9]. These kind of networks are called *bipartite networks* and they are described by the so-called *incidence matrix*

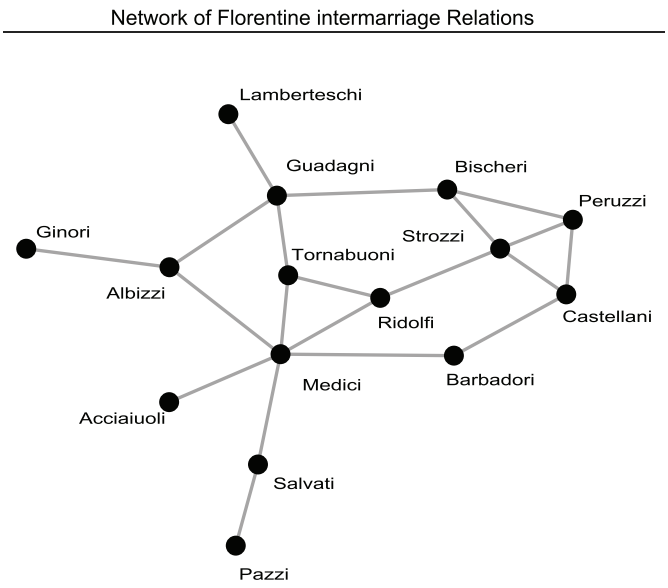

Figure 1.2: The network of the Florentine Intermarriage Relations [13]. The nodes represent the families of the Florentine Renaissance, connected by the presence of marriage contracts among the families. The adjacency matrix of this undirected network is given in Table 1.1.

B. The matrix is rectangular and has dimensions $U \times V$, being U the number of nodes in the first set, V in the second one, instead. Also in this case, the connections may be binary or weighted. A notorious example of a bipartite network in literature is the one from the *"Southern Women Study"* [32], describing the participation of women to social events in the late 30's and through which analyse racial segregation. Another example we will be dealing with in this thesis is the countries – products bipartite network, describing the export baskets of countries [4, 23, 33].

1.2 The Debate around Centrality Metrics

Within network science, a long-standing challenge is to rank the entities for their relevance in the system, i.e., for the centrality of the nodes in its network representation. In fact, centrality is referred to as a tool to quantify the importance of nodes in a network [9, 11]. To make a more realistic example concerning with the World Wide Web, centrality is the reason why when typing "Yesterday" in our web-search engine, among the top results there is a link to a video by The Beatles: there are so many relevant web

Table 1.1: The adjacency matrix of the Florentine Intermarriage Relations [13]. Entries are non-zeros if there was a marriage between the families. Family names are reported.

	Acciaiuoli	Albizzi	Barbadori	Bischeri	Castellani	Ginori	Guadagni	Lamberteschi	Medici	Pazzi	Peruzzi	Ridolfi	Salvati	Strozzi	Tornabuoni
Acciaiuoli	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
Albizzi	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0
Barbadori	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0
Bischeri	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0
Castellani	0	0	1	0	0	0	0	0	0	0	1	0	0	1	0
Ginori	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Guadagni	0	1	0	1	0	0	0	1	0	0	0	0	0	0	1
Lamberteschi	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Medici	1	1	1	0	0	0	0	0	0	0	0	1	1	0	1
Pazzi	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Peruzzi	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0
Ridolfi	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1
Salvati	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
Strozzi	0	0	0	1	1	0	0	0	0	0	1	1	0	0	0
Tornabuoni	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0

pages referring to this URL that made it rank top. A first definition of this property dates back to the 50's, when it was introduced to study the role of nodes in communication patterns [12, 34]. During the following years, progresses in social sciences provided several algorithms to evaluate nodes' centrality. These methods were typically obtained through case-specific considerations about the functioning of social networks, mainly based on reasoning about how information spreads across people in a group [12], and afterwards they were extended to other networks. Examples include the degree centrality [35, 36], Katz centrality [37], eigenvector centrality [38], betweeness [36, 39] and closeness centrality [36], PageRank [40], subgraph [41] and total communicability [42] centrality. Each metric defines node's centrality based on some topological features of the considered node, such as the number of its connections, the connections of its neighbours, the number of walks and paths going across the node, etc., thus testifying its relevance in the network, with possible risks of circular reasoning.

We now introduce three of the most used centrality metrics for undirected networks: the degree, eigenvector and Katz's centrality. Let x_i be a generic centrality metrics. The degree centrality for node i is defined as the sum of all nodes' connection, i.e.,

$$x_i = \sum_j^N A_{ij}. \tag{1.1}$$

The degree is usually referred to as k_i. The eigenvector centrality is defined as the weighted average of all nodes' centrality to which the node i is connected, namely

$$x_i = \alpha \sum_j^N A_{ij} x_j. \tag{1.2}$$

It can be shown that the constant value α is the largest eigenvalue λ_1 of the adjacency matrix and that, the vector $\mathbf{X}$ of the N x_i scores computed

according to Eq. (1.2) are the entries of the corresponding eigenvector [3, 9]. The Katz's centrality modifies the eigenvector centrality by adding a constant value in the equation to avoid divergence to zero of isolated (poorly connected) nodes [9]. This provides:

$$x_i = \alpha \sum_j^N A_{ij} x_j + \beta, \tag{1.3}$$

where $\alpha < 1/\lambda_1$ (λ_1 is the largest eigenvalue of $\mathbf{A}$) for ensuring convergence of the algorithm and the coefficient β is usually set to one [9]. Equivalent measures of the degree and eigenvector centrality exist for directed and bipartite networks, thus accounting for the directionality of the edges and the membership of the nodes, respectively. In particular, for directed networks, we consider i pointing to j, such that the outgoing edges of the node i are described onto the row i of the matrix $\mathbf{A}$. In this kind of networks, nodes are hence characterized by two properties, one concerning with the *outgoing* centrality of the node, x_i^{out}, and the other concerning with the *incoming* centrality, x_j^{in}. For the degree centrality it holds

$$k_i^{out} = \sum_j^N A_{ij}, \qquad k_j^{in} = \sum_i^N A_{ij}; \tag{1.4}$$

for the eigenvector centrality, instead

$$x_i^{out} = \frac{1}{\sqrt{\lambda_1}} \sum_j^N A_{ij} x_j^{in}, \qquad x_j^{in} = \frac{1}{\sqrt{\lambda_1}} \sum_i^N A_{ij} x_i^{out}. \tag{1.5}$$

In this case, the resulting vectors $\mathbf{X}^{out}$ and $\mathbf{X}^{in}$ are the eigenvectors of the matrices $\mathbf{A}\mathbf{A}^\mathbf{T}$ and $\mathbf{A}^\mathbf{T}\mathbf{A}$ associated to the largest eigenvalue λ_1, respectively, with $\lambda_1(\mathbf{A}\mathbf{A}^\mathbf{T}) = \lambda_1(\mathbf{A}^\mathbf{T}\mathbf{A})$ [43] (further details on this topic will be given in Chapter 2). The two eigen-centrality of the nodes are usually referred to as *hub-authority* centrality [9]. For bipartite networks the notations "out" and "in" are simply replaced by the membership to the set, U and V, with the sum term running over V and U, respectively. Notice that, in linear algebra, the eigen-system defined in Eqs. (1.5) also defines the Singular Value Decomposition (SVD) [43]. SVD solves the decomposition of any asymmetric or rectangular matrix: the singular values σ of any matrix $\mathbf{A}$, whether this is asymmetric or rectangular, relates with the eigenvalues λ of the matrices $\mathbf{A}\mathbf{A}^\mathbf{T}$ and $\mathbf{A}^\mathbf{T}\mathbf{A}$ according to the following relation

$$\sigma = \sqrt{\lambda},$$

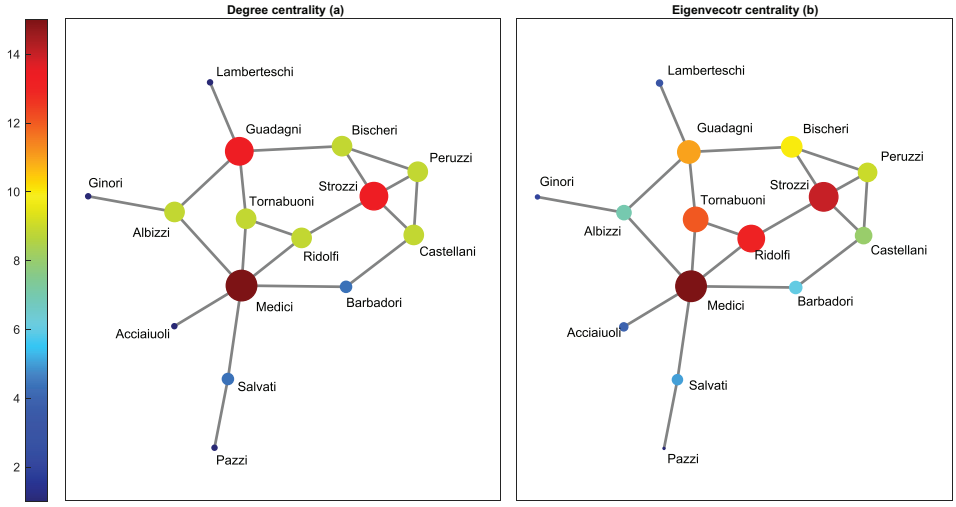

Figure 1.3: Degree vs eigenvector centrality. Scores of the degree and eigenvector centrality computed for the network of the Florentine Intermarriage Relations [13]. Panel (a) grades and sizes the network's nodes according to their degree. Panel (b) grades and sizes the nodes according to the eigenvector centrality, instead. In both panels, colors and sizes of the nodes are proportional to the scores assigned by each metrics, and the ranking is computed according to decreasing centrality values, so that the most central nodes occupies the first position.

and the corresponding right and left singular vectors are the eigenvectors associated to the λ values of $\mathbf{A}\mathbf{A}^{\mathbf{T}}$ and $\mathbf{A}^{\mathbf{T}}\mathbf{A}$, respectively (we refer the reader to [43] for further details). Therefore, for directed or bipartite networks to solve the eigen-centrality entails computing the right and left singular vectors of the adjacency or incidence matrix, respectively.

It is thus evident that different centrality metrics provide different answers to the question *"what does it mean to be central in a network ?"* (see, e.g., [44–46] for a literature review on centrality indexes and definitions). For example, one may assume a node is more central if it has many connections with other nodes, which leads to the degree centrality as the natural measure. However, one may argue that nodes are not all equivalent, and that a weighted version of the degree of the nodes should be adopted, where the weight is the centrality itself: this leads to the eigenvector centrality as the adequate metric. The differences in the outcomes of this two metrics are shown in Fig. 1.3 through the network of the Florentine Intermarriage Relations [13]. Although both metrics agree on ranking the Medici family as the most central, clearly there are some mismatching in the way other families, such as the Guadagni's and Ridolfi's ones, are ranked.

As explained, all these measures have a solid intuitive background.

Nevertheless, one is left without the possibility of comparing the reliability of different measures of centrality, and therefore, of choosing which is the most effective metric – and resulting node ranking – for the specific problem at hand. A possible solution lies in the introduction of a statistical tool to compare the information that each centrality metrics is able to retrieve about the network.

1.3 Centrality within Economics: The Economic Complexity Case

As we have described, centrality is a useful tool to gain information about the nodes in a system, specifically regarding the quantity and quality of connections. Centrality is key for understanding structural features of various systems and it has found application also in economics. In particular, the Economic Complexity (EC) field has introduced new centrality metrics to deal with the complex system of the international trade of goods, with the aim of ranking countries and produts. In fact, EC approaches structure export data in a bipartite network framework of countries and exports and the methodologies measure nodes' centrality within the system [47]. Within this field, the measures of centrality unveil the productive knowledge — standing for capabilities, finance, technology, human capital and resources – owned by countries and required to produce goods and for which economic growth and innovation is determined [4, 48, 49]. These data-based approaches mainly serve as an alternative to more traditional economic growth theories [50–55] which are often blamed for shrinking the intricacy of countries' socio-economic dynamics through simplistic assumptions [56, 57].

The easiest measure of economic complexity in this field is given by the simple degree of countries and products, defined as *diversity* and *ubiquity*, respectively [22, 48, 58]. However, the degree does not account for the compositions and complexity of the export baskets, i.e., for the set of capabilities required to produce and so, export, different kind of products. Although the degree is a necessary information, the currently used EC methodologies aim at improving this measure [48, 59].

The most common methodologies of economic complexity are the well-known *Method of Reflections* (MR) [22] and the *Fitness and Complexity* algorithm (FC) [23]. Both these methodologies ground their rationales on the assumption that only complex countries can export complex products, since they require a wider set of capabilities, such as human and capital

resources, to be produced and exported. Instead, less complex products are ubiquitous and can be found in all countries' baskets. Therefore, countries' complexity is assessed by means of the quality (rather than quantity) of products they export, i.e., of the products complexity, itself determined by the countries exporting them. In spite of their common root, MR and FC radically differ in the conceptual approach to the problem and, as a consequence, in the obtained outcomes. Clearly, these differences pose an issue of practical use of these methodologies and undermine their ability in assess economic and innovation potential. Instead, we argue that the combination of these two metrics would solve this issue, maintaining the advantages of both methods.

1.4 Centrality within the Agenda 2030: the Sustainable Development Goals Case

The United Nations' Agenda 2030 of sustainable development is a call for action to tackle the major challenges the world faces, such as environmental problems, climate change, economic growth, water, food and financial security, poverty and inequalities [60–65] (also recently exacerbated by the Sars-CoV-2 pandemic [66, 67]). In practical terms, the Agenda addresses a more equal, just and sustainable future by introducing the 17 Sustainable Development Goals - SDGs [60]. The 17 Goals are constructed upon 5 pillars: people, prosperity, planet, peace and justice, and partnership, and connections and spillover effects among the Goals are unavoidably present [68–77].

The ensemble of countries and Goals within the Agenda 2030 is a complex system of its own [78] (i.e., characterized by non-trivial and non-random interactions among many entities [8]), which require proper mathematical approaches to be analyzed. In fact, countries exhibit remarkable heterogeneity in the challenges they have to face, an issue which is crucial in global sustainable development [60, 66]. Moreover, the interconnections among SDGs, also define trade-offs and synergies within different sectors of development [69], which are enhanced by the strategies each country implements [79, 80]. These factors unavoidably create different responses at the country level [64, 65, 81, 82].

Indeed, the presence of interconnections among the Goals, and no less, of synergies and trade-offs among development sectors, can be unveiled thanks to the use of complex network theory (see, e.g., Le Blanc [79] and Guerrero *et al.* [75]). At the same time, within the development topic, the strategy

Figure 1.4: The 17 Sustainable Development Goals under the Agenda 2030 [68].

of indexing is often used to rank countries for their performances, thus making the creation of aggregated scores necessary [83] (notable examples are the Human Development Index [84] and the Multidimensional Poverty Index [85]), and the Agenda 2030 makes no exception. To create aggregated scores of performances entails mathematically valuing the contribution of each Goal to the overall countries' output, according to which compute a final score. In the construction of aggregated indices, many options can be pursued to weight these contributions [86–88]. A possible strategy would be to mathematically implement the egalitarian principle of the Agenda (i.e., all Goals must be of equal importance), thus entailing assigning the same weights to all SDGs (see, e.g., the SDG Index by Sachs *et al.*) [82, 89, 90]; nevertheless, other suitable strategies might exist (see, e.g., the Integrated Sustainable Development Index by Biggeri *et al.* [81]).

So far, the complex network analysis of the SDG system and the creation of aggregated scores have been treated in parallel, without relevant overlaps. Instead, we argue that the combination of data and network science may help in disentangling the dynamics of development and in defining data-driven weights for the creation of more objective and comprehensive aggregated scores. More precisely, we will show that aggregated scores can be obtained as the solution of a centrality exercise in the bipartite network representation

of countries and SDGs within the Agenda.

1.5 Outline of the Thesis

The thesis is organized in five chapters and their contents follow.

After a general introduction in Chapter 1, **Chapter 2** addresses the problem concerning with the use of centrality metrics, and the lack of one commonly accepted way to compare the effectiveness and reliability of different metrics. Here, we propose a new perspective where the definition of centrality metrics naturally arises from the most basic feature of a network, its adjacency matrix. In particular, we propose to tackle the centrality problem as a matrix-estimation exercise in which different centrality measures emerge as the result of the least squares estimation of the adjacency matrix. The results include the degree, eigenvector, and hub-authority centrality as natural solutions of the estimation problem at hand. Within this theoretical framework, the effectiveness of different metrics in reconstructing the matrix is evaluated and compared. Tests on a large set of networks show that the standard centrality metrics perform unsatisfactorily, highlighting intrinsic limitations for describing the centrality of nodes in complex networks. More informative *multi-component* centrality metrics are proposed as the natural extension of standard metrics.

Chapter 3 develops upon the results shown in Chapter 2 to reconcile the contrasting methodologies on economic complexity, namely the Method of Reflections and Fitness and Complexity algorithm. We recast the two approaches into a mathematically-sound, multidimensional framework, which allows us to recover and combine the strengths of both methods, still maintaining the relevant feature of providing countries' and products' rankings. The obtained results shed new light on the potential of economic complexity to trace and forecast countries' innovation potential and to interpret the temporal dynamics of economic growth, possibly paving the way to a micro-foundation of the field, in line with what was already proposed by Hausmann *et al.* [4].

Chapter 4 introduces the application of centrality metrics to the topic of sustainable development, and more precisely, to the Agenda 2030. We structure the data regarding countries' performances in the 17 Sustainable Development Goals as the incidence matrix describing a bipartite system of countries and Goals. This representation allows one to high-light and disentangle the intrinsic complexity of this system and to use the centrality metrics tools to obtain aggregated scores of sustainable development, hence

introducing bottom-up and data-driven weighting of the Goals. More importantly, our analysis allows one to take a data-driven picture of the possible current strategies of policy implementation in countries and unveils crucial features of their efficiency in sustainable development.

Chapter 5 presents the final remarks and take home messages of this thesis, highlighting the role of neatness in the methodologies of network theory for an easy track of global performances of innovation and sustainability. This final discussion also suggests future directions of work and applications, and limitations of the current studies.

As final contributions to the literature, in **Annexes** we add comments about the findings illustrated in Chapters 2 – 4. In Annex A we point out that, beyond its economic relevance, the international trade of commodities described in Chapter 3 has also an environmental impact on water resources. By overlapping the information about trade and the water required for food processing, we provide further food for thoughts about the trade-offs between economics and environment, trade-offs that arise also from the results of Chapter 4 concerning with sustainable development. Instead, in Annex B and C the complete ranking of countries in economic complexity (Chapter 3) and sustainable development (Chapter 4), respectively, are provided.

2

A Change of Perspective in Network Centrality

The work described in this chapter has been partially derived from Sciarra et al., Scientific Reports, *2018 [91].*

Centrality measures the importance of the nodes in a network and it plays a crucial role in several fields, ranging from sociology to engineering, and from biology to economics. Many centrality metrics are available. However, as described in Section 1.2, these measures are generally based on *ad hoc* assumptions, and there is no commonly accepted way to compare the effectiveness and reliability of different metrics. Focusing on providing a clearer answer to the question *"what does it mean to be central in a network?"*, we propose to tackle the centrality problem as a matrix-estimation exercise. In this framework, classical centrality measures (degree, eigenvector, Katz, hub-authority centrality) arise as the solution of a least square estimation. This allows to compare different centrality measures by evaluating their performances in terms of their capability to reproduce the network topology and, most importantly, to extend the notion of centrality to a multi-component setting. Although dimensions are added to the centrality metrics, our framework preserves the possibility to rank the nodes according to a scalar value. Our results, within the context of the still ongoing debate on the centrality metrics and the associated rankings (in several fields, see, e.g., [45, 46, 92–94]), provide further proofs that centrality metrics are highly correlated [95–100] and that they provide similar information about the importance of the nodes. Within this new framework, the natural multi-component extension of node centrality emerges as a possible solution to improve the quality of the estimations and, subsequently, of node ranking.

In this chapter, we first introduce the proposed perspective for undirected, unweighted networks, also introducing the multidimensional extension of centrality metrics. We then broad the perspective on directed, weighted and

bipartite networks. Particularly, our results are exemplified through the network of the Florentine Intermarriage Relations [13] and the friendship network of the Zachary's Karate Club members [101].

2.1 Undirected, Unweighted Networks

2.1.1 General Considerations

Let G be an undirected, unweighted graph, with N nodes and E edges. G is mathematically described by the symmetric adjacency matrix $\mathbf{A}$, whose ij-th element is 1 if i and j share an edge, zero otherwise [9]. Let $\hat{\mathbf{A}}$ be an estimator of the adjacency matrix. We expect a good estimator has larger $\hat{A}_{ij}$ values when i and j are connected (i.e., $A_{ij} = 1$), and lower values otherwise (i.e., when $A_{ij} = 0$). Our key idea is that the estimator of the generic element A_{ij} should depend on some emerging property x_i of the node i and x_j of the node j (with i, j=1:N) representing the topological importance of each node, i.e. its centrality. In formulas, $\hat{A}_{ij} = f(x_i, x_j)$ where f is an increasing function of both its arguments, since $\hat{A}_{ij}$ should increase when the nodes i and j are more "central" in the network. Due to the symmetry of the matrix $\mathbf{A}$, the arguments of f should also be exchangeable (i.e., $f(x_i, x_j) = f(x_j, x_i)$). Notice that the estimation process projects the information from N^2 to N as we are estimating a $N \times N$ matrix using the N values of nodes' centrality x_i. By definition, estimation is non exact, and $A_{ij} \neq \hat{A}_{ij}$. We suppose here that the error ϵ_{ij} related to the estimation is in additive form, namely

$$A_{ij} = \hat{A}_{ij} + \epsilon_{ij} = f(x_i, x_j) + \epsilon_{ij}. \tag{2.1}$$

Under this perspective, the centrality measures can be obtained on sound statistical bases, as they arise from the result of a standard estimation problem. Different constraints about the error structure can be considered. The most classical approach – least squares estimation – entails minimizing the sum of the squared errors, i.e., in this case,

$$SE(x_1, x_2, ..., x_N) = \sum_i^N \sum_j^N \epsilon_{ij}^2 = \sum_i \sum_j (A_{ij} - f(x_i, x_j))^2. \tag{2.2}$$

The minimization procedure entails taking the derivative of *SE* with respect to the considered variable (say, x_k), and equaling it to zero. *SE* can be partitioned into two components: a first part which is independent of x_k

(SE_0), and a second part depending on x_k (SE_k) i.e.,

$$SE = SE_0 + SE_k.$$

The derivative of SE with respect to the variable x_k, being SE_0 independent of x_k, corresponds to the derivative of SE_k. Notice that SE_k only depends on the k-th row and column of the two matrices $\mathbf{A}$ and $\hat{\mathbf{A}}$, namely

$$SE_k = \sum_{i\neq k}\Big(A_{ik} - f(x_i, x_k)\Big)^2 + \sum_{j\neq k}\Big(A_{kj} - f(x_k, x_j)\Big)^2 + \Big(A_{kk} - f(x_k, x_k)\Big)^2, \tag{2.3}$$

and the sums over the row and over the column coincide due to the symmetry of the matrix $\mathbf{A}$. By using Eq. (2.3), the derivative of the function SE with respect to the variable x_k is

$$\begin{aligned}\frac{\partial SE}{\partial x_k} = \frac{\partial SE_k}{\partial x_k} = 4\sum_{i\neq k}\Big[A_{ik} - f(x_i, x_k)\Big]\ \frac{\partial f(x_i, x_k)}{\partial x_k} + \\ 2\Big[A_{kk} - f(x_k, x_k)\Big]\ \frac{\partial f(x_k, x_k)}{\partial x_k} = 0.\end{aligned} \tag{2.4}$$

Let us introduce the bound variable z_m which allows one to formalize more concisely the mathematics behind Eq. (2.4). One has that

$$\frac{\partial f(x_i, x_k)}{\partial x_k} = \frac{\partial f(x_i, z_m)}{\partial z_m}\bigg|_{z_m = x_k} \qquad \text{if} \qquad i \neq k \tag{2.5}$$

and, if $i = k$,

$$\begin{aligned}\frac{\partial f(x_k, x_k)}{\partial x_k} &= \frac{\partial f(z_m, x_k)}{\partial z_m}\bigg|_{z_m = x_k} + \frac{\partial f(x_k, z_m)}{\partial z_m}\bigg|_{z_m = x_k} \\ &= 2\frac{\partial f(x_k, z_m)}{\partial z_m}\bigg|_{z_m = x_k}.\end{aligned} \tag{2.6}$$

In Eq. (2.6), the first equality in the second equation can be obtained by invoking the chain rule of derivation, i.e.,

$$\frac{\partial f(x, y(x))}{\partial x} = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\cdot\frac{\partial y}{\partial x},$$

and setting $y = x$ afterwards; instead, the second equality holds because, due to the exchangeability of the arguments of the function f, it holds the following equivalence

$$\frac{\partial f(x_k, z_m)}{\partial z_m} = \frac{\partial f(z_m, x_k)}{\partial z_m}.$$

In this way, by using Eq. (2.6), Eq. (2.4) becomes

$$\frac{\partial SE_k}{\partial x_k} = 4 \sum_i [A_{ik} - f(x_i, x_k)] \cdot \frac{\partial f(x_i, z_m)}{\partial z_m}\Big|_{z_m = x_k}, \tag{2.7}$$

where the index i runs over the range $[1, N]$. From Eq. (2.7), x_k is obtained imposing the equivalence of the derivative to zero. An equation equivalent to Eq. (2.7) is obtained for any centrality value x_i, $(i = 1, ..., N)$, which allows one to estimate the centrality value for all nodes. Notice that the framework can be extended to consider the error term in Eq. (2.1) in multiplicative form, and/or to consider a node-wise unbiased constraint instead of minimizing *SE*. However, in this work, we only consider the error term to be defined as in Eq. (2.1).

2.1.2 The Unique Contribution

Within this statistical framework, the answer to the question *"what does it mean to be central in a network ?"* is given through the analysis of the importance of the nodes in the estimation of A_{ij}: a node i is more central than a node j if the effect of its property x_i on the minimization of *SE* is larger than that of x_j. In a nutshell, if x_i is more "useful" than x_j for estimating $\mathbf{A}$. Put it another way, the node i is more important than the node j if, when removing its property from the estimation of A_{ij}, the change in *SE* recorded is higher than the one provoked by the exclusion of other nodes' property x_j. In order to account for this effect, we borrow the concept of the *unique contribution* from the theory of commonality analysis [102, 103]. The unique contribution is a quantitative measure of the effect a single variable has in the estimation procedure [104]. We define the unique contribution of a generic node k as the gain in the coefficient of determination R^2 induced by considering x_k in the estimation procedure. In formulas

$$UC_k = R^2_N - R^2_{N \setminus k} = \frac{SE_{N \setminus k} - SE_N}{TSS}, \tag{2.8}$$

where

$$R^2 = 1 - \frac{SE}{TSS},$$

with *SE* as in Eq. (2.2). The subscripts N and $N \setminus k$ in Eq. (2.8) refer to the case when all the N x_i values are considered in the estimation (subscript N), or to the case when the k-th property is excluded (subscript $N \setminus k$). In Eq. (2.8), *TSS* is the variance of the adjacency matrix, i.e.,

$$TSS = \sum_i \sum_j (A_{ij} - \bar{A})^2,$$

with $\bar{A}$ the mean of the matrix $\mathbf{A}$, namely

$$\bar{A} = \frac{\sum_i \sum_j A_{ij}}{N^2} = \frac{K_{tot}}{N^2}.$$

K_{tot} is the total degree of the network, i.e., the sum of all nodes' degrees in the network. Therefore it holds

$$\begin{aligned} TSS &= \sum_i \sum_j (A_{ij} - \bar{A}_{ij})^2 \\ &= \sum_i \sum_j A_{ij}^2 - 2\frac{K_{tot}}{N^2} \sum_i \sum_j A_{ij} + \frac{K_{tot}^2}{N^2}. \end{aligned}$$

Since the elements of the adjacency matrix are either 1 or 0, $A_{ij}^2 = A_{ij}$. This yields

$$TSS = K_{tot}\Big(1 - \frac{K_{tot}}{N^2}\Big). \tag{2.9}$$

Therefore, if the UC of node k is larger compared with the one obtained for node j, to exclude x_k from the estimation produces a larger drop in our capacity to estimate the adjacency matrix (i.e., a larger drop in R^2). As a consequence, the larger is UC_k, the most relevant (or central) the node is for reconstructing the adjacency matrix with a limited amount of information (i.e., the N centrality values). This allows one to perform a ranking of the network nodes for their capacity to contribute to the network estimation.

As obvious in Eq. (2.9), the term TSS does not change with the exclusion of x_k. Therefore, in order to evaluate the unique contribution Eq. (2.8), it is hence sufficient to compute the variation

$$\Delta SE = SE_{N\setminus k} - SE_N$$

in Eq. (2.8). According to the commonality analysis [104], the unique contribution should be computed eliminating the k-th node and repeating the estimation procedure with $(N-1)$ variables, in order to compute the determination coefficient $R^2_{N\setminus k}$. However, this approach would entail repeating the estimation for $(N+1)$ times, a potentially cumbersome effort in large networks. To bypass this difficulty, in this work we set a baseline scenario in which the k-th node is not formally excluded from the estimation, but the computation of the UC_k is performed setting to zero the centrality value x_k in the estimation procedure and keeping unchanged the other estimators x_i, $i \neq k$. This also allows one keeping the results in analytical form. As will be clear in the following, the assumption $x_k = 0$ corresponds to

assume a node with the lowest possible centrality value, since the centrality values are positive-valued. This assumption does not necessarily entail that the estimated link between node k and any other of its connection does not exist.

Under these conditions, we can focus our attention on the k-th row and column only. For a generic function $f(x_i, x_k)$, used to estimate A_{ij} under the condition of Eq. (2.1), ΔSE reads

$$\begin{aligned}\Delta SE =& 2\sum_{i\neq k}\Big[\Big(A_{ik} - f(x_i, 0)\Big)^2 - \Big(A_{ik} - f(x_i, x_k)\Big)^2\Big] + \\ & \Big(A_{kk} - f(0, 0)\Big)^2 - \Big(A_{kk} - f(x_k, x_k)\Big)^2,\end{aligned} \tag{2.10}$$

that can be expressed as

$$\begin{aligned}\Delta SE =& 2\sum_{i\neq k}\Big[f(x_i, 0)^2 - f(x_i, x_k)^2 - 2f(x_i, 0)A_{ik} + 2f(x_i, x_k)A_{ik}\Big] + \\ & f(0, 0)^2 - f(x_k, x_k)^2 - 2f(0, 0)A_{kk} + 2f(x_k, x_k)A_{kk},\end{aligned} \tag{2.11}$$

or, equivalently,

$$\begin{aligned}\Delta SE =& 2\sum_{i\neq k}\Big(f(x_i, 0) - f(x_i, x_k)\Big)\Big(f(x_i, 0) + f(x_i, x_k) - 2A_{ik}\Big) + \\ & \Big(f(0, 0) - f(x_k, x_k)\Big)\Big(f(0, 0) + f(x_k, x_k) - 2A_{kk}\Big).\end{aligned} \tag{2.12}$$

Within this work, we consider networks with no self-loops, hence in all formulas holds $A_{kk} = 0$.

Notice that the concept of centrality introduced with the unique contribution in Eq. (2.8) may resemble the definition of the "induced" centrality measures [105] deriving from graph invariants. However, the two approaches are different. The induced centrality is obtained from the contribution that a single node provides into the computation of a given graph invariant. In this framework instead, any node is ranked according to its contribution to the estimation of the adjacency matrix of the graph.

2.1.3 Examples of Estimator Functions

Different definitions of the function f in Eq. (2.1) allow one to obtain different centrality metrics. Table 2.1 summarizes the results concerning three different estimator functions and so, three centrality metrics, namely

the degree, eigenvector [38] and Katz centrality [37] metrics. Details follow on how these variables are obtained by adopting very simple link-estimation functions.

Please, notice that the formal resemblance between our $f(x_i, x_j)$ (see Tab 2.1) and the function used to attribute a probability of link activation based on the nodes' *fitness* [106, 107] is actually just a resemblance. In fact, the perspective is reversed here: differently from the link activation framework, which aims at generating a suitable network structure with a given node property distribution, we are estimating the nodes' properties that best represent a given adjacency matrix.

Degree Centrality

Let us start by considering the estimator f_1 for undirected networks,

$$\hat{A}_{ij} = f_1(x_i, x_k) = a \left[x_i + x_k - \frac{1}{N} \right]. \tag{2.13}$$

The derivative of the function f_1 with respect to x_k is

$$\left. \frac{\partial f_1(x_i, z_m)}{\partial z_m} \right|_{z_m = x_k} = a,$$

Applying Eq. (2.7) one obtains

$$4a \sum_i \Big[A_{ik} - a\Big(x_i + x_k - \frac{1}{N} \Big) \Big] = 0.$$

Since $\sum_i A_{ik} = k_k$ is the degree of the node k, solving the equation for x_k yields $x_k = \frac{k_k}{aN}$. Assuming the vector of centralities to have unitary 1-norm i.e., $\sum_i x_i = 1$, one obtains

$$a = \frac{K_{tot}}{N}, \tag{2.14}$$

finally yielding

$$x_k = \frac{k_k}{K_{tot}}. \tag{2.15}$$

Eq. (2.15) corresponds to rescaling the **degree centrality** by the total degree of the network.

Unique contribution

In order to compute the unique contribution UC_k, we need to re-consider the estimation function f in which x_k is set to zero. Considering the function f in Eq. (2.13), one has

$$f_1(x_i, 0) = ax_i - \frac{a}{N},$$

and

$$f_1(0, 0) = -\frac{a}{N}.$$

Using Eq. (2.12), this provides

$$\begin{aligned}\Delta SE =& 2\sum_{i\neq k}(-ax_k)\Big(2ax_i + ax_k - 2\frac{a}{N} - 2A_{ik}\Big) + (-2ax_k)\Big(2ax_k - 2\frac{a}{N}\Big) \\ =& -2ax_k\sum_i\Big(2ax_i + ax_k - 2\frac{a}{N} + 2A_{ik}\Big) + 2a^2x_k^2.\end{aligned}$$

Some further algebra provides

$$\Delta SE = -2a^2x_k^2N + 4ax_kk_k + 2a^2x_k^2.$$

Substituting the value of x_k as in Eq. (2.15) and $a = K_{tot}/N$ in Eq. (2.14), one obtains

$$\Delta SE = \frac{2(N+1)k_k^2}{N^2}$$

from which the unique contribution for the degree centrality is obtained,

$$UC_k = \frac{2(N+1)k_k^2}{N^2TSS}. \tag{2.16}$$

Since UC_k is a monotonic increasing function of k_k, ranking for increasing UC_k values provides the same ranking as the classical degree centrality.

Eigenvector Centrality

Consider the estimator for undirected network f_2 to be defined as

$$\hat{A}_{ik} = f_2(x_i, x_k) = \gamma x_ix_k. \tag{2.17}$$

The derivative of the function f_2 with respect to x_k is

$$\frac{\partial f_2(x_i, z_m)}{\partial z_m}\bigg|_{z_m=x_k} = \gamma x_i,$$

Applying Eq. (2.7) one obtains

$$4\sum_i \Big(A_{ik} - \gamma x_i x_k\Big)\gamma x_i = 0,$$

that solved for x_k provides

$$x_k = \frac{\sum_i A_{ik} x_i}{\gamma \sum_i x_i^2}.$$

We can assume the centrality vector to have unitary 2-norm (i.e., $\sum_i x_i^2 = 1$). This yields

$$x_k = \frac{1}{\gamma}\sum_i A_{ik} x_i. \tag{2.18}$$

Eq. (2.18) carries the same structure of the **eigenvector centrality** [3, 9], where $\gamma = \lambda_1$ is the largest eigenvalue of $\mathbf{A}$. It is worth to notice that the relation in Eq. (2.2), with the function Eq. (2.17), recalls one of the relations from which Bonacich demonstrates the eigenvector centrality [38]. However, the contexts of these two demonstrations are different; in fact, Bonacich used the Principal Factor Method, assuming $\mathbf{A}$ to be a special correlation matrix and $\mathbf{x}$ to be its first principal factor associated to the largest eigenvalue (see [15, 108] for further details).

Unique contribution

In order to compute the unique contribution UC_k, we need to re-consider the estimation function f in which x_k is set to zero. Defining f_2 as in Eq. (2.17), it holds

$$f_2(x_i, 0) = f_2(0, 0) = 0,$$

from which Eq. (2.11) becomes

$$\begin{aligned}\Delta SE &= 2\sum_{i\neq k}\Big[-\gamma^2 x_i^2 x_k^2 + 2\gamma x_i x_k A_{ik}\Big] - \gamma^2 x_k^4 + 2\gamma x_k^2 A_{kk} \\ &= 2\sum_{i}\Big[-\gamma^2 x_i^2 x_k^2 + 2\gamma x_i x_k A_{ik}\Big] + \gamma^2 x_k^4 - 2\gamma x_k^2 A_{kk}\end{aligned}$$

Since the 2-norm of the vector is unitary, using Eq. (2.18) it holds

$$\sum_i A_{ik} x_i = \gamma x_k,$$

from which, using the assumption $A_{kk} = 0$, one obtains

$$\Delta SE = 2\gamma^2 x_k^2 + \gamma^2 x_k^4.$$

Therefore, the unique contribution of the node k, according to the definition in Eq. (2.8), is given by

$$UC_k = \frac{\gamma x_k^2}{TSS}\left(\gamma x_k^2 + 2\gamma\right). \tag{2.19}$$

Since UC_k is a monotonic increasing function of x_k, ranking for increasing UC_k values provides the same ranking as the classical eigenvector centrality.

Katz Centrality

Consider the estimation function f_3 defined for undirected networks (see Table 1) as

$$\hat{A}_{ik} = f_3(x_i, x_k) = \gamma x_i x_k - B, \tag{2.20}$$

in which we assume the parameter B to be negative. The derivative of the function f_3 with respect to x_k, is

$$\left.\frac{\partial f_3(x_i, z_m)}{\partial x_k}\right|_{z_m = x_k} = \gamma x_i,$$

from which, according to Eq. (2.7), the derivative of the function *SE* is

$$4\sum_i \Big(A_{ik} - \gamma x_i x_k + B\Big)\gamma x_i = \tag{2.21}$$
$$\sum_i A_{ik} x_i - \gamma x_k \sum_i x_i^2 + B\sum_i x_i = 0.$$

Solved for x_k, the minimisation procedure provides

$$x_k = \frac{\sum_i A_{ik} x_i}{\gamma \sum_i x_i^2} + \frac{B\sum_i x_i}{\gamma \sum_i x_i^2}. \tag{2.22}$$

We now introduce the *attenuation factor* α of the Katz centrality [37] and define the equivalences

$$\frac{1}{\gamma \sum_i x_i^2} = \alpha, \qquad \frac{B\sum_i x_i}{\gamma \sum_i x_i^2} = \beta \tag{2.23}$$

obtaining

$$x_k = \alpha \sum_i A_{ik} x_i + \beta. \tag{2.24}$$

Eq. (2.24) corresponds to the definition of the Katz centrality measure [37], in which α is the attenuation factor whose value is $\alpha < 1/\lambda_1$, being λ_1 the

largest eigenvalue of $\mathbf{A}$ and β a constant, whose value is usually set to one [9]. Due to the constraint imposed by the form of the Katz centrality, the N x_i values are always positive and greater than one; hence no assumptions can be made on the norms of the vector $\mathbf{x} = [x_1, ..., x_N]$.

Unique contribution

Again, in order to compute the unique contribution UC_k, we need to re-consider the estimation function f in which x_k is set to zero. Using the function f_3 in Eq. (2.20), one has

$$f_3(x_i, 0) = f_3(0, 0) = -B.$$

Using the form of ΔSE as given in Eq. (2.12) and substituting the values of the functions

$$f_3(x_i, 0) - f_3(x_i, x_k) = -\gamma x_i x_k, \qquad f_3(0, 0) - f_3(x_k, x_k) = -\gamma x_k^2,$$

one obtains

$$\begin{aligned}\Delta SE =& 2\sum_{i\neq k}(-\gamma x_i x_k)(\gamma x_i x_k - 2B - 2A_{ik}) - \gamma x_k^2(\gamma x_k^2 - 2B - 2A_{kk}) \\ =& 2\sum_{i}(-\gamma^2 x_i^2 x_k^2 + 2\gamma B x_i x_k + 2\gamma x_i x_k A_{ik}) + \gamma^2 x_k^4 - 2\gamma B x_k^2,\end{aligned}$$

where the assumption $A_{kk} = 0$ is used. Using the equivalences in Eq. (2.23), and the one deriving from Eq. (2.24), it holds

$$\sum_i A_{ik} x_i = \frac{x_k}{\alpha} - \frac{\beta}{\alpha},$$

from which

$$\begin{aligned}\Delta SE =& -2\gamma^2 x_k^2 \frac{1}{\alpha\gamma} + 4\gamma B x_k \frac{\beta}{\alpha B} + 4\gamma x_k \Big(\frac{x_k}{\alpha} - \frac{\beta}{\alpha}\Big) + \gamma^2 x_k^4 - 2B\gamma x_k^2 \\ =& \; 2\gamma \frac{x_k^2}{\alpha} + \gamma^2 x_k^4 - 2B\gamma x_k^2.\end{aligned}$$

The unique contribution of the node k, according to the definition in Eq. (2.8), is

$$UC_k = \frac{\gamma x_k^2}{TSS}\left(\gamma x_k^2 - 2B + \frac{2}{\alpha}\right). \tag{2.25}$$

Since we have defined B to be a negative value, while γ and α are positive ones, UC_k is a monotonic increasing function of x_k and ranking for increasing UC_k values provides the same ranking as the classical Katz centrality.

Table 2.1: Examples of the one-dimensional estimator functions f to be set in Eq. (2.1) to obtain some commonly-used centrality measures. The unique contribution, which is here used to rank nodes for their centrality, is also reported. In the formulas, $K_{tot} = \sum_i \sum_j A_{ij}$ is the total degree of the network; N is the number of nodes; $k_i = \sum_j A_{ij}$ is the degree of the node i; γ and B are two parameters whose values change according to the estimator function. In case of f_2, γ equals the largest eigenvalue of $\mathbf{A}$. In case of f_3, $\gamma = 1/\alpha \sum_j x_j^2$ and $B = -1/\sum_j x_j$, where α is the *attenuation factor* of the Katz centrality. *TSS* is defined in the text. Further details are given in Section 2.1.3.

Undirected networks			
Estimator function f	**Centrality of node i**	**Unique contribution of node i**	**Corresponding metrics**
$f_1 = \frac{K_{tot}}{N}\ \left(x_i + x_j - \frac{1}{N}\right)$	$x_i = \frac{k_i}{K_{tot}}$	$UC_i = \frac{2(N+1)k_i^2}{N^2\ TSS}$	**Degree centrality**
$f_2 = \gamma x_i x_j$	$x_i = \frac{1}{\gamma}\sum_j A_{ij}x_j$	$UC_i = \frac{\gamma x_i^2}{TSS}\left(\gamma x_i^2 + 2\gamma\right)$	**Eigenvector centrality**
$f_3 = \gamma x_i x_j + B$	$x_i = \frac{\sum_j A_{ij}x_j}{\gamma\sum_j x_j^2} + \frac{B\sum_j x_j}{\gamma\sum_j x_j^2}$	$UC_i = \frac{\gamma x_i^2}{TSS}\left(\gamma x_i^2 - 2B + 2\gamma\sum_j x_j^2\right)$	**Katz centrality**

Multicomponent Estimator and Centrality for Undirected, Unweighted Networks

A natural extension of the *one-component* estimators (resumed in Table 2.1) is to move toward more informative *multi-component* metrics of nodes' centrality. The multi-component centrality considers more facets of the networks, by describing the role of network's nodes through more than one scalar property.

We consider the function f_2, Eq. (2.17), as the starting point for our reasoning. A possible design of the multidimensional estimator is

$$\begin{aligned}\hat{A}_{ij}(s) = f(\mathbf{x}_i, \mathbf{x}_j) &= \gamma_1 x_{i,1} x_{j,1} + ... + \gamma_k x_{i,t} x_{j,t} + ... + \gamma_s x_{i,s} x_{j,s} \\ &= \sum_{t=1}^{s} \gamma_t x_{i,t} x_{j,t}\end{aligned} \tag{2.26}$$

where $\mathbf{x}_i = [x_{i,1}, ..., x_{i,s}]$ (and so is defined $\mathbf{x}_j$) is an s-dimensional vector embedding the s properties of the node i (j) that should be considered for evaluating its importance (for $s = 1$ the one-component metrics are recovered). Using Eq. (2.26), the estimation process projects N^2 data (i.e., the number of entries of the adjacency matrix) to $s \cdot N$, which is the number of independent variables used in the estimation.

The reason why we consider f_2, Eq. (2.17), for the extension is that the additive form in which f_1, Eq. (2.13), is framed does not allow for an increment of information, since the contribution carried by different variables $(x_{i,1}, .., x_{i,s})$ cancels out if one refers to a single variable, ξ_i, which is a linear combination of the different components. In other words, the components beyond the first one cannot bring any additional information into the estimation exercise in additive form. Instead, an extension of f_3, Eq. (2.20), would simply imply to add a constant value to Eq. (2.26), with no added valued information to the estimation.

Steps on how multidimensional centrality metrics are obtained follow.

Let us assume the vector describing $\mathbf{x}_t = [x_{1,t}, ..., x_{N,t}]$, which describes the t node's properties, to have unitary 2-norm, such that

$$\sum_i x_{i,t}^2 = 1.$$

Let there be an orthogonality condition between any two vectors $\mathbf{x}_t$ and $\mathbf{x}_{t^*}$, such that

$$\sum_i x_{i,t} \cdot x_{i,t^*} = 0, \quad \forall\ t \neq t^*. \tag{2.27}$$

The steps described for the one-component centrality can be adapted to the multidimensional setting (see Section 2.1). We compute the multidimensional metrics by considering the contribution to *SE* of a generic variable x_{k,t^*}, i.e., of the t^* properties of node k. As also defined for the one-component centrality, Section 2.1, *SE* is partitioned into a part SE_0, which does not depend on x_{k,t^*}, and a part SE_{k,t^*}, which is a function of x_{k,t^*},

$$SE = SE_{0,t} + SE_{k,t^*}. \tag{2.28}$$

The computation of the centrality values by minimisation of the $SE_{k,t*}$ entails computing Eq. (2.7) accounting for each dimension considered, i.e., $t = 1, ..., s$. The derivative of *SE* has the same form as Eq. (2.7). Using the bound variable $\mathbf{z_m}$ (which, in this case, is a vector), the derivative of the function is

$$\left.\frac{\partial f(\mathbf{x}_i, \mathbf{z}_m)}{\partial z_{k,t^*}}\right|_{\mathbf{z}_m = \mathbf{x}_k} = \gamma_{t^*} x_{i,t^*}.$$

Therefore, in this case, the minimisation of $SE_{k,t*}$ in Eq. (2.7) reads

$$4 \sum_i \Big[A_{ik} - \sum_t \gamma_t x_{i,t} x_{k,t} \Big] \gamma_{t^*} x_{i,t^*} = 0,$$

that is equivalent to

$$\sum_i A_{ik} x_{i,t^*} - \sum_t \gamma_t x_{k,t} \sum_i x_{i,t} \cdot x_{i,t^*} = 0.$$

Due to the orthonormality condition set in Eq. (2.27), it holds

$$\begin{aligned} \sum_t \gamma_t x_{k,t} \sum_i x_{i,t} \cdot x_{i,t^*} &= \gamma_{t^*} x_{k,t^*} \sum_i x_{i,t^*} \cdot x_{i,t^*} \\ &= \gamma_{t^*} x_{k,t^*} \sum_i x_{i,t^*}^2 = \gamma_{t^*} x_{k,t^*}. \end{aligned}$$

Finally, for any component t, the centrality value reads

$$x_{k,t} = \frac{1}{\gamma_t} \sum_i A_{ik} x_{i,t}, \tag{2.29}$$

which corresponds to computing the eigenvector $\mathbf{x}_t$ corresponding to the eigenvalue $\gamma_t = \lambda_t$.

The formal structure of $\hat{\mathbf{A}}$ in Eq. (2.26) corresponds to the *s-order low-rank approximation* of the matrix $\mathbf{A}$ [43]. In fact, under a least squares constraint, and the assumption of orthogonality between the s vectors $\mathbf{x}_t$,

one obtains that $\gamma_t = \lambda_t$ is the t-th eigenvalue of the adjacency matrix $\mathbf{A}$ and $\mathbf{x}_t = [x_{1,t}, ..., x_{N,t}]$ is its corresponding eigenvector. In Eq. (2.26), the eigenvalues γ_t, and hence their corresponding eigenvectors $\mathbf{x}_t$, can be ordered according to their absolute value, from the greatest to the least one, and eigenvectors of increasing order bring a monotonically decreasing amount of information [109]. This solution corresponds to the *Singular Value Decomposition* (SVD) for symmetric matrices [43], being $\hat{\mathbf{A}}(s)$ the *s-order low-rank approximation* of the original adjacency matrix $\mathbf{A}$ (see Chapter 1, Section 1.2). The *Eckhart-Young-Mirsky theorem* [109] proofs that the total amount of explained variance *VE* of the *s-order low-rank approximation* equals the sum of the squares of the s γ_t eigenvalues, when the approximation is truncated at s, namely

$$VE(s) = \sum_{t=1}^{s} \gamma_t^2. \tag{2.30}$$

Different strategies can be pursued to choose a proper value of s (see, e.g., [110] for a review of the criteria). In fact, the choice of the s value entails finding a good balance between the necessity to accurately describe the adjacency matrix and the willingness to have a parsimonious representation of a complex system. Different strategies can be pursued, also borrowing from the wide literature pertaining with the similar problem of deciding where to arrest the eigenvalue decomposition or SVD (see, e.g., [110] for a review). For example, one may choose the s value corresponding to the first gap in the eigenspectrum of the adjacency matrix (see, e.g., [111]). Alternatively, one may stop the expansion in Eq. (2.26) when the explained variance reaches a predefined amount of the total variance of $\mathbf{A}$. This would entail that the remaining amount of variance is attributed to noise. For a given number of components s, at each component t^* added to the estimation, the total amount of explained variance increases by $\gamma_{t^*}^2$. Hence it holds

$$VE(t^*) - VE(t^* - 1) = \gamma_{t^*}^2 \tag{2.31}$$

The total amount of unexplained variance *VU* is

$$\begin{aligned} VU(t^*) &= \sum_i \sum_j \Big(A_{ij} - \hat{A}_{ij}(t^*)\Big)^2 = TSS - VE(t^*) \\ &= TSS - \sum_{t=1}^{t^*} \gamma_{t^*}^2, \end{aligned} \tag{2.32}$$

with *TSS* as in Eq. (2.9).

The ordering of the eigenvalues, however, requires some additional considerations. In fact, Eq. (2.30) ensures that the explained variance with s components is maximised by taking the first s eigenvalues, ordered in absolute values from the largest to the smallest. However, a consistency issue emerges when considering networks with no self loops. For these networks, the elements on the diagonal of $\mathbf{A}$ are zero. The estimated matrix has instead its diagonal elements different from zero, namely

$$\hat{A}_{ii}(s) = \sum_{t=1}^{s} \gamma_t x_{i,t}^2. \tag{2.33}$$

This entails that, in order to provide a good description of the system, the eigenvalues should be ordered according to the total amount of explained variance they bring *off-diagonal*. In fact, Eq. (2.30) can be partitioned in two terms, the one pertaining with the diagonal D and the other with the off-diagonal OD terms, i.e.,

$$VE(s) = VE(s)_D + VE(s)_{OD}. \tag{2.34}$$

We are therefore interested in ordering the eigenvalues so that the value $VE(t^*)_{OD}$ is maximised at each new added component t^*. This allows one to provide a more accurate representation of the system without the need of adding eigenvectors that erroneously approximate the diagonal values.

Consider the term $VE(t^*)_D$. Using Eq. (2.32) and Eq. (2.33), this term reads

$$\begin{aligned} VE(t^*)_D &= TSS - \sum_i \Big(A_{ii} - \hat{A}_{ii}(t^*)\Big)^2 = TSS - \sum_i \Big(\hat{A}_{ii}(t^*)\Big)^2 \\ &= TSS - \sum_i \Big(\sum_{t=1}^{t^*} \gamma_t x_{i,t}^2\Big)^2 \\ &= TSS - \sum_i \Big(\sum_{t=1}^{t^*-1} \gamma_t x_{i,t}^2 + \gamma_{t^*} x_{i,t^*}^2\Big)^2 \\ &= TSS - \sum_i \Big(\sum_{t=1}^{t^*-1} \gamma_t x_{i,t}^2\Big)^2 - \gamma_{t^*}^2 \sum_i x_{i,t^*}^2 - \sum_i 2\gamma_{t^*} x_{i,t^*}^2 \Big(\sum_{t=1}^{t^*-1} \gamma_t x_{i,t}^2\Big). \end{aligned} \tag{2.35}$$

Eq. (2.35) entails that at each new component $t = t^*$ added to the estimation, the increment in the total amount of explained variance on the diagonal $\Delta VE(t^*)_D = VE(t^*)_D - VE(t^* - 1)_D$ equals

$$\Delta VE(t^*)_D = -\gamma_{t^*}^2 \sum_i x_{i,t^*}^4 - 2 \sum_i \gamma_{t^*} x_{i,t^*}^2 \sum_{t=1}^{t^*-1} \gamma_t x_{i,t}^2. \tag{2.36}$$

Considering that the total amount of explained variance by the t^* component is $\gamma_{t^*}^2$ (see Eq. (2.31)), from Eq. (2.34) and Eq. (2.36), one obtains

$$\Delta VE(t^*)_{OD} = \gamma_{t^*}^2 \Big(1 + \sum_i x_{i,t^*}^4\Big) + 2 \sum_i \gamma_{t^*} x_{i,t^*} \sum_{t=1}^{t^*-1} \gamma_t x_{i,t}^2. \tag{2.37}$$

Aiming at choosing the order in which the eigenvalues, and respective eigenvectors, should be embedded into the estimation Eq. (2.26), one should maximise, at each step, the function in Eq. (2.37). For $t = 1$ – i.e., for choosing the first eigenvalue and respective eigenvector – the function to be maximised is

$$\Delta VE(t^* = 1)_{OD} = \gamma_t^2 \Big(1 + \sum_i x_{i,t}^4\Big).$$

When $t = 2$, the second eigenvalue to be embedded into the function Eq. (2.26) is the one that maximises the function

$$\Delta VE(t^* = 2)_{OD} = \gamma_{t=1}^2 \Big(1 + \sum_i x_{i,t=1}^4\Big) + 2\gamma_{t=1}\gamma_t \sum_i x_{i,t}^2 x_{i,t=1}^2.$$

In this work, aiming at improving the reconstruction of the network topology, the eigenvectors have been sorted following the just described procedure and the dimension s has been set to 2.

Unique contribution

In the multi-component setting, the unique contribution of the k-th node, and hence its centrality value, is found accounting for all components $\mathbf{x}_t$, $t = (1, ..., s)$ for each node k. In this case, to exclude the generic node k from the estimation procedure corresponds to set to zero all of its properties $x_{k,t}$, i.e., $x_{k,t} = 0$, for $t = 1 : s$. This yields

$$f(\mathbf{x}_i, 0) = f(0, 0) = 0.$$

Within this multi-component setting, Eq. (2.12) becomes

$$\begin{aligned}
\Delta SE = {} & 2 \sum_{i \neq k} \Big(- \sum_{t=1}^{s} \gamma_t x_{i,t} x_{k,t} \Big) \Big(\sum_{t=1}^{s} \gamma_t x_{i,t} x_{k,t} - 2A_{ik} \Big) - \\
& \Big(\sum_{t=1}^{s} \gamma_t x_{k,t}^2 \Big) \Big(\sum_{t=1}^{s} \gamma_t x_{k,t}^2 - 2A_{kk} \Big) \\
= {} & 2 \sum_i \Big[- \Big(\sum_{t=1}^{s} \gamma_t x_{i,t} x_{k,t} \Big)^2 + 2A_{ik} \sum_{t=1}^{s} \gamma_t x_{i,t} x_{k,t} \Big] + \Big(\sum_{t=1}^{s} \gamma_t x_{k,t}^2 \Big)^2,
\end{aligned}$$

that is equivalent to

$$\Delta SE = -2\sum_{t=1}^{s}\gamma_t^2 x_{k,t}^2 \sum_i x_{i,t}^2 + 4\sum_{t=1}^{s}\gamma_x x_{k,t}\sum_i A_{ik}x_{i,t} + \Big(\sum_{t=1}^{s}\gamma_t x_{k,t}^2\Big)^2$$

Using the orthonormality condition in Eq. (2.27) and Eq. (2.29), the unique contribution of node k in the case of the multi-component estimator is given by

$$UC_k(s) = \frac{1}{TSS}\left[\left(\sum_{t=1}^{s}\gamma_t x_{k,t}^2\right)^2 + 2\sum_{t=1}^{s}\gamma_t^2 x_{k,t}^2\right]. \tag{2.38}$$

The $x_{k,t}$ values in Eq. (2.38) appear in squared form. As a consequence, the sign of $x_{k,t}$ does not affect the UC_k value.

It is clear that, by considering additional dimensions beyond the first, the node centrality ranking may significantly change, revealing node features which were hidden by the one-dimensional assumption. In fact, information on the structure and clustering of the network is contained in the eigenvectors beyond the first one (for more information see, e.g., [3, 111, 112]). In the case $s = N$, through UC one recovers the same ranking given by the degree centrality. In fact, in this case the approximated matrix equals the adjacency matrix, i.e., $\hat{\mathbf{A}} = \mathbf{A}$ and the errors are zero. In contrast, since the k-th row and column of $\hat{\mathbf{A}}$ are zero when excluding the k-th node from the estimation, $R^2_{N\setminus k}$ turns out to be proportional to the squared degree of node k, k_k^2. Therefore, when considered under the perspective of the unique contribution, the expansion with $s = N$ copies the same information of the node degree, in terms of the obtained nodes' ranking. It may be useful to note that the multi-component estimation of centrality, and the subsequent ranking given through UC, entail a two-steps shrinkage of information. Firstly, the estimation projects data from N^2 to $s \cdot N$, and secondly the ranking projects from $s \cdot N$ to N. Therefore, the multi-component centrality acts as an additional pier for the bridge from N^2 to N, a pier which can be essential to pose the centrality estimation problem on more solid grounds. Clearly, both cases $s = 1$ and $s = N$ correspond to limit situations when the additional pier is not in between N^2 and N, but it is on one of the two sides; in fact, in these situations one recovers the eigenvector centrality ($s = 1$) and the degree centrality ($s = N$).

2.1.4 Results on Undirected, Unweighted Networks

To recast the centrality metrics into this new framework allows us to compare their performances, in terms of their ability to approximate the

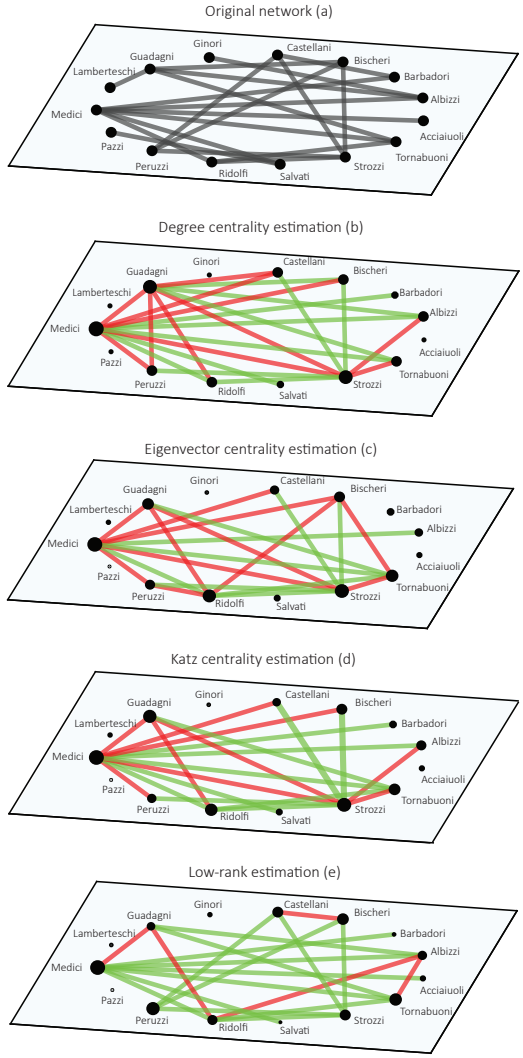

Figure 2.1: Estimation results for the undirected network of Florentine Intermarriage Relations, represented in panel (a). Panels (b) - (d) refer to the topology estimated by the degree, eigenvector, and Katz centrality, respectively. Panel (e) shows the estimated network as given by the multi-component estimator with two components ($s = 2$). In the figure, correctly estimated links are highlighted in green, while spurious links are red colored. Nodes' size in panels (b) - (e) is proportional to the position in the ranking resulting from the unique contribution, ordering the list from least to most central node. We plot in Figure 2.1 only the E larger values of $\hat{A}_{ij}$, thus preserving in all the reconstructed networks the number E of edges of the real network. Exception is made when the E-th larger value of $\hat{\mathbf{A}}$ is a tie, in which case more than E edges are plotted. Rankings are available in Table 2.2.

adjacency matrix. We illustrate our new perspective starting in Figure 2.1 with an analysis of the network of the Florentine Intermarriage Relations [13], already introduced in Chapter 1. Within our framework, the centrality measures have a counterpart in a link-estimation function, which allows one to perform a visual and numerical comparison with the original network. In Figure 2.1, we plot the original network (panel (a)) , and those resulting from the use of the one-component centrality measures in (panels (b) - (d)). The centrality-based estimations are performed using the functions reported in Table 1.1. For the computation of the network estimation based on the Katz centrality, function f_3, we used $\alpha = 0.5/\lambda_1$ following the directions by Benzi *et al.* [95], being λ_1 the principal eigenvalue of $\mathbf{A}$, and $\beta = 1$; the values of γ and B to be used in Eq. (2.20) are straightforward from Eq. (2.23). The network representation in panel (e) shows the result of the estimation provided by the multi-component estimation with $s = 2$. Figure 2.1 highlights the low agreement between the one-dimensional modelled networks and the real one. Several spurious and lacking links appear in the reconstructed graphs. The network representation is significantly improved when using the multi-component estimator ($s = 2$) in panel (e).

Besides the visual inspection, we compute the adjusted coefficient of determination R_a^2 between the original and the estimated matrices, $\mathbf{A}$ and $\hat{\mathbf{A}}$, in order to measure the quality of the estimation. R_a^2 is defined as

$$R_a^2 = 1 - (1 - R^2)\frac{N^2}{N^2 - s \cdot N} = 1 - (1 - R^2)\frac{N}{N - s}. \tag{2.39}$$

The choice of R_a^2 as an error metric is consistent with the concept of unique contribution (see Eq. (2.8)). Moreover, this error measure is applicable to binary variables as well and the "adjusted" version of R^2 allows one to compare the results obtained from distinct estimators and on differently sized networks. Notice that, while using R_a^2 instead of R^2 is formally correct, the term $N/(N - s)$ in Eq. (2.39) rapidly converges to 1 in large networks, making this correction negligible in some practical applications. For the Florentine Intermarriage Relations network, the adjusted determination coefficient for the multi-component estimator is $R_a^2 = 0.30$, while for the other estimators is roughly $R_a^2 = 0.07$, confirming the outcomes of the visual inspection.

The three classical centrality metrics (degree, eigenvector, Katz) produce different rankings of the Florentine families (see Figures 1.3 and 2.1, Table 2.2). While the Medici is always the top-ranked family, other families significantly change their position in the rankings (e.g., the ranking of the Ridolfi family changes from 3 to 8 when different methods are considered,

Table 2.2: Rankings of the Florentine Renaissance Families in the network of intermarriage relations. Rankings are the results of the unique contribution of the centrality based estimation degree, eigenvector, Katz and multi-component one.

Rankings of the Florentine Renaissance Families				
Families	Degree centrality	Eigenvector centrality	Katz centrality	Multi-component centrality ($s = 2$)
Acciaiuoli	13.5	12	12	12
Albizzi	6.5	9	7	8
Barbadori	10.5	10	10	10
Bischeri	6.5	6	6	5
Castellani	6.5	8	9	6
Ginori	13.5	14	14	13
Guadagni	2.5	5	3	9
Lamberteschi	13.5	13	13	14
Medici	1	1	1	1
Pazzi	13.5	15	15	15
Peruzzi	6.5	7	8	3
Ridolfi	6.5	3	4	7
Salvati	10.5	11	11	11
Strozzi	2.5	2	2	2
Tornabuoni	6.5	4	5	4

see Table 2.2). By embracing our new perspective on network centrality it is possible to compare these rankings claiming that, despite the differences, from a statistical point of view the three metrics bring the same information about the topology of the network. The need to extend the centrality concept toward multiple dimensions manifestly emerges from Figure 2.2. The figure plots the contours of the unique contribution as a function of the first two principle eigenvectors, i.e., associated to the two largest eigenvalues of the adjacency matrix. The second eigenvector distinctly identifies the group constituted by the families Strozzi-Peruzzi-Castellani-Bischeri, while highlighting how the Medici family is left alone by these four families. In this case the information brought by the second eigenvector is clearly relevant in determining the ranking of the nodes. In fact, the ranking in the case of Figure 2.2 corresponds to the distance from the axes-origin. If one had considered only the first eigenvector, the *Ridolfi* family would have been ranked in the third position. The additional information carried by the second eigenvector, combined through the unique contribution, downgrades the Ridolfi family to the seventh position, instead.

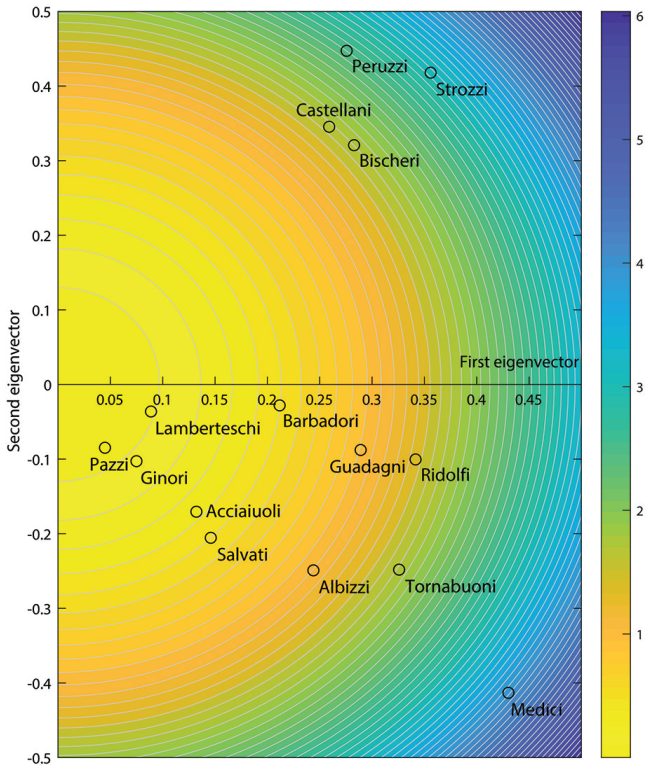

Figure 2.2: Contour plot of the unique contribution resulting from the application of Eq. (2.38) with $s = 2$. The contours range from lower values of unique contribution (in yellow) to larger values (in blue). The $x_{i,1}$ values (corresponding to the components of the first eigenvector) are on the x-axis, while the values of $x_{i,2}$ (related to the components of the eigenvector corresponding to the second eigenvalue, ordered following the method described in Section 2.1.3, are on the y-axis. The open circles correspond to the $x_{i,1}$ and $x_{i,2}$ values for the Florentine Intermarriage Relations network. Nodes with larger unique contribution are found further away from the origin.

Another example we here provide to support the need for multi-dimensional centrality metrics concerns with the Zachary's Karate Club network [101]. The network has 34 nodes representing the club members that, at the time, had interactions outside the context of the club (78 edges), as shown in Figure 2.3, panel (a). The network is a notorious case study in community detection literature [9]. In fact, the fission the club faced after a discussion between the instructor and the president was very well predicted by the

mathematical model in Zachary's work [101]. In Figure 2.3, the nodes 1 and 34 represent the instructor and the president of the club, respectively. We apply the multi-component centrality metric to the network, evaluating the unique contribution of the nodes as in Eq. (2.38), with $s = 2$. In Figure 2.3, the panel (b) plots the contours of the unique contribution as a function of the first two eigenvectors of the adjacency matrix. The figure clearly shows how the second eigenvector is able to identify the fission into the two groups, assigning positive or negative values to the nodes according to the group they belong to. This information is clearly relevant in determining the ranking of the nodes. In fact, if one compares the ranking obtained from the eigenvector centrality only (namely, for $s = 1$), the positions of the nodes change among the charts. While the instructor – node 1 – and the president – node 34 – are stably ranked on the very first two positions, other nodes' positions significantly change. For instance, if only the first eigenvector is considered, the node 9 is positioned at the 6-th place in the ranking, while the information carried by the second eigenvector downgrades the node to the 9-th position.

The outcomes of the analysis of the network of the Florentine Intermarriage Relations and the Karate Club one are fully confirmed by a more extended analysis on 106 undirected networks, all freely available at `https://sparse.tamu.edu/` [113]. Our analysis includes all of the binary symmetric matrices available in the database sized $N \leq 1000$. Other networks included in the analysis are (as named as in the database):

- HB/dwt_1005 - size $N = 1005$;
- HB/dwt_1007 - size $N = 1007$;
- HB/jagmesh2 - size $N = 1009$;
- Arenas/email - size $N = 1133$;
- HB/bcspwr06 - size $N = 1454$;
- Rajat/rajat02 - size $N = 1960$;
- Barabasi/NotreDame_yeast - size $N = 2114$;
- Gleich/minnesota - size $N = 2642$;
- HB/sstmodel - size $N = 3345$;
- AG-Monien/airfoil1 - size $N = 4253$;

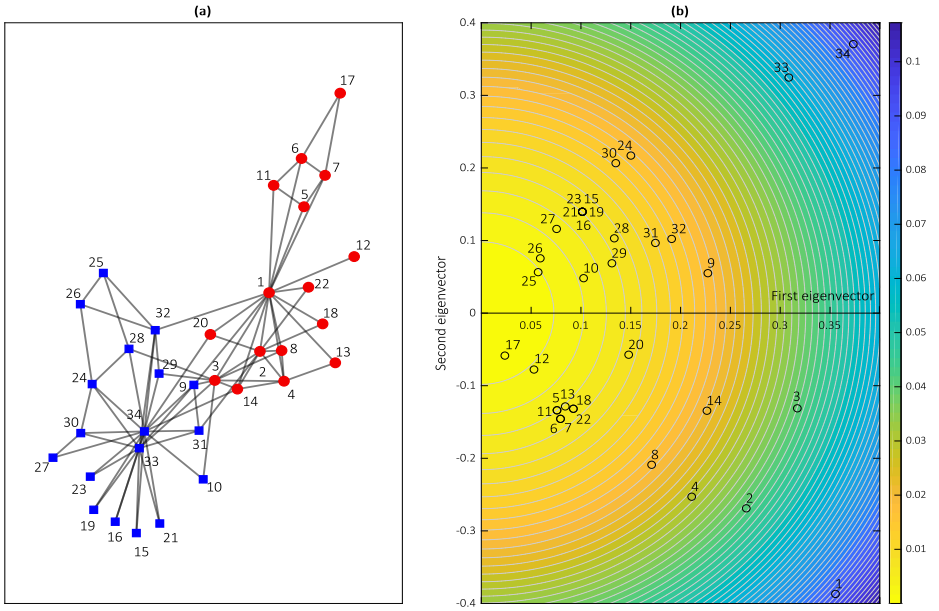

Figure 2.3: **(a)** Friendship network of the Zachary's Karate Club members. The two groups in which the club divided after the fission are highlighted in blue and red. **(b)** Contour plot of the unique contribution resulting from the application of Eq. (2.38) with $s = 2$. Nodes with larger unique contribution are found further away from the origin. The contours range from lower values of unique contribution (in yellow) to larger values (in blue). The $x_{i,1}$ values (corresponding to the components of the first eigenvector) are on the x-axis, while the values of $x_{i,2}$ (related to the components of the eigenvector corresponding to the second eigenvalue) are on the y-axis. The open circles correspond to the $x_{i,1}$ and $x_{i,2}$ values for the Zachary's Karate Club network.

- Newman/power - size $N = 4941$.

The values of R_a^2 obtained from the application of the functions in Table 2.1 are reported in Figure 2.4. Two features clearly emerge. Firstly, the degree, eigenvector and Katz centrality systematically perform poorly when considered under the perspective of estimating the networks topology. This is essentially due to the compression of information from N^2 to N implied by the matrix-estimation exercise, undermining the performance of the estimators. In general, R_a^2 decreases proportionally to the square root of N, following the behaviour of the standard deviation of the centrality-based estimators. Hence, the largest the size, the more information is lost during the estimation. The plot shows systematically higher values of R_a^2 resulting from the application of the two-components estimator Eq. (2.26). As expected, considering more node's properties dramatically improves the

estimation quality. Qualitatively similar results for directed networks are reported in the next section.

A second key feature emerging from Figure 2.4 is that the values of R_a^2 obtained from different one-component estimators are only slightly different from one another, and there is no evidence of one centrality measure outperforming the others. It follows that, despite the different nature of the metrics (i.e., the degree is a *local* measure of nodes' importance, while the eigenvector and the Katz centrality are *global* measures [46]), all the metrics provide very similar and limited information about the topology of the networks. In this case, using different centrality metrics would not add new and divers information, resulting with redundancy of the metrics and therefore providing a further proof of their correlation [96].

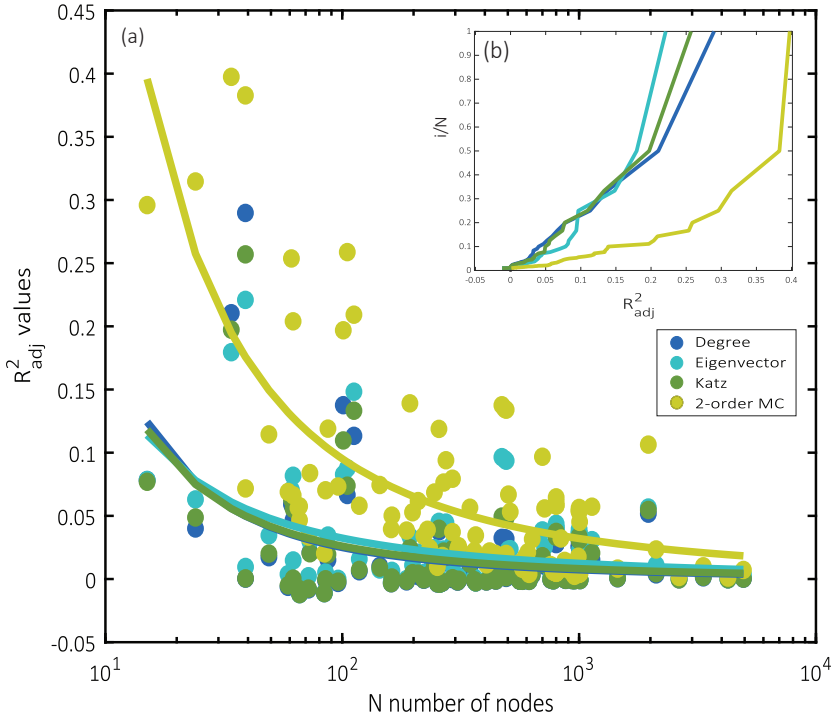

Figure 2.4: **(a)** Values of the coefficient of determination R_a^2, in semi-log scale obtained through the centrality-based estimators degree, eigenvector, Katz and multi-component (MC). Each dot refer to a network in the *Sparse Matrix* database [113]. Power-law curves are fitted to the data to facilitate visual comparison. **(b)** Cumulative frequency curves for the R_a^2 obtained by the four estimators.

2.2 Directed, Unweighted Networks

2.2.1 General Considerations

In directed, unweighted networks, the edges are directed and the elements A_{ij} of the adjacency matrix $\mathbf{A}$ are 1 if the edge points from i to j, and zero otherwise. The adjacency matrix is generally asymmetric [9] (notice that we here consider i pointing to j, i.e., the outgoing edges of the node i are described onto the row i of the matrix $\mathbf{A}$). In this kind of networks, nodes can be characterised by two properties, the one concerning with the *outgoing* centrality of the node, x_i^{out}, and the other concerning with the *incoming* centrality, x_i^{in}. Therefore, the estimator $\hat{A}_{ij}$ should depend on the outgoing centrality of node i and on the incoming centrality of node j, namely

$$\hat{A}_{ij} = f(x_i^{out}, x_j^{in}). \tag{2.40}$$

In the case of directed networks, the arguments of the function are exchangeable only on the diagonal, namely

$$f(x_k^{out}, x_k^{in}) = f(x_k^{in}, x_k^{out}).$$

The steps described for undirected networks to obtain the centrality values and to compute the unique contribution, Section 2.1, can be easily adapted to directed networks. Similarly to what described for undirected networks, the value of the sum of squared errors *SE* (Eq. (3.21)) can be partitioned into two components, a first part which is independent of $x_k^{out/in}$, SE_0, and a second part depending on $x_k^{out/in}$, SE_k, i.e.,

$$SE = SE_0 + SE_k.$$

When deriving *SE* with respect to any bound variable $z_m = x_k^{out/in}$, the contribution of the term SE_0 is zero and the derivative of *SE* equals the derivative of SE_k. The term SE_k is

$$SE_k = \sum_{i \neq k} \Big(A_{ik} - f(x_i^{out}, x_k^{in})\Big)^2 + \sum_{j \neq k} \Big(A_{kj} - f(x_k^{out}, x_j^{in})\Big)^2 + \Big(A_{kk} - f(x_k^{out}, x_k^{in})\Big)^2. \tag{2.41}$$

Notice that this term only depends on the k-th row and column of the two matrices $\mathbf{A}$ and $\hat{\mathbf{A}}$, to which the information of the outgoing and incoming edges of node k is associated. Separating the row and column contribution, using the bound variable z_m, the derivatives of SE_k with respect to the

variables x_k^{out} and x_k^{in} are, respectively,

$$\frac{\partial SE_k}{\partial x_k^{out}} = 2\sum_{j\neq k}\Big[A_{kj} - f(x_k^{out}, x_j^{in})\Big]\cdot \left.\frac{\partial f(z_m, x_j^{in})}{\partial z_m}\right|_{z_m=x_k^{out}} + \tag{2.42}$$
$$2\Big[A_{kk} - f(x_k^{out}, x_k^{in})\Big]\cdot \left.\frac{\partial f(z_m, x_k^{in})}{\partial z_m}\right|_{z_m=x_k^{out}} = 0,$$

and

$$\frac{\partial SE_k}{\partial x_k^{in}} = 2\sum_{i\neq k}\Big[A_{ik} - f(x_i^{out}, x_k^{in})\Big]\cdot \left.\frac{\partial f(x_i^{out}, z_m)}{\partial z_m}\right|_{z_m=x_k^{in}} + \tag{2.43}$$
$$2\Big[A_{kk} - f(x_k^{out}, z_m)\Big]\cdot \left.\frac{\partial f(x_i^{out}, z_m)}{\partial z_m}\right|_{z_m=x_k^{in}} = 0.$$

In Eq. (2.42) and Eq. (2.43), both the terms $i = k$ and $j = k$ can be included into the sums. Hence holds

$$\frac{\partial SE_k}{\partial x_k^{out}} = 2\sum_{j}\Big[A_{kj} - f(x_k^{out}, x_j^{in})\Big]\cdot \left.\frac{\partial f(z_m, x_j^{in})}{\partial z_m}\right|_{z_m=x_k^{out}} = 0, \tag{2.44}$$

and

$$\frac{\partial SE_k}{\partial x_k^{in}} = 2\sum_{i}\Big[A_{ik} - f(x_i^{out}, x_k^{in})\Big]\cdot \left.\frac{\partial f(x_i^{out}, z_m)}{\partial z_m}\right|_{z_m=x_k^{in}} = 0. \tag{2.45}$$

These formulas are used to obtain, from different estimator functions, different centrality metrics.

2.2.2 The Unique Contribution

Following the same reasoning shown for undirected networks, the unique contribution is found using Eq. (2.8), hence computing $\Delta SE = SE_{N\setminus k} - SE_N$. In directed networks, nodes are characterised by two properties. Within this framework, the unique contribution can be computed with respect to one of the properties, or at the need, with respect to both ones. In the first case, one finds the *out*-centrality (or the *in*-centrality) of the node. In the second case the overall centrality of the node is obtained.

If both properties are considered in the computation, we can define ΔSE as

$$\begin{aligned}\Delta SE^{tot} = &\sum_{i\neq k}\Big[\Big(A_{ik} - f(x_i^{out}, 0)\Big)^2 - \Big(A_{ik} - f(x_i^{out}, x_k^{in})\Big)^2\Big] + \\ &\sum_{j\neq k}\Big[\Big(A_{kj} - f(0, x_j^{in})\Big)^2 - \Big(A_{ik} - f(x_k^{out}, x_j^{in})\Big)^2\Big] + \\ &\Big(A_{kk} - f(0,0)\Big)^2 - \Big(A_{kk} - f(x_k^{out}, x_k^{in})\Big)^2,\end{aligned} \tag{2.46}$$

in which we consider the exclusion of the properties x_k^{out} and x_k^{in} to be equivalent to setting $x_k^{out} = x_k^{in} = 0$. Eq. (2.46) can be expressed as

$$\begin{aligned}\Delta SE^{tot} = &\sum_{i\neq k}\Big[f(x_i^{out}, 0)^2 - f(x_i^{out}, x_k^{in})^2 - 2f(x_i^{out}, 0)A_{ik} + 2f(x_i^{out}, x_k^{in})A_{ik}\Big] + \\ &\sum_{j\neq k}\Big[f(0, x_j^{in})^2 - f(x_k^{out}, x_j^{in})^2 - 2f(0, x_j^{in})A_{kj} + 2f(x_k^{out}, x_j^{in})A_{kj}\Big] + \\ &f(0,0)^2 - f(x_k^{out}, x_k^{in})^2 - 2f(0,0)A_{kk} + 2f(x_k^{out}, x_k^{in})A_{kk},\end{aligned} \tag{2.47}$$

or

$$\begin{aligned}\Delta SE^{tot} = &\sum_{i\neq k}\Big(f(x_i^{out}, 0) - f(x_i^{out}, x_k^{in})\Big)\Big(f(x_i^{out}, 0) + f(x_i^{out}, x_k^{in}) - 2A_{ik}\Big) + \\ &\sum_{j\neq k}\Big(f(0, x_j^{in}) - f(x_k^{out}, x_j^{in})\Big)\Big(f(0, x_j^{in}) + f(x_k^{out}, x_j^{in}) - 2A_{kj}\Big) + \\ &\Big(f(0,0) - f(x_k^{out}, x_k^{in})\Big)\cdot\Big(f(0,0) - f(x_k^{out}, x_k^{in}) - 2A_{kk}\Big)\end{aligned} \tag{2.48}$$

The unique contribution is then found deploying the expression in Eq. (2.47) or Eq. (2.48), and applying the definition in Eq. (2.8).

Instead, to compute the unique contribution with respect to one of the two properties entails considering, in Eq. (2.47) or Eq. (2.48), only the terms on the dimension related to the specific property at hand; hence, the k-th row (sum over j) for the *out* centrality of the node k and the k-th column (sum over i) for its *in* centrality. In formulas

$$\begin{aligned}\Delta SE^{out} = &\sum_{j}\Big[f(0, x_j^{in})^2 - f(x_k^{out}, x_j^{in})^2 - 2f(0, x_j^{in})A_{kj} + 2f(x_k^{out}, x_j^{in})A_{kj}\Big] \\ = &\sum_{j}\Big(f(0, x_j^{in}) - f(x_k^{out}, x_j^{in})\Big)\Big(f(0, x_j^{in}) + f(x_k^{out}, x_j^{in}) - 2A_{kj}\Big),\end{aligned} \tag{2.49}$$

and

$$\Delta SE^{in} = \sum_i \Big[f(x_i^{out}, 0)^2 - f(x_i^{out}, x_k^{in})^2 - 2f(x_i^{out}, 0)A_{ik} + 2f(x_i^{out}, x_k^{in})A_{ik} \Big] \tag{2.50}$$

$$= \sum_i \Big(f(x_i^{out}, 0) - f(x_i^{out}, x_k^{in}) \Big) \Big(f(x_i^{out}, 0) + f(x_i^{out}, x_k^{in}) - 2A_{ik} \Big).$$

In the following, we consider networks with no self-loops, hence $A_{kk} = 0$ in all formulas.

2.2.3 Examples of Estimator Functions

Again, different definitions of the function f in Eq. (2.40) allow one to obtain different centrality metrics for directed networks. Examples of the *out* and *in* centrality of the nodes that we are able to recover in this statistical framework are the degree and the hub-authority centrality [114]. Details follow (see Table 2.3 for a resume of the mathematical results).

Degree Centrality

Consider the function f_1 to be defined as

$$\hat{A}_{ij} = f_1(x_i^{out}, x_k^{in}) = a \Big[x_i^{out} + x_k^{in} - \frac{1}{N} \Big]. \tag{2.51}$$

The derivatives of the function f_1 with respect to both properties x_k^{out} and x_k^{in} are

$$\frac{\partial f_1}{\partial x_k^{out}} = \frac{\partial f_1}{\partial x_k^{in}} = a.$$

Applying Eq. (2.44) and Eq. (2.45) one obtains

$$2a \sum_i \Big[A_{ik} - a \Big(x_i^{out} + x_k^{in} - \frac{1}{N} \Big) \Big] = 0,$$

and

$$2a \sum_j \Big[A_{kj} - a \Big(x_k^{out} + x_j^{in} - \frac{1}{N} \Big) \Big] = 0,$$

in which $\sum_i A_{ik} = k_k^{in}$ is the in-degree of the node k and $\sum_j A_{kj} = k_k^{out}$ is its out-degree. Solving both equations for the properties x_k^{out} and x_k^{in} yields

$$x_k^{in} = \frac{k_k^{in}}{aN}$$

and

$$x_k^{out} = \frac{k_k^{out}}{aN}.$$

Assuming the vectors of centralities $\mathbf{x}^{out}$ and $\mathbf{x}^{in}$ to have unitary 1-norm, i.e., $\sum_i x_i^{out} = \sum_i x_i^{in} = 1$, one obtains $a = K_{tot}/N$ as in Eq. (2.14), finally yielding the values

$$x_k^{in} = \frac{k_k^{in}}{K_{tot}}, \tag{2.52a}$$

$$x_k^{out} = \frac{k_k^{out}}{K_{tot}}. \tag{2.52b}$$

Eq. (2.52b) – Eq. (2.52a) correspond to scale the **out-degree** and **in-degree** by the total degree of the network.

Unique contribution

Let us start from the computation of the total unique contribution i.e., the UC of the node k when its properties *out* and *in* are considered together. From Eq. (2.51), to exclude the properties x_k^{out} and x_k^{in} from the estimation procedure provides the following function values

$$\begin{aligned} f_1(x_i^{out}, 0) &= ax_i^{out} - \frac{a}{N}; \\ f_1(0, x_j^{in}) &= ax_j^{in} - \frac{a}{N}; \\ f_1(0,0) &= -\frac{a}{N}. \end{aligned}$$

Using Eq. (2.48), the variation induced in the sum of squared errors SE^{tot} is

$$\begin{aligned} \Delta SE^{tot} =& \sum_{i \neq k} (-ax_k^{in}) \Big(2ax_i^{out} + ax_k^{in} - 2\frac{a}{N} - 2A_{ik} \Big) + \\ & \sum_{j \neq k} (-ax_k^{out}) \Big(2ax_j^{in} + ax_k^{out} - 2\frac{a}{N} - 2A_{kj} \Big) \\ & + (-ax_k^{out} - ax_k^{in}) \Big(ax_k^{out} + ax_k^{in} - 2\frac{a}{N} - 2A_{kk} \Big) \\ =& - ax_k^{in} \sum_i \Big(-2ax_i^{out} - ax_k^{in} + 2\frac{a}{N} + 2A_{ik} \Big) - \\ & ax_k^{out} \sum_j \Big(-2ax_j^{in} - ax_k^{out} + 2\frac{a}{N} + 2A_{kj} \Big) + 2a^2 x_k^{out} x_k^{in}, \end{aligned}$$

in which the assumption $A_{kk} = 0$ is used. Substituting the values of x_k^{out} and x_k^{in} according to Eqs. (2.52), and considering $a = K_{tot}/N$, some algebra gives

$$\Delta SE^{tot} = \frac{(k_k^{in})^2 + (k_k^{out})^2}{N} + \frac{2k_k^{in}k_k^{out}}{N^2},$$

from which the total unique contribution is obtained

$$UC_k^{tot} = \frac{1}{TSS}\Big[\frac{(k_k^{in})^2 + (k_k^{out})^2}{N} + \frac{2k_k^{in}k_k^{out}}{N^2}\Big]. \tag{2.53}$$

The unique contribution for separately considering the property *out* or *in* is obtained by applying the definitions in Eq. (2.49) - Eq. (2.50), respectively. In this case it holds

$$UC_k^{out} = \frac{(k_k^{out})^2}{NTSS}, \tag{2.54}$$

$$UC_k^{in} = \frac{(k_k^{in})^2}{NTSS}. \tag{2.55}$$

Both the formulations in Eq. (2.54) and Eq. (2.55) are monotonic increasing function of x_k^{out} and x_k^{in}, respectively. Hence, ranking for increasing UC_k^{out} and UC_k^{in} values provide the same ranking as the classical in and out degree centrality.

Hub-Authority Centrality

Consider the estimator for directed network f_2 to be defined as follows

$$\hat{A}_{ik} = f_2(x_i^{out}, x_k^{in}) = \gamma x_i^{out} x_k^{in}. \tag{2.56}$$

Clearly,

$$\frac{\partial f_2}{\partial x_k^{out}} = \gamma x_j^{out}, \qquad \frac{\partial f_2}{\partial x_k^{in}} = \gamma x_i^{in}.$$

By applying the minimisation procedure defined by Eq. (2.44) and Eq. (2.45), one obtains

$$\begin{cases} \frac{\partial SE}{\partial x_k^{out}} = 2\sum_j(\gamma A_{kj}x_j^{in} - \gamma^2 x_k^{out}(x_j^{in})^2) = 0, \\ \frac{\partial SE}{\partial x_k^{in}} = 2\sum_i(\gamma A_{ik}x_i^{out} - \gamma^2 (x_i^{out})^2 x_k^{in}) = 0. \end{cases}$$

that, solved with respect to the properties x_k^{out} and x_k^{in}, within the assumption of unitary 2-norm of the vectors (i.e., $\sum_i (x_i^{out})^2 = 1$ and $\sum_j (x_j^{in})^2 = 1$) yields

$$\begin{cases} x_k^{out} = \frac{1}{\gamma} \sum_j A_{kj} x_j^{in}, \\ x_k^{in} = \frac{1}{\gamma} \sum_i A_{ik} x_i^{out}. \end{cases} \tag{2.57}$$

In matrix form,

$$\begin{cases} \gamma \mathbf{x}^{out} = \mathbf{A}\mathbf{x}^{in}, \\ \gamma \mathbf{x}^{in} = \mathbf{A}^{\mathbf{T}}\mathbf{x}^{out}. \end{cases}$$

Some algebra provides

$$\begin{cases} \gamma^2 \mathbf{x}^{out} = \mathbf{A}\mathbf{A}^{\mathbf{T}}\mathbf{x}^{out}, \\ \gamma^2 \mathbf{x}^{in} = \mathbf{A}^{\mathbf{T}}\mathbf{A}\mathbf{x}^{in}. \end{cases}$$

Introducing the matrices $\mathbf{C} = \mathbf{A}^{\mathbf{T}}\mathbf{A}$ and $\mathbf{D} = \mathbf{A}\mathbf{A}^{\mathbf{T}}$, one has

$$\gamma^2 \mathbf{x}^{out} = \mathbf{D}\mathbf{x}^{out}, \tag{2.58a}$$

$$\gamma^2 \mathbf{x}^{in} = \mathbf{C}\mathbf{x}^{in}. \tag{2.58b}$$

Eq. (2.58) states that $\mathbf{x}^{out}$ and $\mathbf{x}^{in}$ are the dominant eigenvectors of the matrices $\mathbf{D}$ and $\mathbf{C}$, respectively, associated to the principal eigenvalue of the two matrices, such that $\gamma^2 = \lambda_1(\mathbf{C}) = \lambda_1(\mathbf{D}) = \sigma_1^2(\mathbf{A})$ [43, 115], being σ_1 the principal singular value of the matrix $\mathbf{A}$ (see Chapter 1, Section 1.2). The formulation in Eq. (2.58) matches the **HITS algorithm** [114], used to identify *hubs* and *authorities* in networks.

Unique contribution

First, consider the unique contribution to be computed with respect to both the properties. Using Eq. (2.56), the function f_2 takes the zero values when one or both properties of x_k are set to zero; namely,

$$f_2(x_i^{out}, 0) = f_2(0, x_j^{in}) = f_2(0, 0) = 0.$$

It follows, from Eq. (2.47) that the variation in SE^{tot} induced by excluding

both properties x_k^{out} and x_k^{in} is

$$\begin{aligned}\Delta SE^{tot} =& \sum_{i\neq k}\Big[-(\gamma x_i^{out}x_k^{in})^2 + 2\gamma x_i^{out}x_k^{in}A_{ik}\Big] + \\ & \sum_{j\neq k}\Big[-(\gamma x_k^{out}x_j^{in})^2 + 2\gamma x_k^{out}x_j^{in}A_{kj}\Big] + \\ & \Big[-(\gamma x_k^{out}x_k^{in})^2 + 2\gamma x_k^{out}x_k^{in}A_{kk}\Big] \\ =& \sum_{i}\Big[-(\gamma x_i^{out}x_k^{in})^2 + 2\gamma x_i^{out}x_k^{in}A_{ik}\Big] + \\ & \sum_{j}\Big[-(\gamma x_k^{out}x_j^{in})^2 + 2\gamma x_k^{out}x_j^{in}A_{kj}\Big] - \Big[-(\gamma x_k^{out}x_k^{in})^2\Big],\end{aligned}$$

in which the assumption $A_{kk} = 0$ is used. Some algebra provides

$$\begin{aligned}\Delta SE^{tot} =& -\gamma(x_k^{in})^2\sum_i (x_i^{out})^2 + 2\gamma x_k^{in}\sum_i x_i^{out}A_{ik} - \gamma(x_k^{out})^2\sum_j (x_j^{in})^2 + \\ & 2\gamma x_k^{out}\sum_j A_{kj}x_j^{in} + (\gamma x_k^{out}x_k^{in})^2.\end{aligned} \tag{2.59}$$

Since the 2-norm of the vectors $\mathbf{x}^{out}$ and $\mathbf{x}^{in}$ is unitary and using Eq. (2.57), one has

$$\Delta SE^{tot} = \gamma^2(x_k^{out})^2 + \gamma^2(x_k^{in})^2 + (\gamma x_k^{out}x_k^{in})^2.$$

The total unique contribution of the node k applying the definition Eq. (2.8) is

$$UC_k^{tot} = \frac{\gamma^2(x_k^{out})^2 + \gamma^2(x_k^{in})^2 + (\gamma x_k^{out}x_k^{in})^2}{TSS}. \tag{2.60}$$

Instead, to define the unique contribution accounting separately for the properties *out* or *in* the partial variations in *SE* should be computed. Using Eq. (2.49), for the *out*-property it holds

$$\Delta SE^{out} = \sum_j\Big[-(\gamma x_k^{out}x_j^{in})^2 + 2\gamma x_k^{out}x_j^{in}A_{kj}\Big],$$

while, for the *in*-property, using Eq. (2.50), it holds

$$\Delta SE^{in} = \sum_i\Big[-(\gamma x_i^{out}x_k^{in})^2 + 2\gamma x_i^{out}x_k^{in}A_{ik}\Big].$$

Going through the same algebra as done for Eq. (2.59) and applying the definition of unique contribution, one obtains

$$UC_k^{out} = \frac{\gamma^2 (x_k^{out})^2}{TSS}. \tag{2.61}$$

and

$$UC_k^{in} = \frac{\gamma^2 (x_k^{in})^2}{TSS}. \tag{2.62}$$

Both the formulations in Eq. (2.61) and Eq. (2.62) are monotonic increasing function of x_k^{out} and x_k^{in}, respectively. Hence, ranking for increasing UC_k^{out} and UC_k^{in} values provide the same ranking as the classical hub-authority algorithm.

Multicomponent Estimator and Centrality for Directed, Unweighted Networks

In the case of directed networks, the multi-component estimator is a function of the s-dimensional vectors $\mathbf{x}_i^{out}$ and $\mathbf{x}_j^{in}$ considered for evaluating node's importance, namely

$$\hat{A}_{ij} = f(\mathbf{x}_i^{out}, \mathbf{x}_j^{in}),$$

where $\mathbf{x}_i^{out} = [x_{i,1}^{out}, ..., x_{i,s}^{out}]$ and $\mathbf{x}_j^{in} = [x_{j,1}^{in}, ..., x_{i,s}^{in}]$. Within this framework, the multidimensional estimator is

$$\begin{aligned} \hat{A}_{ij}(s) &= \gamma_1 x_{i,1}^{out} x_{j,1}^{in} + \gamma_2 x_{i,2}^{out} x_{j,2}^{in} + ... + \gamma_s x_{i,s}^{out} x_{j,s}^{in} \\ &= \sum_{t=1}^{s} \gamma_s x_{i,t}^{out} x_{j,t}^{in}. \end{aligned} \tag{2.63}$$

We assume the 2-norm of each vector $\mathbf{x}_t^{out} = [x_{1,t}^{out}, ..., x_{N,t}^{out}]$ and $\mathbf{x}_{i,t}^{in} = [x_{1,t}^{in}, ..., x_{N,t}^{in}]$ is unitary i.e., $\sum_i (x_{i,t}^{out})^2 = \sum_i (x_{i,t}^{in})^2 = 1$. Moreover, we set an orthogonality condition between any two vectors $\mathbf{x}_t^{out/in}$ and $\mathbf{x}_{t^*}^{out/in}$, i.e.

$$\sum_i x_{i,t}^{out} \cdot x_{i,t^*}^{out} = 0, \qquad \forall\ t \neq t^*, \tag{2.64}$$

$$\sum_i x_{i,t}^{in} \cdot x_{i,t^*}^{in} = 0, \qquad \forall\ t \neq t^*. \tag{2.65}$$

Similarly to Section 2.1, in this multi-component setting the function *SE* of the squared errors is expressed as Eq. (2.28). In order to compute the

Table 2.3: Estimator functions used for directed networks. In the formulas, K_{tot} is the total degree of the network; N is the number of nodes; k_i^{out} and k_i^{in} are the *out* degree and *in* degree of the node i; γ is a parameter whose value equals the principal singular value σ_1 of $\mathbf{A}$. *TSS* is defined in the text. The equations for the unique contribution are reported for the cases when outgoing and incoming properties of the node are separately considered (superscripts *out* and *in*), or for the case when they are considered together (superscript *tot*). Further details are given in Section 2.2.3

Directed networks			
Estimator function f	**Out, in and total centrality of node i**	**Out, in and total unique contribution of node i**	**Corresponding metrics**
$f_1 = \frac{K_{tot}}{N}\ \left(x_i^{out} + x_j^{in} - \frac{1}{N}\ \right)$	$x_i^{out} = \frac{k_i^{out}}{K_{tot}}$ $x_j^{in} = \frac{k_j^{in}}{K_{tot}}$	$UC_i^{out} = \frac{(k_i^{out})^2}{N\ TSS}$, $UC_i^{in} = \frac{(k_i^{in})^2}{N\ TSS}$ $UC_i^{tot} = \frac{1}{TSS}\left(\frac{(k_i^{out})^2+(k_i^{in})^2}{N} + \frac{2k_i^{out}k_i^{in}}{N^2}\right)$	**Degree centrality**
$f_2 = \gamma x_i^{out} x_j^{in}$	$\begin{cases} x_i^{out} = \frac{1}{\gamma}\sum_j A_{ij}x_j^{in} \\ x_j^{in} = \frac{1}{\gamma}\sum_i A_{ij}x_i^{out} \end{cases}$	$UC_i^{out} = \frac{(\gamma x_i^{out})^2}{TSS}$, $UC_i^{in} = \frac{(\gamma x_i^{in})^2}{TSS}$ $UC_i^{tot} = \frac{1}{TSS}\left[\gamma^2\left((x_i^{out})^2 + (x_i^{in})^2\right) + (\gamma x_i^{out} x_i^{in})^2\right]$	**Hub-authority centrality**

centrality values, it is necessary to derive the function $SE_{k,t}$ accounting for the s dimensions embedded in the estimators. We use the bound variable $\mathbf{z}_m$ to define the derivatives of the multi-component estimator Eq. (2.63) with respect to the variables $\mathbf{x}^{out}_{k,t^*}$ and $\mathbf{x}^{in}_{k,t^*}$ at any order t^*; namely,

$$\left.\frac{\partial f(\mathbf{x}^{out}_i, \mathbf{z}_m)}{\partial z_{k,t^*}}\right|_{\mathbf{z}_m=\mathbf{x}^{in}_k} = \gamma_{t^*} x^{out}_{i,t^*},$$

and

$$\left.\frac{\partial f(\mathbf{z}_m, \mathbf{x}^{in}_j)}{\partial z^{in}_{k,t^*}}\right|_{\mathbf{z}_m=\mathbf{x}^{out}_k} = \gamma_{t^*} x^{in}_{j,t^*},$$

that, introduced in Eq. (2.44) and Eq. (2.45), provide the following minimisation conditions

$$\begin{aligned} &2\sum_i \Big[A_{ik} - \sum_t \gamma_t x^{out}_{i,t} x^{in}_{k,t}\Big]\gamma_{t^*} x^{out}_{i,t^*} = \\ &\sum_i A_{ik} x^{out}_{i,t^*} - \sum_t \gamma_t x^{in}_{k,t} \sum_i x^{out}_{i,t^*} x^{out}_{i,t} = 0 \end{aligned}$$

and

$$\begin{aligned} &2\sum_j \Big[A_{kj} - \sum_t \gamma_t x^{out}_{k,t} x^{in}_{j,t}\Big]\gamma_{t^*} x^{in}_{j,t^*} = \\ &\sum_j A_{kj} x^{in}_{j,t^*} - \sum_t \gamma_t x^{out}_{k,t} \sum_j x^{in}_{j,t^*} x^{in}_{j,t} = 0 \end{aligned}$$

Using the conditions of othonormality, Eq. (2.64) - Eq. (2.65), some algebra provides

$$\begin{cases} x^{out}_{k,t} = \frac{1}{\gamma_t} \sum_j A_{kj} x^{in}_{j,t}, \\ x^{in}_{k,t} = \frac{1}{\gamma_t} \sum_i A_{ik} x^{out}_{i,t}. \end{cases} \tag{2.66}$$

Eq. (2.66) states that at any order t, the vectors $\mathbf{x}^{out}_t = [x^{out}_{1,t}, ..., x^{out}_{N,t}]$ and $\mathbf{x}^{in}_t = [x^{in}_{1,t}, ..., x^{in}_{N,t}]$ are the left and right singular vectors associated to the singular value $\gamma_t = \sigma_t$, respectively (see Chapter 1, Section 1.2). Therefore, the estimation provided in Eq. (2.63) is the *s-order low-rank approximation* of the original adjacency matrix $\hat{A}$. In fact, Eq. (2.63) coincides with the Singular Value Decomposition (SVD) [43, 116], being γ_t the singular values and $\mathbf{x}^{out}_t$ and $\mathbf{x}^{in}_t$ the related singular vectors, as defined by Eqs. (2.66).

Unique contribution

In the multi-component setting for directed networks, the unique contribution is found accounting for the s dimensions embedded in the estimator function f (see Eq. (2.63)). In this case, when excluding the generic node k from the estimation, all the properties $x_{k,t}^{out}$ and $x_{k,t}^{in}$, with $t = (1, ..., s)$, are set to zero. This yields

$$f(\mathbf{x}_i^{out}, 0) = f(0, \mathbf{x}_j^{in}) = f(0, 0) = 0.$$

Within this multi-component setting, the unique contribution can be computed with respect to both the properties $\mathbf{x}_{k,t}^{out}$ and $\mathbf{x}_{k,t}^{in}$, or with respect to one of the two.

If both the properties are considered, Eq. (2.48) holds, providing the value ΔSE^{tot}, namely

$$\begin{aligned}\Delta SE^{tot} = &\sum_{i\neq k}\Big(-\sum_{t=1}^{s}\gamma_t x_{i,t}^{out}x_{k,t}^{in}\Big)\Big(\sum_{t=1}^{s}\gamma_t x_{i,t}^{out}x_{k,t}^{in} - 2A_{ik}\Big) + \\ &\sum_{j\neq k}\Big(-\sum_{t=1}^{s}\gamma_t x_{k,t}^{out}x_{j,t}^{in}\Big)\Big(\sum_{t=1}^{s}\gamma_t x_{k,t}^{out}x_{j,t}^{in} - 2A_{kj}\Big) + \\ &\Big(-\sum_{t=1}^{s}\gamma_t x_{k,t}^{out}x_{k,t}^{in}\Big)\Big(\sum_{t=1}^{s}\gamma_t x_{k,t}^{out}x_{k,t}^{in} - 2A_{kk}\Big)\end{aligned}$$

that is equivalent to

$$\begin{aligned}\Delta SE^{tot} = &\sum_{i}\Big[-\Big(\sum_{t=1}^{s}\gamma_t x_{i,t}^{out}x_{k,t}^{in}\Big)^2 + 2A_{ik}\sum_{t=1}^{s}\gamma_t x_{i,t}^{out}x_{k,t}^{in}\Big] + \\ &\sum_{j}\Big[-\Big(\sum_{t=1}^{s}\gamma_t x_{k,t}^{out}x_{j,t}^{in}\Big)^2 - 2A_{kj}\sum_{t=1}^{s}\gamma_t x_{k,t}^{out}x_{j,t}^{in}\Big] + \Big(\sum_{t=1}^{s}\gamma_t x_{k,t}^{out}x_{k,t}^{in}\Big)^2.\end{aligned}$$

Some algebra provides

$$\begin{aligned}\Delta SE^{tot} = &-\sum_{t=1}^{s}\gamma_t(x_{k,t}^{in})^2\sum_{i}(x_{i,t}^{out})^2 + 2\sum_{t=1}^{s}\gamma_t x_{k,t}^{in}\sum_{i}A_{ik}x_{i,t}^{out} - \\ &\sum_{t=1}^{s}\gamma_t(x_{k,t}^{out})^2\sum_{j}(x_{j,t}^{in})^2 + 2\sum_{t=1}^{s}\gamma_t x_{k,t}^{out}\sum_{j}A_{kj}x_{j,t}^{in} \\ &+\Big(\sum_{t=1}^{s}\gamma_t x_{k,t}^{out}x_{k,t}^{in}\Big)^2.\end{aligned} \tag{2.67}$$

Using the orthonormality conditions, Eq. (2.64) – Eq. (2.65), and the formulation in Eq. (2.66), the unique contribution in the case of the multi-component estimator in directed networks is obtained

$$UC(s)_k^{tot} = \frac{1}{TSS}\sum_{t=1}^{s}\gamma_t^2\Big((x_{k,t}^{in})^2 + (x_{k,t}^{out})^2\Big) + \Big(\sum_{t=1}^{s}\gamma_t x_{k,t}^{out}x_{k,t}^{in}\Big)^2. \tag{2.68}$$

The unique contribution when accounting separately for the *out* and *in* properties, applying Eq. (2.49) and Eq. (2.50), requires defining

$$\Delta SE^{out} = \sum_j\Big[-\Big(\sum_{t=1}^{s}\gamma_t x_{k,t}^{out}x_{j,t}^{in}\Big)^2 - 2A_{kj}\sum_{t=1}^{s}\gamma_t x_{k,t}^{out}x_{j,t}^{in}\Big]$$

and

$$\Delta SE^{in} = \sum_i\Big[-\Big(\sum_{t=1}^{s}\gamma_t x_{i,t}^{out}x_{k,t}^{in}\Big)^2 + 2A_{ik}\sum_{t=1}^{s}\gamma_t x_{i,t}^{out}x_{k,t}^{in}\Big].$$

Going through some algebra and applying the definition in Eq. (2.8), one obtains

$$UC(s)_k^{out} = \frac{1}{TSS}\sum_{t=1}^{s}\gamma_t^2(x_{k,t}^{out})^2, \tag{2.69}$$

and

$$UC(s)_k^{in} = \frac{1}{TSS}\sum_{t=1}^{s}\gamma_t^2(x_{k,t}^{in})^2. \tag{2.70}$$

2.2.4 Results on Directed, Unweighted Networks

We tested our framework on 36 networks freely available on the *Suite Sparse Matrix Collection* [113]. Our analysis includes all of the binary asymmetric matrices collected in the database, sized $N \leq 2000$. Other networks included in the analysis are (as named as in the database):

- Pajek/Kohonen - size $N = 4470$;
- Rajat/rajat01 - size $N = 6833$;
- SNAP/p2p-Gnutella09 - size $N = 8114$;
- Gleich/wb-cs-stanford - size $N = 9914$;
- SNAP/p2p-Gnutella04 - size $N = 10879$.

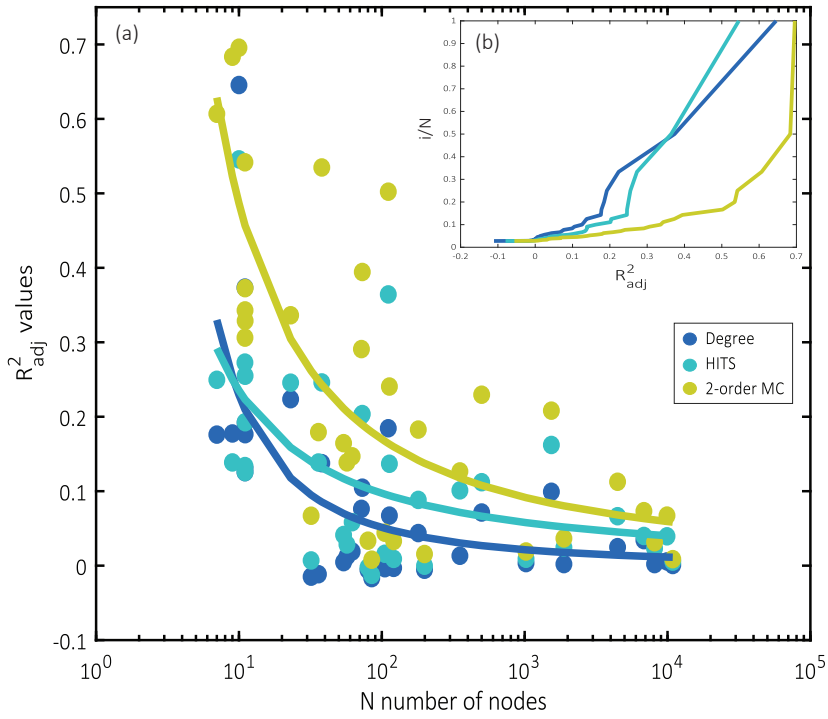

Figure 2.5: **(a)** Values of the coefficient of determination R^2_a in semi-log scale obtained through the centrality-based estimators degree, hub-authority and multi-component (MC). Each dot refer to a directed network in the *Sparse Matrix* database [113]. Power-law curves are fitted to the data to facilitate visual comparison. **(b)** Cumulative frequency curves for the R^2_a obtained by the three estimators.

The results obtained from our tests are shown in Figure 2.5.

The values in Figure 2.5 of the adjusted coefficient of determination, R^2_a, are higher than those shown in Figure 2.4, which were obtained from the application of our framework to undirected networks. This is mainly due to the fact that we are using two properties to characterise each node. As a consequence, the estimators applied in case of directed networks (see Table 2.3) project the information of the adjacency matrix from N^2 to $2N$, reducing the information gap. Also for directed networks, the one-component estimators perform poorly with respect to the two-component estimator. The hub-authority algorithm, however, has better performances than the degree, in particular when considering larger networks.

2.3 Weighted Networks

To extend our approach to weighted networks, one has to replace in all formulas the adjacency matrix $\mathbf{A}$ with the matrix of the weights $\mathbf{W}$, whose elements are defined as $W_{ij} > 0$ if there is a flux connecting i to j, zero otherwise. All the centrality measures in their weighted version are obtained as the solution of the same matrix estimation exercises we have detailed for undirected and directed networks.

2.4 Bipartite Networks

As described in Section 1.1, bipartite networks constitutes of two sets of nodes - $\mathbf{U}$ and $\mathbf{V}$ - with E edges connecting nodes between the two ensembles. These networks are described by the incidence matrix [9] $\mathbf{B}$ whose elements B_{ij} define the relationship between the nodes $i \in \mathbf{U}$ and the nodes $j \in \mathbf{V}$. In this case, a proper estimator $\hat{B}_{ij}$ would be a function of a property x_i of the nodes in the ensamble $\mathbf{U}$ and property y_j of the nodes in the ensemble $\mathbf{V}$ i.e., $\hat{B}_{ij} = f(x_i, y_j)$. Therefore, the asymmetry of this kind of systems is similar to the one concerning with directed metrics: it is hence sufficient to substitute the adjacency matrix $\mathbf{A}$ with the incidence matrix $\mathbf{B}$ and the results presented in Table 2.3 can be straightforward extended to bipartite networks. Nevertheless, a special focus on bipartite networks and their centrality metrics follows in Chapter 3, for the economic bipartite network of trade, and Chapter 4 for the network of sustainable development.

2.5 Concluding Remarks

Aiming at addressing the debate on the use of centrality metrics, we introduced a different point of view about centrality in which the evaluation of the importance of nodes is recast as a statistical exercise. Here, centrality becomes the node-property through which one estimates the adjacency matrix of the network, breaking new ground in the way we understand node centrality. While the extensive literature on network reconstruction (see, e.g., [117] and applications [118]) consider the degree to be suitably used for estimating the network topology – also in combination with other statistically-sound methodology as the maximum likelihood or maximum-entropy inference – our approach provides a framework in which many of the most commonly used centrality metrics can be deduced within this theoretical framework, thus paving the way for an unprecedented chance to

quantitatively compare the performances of different centrality measures. In particular, we have shown the innovative power of our statistical perspective on centrality metrics by focusing on the application on monopartite networks and paying attention to the degree and eigenvector-based centrality measures. However, we stress that our approach is very general and should not be restricted to the examples reported above. In fact, this approach can be extended to other centrality measures, by changing the estimator function in Eq. (2.1), and/or the error structure – additive or multiplicative – and/or the matrix whereon the estimation procedure is carried out (either the adjacency matrix or a transformation of this one). One example of this extension is the Freeman closeness [36], which is recovered in this framework by simply substituting the adjacency matrix with the *geodesic distance matrix* $\mathbf{D}$ [119] – using the function $f(x_i, x_j) = a[x_i + x_j - 1/N]$, since the measure of node's closeness corresponds to the degree of the node computed on the matrix $\mathbf{D}$ [119].

The application of the estimators could also explain the ability of the various algorithms to account for the node-node interactions and so, in the reconstruction of the network topology. Tests on a large number of networks show that there are no outperforming one-dimensional, centrality-based estimators and that all the metrics provide poor information regarding networks' topology. Our results, within the context of the still ongoing debate on the centrality metrics and the associated ranking (in several fields, see, e.g., [45, 46, 92–94]), provide further proofs that centrality metrics are highly correlated [95–100] and that they provide similar information about the importance of the nodes. Within this new framework, the natural multi-component extension of node centrality emerges as a possible solution to improve the quality of the estimations and, subsequently, of node ranking.

This work therefore provides a possible quantitative answer to the long-standing question *"what does it mean to be central in a network ?"*

As will be better detailed and explained in Chapters 3 and 4, the application of this statistical framework and of its multi-component extension also helps in shed new light on the mathematical nature of the algorithms used to evaluate node centrality and so, on the nature of the nodes interactions of a given system especially in bipartite systems.

3

Reconciling Contrasting Views on Economic Complexity

The work described in this chapter has been partially derived from Sciarra et al., Nature Communications, *2020 [120].*

Within economics, the Gross Domestic Product stands as the preferred indicator for the assessment of the status of a country's economy. However, Economic Complexity (EC) methodologies arising within network science have been stirring the pot in the last decade. In fact, economic complexity metrics, framed within a centrality exercise, aim at defining the socio-economic status of countries grounded on the data on their export baskets, structured as a bipartite network [4, 48]. The methodologies improve the basic information of the degree of countries and products by exploiting the information related to the sophistication of the exports and the capabilities required to produce and export a given good: countries with low productive knowledge (see Chapter 1, Section 1.3), only produce and export fewer and less sophisticated products, resulting in lower stages of competition [23, 48]; while more competitive countries exploit their know-how and resources to diversify their export baskets [23, 48]. By reversing this reasoning, it is thus expected that the diversification and composition of the export basket can be used to measure the countries' and products' economic complexity, thus posing the bases for a data-based (bottom-up) ranking of countries and products. This rationale lies at the base of the commonly used methodologies to measure the economic complexity of countries and products, namely the Method of Reflections (MR) [22] and the Fitness and Complexity algorithm (FC) [23]. In spite of their common root, these two centrality metrics radically differ in the conceptual approach to the problem and, as a consequence, in the obtained outcomes, posing an issue of practical use.

Here, we show that the MR and FC approaches can be reconciled by

recasting them into a mathematically-sound, multidimensional framework, which allows us to recover and combine the strengths of both methods, still maintaining the relevant feature of providing countries' and products' rankings. This is obtained thanks to the results on centrality metrics we have described in Chapter 2, in particular the one regarding multi-dimensional centrality metrics. In this chapter, after a brief description of the MR and FC methods, we first introduce a general framework in which recast the economic complexity metrics provided by the two algorithms. Successively, we provide a unique measure of complexity, called the *GENeralized Economic comPlexitY* index, which allows us to recover and combine the strengths of both MR and FC methods, still maintaining the relevant feature of providing countries' and products' rankings. Finally, we discuss on how the obtained results shed new light on the potential of economic complexity to trace and forecast countries' innovation potential and to interpret the temporal dynamics of economic growth, possibly paving the way to a micro-foundation of the field.

3.1 The MR and FC Metrics of Economic Complexity

Economic complexity approaches are grounded on the trade data collected into a bipartite network, defining exporters and products, and detailing whether and how much (in monetary value) a country exports a given product. The bipartite network is interpreted as the compact representation of the tripartite network constituted by countries-capabilities-products [22, 23], according to which countries are only able to export products for which they own the required capabilities (Figure 3.1). These capabilities are intended as capabilities to innovate and they are intrinsic characteristics of countries and products; thus, they can only be unveiled by the analysis of the exports. Most applications [22, 23] take into account only the relevant exporters in the network, where the relevance is computed according to the Relative Comparative Advantage (RCA) [121]. The Relative Comparative Advantage procedure is used to construct the incidence binary matrix $\mathbf{M}$, setting the threshold of RCA to 1 in line with the economic complexity framework [22]. RCA weights how much a product p counts within the export basket of the country c. This fraction is weighted by the ratio of the total monetary flux globally generated by the same product p, and the total monetary flux of all products traded worldwide during the reference year.

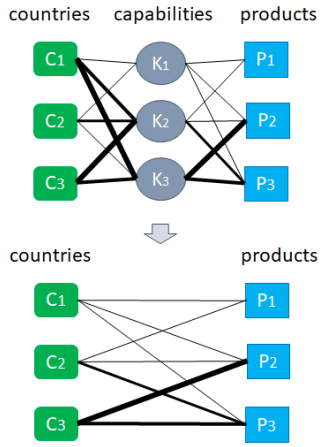

Figure 3.1: Qualitative representation of the tripartite and bipartite network of trade according to EC. The bipartite network connecting countries to products is the compact representation of the tripartite network connecting countries to their available capabilities, from which connections to products are determined. According to EC theory, countries are only able to export the goods for which they have the required capabilities to produce. Thickness of the links exemplifies the quantification of capabilities, in the tripartite network, and the observable monetary value in the bipartite network.

In formulas,

$$RCA_{cp} = \frac{\frac{D_{cp}}{\sum_p D_{cp}}}{\frac{\sum_c D_{cp}}{\sum_{cp} D_{cp}}}, \tag{3.1}$$

where D_{cp} is the return in dollars of a country c through the export of product p. The input matrix $\mathbf{M}$ is given by $M_{cp} = 1$ if $RCA_{cp} \geq 1$, implying that the country c is a relevant exporter of the product p, and 0 otherwise [121] [1].

The MR approach by Hidalgo & Hausmann [22] measures a country's economic complexity as the average of the complexities of the products in its export basket. In a specular manner – from which the name "Reflections" –, a product's complexity is obtained as the average of the complexities of the countries exporting it. The equations defining the two averages are coupled to obtain the Economic Complexity Index, ECI, and the Product Complexity Index, PCI. Namely,

[1] In Annex A, we provide insights about the implications of becoming a relevant exporter on countries' water resources. This analysis bridges two facets of the international trade of food commodities, the economic and environmental ones, both framed within the CWASI project, ERC-2014-CoG, project 647473, which has supported this Thesis work.

$$\begin{cases} ECI_c = \frac{1}{k_c} \sum_p M_{cp} PCI_p, \\ PCI_p = \frac{1}{k_p} \sum_c M_{cp} ECI_c. \end{cases} \tag{3.2}$$

In Eqs. (3.2), $k_c = \sum_p M_{cp}$ is the degree of country c (i.e., the number of products the country exports with RCA, a.k.a. diversity) and $k_p = \sum_c M_{cp}$ is the degree of product p (i.e., the number of exporters of a given product, a.k.a. ubiquity). The rationale in Eqs. (3.2) is the result of a linear algebra exercise [48, 122, 123]. However, as an effect of taking the averages, the obtained measures turn out to lose information about countries' diversification and products' ubiquity [124]. As we have already discussed, even if the degree is insufficient for a complete assessment of the complexity, it still remains a necessary and relevant information to understand the trading competitiveness of countries [48, 59]. Therefore, correlation with k_c is desirable in a limited way.

In contrast to MR, Tacchella *et al.* [23] counter on the assumption of a linear relation between the products' and countries' complexities. In their view, the fact that a less competitive country exports a given product should unavoidably downgrade the product's complexity, an effect that the Authors argue could only be obtained through the use of a non-linear relation. As a consequence, these Authors introduce two metrics, the Fitness of countries F_c and the Quality of products Q_p, where products' Quality non-linearly depends on the Fitness of the exporting countries (see Eqs. (3.3)); in contrast, the Fitness is obtained as the sum of the Qualities of the exported products. In this approach, contrarily to MR, the countries' Fitness preserves the information on the diversification of the export baskets [47, 59]. The measures are computed by iteration as defined in the following system

$$\begin{cases} \widetilde{F_c}^{(n+1)} = \sum_p M_{cp} Q_p^{(n)}, \quad F_c^{(n+1)} = \dfrac{\widetilde{F_c}^{(n+1)}}{\left(\sum_c \widetilde{F_c}^{(n+1)}\right)/C}; \\ \widetilde{Q_p}^{(n+1)} = \dfrac{1}{\sum_c M_{cp} \frac{1}{F_c^{(n)}}}, \quad Q_p^{(n+1)} = \dfrac{\widetilde{Q_p}^{(n+1)}}{\left(\sum_p \widetilde{Q_p}^{(n+1)}\right)/P}; \end{cases} \tag{3.3}$$

where C and P are the number of exporting countries and exported products in the system, respectively. In Eqs. (3.3), $\widetilde{F_c}^{(n+1)}$ and $\widetilde{Q_p}^{(n+1)}$ are the intermediate values of $F_c^{(n+1)}$ and $Q_p^{(n+1)}$ obtained at each iteration $(n+1)$ [23]. At each step, the intermediate values are normalised by their algebraic means, in this way providing the final values $F_c^{(n+1)}$ and $Q_p^{(n+1)}$. The

normalisation is required for the stabilisation of the non-linear map in Eqs. (3.3) [125]. For the assessment of the final values of the metrics, ranking convergence has been proposed by the Authors as a possible solution [126]. This entails taking as a solution of the algorithm the one that ensures stable rankings of the values among successive iterations. In this case, a valid option is given by Lin *et al.*, [127], which propose to stop the algorithm at the iteration N, when the rankings between step N and step $N + \Delta N$, have a Spearman's correlation coefficient larger than 0.999. Note that, for the computation of the results in this work we have followed this criteria assuming $\Delta N = 10$.

It is not only the mathematics of the two approaches which is different, but also the obtained outcomes significantly diverge: as shown in Figure 3.2, the countries' rankings obtained with ECI_c and F_c widely scatter. This poses an issue of practical use of the economic complexity measures, potentially undermining the very essence of the economic complexity theory. We argue that the role played by EC measures in economics and policy making (see, e.g., [128–132]) requires more consistency in the outcomes of different methods, which we address later in this Chapter.

3.2 A General Framework for Economic Complexity

The introduction of a general framework of EC is necessary to set the ground upon which to build the reconciling of these two metrics. In a general framework, economic complexity theories aim at determining two properties X_c and Y_p – describing the complexity of country c and product p, respectively – by a system of coupled equations

$$\begin{cases} X_c = f(Y_1, Y_2, ..., Y_p, M_{cp}), & p = [1, ..., P], \\ Y_p = g(X_1, X_2, ..., X_c, M_{cp}), & c = [1, ..., C], \end{cases} \tag{3.4}$$

where f and g are linear functions and C and P are the number of countries and products considered in the analysis, respectively. To consider f and g as linear functions allows one to recast the determination of X_c and Y_p as the solutions of an eigen-problem of a suitable (approach dependent) transformation matrix $\mathbf{W}$, whose elements W_{cp} are derived from $\mathbf{M}$. In this case, these properties' values are obtained from the following coupled linear

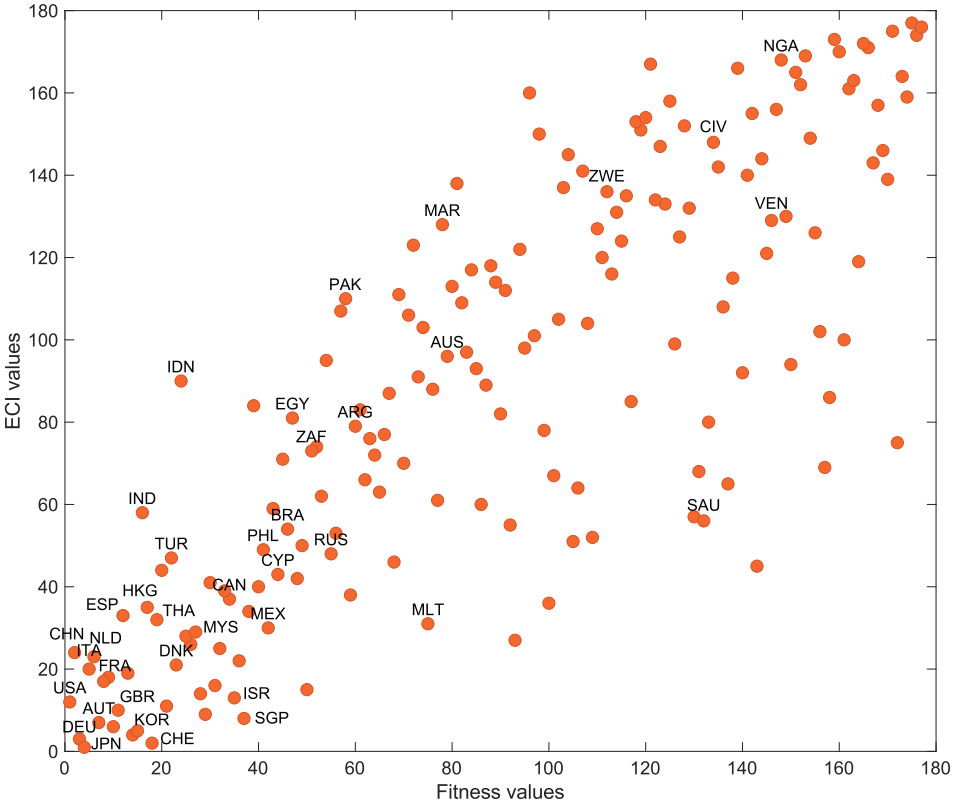

Figure 3.2: Comparison between the rankings provided by the Fitness and ECI values. The rankings sort countries according to decreasing complexity, as computed by the two metrics. Results refer to the trade data of year 2017.

equations:

$$\begin{cases} X_c = \frac{1}{\sqrt{\lambda}} \sum_p W_{cp} Y_p, \\ Y_p = \frac{1}{\sqrt{\lambda}} \sum_c W_{cp} X_c, \end{cases} \tag{3.5}$$

being λ the eigenvalue of the equivalent eigen-problem, such that the following relations hold

$$X_c = \frac{1}{\lambda} \sum_p \sum_{c^*} W_{cp} W_{c^*p} X_{c^*} = \frac{1}{\lambda} \sum_{c^*} N_{cc^*} X_{c^*}, \tag{3.6}$$

and

$$Y_p = \frac{1}{\lambda} \sum_c \sum_{p^*} W_{cp^*} W_{cp} Y_{p^*} = \frac{1}{\lambda} \sum_{p^*} G_{pp^*} Y_{p^*}. \tag{3.7}$$

A by-product of Eqs. (3.6) – (3.7) is that the squared, symmetric matrices

$$\mathbf{N} = \mathbf{W}\mathbf{W}^{\mathbf{T}} \tag{3.8}$$

and

$$\mathbf{G} = \mathbf{W}^{\mathbf{T}}\mathbf{W} \tag{3.9}$$

can be interpreted as proximity matrices for nations and products, respectively, where proximity defines similarity. In fact, the set of equations in Eqs. (3.5) involves the same transformation matrix $\mathbf{W}$. This entails that: the matrix $\mathbf{W}$ can be interpreted as the weighted incidence matrix of an undirected bipartite network uniquely describing the relations between countries and products according to the EC rationale. This would no longer be true if two different matrices were used for the mutual computation of X_c and Y_p. This also entails that we can interpret the symmetric squared matrices $\mathbf{N}$, for countries, and $\mathbf{G}$, for products, as the mathematical description of the weighted topology of two distinct, yet related, undirected networks [9, 116]. For what concerns countries, the network describes the connections among countries by weighting the similarities between their export baskets, and so, in their productive knowledge (i.e., $N_{cc^*} = N_{c^*c}$ describes the similarity in the export baskets between countries c and c^*). In case of products, the network describes similar requirements of productive knowledge. We stress that the feature of symmetry for the matrices $\mathbf{N}$ and $\mathbf{G}$ is essential to interpret them as proximity matrices, thus defining bijective functions of connections. Clearly, the grounding hypotheses about the hidden capabilities of countries – and on how these can be deducted by looking at the export baskets of countries upon which the EC algorithms are built – are preserved through these interpretations.

The eigen-problems in Eqs. (3.6) – (3.7) have multiple solutions, provided by the eigenvalues λ_t and the corresponding eigenvectors of the matrices $\mathbf{N}$ and $\mathbf{G}$, respectively [9]. In most situations, the eigenvector corresponding to the largest eigenvalue λ_1 carries the maximum amount of information [43] and it is thus taken as solution (although we will demonstrate the potential of combining more eigenvectors). In complex network jargon, X_c and Y_p are (eigen-)centrality metrics in the bipartite network of countries and products [47, 91] (Chapter 2, Section 2.4).

We now provide two examples of application of this general framework pertaining with the two aforementioned EC metrics, MR and FC, referring to these examples by using the superscripts A and B, respectively.

3.2.1 Recast of MR Metrics

The equation for the computation of ECI and PCI can be mapped in our general framework by using in Eqs. (3.5) the following transformation matrix

$$W^A_{cp} = \frac{M_{cp}}{\sqrt{k_c k_p}}, \tag{3.10}$$

in which, again, k_c and k_p are the degrees of countries and products, respectively, computed from the binary matrix $\mathbf{M}$. By solving the eigenvectors of the matrices

$$N^A_{cc^*} = \sum_p \frac{M_{cp} M_{c^*p}}{\sqrt{k_c}\sqrt{k_{c^*}} k_p}, \tag{3.11}$$

for countries, and

$$G^A_{pp^*} = \sum_c \frac{M_{cp} M_{cp^*}}{k_c \sqrt{k_p}\sqrt{k_{p^*}}}, \tag{3.12}$$

for products, the results of the Method of Reflections are provided using the formulas

$$\begin{cases} X^A_c = ECI_c \sqrt{k_c}, \\ Y^A_p = PCI_p \sqrt{k_p}, \end{cases} \tag{3.13}$$

being X^A_c and Y^A_p the eigenvector solutions of the matrices $\mathbf{N}^A$, Eq. (3.11), and $\mathbf{G}^A$, Eq. (3.12), respectively. Naturally, for any matrix there exist as many eigenvector solution as its dimensions, and the first eigenvector is usually taken for being the most informative [43]. In this case, the first eigenvectors $X^A_{c,1}$ and $Y^A_{p,1}$ carry a trivial information, since they equal the square roots of the degrees, k_c and k_p, thus leading to unitary ECI_c and PCI_p values, discarded in the original works for being uninformative [22, 48]. For this reason the eigenvectors $X^A_{c,2}$ and $Y^A_{p,2}$, corresponding to the

second largest eigenvalue, are taken by the Authors as the solution of Eqs. (3.6) – (3.7) [124]. We stress that the matrices $\mathbf{N}^A$ and $\mathbf{G}^A$ are symmetric ones thanks to the presence of the square roots of the degrees k_c and k_p, respectively, for which they differ from the corresponding asymmetric matrices that one would obtain directly from the original MR formulation [48]. Nevertheless, the mapping $\{X_{c,2}, Y_{p,2}\} \Leftrightarrow \{ECI_c, PCI_p\}$ completely preserves the information provided by the original methodology.

3.2.2 Recast of FC Metrics

The recast of the FC metrics onto our general framework requires a non-trivial linearisation of the relation between Quality and Fitness values.

Let us consider the system in Eqs. (3.3). This one can be written in closed and non-iterative form as

$$\begin{cases} F_c = c_F \sum_p M_{cp} Q_p, \\ Q_p = c_Q \frac{1}{\sum_c M_{cp} \frac{1}{F_c}}, \end{cases} \tag{3.14}$$

in which we have embedded the normalisation procedure by introducing the parameters c_F and c_Q. Eqs. (3.3) can be seen as the simplest numerical solution of Eqs. (3.14). The values of the coefficients c_F and c_Q can be obtained by simple considerations. As a consequence of dividing by their mean value, see Eqs. (3.3), the values of Fitness and Quality are normalised so that

$$\sum_c F_c = C; \quad \text{and} \quad \sum_p Q_p = P,$$

where C is the number of countries and P of products. Using the definition of F_c according to Eqs. (3.14), and its normalisation condition, clearly holds

$$c_F = \frac{C}{\sum_p Q_p k_p}.$$

From the definition of Q_p according to Eqs. (3.14), one has

$$c_Q = Q_p \sum_c \frac{M_{cp}}{F_c},$$

and by summing on all Q_p values holds

$$c_Q P = \sum_c \frac{1}{F_c} \sum_p Q_p M_{cp}.$$

Nevertheless, by recalling that F_c and Q_p are related through Eqs. (3.14), a further condition to be satisfied is

$$\sum_p Q_p M_{cp} = C/c_F,$$

and it must hold

$$c_F c_Q = \frac{C}{P}.$$

Therefore, the value c_Q is defined as

$$c_Q = \frac{\sum_p Q_p k_p}{P}.$$

Eqs. (3.14) represents a functional relationship between the vectors of values F_c and Q_p and, in particular, the Quality values can be formally expressed as

$$Q_p = h(F_1, F_2, ..., F_c), \quad c = [1, ..., C],$$

where $h(F_1, F_2, ..., F_c)$ is a non-linear function of the C–Fitness values. In order to map the FC algorithm onto the linear X_c – Y_p framework, we linearize the function $h(F_c)$ using the Taylor's series and expanding the function around the value $F_c = k_c$, which is known to dominate the information contained in F_c [47]. Moreover, k_c is the first result of the map at iteration $n = 1$ for the Fitness values [2]. In their closed, non linear form, the Q_p values are defined as (see Eqs. (3.14))

$$Q_p = c_Q \frac{1}{\sum_c \frac{M_{cp}}{F_c}}.$$

Its derivative in a generic point F_c^* is

$$\frac{dQ_p}{dF_c^*} = -\frac{c_Q}{(\sum_c \frac{M_{cp}}{F_c})^2} \cdot \Big(-\frac{M_{c^*p}}{F_{c^*}^2}\Big),$$

and substituting $F_c = k_c$ holds

$$\frac{dQ_p}{dF_c^*}\Big|_{k_c} = -\frac{c_Q}{(\sum_c \frac{M_{cp}}{k_c})^2} \cdot \Big(\frac{M_{c^*p}}{k_{c^*}^2}\Big),$$

[2] Since the values F_c are normalised so to have unitary mean, so should the values of the degree around which to expand the function $h(F_c)$. Therefore, we introduce the normalised degree of countries $k_c^+ = \frac{k_c}{K_{tot}} C$, in which $K_{tot} = \sum_c k_c$ and C is the number of countries. Nevertheless, we can omit this normalisation from the linearisation procedure, since it does not affect the results.

The complete expression of the Taylor's expansion of Q_p in $F_c = k_c$ is

$$Q_p = \frac{c_Q}{\frac{M_{cp}}{k_c}} + \sum_c \frac{dQ}{dF_c} \cdot (F_c - k_c) = \frac{c_Q}{\left(\sum_c \frac{M_{c^*p}}{k_{c^*}}\right)^2} \sum_c \frac{M_{cp} F_c}{k_c^2}.$$

Therefore, the linear expression to evaluate the Fitness of countries and the Quality of the products, is

$$\begin{cases} F_c \simeq c_F \sum_p M_{cp} Q_p, \\ Q_p \simeq \frac{c_Q}{(k'_p)^2} \sum_c \frac{M_{cp} F_c}{k_c^2}, \end{cases} \tag{3.15}$$

where $k'_p = \sum_c M_{cp}/k_c$. Notice that the system in Eqs. (3.15) is an eigenproblem, thus it can be solved without the use of iterative algorithms. This avoids the convergence problem which is known to affect the system in Eqs. (3.3)) [126], due to the hyperbolic nature of the second equation [125].

Taking the linearised equations in Eqs. (3.15) as the starting point, the mapping of FC metrics within our framework is given by setting in Eqs. (3.5) the value

$$W_{cp}^B = \frac{M_{cp}}{k_c k'_p}, \tag{3.16}$$

with $k'_p = \sum_c M_{cp}/k_c$. By solving the eigenvectors of the matrices

$$N_{cc^*}^B = \sum_p \frac{M_{cp} M_{c^*p}}{k_c k_c^* (k'_p)^2}, \tag{3.17}$$

for countries, and

$$G_{pp^*}^B = \sum_c \frac{M_{cp} M_{cp^*}}{k_c^2 k'_p k'_{p^*}}, \tag{3.18}$$

for products, the results of the Fitness and Complexity algorithm are obtained using the formulas

$$\begin{cases} X_c^B = F_c/k_c, \\ Y_p^B = Q_p \cdot k'_p, \end{cases} \tag{3.19}$$

where we neglected the scaling factors c_F and c_Q, since their role of stabilising the numerical values is not anymore required due to linearity, thus reducing the number of unknowns in the system. The linearised values for Fitness and Quality are recovered from the eigenvectors of the proximity matrices $\mathbf{N}$ and $\mathbf{G}$ associated to the largest eigenvalue λ_1, from which holds $X_{c,1} = F_c/k_c$ and

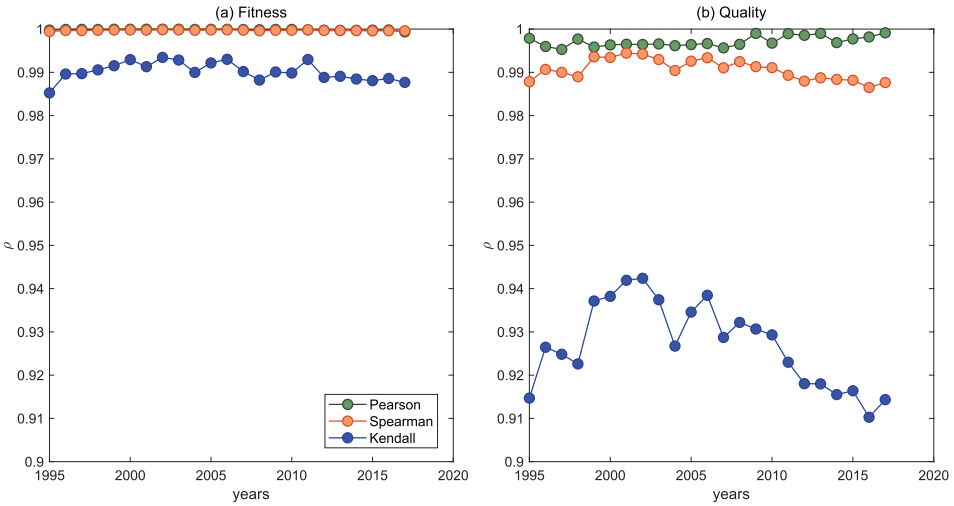

Figure 3.3: Correlation coefficients among the non-linearly and the linearly computed values of Fitness, panel (a), and Quality, panel (b). In green is the Pearson's correlation coefficient, while the Spearman's one is in orange and Kendall's in blue. The Spearman's and Kendall's coefficients are ranking-based, while the Pearson's one compares the values between the two vectors. As the coefficients witness, the linearised algorithm almost perfectly reproduces the outcomes of the non-linear one.

$Y_{p,1} = Q_p k'_p$, respectively. In Figure 3.3, we plot the correlation coefficients obtained by comparing the terms $X^B_{c,1}$ and F_c/k_c (or, equivalently, $X^B_{c,1}k_c$ and F_c) for the Fitness values – panel (a) –, and $Y^B_{p,1}$ and $Q_p k'_p$ (or $Y^B_{p,1}/k'_p$ and Q_p) for the Quality values – panel (b) –. As the Figure shows, this linearization almost entirely preserves the information of the non-linearly computed values, also independently of the kind of indicator used to measure correlation. Slightly smaller values of correlation are computed using the Spearman's correlation between the Quality values. A possible explication may lie in the obscillation of the iteratively computed values, which may change the ranking of the Quality values due to the choice of the convergence criteria. Instead, our linear formulation does not suffer from the well-known convergence problems of the iterative FC algorithm [126] and provides more regular solutions, lacking the necessity of defining any convergence criteria.

Some more comments about this mapping are necessary. The fact of having found very similar results between the linear and the non-linear versions of the FC algorithm (on average, 99.5% Pearson's correlation, see Figure 3.3) cannot be systematically generalised to other cases: in fact, some bipartite systems may require a genuine non-linear approach to let their nested nature emerge. Nestedness describes the pyramidal

structure of some complex systems, especially the ecological ones as the plants-pollinators network, in which the neighbours of lower-degree nodes are the neighbours of higher-degree nodes [133]. Non-linearity has been shown to be an important feature of the algorithms for packing the entries of the incidence matrix describing such systems, thus reducing their disorder (i.e., the so-called *nestedness temperature* [134]) and letting their nested nature emerge [127]. As such, the original FC algorithm has shown very good potential in minimizing the nestedness temperature of ecological networks [127]. Therefore, we have tested the packing performance of the linearised form of the FC algorithm, similarly to the comparison showed in [127]. We exemplify the results through the analysis of two pollination networks provided by The Web of Life project and available at www.web-of-life.es (network IDs M_PL_062 and M_PL_015). The networks describe the pollination phenomena among plants and pollinators in two different sites. As Figure 3.4 shows, the non-linear algorithm outperforms the linearised form in its capability of maximising the nestedness of the incidence matrices of the two systems. Instead, there are no significant differences between the non-linear and the linear algorithm for maximising the data-packing of the trade matrix, suggesting that there exist bipartite systems where non-linearity plays a minor role for nestedness evaluation. In the case of trade, we speculate that this might be related to the differences in the decision making processes ruling these systems. On the one hand, e.g., nested ecological networks self-organise following ecological rules of non-linear population dynamics [135]. These systems are thus driven by more rigid, evolutionary decision making processes. On the other hand, the plastic human decision-making process – which is of course at the base of the trade network self-organisation – may give rise to less nested network structures: for a given productive knowledge, trade may follow a simpler sum rule, i.e., "the more, the better", as trade enhances growth [136]; thus clarifying the reason why the diversity of a country is used as a first proxy of the productive knowledge itself.

3.2.3 Comments on the General Framework

The original ECI_c, PCI_c, F_c and Q_p variables are recovered within our general framework through simple (but non-trivial) mappings from X_c and Y_p. The use of the variables X_c and Y_p allows one to gain neatness in the mathematics, reflected by the fact that the matrices $\mathbf{N}$ and $\mathbf{G}$ can be considered as suitable proximity matrices containing information about the similarities among countries and products, respectively. As we will discuss

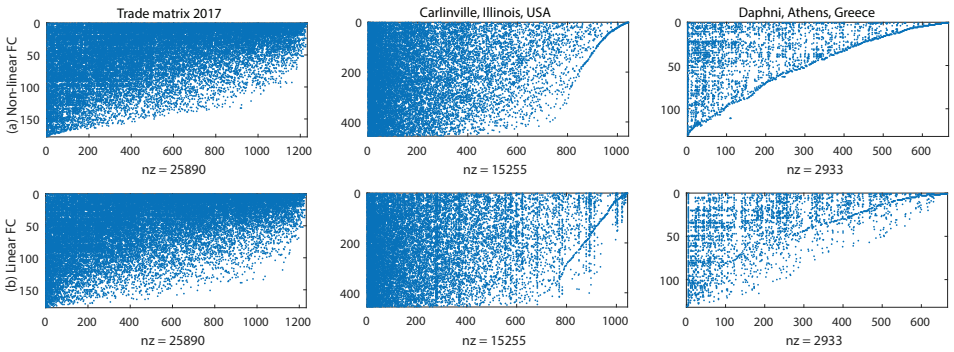

Figure 3.4: Nestedness maximization performances. Visual comparison among the performances of the non-linear FC algorithm (panels (a)) and its linearized form, (panels (b)) in maximizing nestedness of matrices. The left panels refer to the countries-products bipartite network during 2017. The central panels refer to the network of pollination in Carlinville, Illinois, USA (network ID: M_PL_062); on the right the one referring to the pollination in Daphní, Athens, Greece (network ID: M_PL_015). Data for the pollination networks are freely available at www.web-of-life.es.

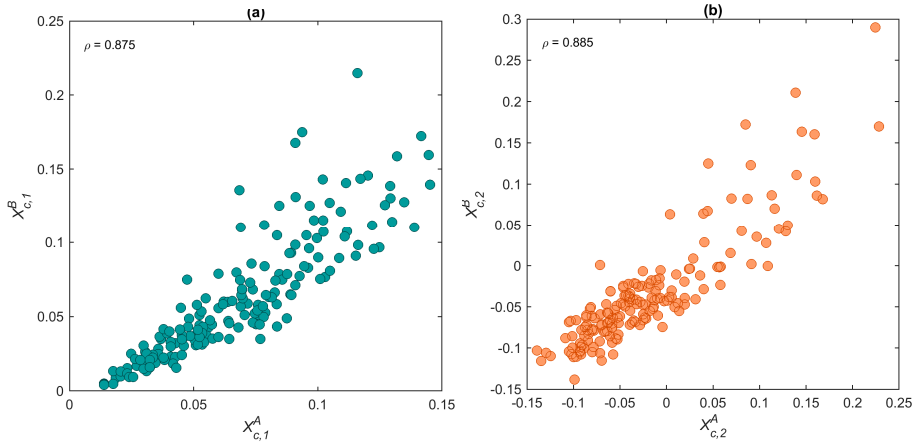

Figure 3.5: Scatter plots comparing the eigenvectors of the proximity matrices N^A and N^B. The matrices are computed using the transformation matrices $W^A_{cp} = M_{cp}/\sqrt{(k_c k_p)}$, Eq. (3.10), and $W^B_{cp} = M_{cp}/(k_c k_p^{'})$, Eq. (3.16), respectively, as the starting point for the computation of the proximity matrices $\mathbf{N}$, Eqs. (3.11) – (3.17). The left plot (a) compares the eigenvectors $X^A_{c,1}$ and $X^B_{c,1}$ associated to the largest eigenvalues λ^A_1 and λ^B_1 of the two proximity matrices, $\mathbf{N}^A$ and $\mathbf{N}^B$. The right plot (b) compares the eigenvectors corresponding to the second largest eigenvalues λ^A_2 and λ^B_2, namely $X^A_{c,2}$ and $X^B_{c,2}$. The eigenvectors are normalised such that the 2-norm is unitary, i.e., $\sqrt{\sum_c X^2_{c,i}} = 1$, with $i = [1, 2]$. As the correlation coefficients highlight, the eigenvectors from the two matrices carry similar information. The correlation coefficients are of the Pearson's kind.

further in this Chapter, this aspect may have important consequences on the interpretation of the economic significance of these metrics.

Moreover, the matrices $\mathbf{W}^A$ and $\mathbf{W}^B$, Eqs. (3.10) – (3.16), respectively, differ for the specific scaling factors adopted on the matrix $\mathbf{M}$. It is hard to recognise an economic (or a mathematical) basis on how the factors are determined, and this leaves no solid ground for a potential user to decide which approach, between MR and FC, to follow. Notwithstanding the differences among $\mathbf{W}^A$ and $\mathbf{W}^B$, the eigenvectors $X^A_{c,1}$ and $X^B_{c,1}$, and also $X^A_{c,2}$ and $X^B_{c,2}$, carry similar information, as witnessed by the scatter plot between the two pairs in Figure 3.5 (this is also partially true for Y_p, see Section 3.3). Therefore, the divergences between F_c and ECI_c – and corresponding outcomes – shown in Figure 3.2 should be mainly attributed to the fact that eigenvectors of different order are considered in the two approaches. Hence, the two metrics bring different information; albeit different, this information is relevant for both metrics, as demonstrated by numerous practical applications of the two approaches [2, 127–130, 132, 137, 138]. Therefore, an integrated measure would help in uniquely defining the complexity of countries and products exploiting both information of the MR and FC algorithms.

3.3 The Generalized Economic Complexity Index

In light of the analytical results shown in Section 3.2, we propose to distil the information that both methodologies of economic complexity provide into a GENeralised Economic comPlexitY index, GENEPY (on behalf of the notorious herb-based distillate typical of the north-western part of Italy). Through this integrated measure of complexity, we aim at identifying central nodes in the bipartite network of countries and products, in line with the general framework presented in Section 3.2.

In order to construct an integrated measure of complexity, one has to choose which mapping, between the MR and FC one, to take as input of the general framework. Since the information carried by the eigenvectors are similar (Figure 3.5), either using $\mathbf{W}^A$ or $\mathbf{W}^B$, Eqs. (3.10) – (3.16), respectively, to develop the new integrated measure of complexity would lead to reliable and comparable results. We lean toward the use of $\mathbf{W}^B$, the one related to the FC method, for the following reasons: on the one hand, the first eigenvector $X^A_{c,1}$ – from which, using Eqs. (3.13), the unitary first eigenvector of MR is recovered – equals $\sqrt{k_c}$, thus carrying no added information beyond diversity (and the same holds for products); on the

other hand, the last update on the MR method, named ECI+ [139], has been shown to be equivalent to the non-linear FC algorithm [125], thus implicitly supporting the idea that FC carries more information then MR. From here on, we will thus use the matrix $\mathbf{W}^B$ in Eqs. (3.5) and drop the superscript B in the mathematical notation.

The GENEPY index exploits the neatness of the proposed general framework on economic complexity and it is computed by employing the statistical setting introduced in Chapter 2. The interpretation of the matrices $\mathbf{N}$ and $\mathbf{G}$ as adjacency matrices of undirected, weighted networks (describing the proximity among countries and products, respectively) allows us to consider *multi-dimensional* metrics as the solution of the centrality problem at hand, which are able to capture different facets of the network structure.

We detail the procedure for the derivation of the GENEPY index for countries, which also applies for the computation of the GENEPY index for products. In fact, it is sufficient to replace in the following the terms $X_{c,1}$, $X_{c,2}$ and $\mathbf{N}$ with $Y_{p,1}$, $Y_{p,2}$ and $\mathbf{G}$, respectively, to obtain the GENEPY index for products.

Let us consider the network of similarities among countries described by $\mathbf{N}$. Let ζ be a centrality-dependent estimator function. In the case of the eigenvector centrality, better estimate results are found when the function linearly depends on the eigenvectors $X_{c,1}$ and $X_{c,2}$, corresponding to the two largest eigenvalues λ_1 and λ_2 of the matrix $\mathbf{N}$ we aim at estimating [38, 43, 91] (see Chapter 2, Section 2.1.3 on multicomponent estimators). In formulas

$$\zeta(\lambda_t, \mathbf{X}_{c,t}, \mathbf{X}_{c^*,t}) = \sum_{t=1}^{2} \lambda_t X_{c,t} X_{c^*,t}, \tag{3.20}$$

where $t = [1, 2]$ and c and c^* run in the range $[1, C]$, being C the number of nodes, i.e., countries, in the network. The function ζ minimizes the squared errors between the matrix elements and the corresponding estimates; namely

$$SE = \sum_{c}^{C} \sum_{c^*}^{C} \Big(N_{cj} - \zeta(\lambda_t, X_{c,t}, X_{c^*,t}) \Big)^2. \tag{3.21}$$

As shown in Chapter 1, Section 2.1.3 on multicomponent estimators, at a fixed t^*, each eigenvector $\mathbf{X}_{t^*}$ solves the minimisation problem

$$\frac{\partial SE}{\partial \mathbf{X}_{t^*}} = 0.$$

Again, in this muldimensional setting on eigenvector centrality, the ranking of the network' nodes (i.e., countries in the newtork described by

N, products if **G** is considered, instead) is given by the adoption, from the commonality analysis, of the concept of the unique contribution of the $X_{c,t}$ variables. The unique contribution is defined as the drop in the coefficient of determination R^2 induced by excluding the variables $X_{c,t}$ ($t = [1, 2]$) considered in the estimator function ζ, in Eq. (3.20), from the estimation procedure (see Section 2.1.2). The core concept of the unique contribution is that, the larger the drop, the larger is the contribution of the c-th values in the reconstruction of the matrix **N** and, in this application, the more central – thus complex – is the c-th node in the network topology under analysis. Hence, according to this approach, we define the GENeralised Economic comPlexitY index (GENEPY) for the country c as the unique contribution of its complexity values $X_{c,t}$ as

$$GENEPY_c = \Big(\sum_{t=1}^{2} \lambda_t X_{c,t}^2\Big)^2 + 2\sum_{t=1}^{2} \lambda_t^2 X_{c,t}^2, \tag{3.22}$$

where $X_{c,1}$ and $X_{c,2}$ are the eigenvectors corresponding to the first two largest eigenvalues λ_1 and λ_2 of the proximity matrix

$$\begin{cases} N_{cc^*} = \sum_p W_{cp} W_{c^*p} = \sum_p \frac{M_{cp} M_{c^*p}}{k_c k_{c^*} (k'_p)^2}, & \text{if } c \neq c^*, \\ N_{cc^*} = 0, & \text{if } c = c^*, \end{cases} \tag{3.23}$$

in which the redundant information of the self-proximity is deleted setting all diagonal elements to an arbitrary constant value (we set this value to zero).

Notice that, differently from the equation to compute the unique contribution for multi-component centrality, Eq. (2.38), in here we neglect the constant term TSS without affecting the results.

3.4 Results

We exemplify the use of the GENEPY index by considering the international trade of goods during the years 1995 – 2017. Import-export data during this period are extracted from the BACI-CEPII dataset [140], which classifies goods according to the Harmonised System Codes 1992 (HS-1992) at the 6-digits level. To allow comparability with previously published results, we downscale the classification of traded goods to the 4-digits level (thus aggregating products according to their 4-digits categories). Our data include all the countries whose export share is worth at least 10^{-5} of the total flux traded during the year (i.e., the total amount of dollars

exported worldwide). This filters the noise arising by small export baskets. The Relative Comparative Advantage procedure is used to construct the incidence binary matrix **M**, setting the threshold of RCA to 1 in line with the economic complexity framework [22] (Section 3.1).

3.4.1 Countries' GENEPY

In Figure 3.6, we show the GENEPY index for countries and the results are processed for the 2017 trade. Figure 3.6a displays the position of countries on the $\{X_{c,1}\ ,\ X_{c,2}\}$ plane. Most economies with a high drive for innovation and technology [141] – such as the UE-28 countries, Switzerland (CHE), China (CHN), Japan (JPN), Singapore (SGP) and the United States of America (USA) – are found far from the origin. This entails the presence of top-quality products among their exports and, therefore, of relevant productive knowledge. Less economically-stable economies, such as those of many African and South-American countries, are located in the bottom left part of the graph. The GENEPY index also identifies potentially top-competitive countries, such as Australia (AUS) and Canada (CAN), struggling to boost their complexity due to remoteness and resources-dependency, well-known factors for affecting trade and economic growth [142–144]. The information distilled through the GENEPY index can be better understood by considering the meaning of its components, i.e., the two eigenvectors $X_{c,1}$ and $X_{c,2}$, as contextualised in complex network theory [9]. In fact, the elements of the first eigenvector represent the eigenvector centrality of the countries as obtained from the proximity matrix **N**, interpreting the matrix as the weighted, adjacency matrix of an undirected network connecting the countries for the similarities in their export baskets. Instead, the values of $X_{c,2}$ cluster countries according to the similarities in their export baskets. In fact, the strict nexus between $X_{c,2}$ and ECI_c recalls the results provided in [145], where the Authors proved that ECI perfectly solves a spectral clustering algorithm. Interpreting this result within the network of similarities designed by **N**, the GENEPY centrality identifies that set of capabilities (contributing to the productive knowledge) a country owns and shares with others. In this sense, more central nodes are found within a cluster of highly competitive countries, while less complex countries are found moving towards the borders of the graph. This result is confirmed by the reordering of the matrix **N** according to the GENEPY values (see Figure 3.7, Table B.1), showing that countries with higher complexity share similar sets of capabilities, as their export baskets are similar.

As mentioned, our framework combines the advantages – and information

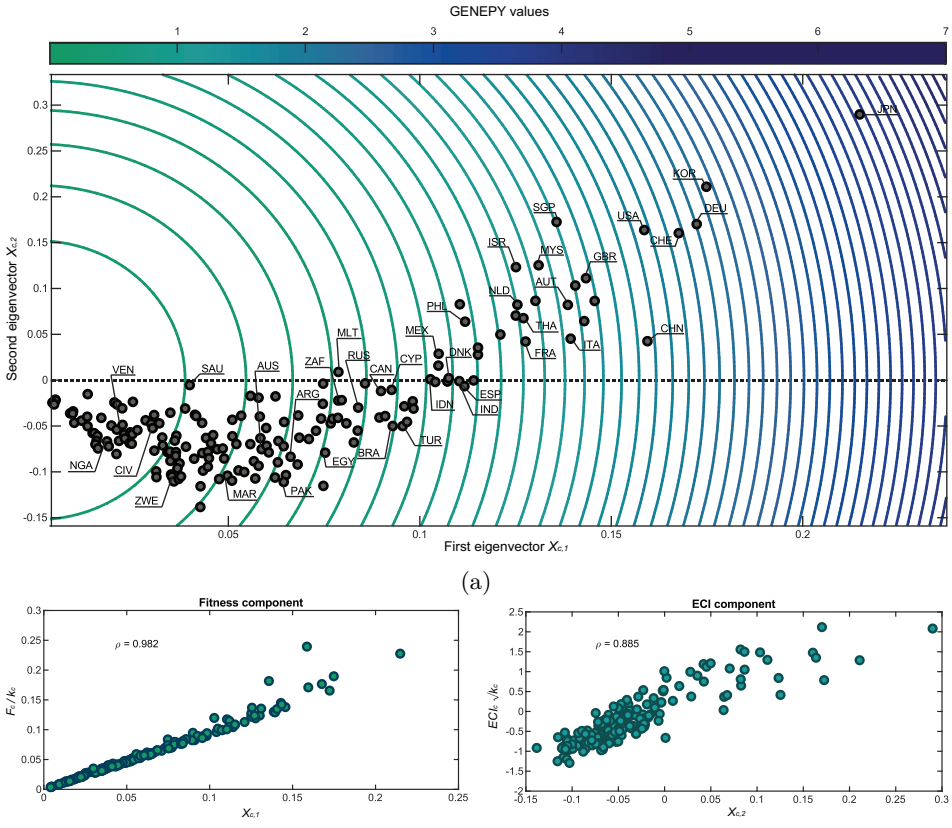

Figure 3.6: The GENEPY index and its components. **(a)** $\{X_{c,1}\ ,\ X_{c,2}\}$ plane and $GENEPY_c$ from the data of 2017 international products' trade. The x-axis reports the components of the first eigenvector $X_{c,1}$, whilst the y-axis the components of the second eigenvector $X_{c,2}$. The eigenvectors are normalised such that their 2-norm is unitary, i.e., $\sum_c X_{c,1}^2 = \sum_c X_{c,2}^2 = 1$. Contours range from lower $GENEPY_c$ values (green) to higher ones (blue). **(b)** Fitness component. Scatter plot of the first component $X_{c,1}$ compared with the values of the Fitness values F_c rescaled by the countries degree k_c (see Section 3.2.2, Eqs. (3.19)). **(c)** ECI component. Scatter plot of the second component $X_{c,2}$ compared with ECI_c values rescaled by the term $\sqrt{k_c}$ (see Section 3.2.1, Eqs. (3.13)). The correlation coefficient in the plots **(b)** and **(c)** is of the Pearson's kind. Figures have been produced with MATLAB 2019b.

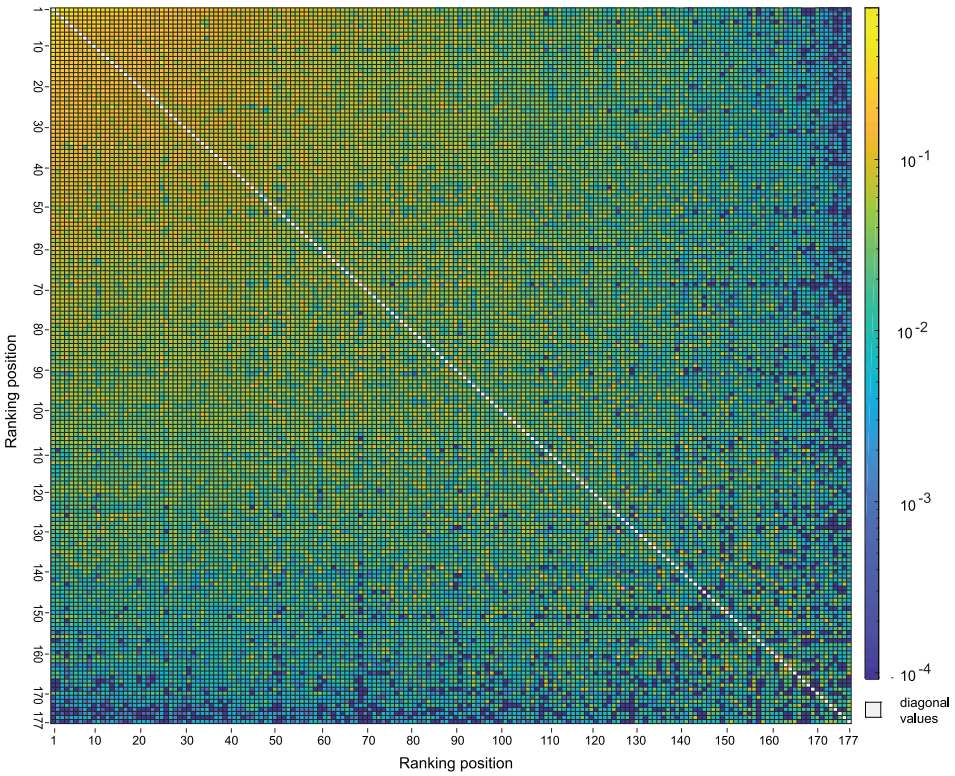

Figure 3.7: The elements N_{cc^*} of the similarity matrix **N** for the 2017 trade. Rows and columns reordered top-to-bottom, left-to-right, according to decreasing values of GENEPY complexity. More complex countries are found on the top (left) of the matrix. Correspondence among ranking positions and countries are defined in Annex B, Table B.1.

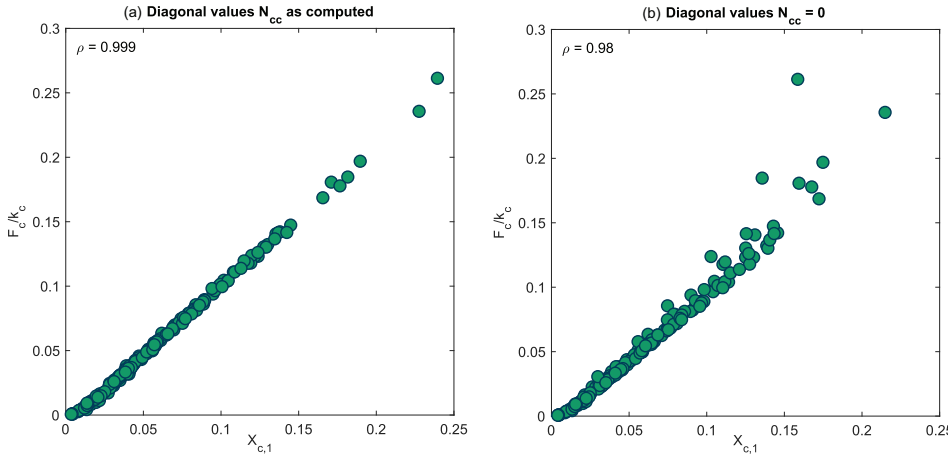

Figure 3.8: Scatter plots of the eigenvectors $X_{c,1}$ and F_c/k_c for different interpretations of the matrix $\mathbf{N}$. On the left, the eigenvector $X_{c,1}$ belongs to the matrix N with diagonal values set to zero. On the right, $X_{c,1}$ is the eigenvector of the matrix N in which we left the diagonal values as computed, i.e., $N_{cc} = \sum_p M_{cp}M_{c^*p}/k_c k_c (k'_p)^2$ (see Eq. (3.23)). Data refer to year 2017.

– of the two existing metrics of economic complexity, ECI and Fitness. On the one hand, the countries' Fitness values obtained with the iterative FC method are recovered, with great accuracy, from the product of the first eigenvector $X_{c,1}$ with k_c (see Figure 3.6b and Eqs. (3.19)). The very small deviations from the 1:1 line shown in Figure 3.6b are not induced by the linearisation procedure. In fact, they disappear when the equation

$$N_{cc^*} = \sum_p \frac{M_{cp}M_{c^*p}}{k_c k_{c^*}(k'_p)^2}$$

is used also for $c = c^*$, i.e., when the matrix $\mathbf{N}$ is not interpreted as a proximity matrix, thus leaving the diagonal values as computed (see Eq. (3.17) and Figure 3.8). However, this would imply inflating the F_c (or $X_{c,1}$) values for countries with large self-interactions, which, in our opinion, induces an undesired bias in the results. On the other hand, a good proxy of the ECI_c values is obtained by dividing the values of the second eigenvector $X_{c,2}$ by $\sqrt{k_c}$, as shown in Figure 3.6c (see Eqs. (3.13)). In this case, the scatter of the plot is due to the differences in the matrices $\mathbf{N}^A$ and $\mathbf{N}^B$, Eq. (3.11) and Eq. (3.17), respectively. The same reasoning applies for the Quality values, as will be further discussed later in this Chapter.

Being the GENEPY framework grounded on both existing indicators of economic complexity (the FC and MR algorithms), it inherits the intuitions

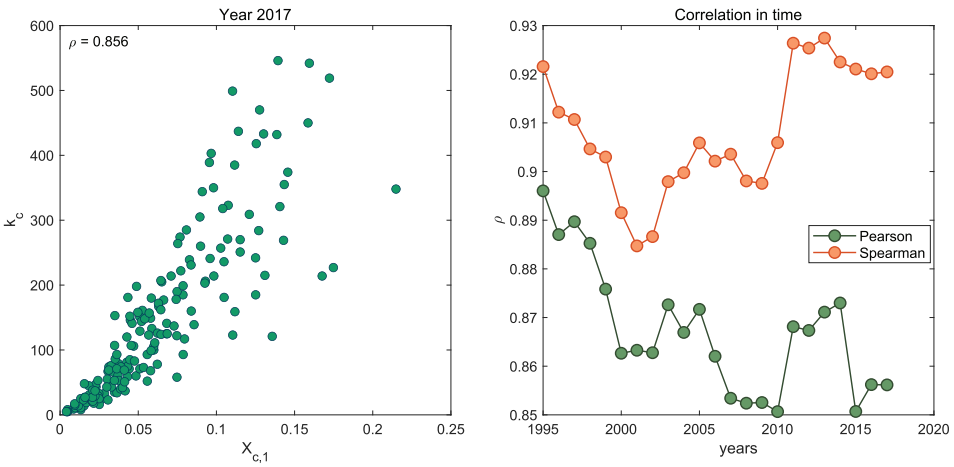

Figure 3.9: Correlation between $X_{c,1}$ and k_c. In panel (a), the values of $X_{c,1}$ and k_c obtained from the trade data of year 2017 are scattered. In panel (b), the plot shows the values of the correlation coefficients between the two vectors during time. The Pearson's correlation is green colored, the Spearman's one is in orange, instead.

and rationales upon which these two metrics are built: the capabilities of countries to export diversely complex goods are hidden within the bipartite network of countries and exports, under which they combine to maximise the complexity of the goods. Also, since $X_{c,1}$ maintains a very high correlation with k_c (see Figure 3.9), our framework preserves the information on diversification, which is a relevant one to understand how export capabilities are exploited by countries.

3.4.2 The Knee-Like Shape

The positions of countries in the plane $X_{c,1} - X_{c,2}$ arrange in a knee-like shape, as displayed in Fig 3.6a. The presence of this shape, recurrent in all the years of analysis, shows the existence of a functional relationship between the two eigenvectors, which reasons of existence are related to linear algebra and network science. Let us define a functional relationship f between $X_{c,1}$ and $X_{c,2}$ such that

$$X_{c,2} = f(X_{c,1}) + \epsilon_c, \tag{3.24}$$

where ϵ_c are the errors. We assume the errors to have null expected value, i.e., $E(\epsilon_c) = 0$ and to be orthogonal to $X_{c,1}$, s.t., $\sum_c X_{c,1}\epsilon_c = 0$. There exist some constraints related to the existence of the eigenvectors of a symmetric matrix, which any functional relationship should respect:

1. the eigenvectors corresponding to distinct eigenvalues of a symmetric squared matrix are, by definition, orthogonal and this entails that the inner product of the vectors is zero, i.e., $\sum_c X_{c,1}X_{c,2} = 0$ [43];

2. for the Perron-Frobenius theorem [43], the eigenvector corresponding to the largest eigenvalue is strictly positive, such that $X_{c,1} \geq 0,\ \forall\ c = [1, \ldots, C]$ (number of countries);

3. we can normalise the eigenvectors such that the 2-norm is unitary, i.e., $\sum_C X_{c,1}^2 = \sum_c X_{c,2}^2 = 1$;

4. if any element c^* of the eigenvector corresponding to the first (largest) eigenvalue λ_1 is zero, the same element is null also within the successive eigenvectors. In fact, the eigen-equation for the matrix $\mathbf{N}$ is [43]:

$$X_{c^*,1}\lambda_1 = \sum_c N_{cc^*}X_{c,1};$$

and, because of condition (2), it holds that $X_{c^*,1} = 0$ iff $\sum_c N_{cc^*} = 0$, i.e., if the matrix has null elements all along the column (or row) c^*.

Interpreting this result through network science lenses, the node to which the null element of the eigenvector refers is disconnected in the network [9]. Therefore, in the hypothesis of existence of any functional relationship between two eigenvectors as in Eq. (3.24), it must hold $f(0) = 0$. The only function able to satisfy the requirements above described is a squared one. We can prove it by exploring two different scenarios of possible functional relationship for Eq. (3.24).

SCENARIO 1: The simplest form considers f as a linear function, i.e.,

$$X_{c,2} = aX_{c,1} + \epsilon_c. \tag{3.25}$$

By imposing the orthogonality condition (1) to Eq. (3.25) one obtains:

$$\sum_C X_{c,1}X_{c,2} = \sum_c X_{c,1}(aX_{c,1} + \epsilon_c) = a\sum_c X_{c,1}^2 + \sum_c X_{c,1}\epsilon_c = 0.$$

Since the errors are orthogonal to $X_{c,1}$ and the 2-norm of the vector is unitary for condition (3), the solution is $a = 0$, which entails no functional relationship exists between $X_{c,1}$ and $X_{c,2}$.

SCENARIO 2: In this case, the function is a polynomial of the second order one, namely

$$X_{c,2} = aX_{c,1} + bX_{c,1}^2 + \epsilon_c. \tag{3.26}$$

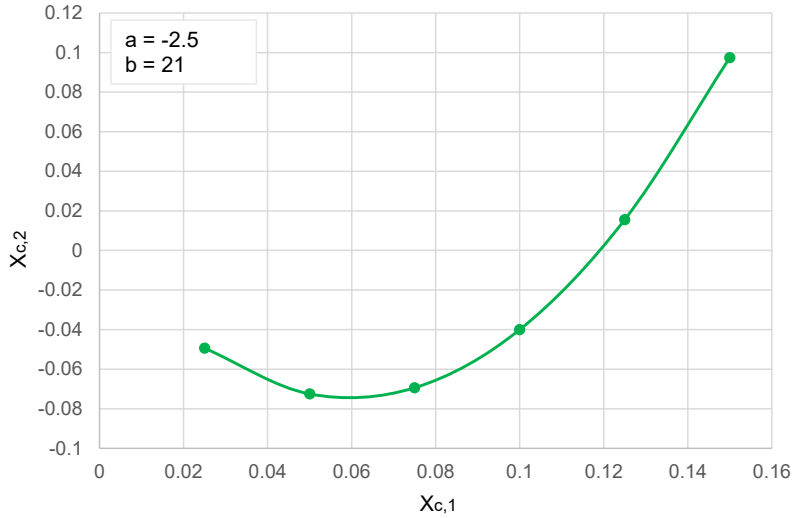

Figure 3.10: Possible values of the parameters a and b of the knee-shape function. These parameters reproduce the existence of a minimum point in $X_{c,1} = 0.05$ and an upward belly, as the one showns in Fig. 3.6a.

Again, by applying the orthogonality condition (1), one has

$$\begin{aligned}\sum_c X_{c,1}X_{c,2} &= \sum_c X_{c,1}(aX_{c,1} + bX_{c,1}^2 + \epsilon_c) \\ &= a\sum_c X_{c,1}^2 + b\sum_c X_{c,1}^3 + \sum_c X_{c,1}\epsilon_c = 0,\end{aligned}$$

which leads to

$$a + b\sum_c X_{c,1}^3 = 0. \tag{3.27}$$

Because of condition (2), the term $\sum_c X_{c,1}^3$ is strictly positive and in order to respect Eq. (3.27), the values of the parameter a and b should have different signs, thus justifying the existence of the knee-like shape. In particular, the upward belly of the relation is given for negative values of the parameter a and positive values of b. In this sense, the minimum point depends on the parameters. One possible fit of the quadratic equation of the knee-shape in Fig 3.6a is given in Fig 3.10, created using the values $a = -2.5$ and $b = 21$ and simulating the ranges of the eigenvectors as obtained by the GENEPY framework on 2017 trade.

We can interpret this result through the meaning of the matrix and its eigenvectors in the context of network theory. In fact, in this case the eigenvectors of the matrix describe the structural properties of the network

[146] and are related to the similarity of the network among the countries. In fact, the shape of the matrix **N**, Fig 3.7, represents a connected network in which a stronger connected component can be spotted, constituted by the top-GENEPY countries, while weaker connections characterize the countries at the periphery. In this weak connection component, the correlation between the two eigenvectors is positive [147]. Also, the mutual signs of the elements of the eigenvectors corresponding to the two largest eigenvalues – whether these are positive or negative in the second eigenvector – acquires a meaning [147], thus justifying the presence of three areas (or groups) in which the points can stand:

- both values $X_{c,1}$ and $X_{c,2}$ are low; these nodes belong to the weaker component and they have no important connections with the stronger connected component;
- both values $X_{c,1}$ and $X_{c,2}$ are high; these nodes belong to the stronger and more connected core of the network, thus defining the area of complexity in which these points (countries) are competitors, also in the sense of collecting most of the links in terms of similarities;
- low values of $X_{c,1}$, high values of $X_{c,2}$, or *vice-versa*: this situation identifies the presence of some "outliers" of the core and periphery components. These nodes connect the stronger and the weaker components and have a role in bridging the gaps across the network.

These three arrangements of the points along the knee-like shape can evolve in time, letting the dynamical regimes of growth emerge as we analyse the aggregated dynamics of countries in time.

3.4.3 The Trajectories of Economic Growth

The ability of the proposed multidimensional index to assess the sophistication of countries' export-baskets and, simultaneously, define clusters of economic growth can be exploited to track the path toward prosperity of countries as driven by economic complexity. In fact, according to the economic complexity theory, a country's acquisition of capabilities, employed in the production – and hence export – of goods [4, 48, 49] is a determining factor for its economic growth. Any country at a lower stage of growth uses its increasing capabilities to fill its export basket with higher-quality goods, possibly similar to those traded by countries at higher stages of growth. This entails the creation of a wider export basket allowing the country to gain momentum in the market. Also, in order to boost its economic complexity

– and growth – such a country may enlarge its offer including products for which it can be considered the only relevant exporter, hence gaining advantage [51]. By analysing the aggregated displacements of countries in time from 1995 to 2017, it is possible to identify three regimes of growth. In fact, for each country whose continuous data in time are available (154 countries), we defined the main displacement recorded by the country during the period of analysis. To identify the starting and ending points of such displacements, we compute the center of mass of $X_{c,1}$ and $X_{c,2}$ during the first 3 years of analysis (1995 – 1998) and their center of mass during the last 3 years (2014 – 2017), respectively. These points are then connected for each country to identify their displacements. In order to make the overall dynamics clearer, we defined overlapping classes of countries using a moving window of 20 countries per each class. Firstly, we ordered the countries (and respective scores of the eigenvectors in time) for increasing starting $X_{c,1}$ values. Secondly, by defining each class through a window of 20 countries, we computed the resultant vector of the displacements of the countries falling in that class. Lastly, we applied the resultant vectors to the barycenter of the starting points of the single vectors that fall into the class. In Figure 3.11, we show the aggregated dynamics of countries along the knee-shape. The colours sort the vectors for their length, as normalised for the longest vector recorded (light blue identifies the shorter ones, light purple the longer ones). The light blue vectors on the bottom left part of the knee identify the *Impasse*: the dynamics of the countries in this area are here tangled, as shown by the horizontal displacement of the vectors, within the borders delimited by low values of $X_{c,1}$ and negative values of $X_{c,2}$. Notwithstanding the presence of some uplift movements of the classes around the minimum point in $X_{c,1} = 0.05$, the countries within this area may suffer from lack of skills, human and capital investments and resources, thus resulting in low productive knowledge and, consequently, reduced diversification and complexity [51]. These countries hence face an *impasse* condition, resulting in a saddle point of growth and poor growth potential. As soon as countries reinforce their knowledge, they experience higher values of $X_{c,2}$ complexity until these values approach to zero: here it starts the *Bounce*, where countries boost their diversification and complexity, turning cluster membership by joining the more economically grown countries club and thus increasing the similarity in the export basket with them. The bounce is marked by the crossing of the zero value of the y-axis and this area defines the increment in quantity and quality of the exports. Longer vectors in violet and light purple highlight the jump. Once the

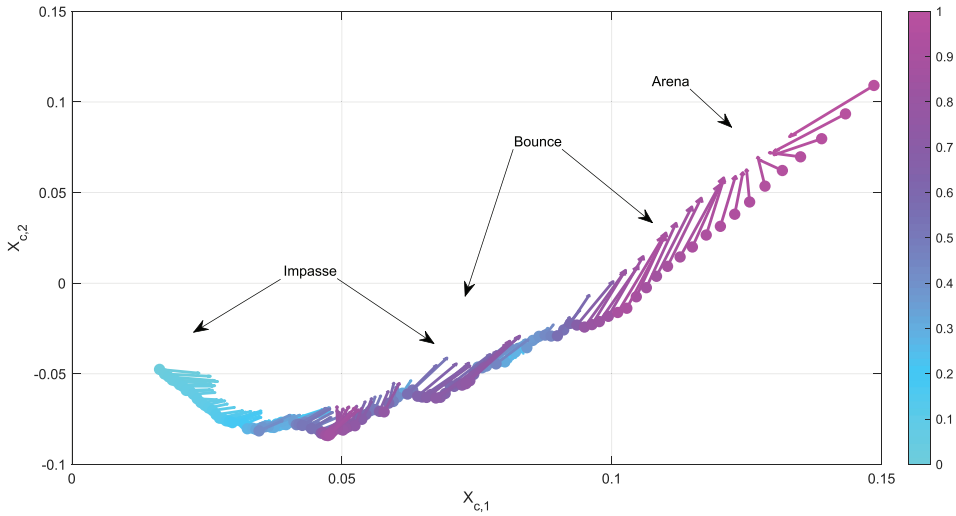

Figure 3.11: The time regimes of economic growth according to the two contributions $X_{c,1}$ and $X_{c,2}$. During time, countries move along the knee-like shape designed by the arrows.

economies have experienced the boost, they join the *Arena* of competition, which enhances growth. It is interesting to observe divergent directions of the arrows in the highest part of the Arena (high values of $X_{c,1}$ and $X_{c,2}$). In fact, in this area countries aim at increasing quantity and sophistication their exports which contribute to the increase of the $X_{c,1}$ values; at the same time, countries compete to become leaders in the economically-grown group, hence trying to earn scores in $X_{c,2}$. Therefore, the entrance of new countries in the competitive market is likely to affect other countries' growth.

The regimes are better understood by connecting the $GENEPY_c$ values of countries in time. This allows one to draw the path along the growth process, as shown in Figure 3.12, in which we show some economic complexity growth paths such as the ones of China (CHN), Germany (DEU), Japan (JPN), Nigeria (NGA) and Philippines (PHL). In Figure 3.12, Nigeria (NGA) and Venezuela (VEN) are tangled in the Impasse zone, in the bottom-left part of the graph. The Bounce area shows average uplifting dynamics of the countries towards highers stages of growth, instead. Countries such as China (CHN), India (IND) and Singapore (SGP) have clearly boosted their complexity to higher levels during the last years, joining the rich countries cluster ($X_{c,2} > 0$) during the period of observation (1995 – 2017). The Arena includes Japan (JPN), USA, Germany (DEU) and Switzerland (CHE) as paradigmatic examples. In Figure 3.12, the interactions among

countries are also evident. The rapid growth of a country, such as the dynamics shown by China [141, 149], naturally impacts other economies, whose $GENEPY_c$ values change according to the increased complexity of the competitor. An example is given by the nested trajectories of the arena-countries, such as Germany, Japan and USA, concurrent with the raise of China and Singapore. Some steadiness points in the trajectories can also be explained by the economic history of the countries. For example, the reduced trade capacity of countries, as a consequence of the 2008 financial crisis [150], produces a drop in complexity as shown by Germany, Italy and USA among the others. Instead, the Chinese last downgrading points of 2016 – 2017 may be explained as spillover effects of the 2015 stock market crash [151] and could also be related to the largely debated hard landing of the Chinese economy of the last years [152].

Therefore, during their economic growth process, countries tend to move from lower stages of complexity, delimited within the bottom-left quadrant, to higher ones, framed into the top-right quadrant. The former stage is associated with low productive knowledge and, consequently, low diversity in the exports. Contrarily, the latter is characterised by gain in skills and capital's investments, for which competition and growth are determined.

To collapse the information on how countries' rankings evolve in time we compute, for each year in the period 1995 – 2017, the world's centre of GENEPY by weighting all countries' geographical barycentres by their $GENEPY_c$ values. This computation has been executed according to the procedure defined by the McKinsey Global Institute in [148] to compute the shift during hystory of the Gross Domestic Product (GDP); the outcomes are shown in Figure 3.13 in yellow. For comparison, in Figure 3.13 we replicate the same procedure to compute the trajectories of the world's barycentre by weighting the countries' barycentres by their GDP at Purchasing Power Parity (GDP PPP in blue) – and, alternatively, their population (in purple). Since the economic complexity metrics are intensive ones (i.e., their values are "per capita" ones [22, 23, 124, 145]), the shifting in the world's centre of GENEPY has been computed by multiplying each country's GENEPY index for its population value in time, thus allowing for a fair comparison with the path followed by the GDP (in absolute value) in time. As the figure shows, the trajectories of the GDP and GENEPY index, differently from the population path, move toward East. The world's centre according to population, although clearly centred in the middle of Asia (as it would have been expected due to the high density of population this area has always recorded [153]), curves toward West as provoked by the increasing

population in Africa [153]. The differences in the world's GENEPY, GDP and population paths confirm that, year by year, the economy is more centred in the East [148] and that increasing population poorly impacts the ability of countries to economically grow. The distance between the current position of the barycenter of GDP and GENEPY may also imply that Asian countries (China included) still have a strong potential for economic growth, as also stated in [131]. Also, the trajectory drawn using GDP differs from the one drawn using the GENEPY index of countries as weights, because of the ability of the latter to capture both the productive knowledge of countries and the aforementioned dynamics of growth and competition between the actors in the trade.

3.4.4 Relation Between GENEPY and EXPY Frameworks

One of the main virtue of this study is the realization that, at least for what concerns the field of economic complexity, non-linearity is a non-necessary feature of the algorithms to rank nodes in a bipartite network (see Section 3.2.2). In fact, in the FC algorithm the Quality of a product is mainly determined by the least fit country exporting it, a crucial property accomplished by the non-linearity of the FC approach. In our linear framework, this property is maintained through the term $k'_p = \sum_c M_{cp}/k_c$, occurring in $W_{cp} = M_{cp}/k_c k'_p$, Eq. (3.16). This term in fact represents the degree of a product adjusted by how easily it is found within the network. Its inverse $1/k'_p$ is an anti-centrality score for the product, determining how limited is its presence within the producers' baskets and thus suggesting the need for higher productive knowledge in its production process. Notice that, by substituting the incidence matrix $\mathbf{M}$ with the traded monetary values, the term k'_p also recurs in the so-called EXPY rationale by Hausmann *et al.* [4]. Based on a decision-making model of firms' investment choices, the Authors in [4] defined an index of economic growth potential of countries, assessed through the required productive level of the exported products, i.e., EXPY. It is easy to verify the similarity of the relation to compute the X_c values in Eq. (3.6) (with the elements N_{cc^*} as given in Eq. (3.17)) with the expression to compute the productivity of a country according to EXPY [4]. In fact, by recalling the weighted incidence matrix of the export volumes in dollars, D_{cp}, and the strengths of countries and products such that

$$k_c = \sum_p D_{cp}, \quad k_p = \sum_c D_{cp}, \quad k'_p = \sum_c \frac{D_{cp}}{k_c};$$

the productivity level of a product, named PRODY, is given as

$$PRODY_p = \sum_c \frac{D_{cp}}{k_c k'_p} R_c,$$

being R_c the GDP per capita of the country c. EXPY, as a function of the PRODY, is computed as

$$\begin{aligned} EXPY = \sum_p \frac{D_{cp}}{k_c} PRODY_p = \sum_p \frac{D_{cp}}{k_c} \sum_{c^*} \frac{D_{c^*p}}{k_{c^*} k'_p} R_c \\ = \sum_{c^*} \sum_p \frac{D_{cp} D_{c^*p}}{k_c k_{c^*} k'_p} R_c. \end{aligned} \tag{3.28}$$

Apart from a rescaling factor k'_p (see Eq. (3.23)), the formal similarity of GENEPY with EXPY is striking. Notice that, this similarity is a result of the application of our framework, and not an "a priori" construction. Clearly, EXPY has been defined from a different deductive rationale, which considers the trade as described by the weighted incidence matrix of the monetary fluxes (thus providing different input information) and embeds exogenous information such as the GDP per capita.

3.4.5 Robustness of the Metrics

We have tested the robustness of the GENEPY index by changing the input matrix onto which perform the computation. In fact, when conceiving the bipartite network of countries and products, the commonly used binarisation procedure of the **RCA** matrix, Eq. (3.1), is adopted, aiming at capturing the network topology. However, a different (but possibly relevant) application of EC method is the one obtained by using the **RCA** matrix as input, thus preserving the information about trading competitiveness. In Figure 3.14 we show that, also if the **RCA** matrix is used as input of the calculation, the GENEPY results remain coherent with respect to changes in the incidence matrix of the network.

3.4.6 Products' GENEPY

The same results shown in Fig 3.6 for countries are given for products in Fig 3.16. In panel (a), we plot the position of products on the $\{Y_{p,1}\ ,\ Y_{p,2}\}$ plane. The knee-shape of the points in the plane of the two eigenvectors is recurrent, due to the algebraic reasons detailed in Section 3.4.2. With respect to the results shown in Fig 3.6a, the points in the plane $\{Y_{p,1}\ ,\ Y_{p,2}\}$

(panel (a)) are more compressed, due to the higher number of nodes in the network described by $\mathbf{G}$, Eq. (3.18), together with the normalisation condition of the two eigenvectors, such that $\sqrt{\sum_p Y_{p,i}^2} = 1$, with $(i = 1, 2)$ (this year totaled 1232 traded commodities). Panel (b) and (c) show how the two components of GENEPY correlate with the output of the FC and MR algorithms. In particular, panel (b) scatters the values of the first eigenvector $Y_{p,1}$ with the linearized values of Quality, $Q_p k'_p$, using Eq. (3.19). This plot testifies that, even when the matrix $\mathbf{G}$ is interpreted as a proximity matrix, thus deleting the self-interactions and over-weights of similarities, our linear recast of the FC metrics still performs well in reproducing the outcomes of the non-linear equations, Eqs. (3.3), as also shown in Figure 3.15. Instead, panel (c) scatters the values of the second eigenvectors $Y_{p,2}$ with the values $PCI_p\sqrt{k_p}$. In this case, the correlation is lower due to the differences among the matrix $W_{cp}^A = M_{cp}/\sqrt{k_c k_p}$, Eq. (3.10), from which we are able to recover the exact metrics ECI and PCI, and the matrix $W_{cp}^B = M_{cp}/k_c k'_p$, Eq. (3.16), which are here used to compute the GENEPY (hence $Y_{p,1}$ and $Y_{p,2}$) values. In Figure 3.17, we aggregate products in categories to resume, through a boxplot, the results the GENEPY index provides about products complexity. As the figure shows, differences in the complexity computed according to the GENEPY index among the categories clearly emerge: commodities into the *Machinery* or *Electrical* categories naturally require different and more sophisticated knowledge in order to be produced, while resource-based commodities, such as Animals or *Foodproducts* do not need special knowledge requirements in order to be produced or traded. In addition, the GENEPY values may vary widely in some categories such as *Chemicals*, where the natural availability of natural resources and the requirements for their extraction may define the need for more complex technologies for making these available for trade.

The variability of products' complexity shown in the boxplot is responsible for the results on countries' complexity, due to the combination of the metrics within the economic complexity framework (see Section 3.2). At completion of Figure 3.12, Figure 3.18 plots the time series of the complexity of the baskets of three countries, Japan, China and Nigeria, iconic examples of the regimes of growth described in Section 3.4.3 [3]. For each of these three countries, we divided the GENEPY value of each product by the total complexity of the export basket (i.e., the sum of complexity).

[3]This work has been conducted by the Master's Degree student Luciano Saraceno under the supervision of the candidate Carla Sciarra and the tutors Francesco Laio and Luca Ridolfi. Further results are available in the thesis work [154].

Products' complexity is then aggregated in sectors. As the Figure shows, in more complex economy such as Japan, more complex products are the ones belonging to the *Machinery or Electrics* categories. China has increased the complexity of its basket in time, investing in products belonging to this category or to the *Metals* one. Countries as Japan and China are poorly characterized by lower complex products, such as the ones in *Animals* and *Vegetables* sectors. These two last sectors are clues in the export baskets of Nigeria, a less complex country. These spectrum of complexity confirms the outputs discussed in Section 3.4.3 regarding the acquisition of capabilities and the investments in time. More developed economies show more stable trend, while less developed economies are characterized by more irregular composition of exports and complexity over time.

3.5 Concluding Remarks

We have introduced the GENEPY, a GENeralised Economic comPlexitY index, which provides a multidimensional metrics of countries' (and products') complexity. GENEPY arises from the eigenvectors of a symmetric proximity matrix, describing the similarities in the export baskets of countries. These eigenvectors combine in a multidimensional fashion the information obtained from MR and FC metrics thanks to a mapping (and linearisation for FC) of the original metrics to reduce the problem of finding these metrics to an eigen-centrality problem. GENEPY ranks countries for their multidimensional complexity, squeezing the eigenvectors through the adoption of a statistical framework on centrality metrics [91]. Moreover, the multidimensionality of our approach can be exploited to trace the economic growth process of countries in time.

A key point is that the proximity matrix $\mathbf{N}$ among countries is symmetric; as a consequence, the left and right eigenvectors coincide and the eigenvector centrality, whereupon our metrics are grounded, is distinctly defined [9, 43]. In contrast, by adopting the mathematical approaches of MR or FC, asymmetric matrices are recovered to map countries' Economic Complexity (Eqs. (3.2) for the MR case) – or Fitness (Eqs. (3.3) for the FC case) – onto itself (a mirror argument holds for products). In this case, the eigen-problem can be formulated by considering either right or left eigenvectors, thus posing the question of how the problem should be tackled. This is not just a matter of mathematical formalism: in fact, the eigenvector centrality for directed networks – whose adjacency matrices are asymmetric – typically considers the right and left eigenvectors for determining the out and in centralities of

the nodes, respectively, as caused by directionality of the edges [9]. In the same vein, the well known PageRank [40] centrality algorithm for directed networks considers the left eigenvector to assess only the in-centrality of the nodes. For bipartite networks, the most basic and simple case to rank nodes would be to set $M_{cp} = W_{cp}$, thus providing two symmetric proximity matrices $\mathbf{MM^T}$ in Eq. (3.6) and $\mathbf{M^TM}$ in Eq. (3.7)[116]. Contrarily, although set in a bipartite network framework, economic complexity methods as MR and FC generate artificial asymmetry by rescaling this symmetric matrices (using the countries' degree or some of its transforms) without taking care of preserving the feature of symmetry; thus leaving almost arbitrary choice to the solution of the eigen-problem. The symmetry of the transition matrices, also in terms of the adherence to the original symmetric structure of the problem, represents an added value of our framework. Moreover, the bilateral information of the proximity matrix can be used to understand the structure of the export baskets of countries and how these are related through shared common capabilities (see Figure 3.7).

We have also shown how GENEPY can be used to track the economic growth of countries during the years as driven by their economic complexity. Even though economic complexity metrics have already been used to draw these paths [49, 155–157], our innovative multidimensional approach allows one to draw these trajectories without the need of embedding the exogenous information on the GDP per capita that most applications require. As such, the chance of maintaining the simplicity of a data-driven approach endows the GENEPY framework with the main founding-reason for which economic complexity theory was born, i.e., to provide the ground for a more quantitative, data-driven approach to the assessment of the potential economic growth of countries as factored by the productive knowledge [158].

A further advantage of the GENEPY index is given by its robustness, since results do not suffer from the change of the input matrix.

Our approach also brings the advantage of having introduced a linear version of the FC algorithm, which provides with the same results of the non-linear method (on average, 99.5% Pearson's correlation). We stress that, as we have shown, linearity may only be a characteristics of the trade, while other systems may require a non-linear approach to be studied. Nevertheless, the linear form we have found and adopted in this framework is also recurrent in other economic complexity frameworks, such as the EXPY one. The term k'_p which, in our rationale, addresses the assumption, in the FC algorithm, that the Quality of a product is mainly determined by the least fit country exporting it. This term also recurs in the EXPY

rationale by Hausmann *et al.* [4]. As we have shown, (see Eq. (3.28)), the equations to compute X_c in the GENEPY framework are similar to those defining the EXPY scores of countries [4]. This similarity is a result of the application of our framework, and not an "a priori" construction: in a sense, the economic concepts are self-emerging, with some significant variations with respect to the original EC framework we here reconcile [22, 23]. In our view, this similarity represents a possible micro-economically sounded bases for the economic complexity theory, toward which we address future work.

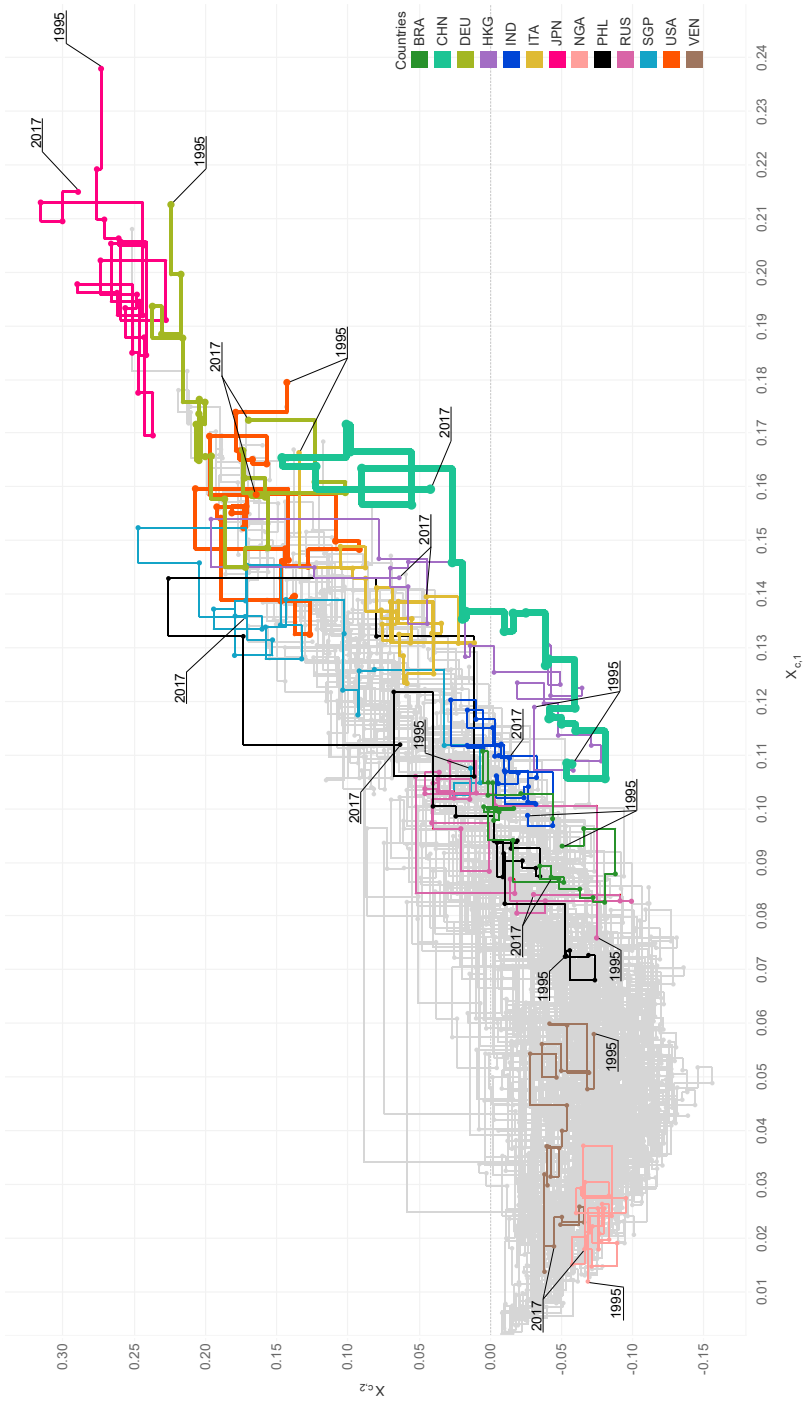

Figure 3.12: Countries' trajectories in the GENEPY plane. The values of the first eigenvector $X_{c,1}$ are on the x-axis, whilst on the y-axis the values of the second eigenvector $X_{c,2}$ are found. The eigenvectors are normalised such that their 2-norm is unitary, i.e., $\sum_c X_{c,1}^2 = \sum_c X_{c,2}^2 = 1$. We highlight the trajectories of Brasil (BRA), China (CHN), Germany (DEU), Hong Kong (HKG), India (IND), Italy (ITA), Japan (JPN), Nigeria (NGA), Philippines (PHL), Russia (RUS), Singapore (SGP), United States of America (USA) and Venezuela (VEN), against the background created by trajectories of all other countries in grey. Line width reflects the countries' share of world exports in monetary value during 2017. To improve the readability of the plot, the paths from one point to another were forced to follow right-angled movements. The figure has been produced with Tableau Public 2019.4.

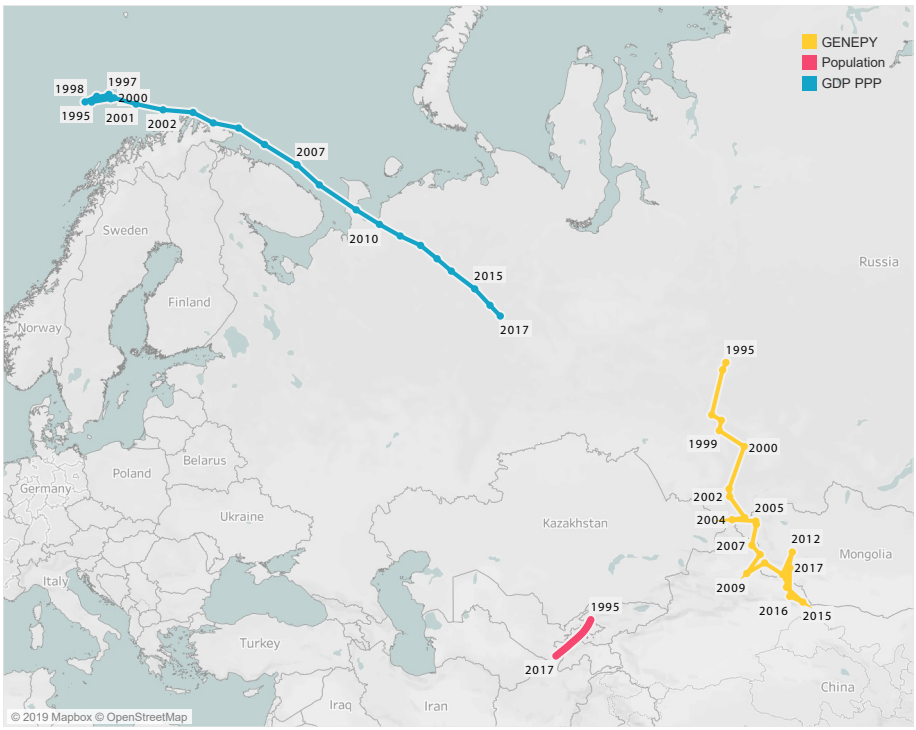

Figure 3.13: The world's economic and demographic barycentre, 1995 - 2017. The trajectories are computed by weighting the countries' geographical centres by their GENEPY index, in yellow, the Gross Domestic Product at Purchasing Power Parity (GDP PPP), in blue, and the population size, in purple. The GDP trajectory is consistent with the one shown by the McKinsey Global Institute [148] taking as reference the path in there shown from 1990 – 2025. Data for the GDP PPP and the population of countries are provided by the World Bank. The coordinates of countries are provided by the Portland State University and defined according to the georeference system WGS 1984. The figure has been produced with Tableau Public 2019.4.

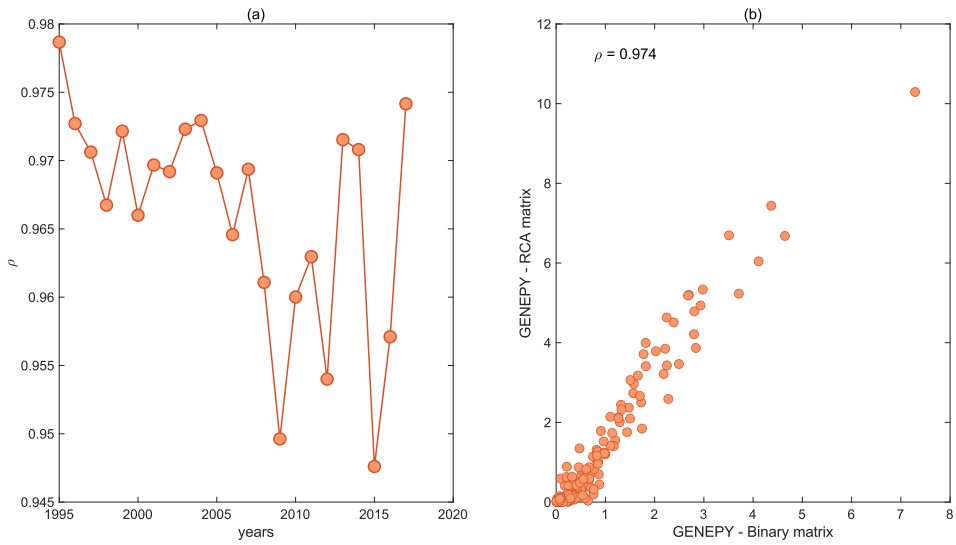

Figure 3.14: Comparison of the GENEPY of countries computed using either binary or RCA matrix. (a) Scatter plot of the GENEPY index as obtained from the use of the binary matrix **M** – on the x-axis – and from the **RCA** matrix – on the y-axis – as input for the computation of the GENEPY values. Values refer to year 2017. In panel (b), time series of the correlation coefficients among the GENEPY values computed using as input for the algorithm the binary matrix **M** and the ones obtained using as input the **RCA** matrix. The correlation coefficients are of the Pearson's kind.

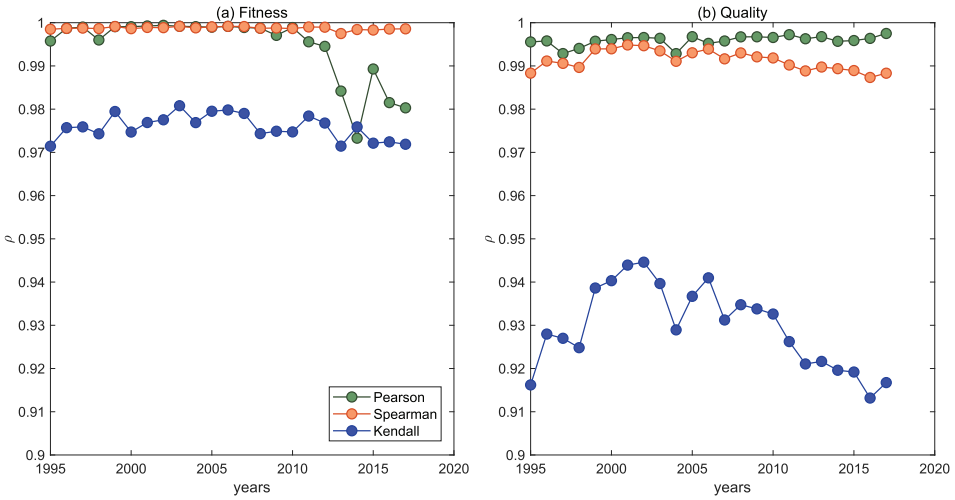

Figure 3.15: Correlation coefficients among the non-linearly and the linearly computed values of Fitness, panel (a), and Quality, panel (b), obtained by setting the diagonal values of the matrix $\mathbf{N}$ and $\mathbf{G}$, respectively, to zero. In green is Pearson's correlation coefficient, while Spearman's one is in orange and Kendall's in blue. The Spearman's and Kendall's coefficients are ranking-based, while the Pearson's one compares the values between the two vectors. As the coefficients witness, the linearised algorithm almost perfectly reproduces the outcomes of the non-linear one even when the diagonal values are set to zero.

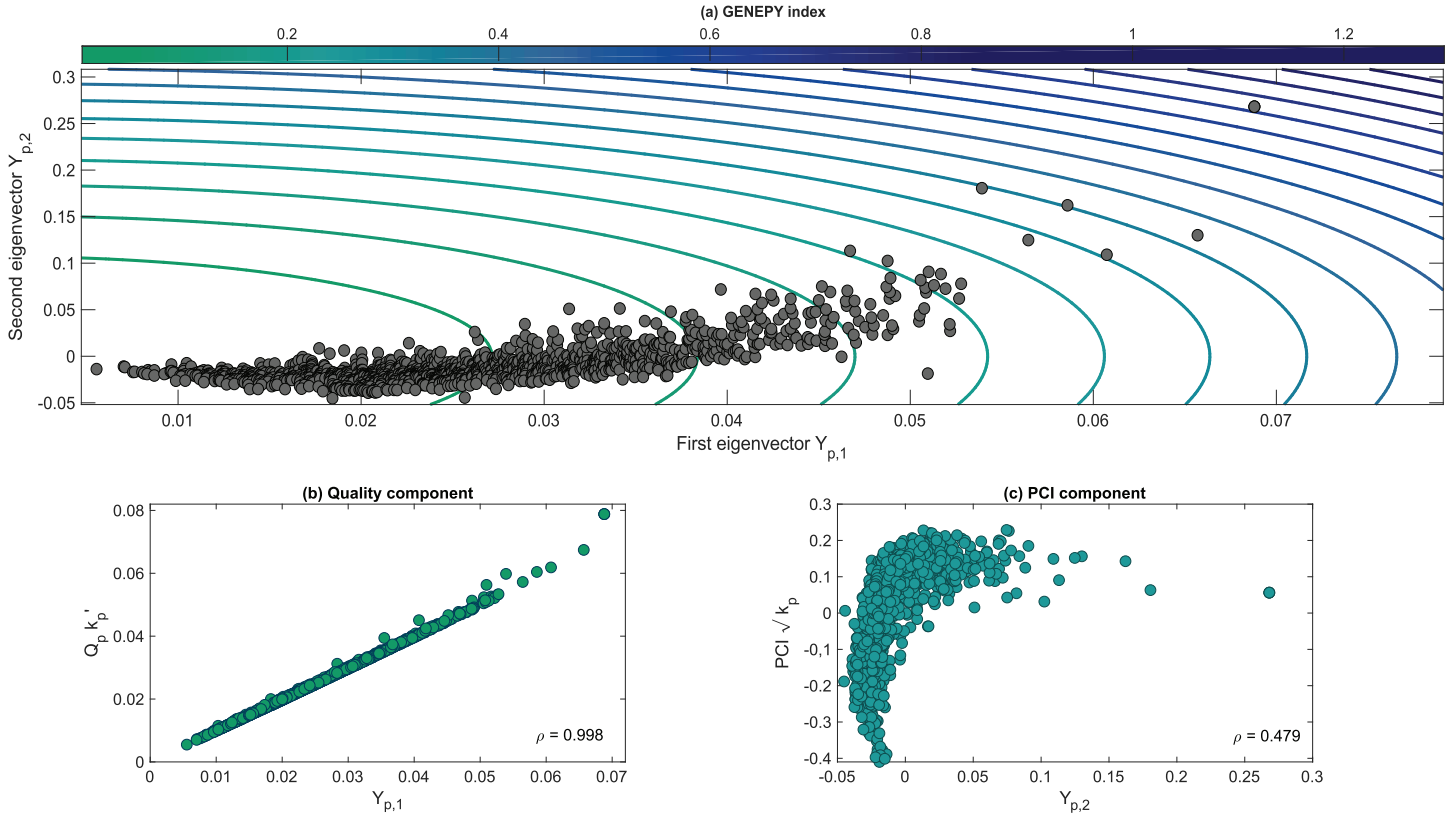

Figure 3.16: Products' GENEPY results and components as referred to the 2017 international commodities trade. (a) Contour plot of the GENEPY values. We interpret increasing economic complexity with increasing radial distance from the origin. The x-axis reports the values of the first eigenvector $Y_{p,1}$ whilst the y-axis the values of the second eigenvector $Y_{p,2}$. The eigenvectors are normalised such that their 2-norm is unitary, i.e., $\sqrt{\sum_p Y_{p,i}^2} = 1$, with $(i = 1, 2)$. Contours range from lower GENEPY values (green) to higher ones (blue). (b) Scatter plot of the first component $Y_{p,1}$ of GENEPY values compared with the values of the Quality values Q_p rescaled by the products adjusted degree k'_p. (c) Scatter plot of the second component $Y_{p,2}$ of GENEPY compared with PCI values, rescaled by the term $\sqrt{k_p}$. The correlation coefficients are of the Pearson's kind.

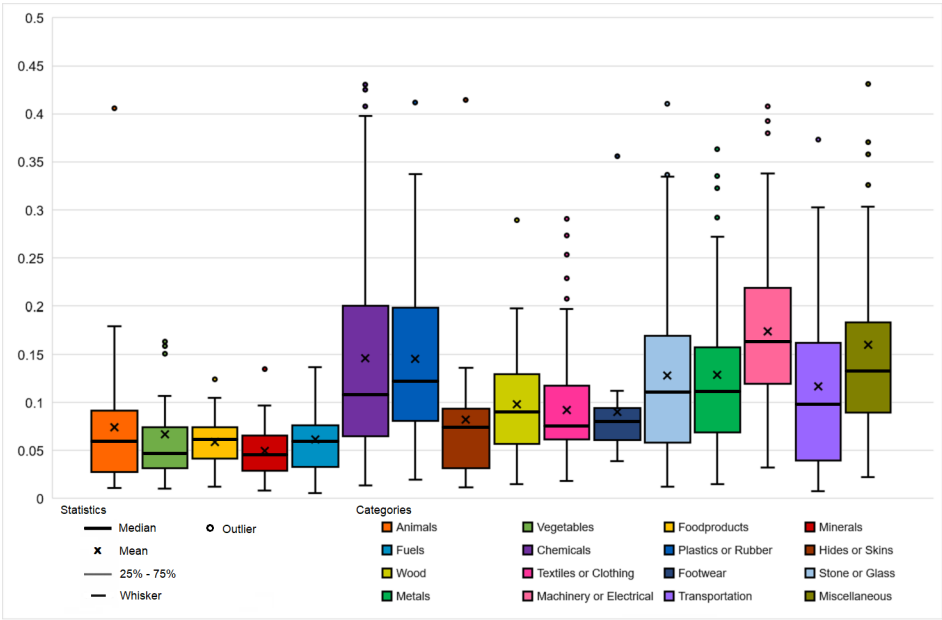

Figure 3.17: Boxplots of the GENEPY values for products aggregated into categories. Categories are defined according to the World Integrated Trade Solutions, WITS, by the World Bank, available at `wits.worldbank.org`. In the boxplot the cross is the mean, the thick bar is the median, the bars define the interquartile range (IQR) 25% - 75%, the shorter bars are the whiskers and the dots are outliers. From above the upper quartile, a distance of 1.5 times the IQR is measured out and a whisker is drawn up to the largest observed point from the dataset that falls within this distance. Similarly, a distance of 1.5 times the IQR is measured out below the lower quartile and a whisker is drawn up to the lower observed point from the dataset that falls within this distance. All other observed points are plotted as outliers.

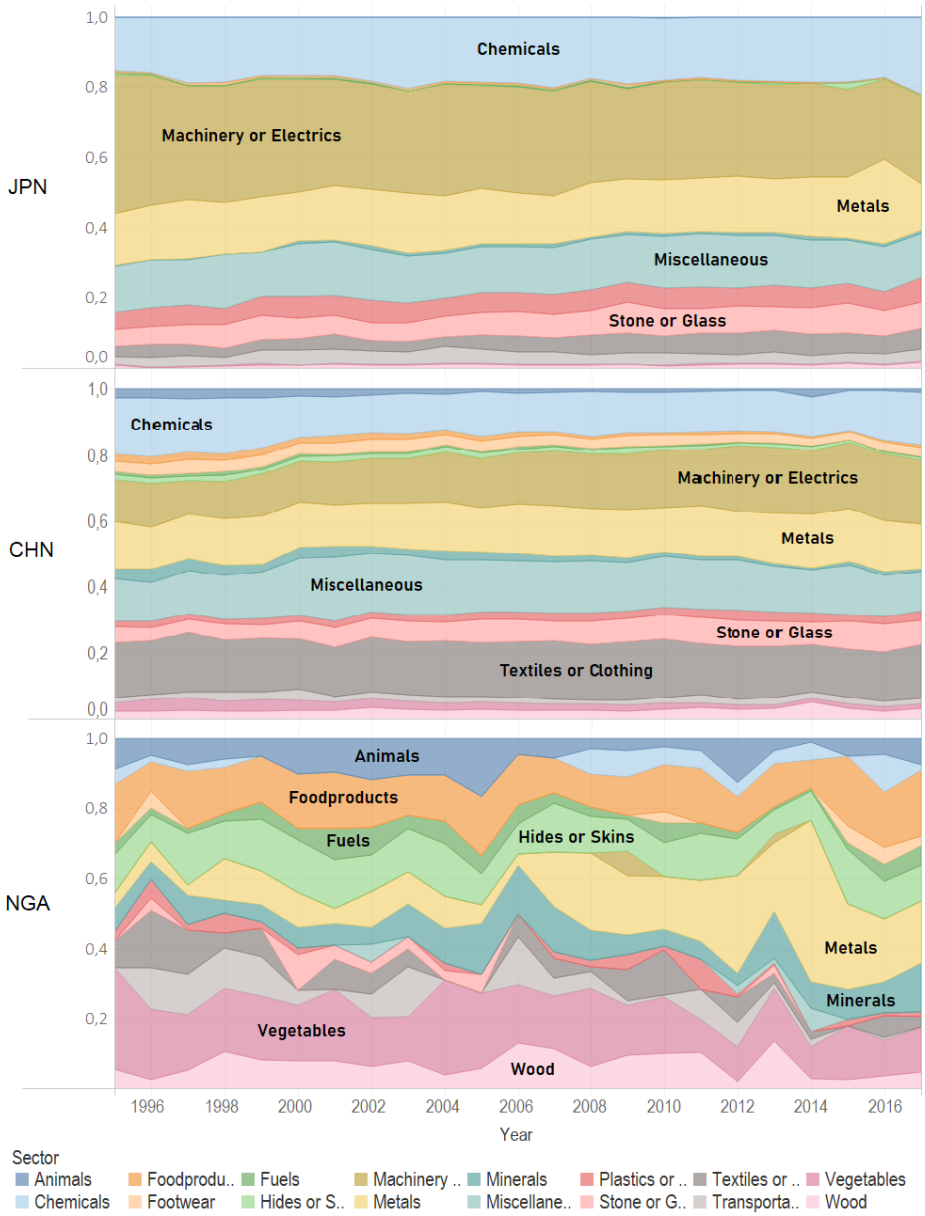

Figure 3.18: Complexity in the export basket composition of Japan, China and Nigeria computed using the GENEPY index for products during the period 1995 – 2017. Products are aggregated in categories, defined according to the World Integrated Trade Solutions, WITS, by the World Bank, available at `wits.worldbank.org`. Each category percentually contributes to the total complexity of the country at hand and evolves in time.

4

Network-Driven Rankings of Countries' Status in SDGs

The work described in this chapter has been partially derived from Sciarra et al., submitted to Scientific Reports, *2021 [159].*

Universality, integration and inclusion: these are the principles and cornerstones upon which the United Nations (UN) have constructed, in 2015, the Agenda 2030 of sustainable development [60, 61]. The world is not new to the request of 'a global agenda for change'. Back in 1987, the report "Our Common Future" already introduced the key idea of a common action plan to address economic growth in equilibrium with the people and environment, thus preserving our world to meet human needs for today's and future generations [160]. The beginning of the XXI century marked a shift in the way countries started being actively engaged in the implementation of sustainable development, with the establishment of the Agenda 2015, allowing the joint forces of UN and governments to achieve significant milestones in poverty and inequalities reduction, as well as in improved water access [161, 162]. In light of these achievements, and also of the limitations and gaps of such experience, the Agenda 2030 inherits and enlarges the views and objectives of the Agenda 2015 [162] introducing the 17 Sustainable Development Goals - SDGs [60]. The 17 Goals are strongly interconnected [68–77] and all have the same importance in being achieved. In fact, in line with the Charter of the United Nations, the Agenda and its principles, the Sustainable Development Goals have no pyramidal structure and there is no Goal prioritized with respect to the others, thus advocating for equal efforts in the designing of proper policies to meet these goals (Art. 40 of the Agenda) [60]. Each Goal targets the implementation of policies, totalling 169 targets across the 17 Goals [75]. Targets also mark the need for data and measurements of the status of countries with respect to the achievement of the Goals. Countries ratifying the Agenda are encouraged to

pursue sustainable development by defining national strategies with a global vision of their actions [60, 61], thus contributing to the common action plan necessary to foster change [60, 74, 163, 164] and embracing the cornerstones of the Agenda.

Since 2015, 5 years have already passed and countries are only left with 10 years to meet all the targets within the Agenda. To monitor the progresses of countries is a necessary step [165], a required one to define responsibilities and identify possible structural limitations and difficulties toward sustainability [81, 89, 166]. In fact, due to the heterogeneity of countries and the differences in the challenges they face [60, 66], it could be well expected that some SDGs are reached first by some countries with respect to others, a fact that calls for metrics in which this dynamics is taken into account and unveiled by the analysis of the data. In this chapter, we propose to tackle the definition of rankings of countries by recasting the system of countries and SDGs within a network science framework. Firstly, we define a representation of the Agenda 2030 as a bipartite network of countries and Sustainable Development Goals, connected via countries' recorded performances. Secondly, we show that aggregated index of countries' status with respect to their achievements in the Agenda 2030 can be cast as a centrality exercise. Within this framework, we demonstrate that the most used centrality metrics, i.e., the degree and eigenvector centrality, are highly correlated and so do not take into account neither the synergies and trade-offs among the Goals, nor the heterogeneity of countries' challenges in sustainable development. We propose to tackle the definition of an aggregated score of countries' status by adopting the GENEPY centrality framework, introduced in Chapter 3. Our framework provides novel insights about countries' efficiency in sustainable development and possibly addresses new directions to boost their activities in time for the 2030 deadline.

4.1 Unveiling the Hidden Network of Countries and Goals

As established by the United Nations [68], progresses in the Sustainable Development Goals (and so, targets) are estimated using a set of indicators providing quantitative information about countries performances; each indicator measures the attainment of certain targets across the 17 SDGs. Let I_{cgk} be the k-th value of the indicator I within Goal g recorded in country c. For the sake of comparison across indicators and Goals, most applications consider the I_{cgk} values to be normalised according to least and

optimal indicator values, resulting in a percentage of achievement of the indicator ranging from 0 to 100 [81, 90, 167] (see Section 4.4). Moreover, per each country c, one single value of achievement P in each Goal g is obtained as the average of the recorded and available values of the indicator I_{cgk} within the Goal. Namely,

$$P_{cg} = \frac{1}{N_{cg}} \sum_{k=1}^{N_{cg}} I_{cgk}, \tag{4.1}$$

where N_{cg} is the number of indicators in Goal g for country c (see Section 4.4). As introduced, to compute any aggregated score S_c for countries status entails defining the mathematical weights w_g of each Goal's performance. In its general formulation, S_c is defined as

$$S_c \propto \sum_g P_{cg} \cdot w_g, \tag{4.2}$$

in which we consider the presence of any possible scaling factors.

Within this framework, our aim is to cast the computation of aggregated scores of SDGs through the use of network science, so to exploit and unveil the complex structure of the Agenda. Let us consider the values P_{cg} as the starting point for our reasoning. We consider these values to be structured as a matrix $\mathbf{P}$ with C rows, i.e., the number of countries in the analysis, and 17 columns, as many as the Goals. Seen through network science lenses, the matrix $\mathbf{P}$ reveals the presence of a bipartite system in which countries and Goals are connected via recorded performances. In network theory, the matrix $\mathbf{P}$ describing these links is denominated as incidence matrix [9] (see Chapter 1). We consider the network structure of countries and Goals emerging from the data taken from the latest SDG Index and Dashboard, referring to year 2020 [82] (see Section 4.4), as exemplified in Figure 4.1.

The bipartite network representation offers the chance to borrow mathematical tools from network centrality science to define the importance of the nodes in the system and rank them accordingly [9]. Bipartite networks are characterized by the existence of two different sets of nodes, as in this case countries and Goals (Figure 4.1), and one centrality score can be computed for each set. The simplest measure of centrality, the nodes' degree, assumes the importance of the node to be described by the number and strength of its connections [35]. This provides the value $k_c = \sum_{g=1}^{17} P_{cg}$ for countries [35], thus implicitly setting $w_g = 1$ for all 17 Goals in the computation of the score S_c in Eq. (4.2). Notice that, in this countries-SDGs network, the link between the nodes describes the existence of a connection between a country

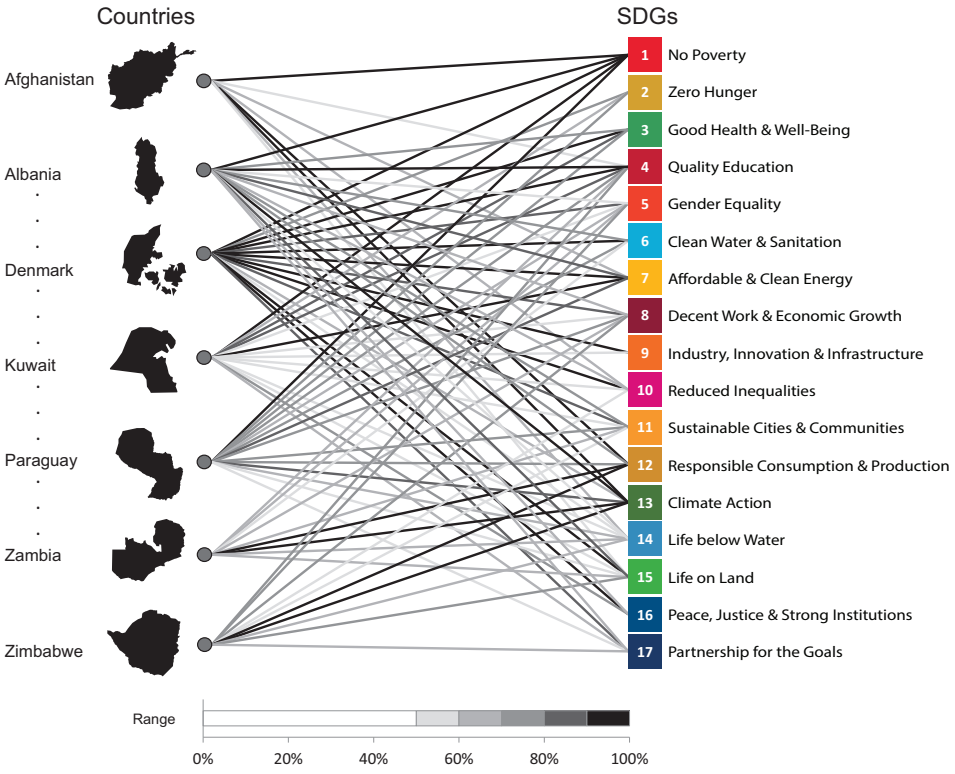

Figure 4.1: **The bipartite network of countries and Goals.** Qualitative representation of the bipartite network constituted by countries and Goals. On the left, we list seven of the countries that can be found by browsing the 2020 Dashboard, as sorted in alphabetical order [82] (the first and last two countries and the ones found at first, second and third quarter of the list). On the right, the 17 SDGs are reported. For each country, we connect the SDGs via the performance values P_{cg} in each Goal, according to the 2020 Dashboard data [82]. The values P_{cg} are intended to be readable as percentage of achievement of the Goals. We have classed these values in ranges of 10% of performances and color-coded, in grey scale, accordingly: the darker the links, the better the performances of the country within the Goals. Countries' performances smaller than 50% have been left blank.

and a Goal but also the magnitude of this connection, represented by the recorded performance of the country in that SDG (as plot in Figure 4.1). Therefore, according to the degree, countries having an higher percentage of achievement across SDGs have better chances of being central, no matter the Goal. This rationale reflects the egalitarian principle of the Agenda, for which all SDGs have equal importance in being achieved [60]. In fact, by recalling that, in light of this principle, the SDG Index by Sachs *et al.* [90] is defined as

$$SDG\ Index = \frac{1}{17}\sum_{g=1}^{17} P_{cg},$$

one recognizes that the SDG Index is the degree centrality of countries ($k_c = \sum_g P_{cg}$) scaled by a factor 17.

The degree only measures the local information of nodes' connections and so it does not depict the global structure of the network (for further details see, e.g., [15, 168]). Therefore, although in line with the principle of equal importance of SDGs, to rank countries with equal Goal weights entails not accounting for the complex behavior in sustainable development we aforementioned. Such behavior can be highlighted introducing network-comprehensive measures of centrality. Thanks to their global outlook on the network, these kind of metrics explore different dimensions of the SDG topic (and consequently, countries' status) and allows one to naturally define bottom-up weighting approaches.

The need for global centrality metrics to measure the complexity of the system clearly arises when considering the heterogeneity of countries' performances across the Goals, as we address in Figure 4.2. The figure plots countries' performances as defined by the 2020 SDG Index and Dashboard [82] (see Section 4.4). In Figure 4.2, countries are ordered according to their ranking position as defined by their degree (or, equivalently, the SDG Index). These countries' performances (which from hereon we define as 'spectra') are relative ones, as they are obtained by subtracting the average performance of the countries, $k_c/17$ (i.e., their SDG Index), from the Goal-specific performance, P_{cg}. This allows one to compare relative Goal performances of all countries according to their efforts in sustainable development, thus identifying areas where countries are investing more/less efforts and disclosing differences in their strategies. At glance, the heterogeneity of the spectra stands out. Countries exhibit very contrasting behaviours among them and across the Goals, witnessing the fact that the world is not moving as a unique ensemble toward the achievement of sustainable development. As mentioned, this is possibly due to the heterogeneity of countries contexts

and challenges, as well as the differences in national strategies that possibly enhance such heterogeneity across SDGs. To group countries according to their degree k_c can help understanding these differences. In fact, Figure 4.2 shows the existence of two limit behaviors of the 28 top and the 28 bottom performing countries according to the SDG Index (or degree), i.e., of classes 1 and 6, whose spectra are almost completely out of phase. These dynamics are more evident within Goals of environmental performances and exploitation, from Goal 12 to 15. As the spectra clearly show, the first 28 best countries in degree (class 1 in light blue) are poorly engaging toward the achievement of SDG 12 and 13. In particular, Norway is the relative worst performer in Climate Action, a Goal in which the country performs almost -60% with respect to its SDG Index. Instead, there are many low-degree countries (class 6 in violet) whose relative performances in Climate Action are higher, with Central African Republic (CAF) recording $+60\%$ of performance with respect to its SDG Index. Even if less accentuated, the spectra of top and bottom degree countries are also out of phase in SDG 17, the one invoking partnership, in which countries nearer to fulfil most of the Agenda are actually the worst relative performers (e.g., Latvia – LVA). Other examples of this out of phase behavior of the countries in class 1 and 6 figure in correspondence of Goals 1, 2, 7 and 14 (Zero Poverty, Zero Hunger, Clean Energy and Life Below Water, respectively). Drops of performances occur for top-degree countries in Goals 2 and 14, while for bottom-degree countries in Goals 1 and 7. For example, Singapore attainment of SDG 14 is -60% with respect to its average performance in sustainable development. Yemen stands as an exception of such pattern since, in Goal 1, this country performs 40% better than its average value.

The spectra depict the complexity of the variety of approaches toward sustainable development, in which the specificity of countries' characteristics has its role in determining the attainment of the Goals. Therefore, we argue that analyses designed to consider and embed this complexity can shed new light about the state of the art in sustainable development. The introduction of network theory is a first step toward this direction and allows us to define novel aggregated scores based on data-driven definition of the weights w_g in Eq (4.2). In particular, the introduction of network-comprehensive measures centrality may help in exploring different dimensions of the SDGs topic (and consequently, countries' status) and allows one to naturally define bottom-up weighting approaches.

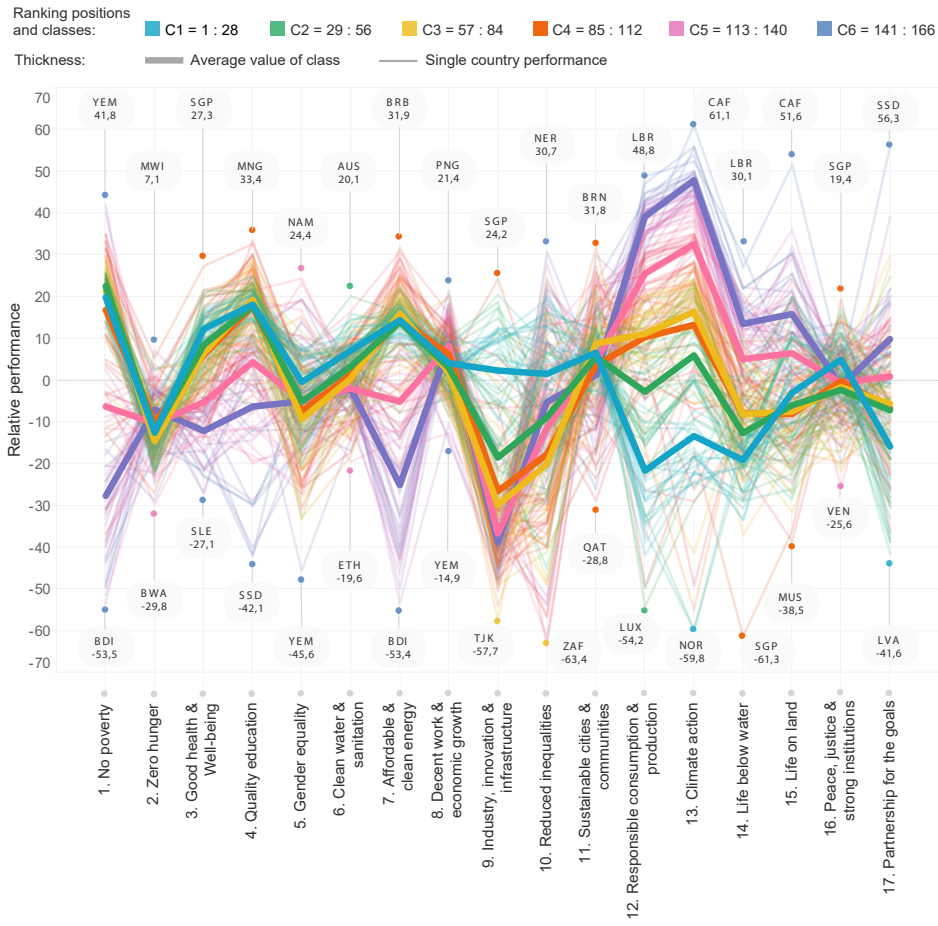

Figure 4.2: The spectra of countries' relative performances as obtained by $P_{cg} - (k_c/17)$. Countries are first ranked and then clustered according to their average performance (i.e., the SDG index or, equivalently, their degree). Based on the ranking positions, we define six classes of performance: light blue (countries in positions $1-28$), green ($29-56$), yellow ($57-84$), magenta ($85-112$), pink ($113-140$) and violet ($141-166$). The classes' average values of relative performances are shown in thicker lines. Top and bottom relative performers in each Goal are pointed out, and their performance value is color-coded as their corresponding class.

4.2 A Data-Driven Weighting of Countries

A first revision of the degree centrality in bipartite networks consists in weighting the connection of the node proportionally to the centrality value of the node at the other edge, as we have introduced in Chapter 1. Therefore, countries connected to more central SDGs obtain a higher scoring value, and *vice versa*. According to this rationale, the weights to define the aggregated score S_c in Eq. (4.2) are $w_g = v_g$, where v_g is the centrality value for Goal g. This entails solving the system of coupled equations

$$\begin{cases} S_c \propto \sum_g P_{cg} v_g, \\ v_g \propto \sum_c P_{cg} S_c. \end{cases} \tag{4.3}$$

Mathematically, the solution of this system is obtained computing the singular vectors of the matrix $\mathbf{P}$ which determine the eigen-centrality vectors $\mathbf{u}$ and $\mathbf{v}$ for countries and Goals, respectively [116] (see Chapter 1, Section 1.2). While the degree is a local measure of centrality, the eigenvector is a global one, as it considers for the computation of the scores all possible links and strengths in the network [9, 15, 168]. However, as we show in Figure 4.3, the eigenvector centrality brings no further information in terms of rankings than the one by the degree centrality (99.9% in both Pearson's and Spearman's correlation measures). This lack of added value is due to the intrinsic correlation that the degree and eigenvector centrality show when the spectral gap – i.e., the delta between the first and second largest singular vectors of the incidence matrix – is large [169]. For this particular bipartite network, the second largest singular value is roughly one fourth of the principal singular value, implying high correlation between the degree and eigenvector centrality [170]. Therefore, in the countries-SDGs network, the use of non-uniform weights as in Eq. (4.3) is almost ineffective in changing the point of view about the state of the art in sustainable development, and other rationale about countries inter-plays with Goals must be introduced to remove the degree-bias that characterizes the eigenvector centrality [170].

The use of the centrality metrics defined within the field of Economic Complexity (EC) [22, 23, 120] – that we defined in Chapter 3 – can help in the characterization of more complex inter-plays between countries and Goals. In fact, the idea upon which EC theory is constructed is that, in a looping system, if a product is only exported by few countries, this item is more knowledge-intensive than other items exported by many other countries. (In EC, the word 'knowledge' intends knowledge of production, resources, human and capital investments, eventually [4].) This determines

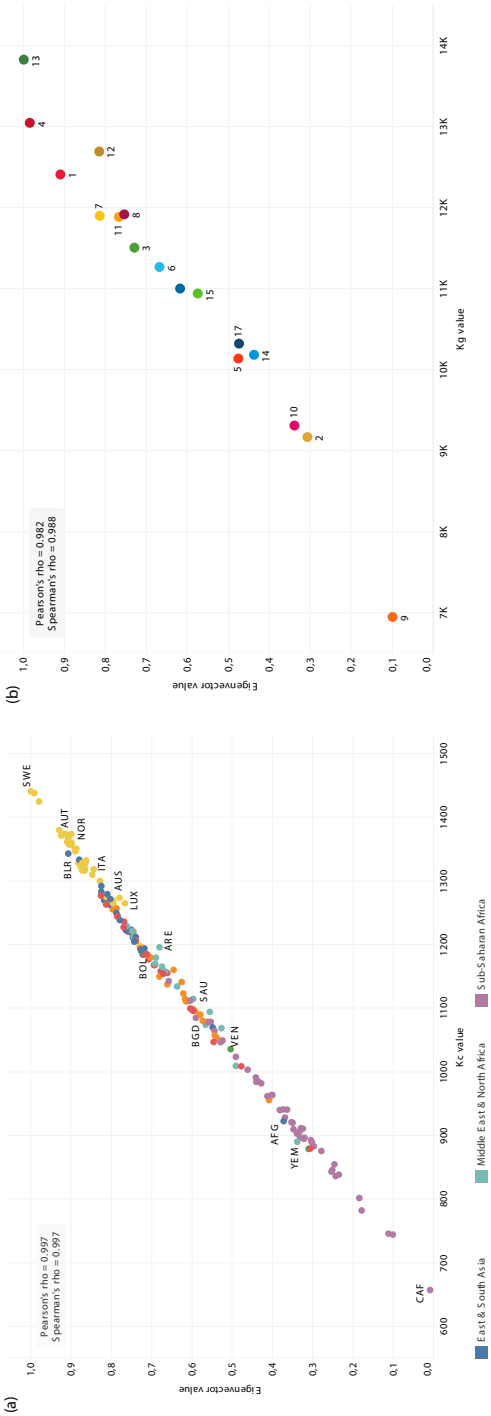

Figure 4.3: Scatter plot of the centrality values obtained by the degree and eigenvector measures. Panel (a) scatters the degree and eigenvector values of countries; panel (b) of Goals, instead. In panel (a), countries are color-coded according to their Region, as defined in the SDG Dashboard 2020 [82]). In both panels, the eigenvector values are normalised between zero and one; the Pearson's and Spearman's correlation coefficients are specified (notice that, the Spearman's one is a ranking-based correlation, thus it also provides information about the correlation between the two rankings.

higher EC scores of more knowledge-intensive goods [22, 23, 118, 120]. Clearly, weights are self-emerging from the methodology and its grounding rationale.

In a similar manner, we can adapt the EC theory and methods to the network of countries and SDGs, therefore introducing new reasoning about how countries act in sustainable development. In tailoring the EC framework to the SDGs one, we assume that, if within a Goal only few countries record near to optimal performance values, this Goal is more knowledge-intensive than the others, thus resulting in a higher EC score. Countries recording such optimal performances are those ones in more favorable conditions to attain the Goal. In fact, in here, we translate 'knowledge' into policy and intervention designs and implementations; awareness and preparedness to face the challenges, all well known factors for affecting countries performances in sustainable development [62, 75, 82, 171–173].

We adopt the GENEPY framework, introduced in Chapter 3, since it reconciles the contrasting methodologies on economic complexity and it is also a reliable method for processing non-binary incidence matrices as the one of the countries-SDGs bipartite system [120]. For the sake of clarity, in the following, the adaptation of the GENEPY framework to the context of the Agenda 2030 is defined as SDGs-GENEPY. To the best of our knowledge, Cho *et al.* [167] is the only existing example in literature proposing to adapt EC methodologies and centrality metrics to score countries performances within the Agenda 2030. However, our work differs from that one in both methodology (the *Method of Reflection* from Hidalgo *et al.* [22] is used, instead) and data, since that work is limited to the Asian region. The work by Cho *et al.* [167] also inherits the conceptual scheme of combining capabilities for driving innovation (human and capital resources, investments, policies [120]), which is typical of the economic complexity. While this conceptual scheme is reasonably suitable for the productive system, it is not in the field of sustainable development. As we discuss, this latter area is mainly characterized by countries' historical phases and challenges, followed by the ensemble of decisions, planning, strategies and willingness that nations experience along their path toward sustainable development [82, 174, 175].

The SDGs-GENEPY rationale defines two centrality properties, S_c for countries and Y_g for SDGs, that can account for the EC rationale and so embed the interplay between countries and Goals. The properties are

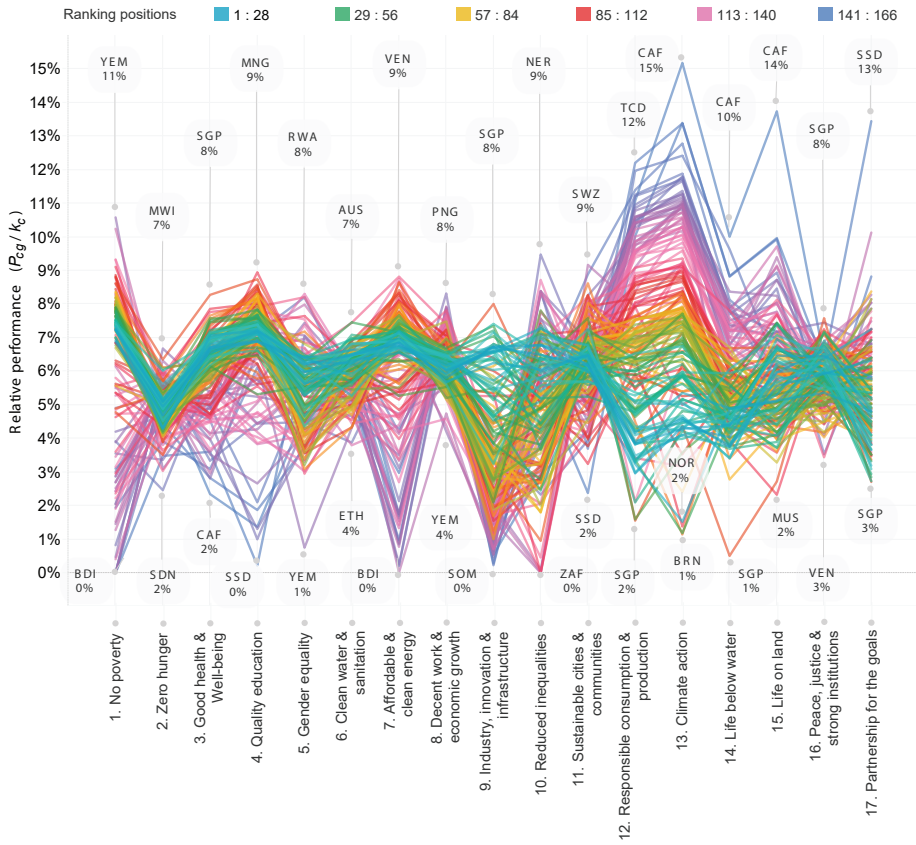

Figure 4.4: The spectra of countries' relative performances as obtained by P_{cg}/k_c. Countries are first ranked and then clustered according to their average performance (i.e., the SDG index or, equivalently, their degree). Based on the ranking positions, we define six classes of performance: light blue (countries in positions $1-28$), green ($29-56$), yellow ($57-84$), magenta ($85-112$), pink ($113-140$) and violet ($141-166$). Top and bottom relative performers in each Goal are pointed out.

defined through the following system (see Chapter 3, Section 3.3)

$$\begin{cases} S_c \propto \frac{1}{k_c} \sum_g P_{cg} \frac{Y_g}{k'_g}, \\ Y_g \propto \frac{1}{k'_g} \sum_c P_{cg} \frac{S_c}{k_c} \end{cases} \tag{4.4}$$

in which $k_c = \sum_g P_{cg}$ is the degree of the countries, therefore the sum of all Goals' performances (i.e., the value of the aggregated score supposing $w_g = 1$ for all SDGs). The term $k'_g = \sum_c P_{cg}/k_c$, that we define as 'adjusted Goal's degree', is the degree of Goal g accounting for the relative performances of countries within it. In fact, relative performances of countries can either be computed as the subtraction of the average performances, as in Fig 4.2, or using the ratio P_{cg}/k_c, and the same results hold, see Figure 4.4. Therefore, to evaluate the aggregated score of countries' status S_c according to the SDGs-GENEPY entails assuming $w_g = Y_g/k'_g$ in Eq (4.2). As introduced in Chapter 3, a closed solution for this system is provided by solving the coupled singular vectors $\mathbf{X}$ and $\mathbf{Y}$ associated to the largest singular value σ_1 of the matrix W defined as

$$W_{cg} = \frac{P_{cg}}{k_c k'_g}.$$

The matrix W helps in defining the EC rationale and in providing a symmetric representation of the bipartite system for which the $\mathbf{X}$ and $\mathbf{Y}$ are determined. In fact, the vector of scores S_c according to Eq.(4.4) is the eigenvector of values $X_{c,1}$ associated to the largest eigenvalue of the matrix $\mathbf{N}$ defined as

$$N_{cc^*} = \mathbf{W}\mathbf{W}' = \sum_g \frac{P_{cg} P_{c^* g}}{k_c k_{c^*} (k'_g)^2}; \tag{4.5}$$

the vector $\mathbf{Y}$ for SDGs is the eigenvector of the largest eigenvalue of the matrix $\mathbf{Z}$ defined as

$$Z_{gg^*} = \mathbf{W}'\mathbf{W} = \sum_c \frac{P_{cg} P_{cg^*}}{k_c^2 k'_g k'_{g^*}}. \tag{4.6}$$

The analytical solution we here provide regarding the SDGs-GENEPY metrics slightly differs from the one defined in Chapter 3, Section 3.3. Thanks to the differences in the bipartite system, to adapt the GENEPY framework to the Agenda 2030 provides a simpler mathematical rationale. In fact, if built upon the export data, the GENEPY index is a multidimensional centrality score for economic complexity in which two eigenvectors of the matrix $\mathbf{N}$ for countries are combined (or $\mathbf{G}$ for products, which here has

its counterpart in **Z** for SDGs); moreover, in the economic complexity case, the self-similarities along the diagonal values of the matrices **N** and **G** inflate the values of the eigenvectors **X** and **Y**, for countries and products, respectively. Without any loss of information, we limit our analysis to the first eigenvectors of the matrices **N** and **Z**, for countries and Goals, respectively. In fact, the eigenvectors associated to smaller eigenvalues – than the principal one – bring no relevant added information and the quadratic terms in the formulation of the GENEPY index (Eq. (3.22) in Chapter 3) can be neglected. In Figure 4.5 we scatter the values $X_{c,1}$ of the first eigenvector of the matrix **N**, Eq. (4.5), in this chapter proposed as the solution of the SDGs-GENEPY S_c values in Eqs. (4.4), and the SDGs-GENEPY index computed according to the quadratic formula in Eq. (3.22) and adapted to the countries-SDGs system. Moreover, differently from the economic complexity case, setting the diagonal values of the matrices **N** and **Z** to zero or leaving these as obtained by Eq. (4.5) and Eq. (4.6) does not affect the eigenvector values $X_{c,1}$ because of the high correlation between the eigenvectors computed in the two different ways (see Figure 4.6).

Notice that, similarly to the eigenvector centrality, the metrics provided by the SDGs-GENEPY framework are also global ones since they account for the overall structure of the network [120]. Nevertheless, although the mathematical structure of Eq. (4.4) is an eigenvector one, the resulting S_c centrality metrics is no longer degree-dominated due to the division of the S_c values by the degree k_c.

A resume of the different weighting approaches for the Sustainable Development Goals we show in this work is given in Table 4.1.

Table 4.1: Weighting approaches through different centrality metrics. In the formulas: S_c is the aggregate score for country c, generally defined according to Eq (4.2); P_{cg} is the value of countries' performances in Goal g; w_g is the weighting value defined in Eq (4.2); v_g is the centrality score for SDGs according to the eigenvector centrality; Y_g is the centrality score for SDGs according to the SDGs-GENEPY framework, Eq (4.4) and $k'_g = \sum_c P_{cg}/k_c$ is the adjusted Goals' degree.

Centrality measure	**Aggregate score**	**Weighting value**
Degree	$S_c = \sum_g P_{cg}$	$w_g = 1$
Eigenvector	$S_c \propto \sum_g P_{cg} v_g$	$w_g = v_g$
GENEPY	$S_c \propto \sum_g P_{cg} \frac{Y_g}{kg'}$	$w_g = \frac{Y_g}{k'_g}$

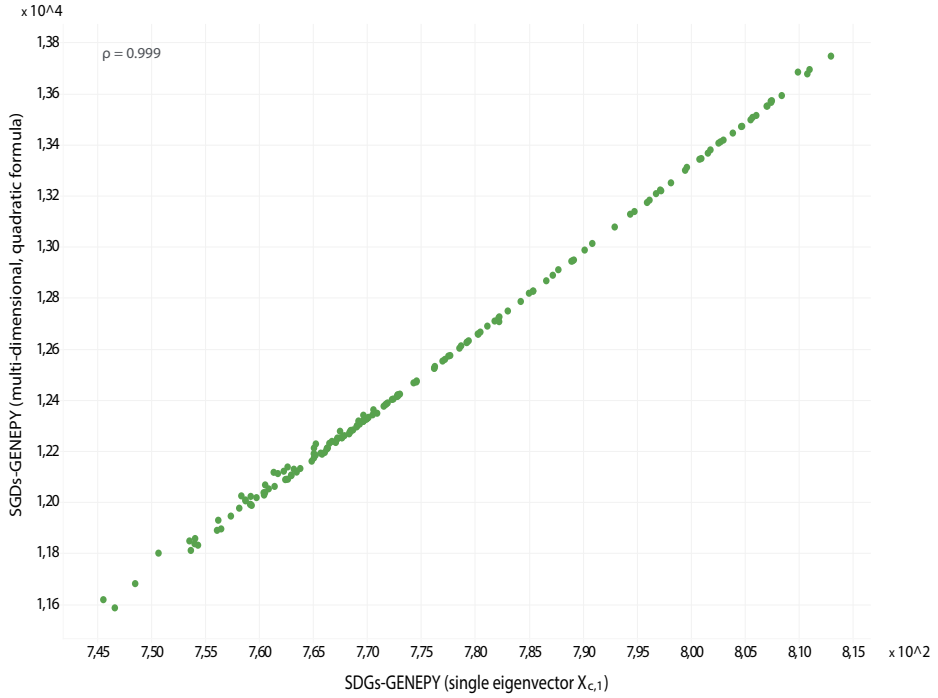

Figure 4.5: Scatter plot between the quadratic and non-quadratic SDGs-GENEPY values of countries in the bipartite countries-SDGs system. On the x-axis, the SDGs-GENEPY values are computing by solving the system in Eqs. (4.4), and computing the first eigenvector $X_{c,1}$ associated to the largest eigenvalue of the matrix N, Eq. (4.5). On the y-axis, the SDGs-GENEPY values are computed by adapting the quadratic formula in Eq. (3.22) provided in Chapter 3 to the the bipartite countries-SDGs system.

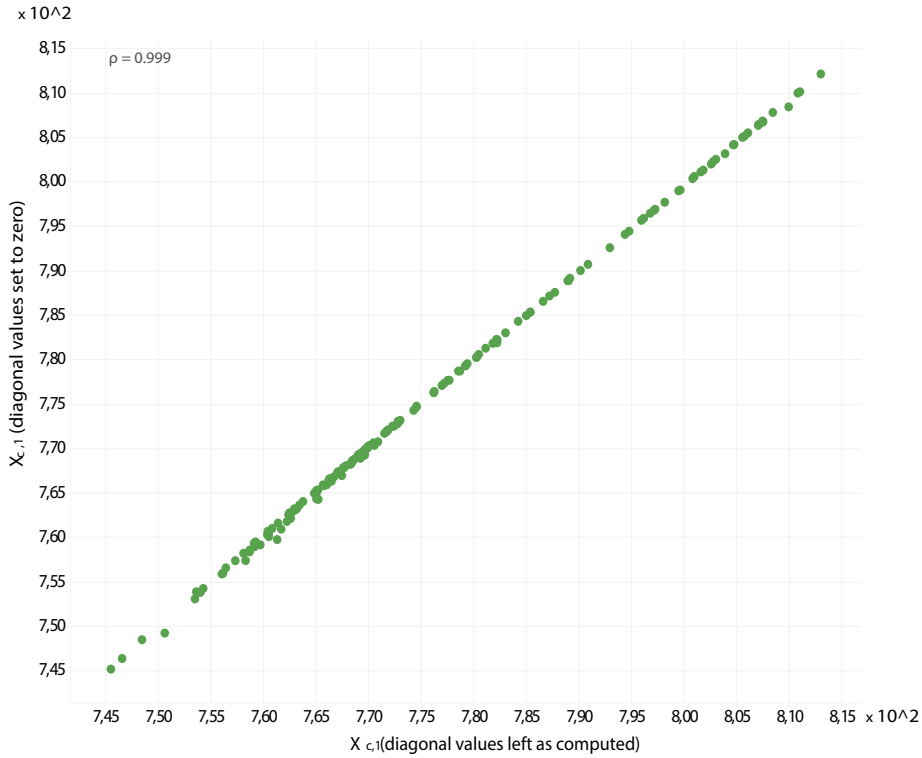

Figure 4.6: Scatter plot between the $X_{c,1}$ values of countries in the bipartite countries-SDGs system computed either or not setting the diagonal values of the matrix $\mathbf{N}$, Eq. (4.5), to zero.

4.3 A Picture of Global Responses in Sustainable Development

The application of the economic complexity theory to the bipartite network of countries and SDGs provides useful insights about how countries are currently responding to the call for actions toward a more equitable, just and sustainable future. We exemplify these results through the application of the SDGs-GENEPY framework on the data from the 2020 Dashboard by Sachs *et al.* [82] (see Section 4.4). Let us start from the results obtained from the computation of the SDGs-GENEPY values for Goals, and, consequently, of the weights Y_p/k'_g. In Figure 4.7 the weighting values Y_p/k'_g are shown. The top-weighted Goal is SDG 9 pertaining with innovation, followed by Zero Hunger and Reduced Inequalities, SDG 2 and 10, respectively. Climate Action (SDG 13) is the least weighted, preceded by SDG 12 and 4, pertaining with sustainable consumption and education, respectively. The wide differences among the weights demonstrate that the SDGs-GENEPY framework is able to capture the contrasting performances among top ranked countries, shown in Fig 4.2. In fact, this weighting of Goals reflects the poor performances by (generally) high performing countries in some SDGs (e.g., Norway in SDG 9, as will be further detailed). Moreover, these results provide a further evidence that the SDGs are not equally integrated in national strategies all around the world. As a consequence, the SDGs-GENEPY weighting values of less prioritized Goals is lower than that of more prioritized ones. (For sake of completeness, the Y_p values are given in Figure 4.8.)

Such a weighting approach determines the ranking of countries according to SDGs-GENEPY score, which differs from the one by the degree centrality. In Figure 4.9, we map countries' rankings according to the SDGs-GENEPY index and the degree value (panels (a) and (b), respectively); panel (c) resumes the differences between the two by scattering the ranking values, with countries color-coded according to Regions, as defined in the 2020 Dashboard [82] (see Section 4.4). As the Figure shows, although the two rankings are mostly aligned (Pearson's correlation coefficient 0.81), significant differences arise. As most remarkable examples, we cite here Singapore (SGP), which jumps from the lower half of the chart to the top of it, moving from position 93 in degree to position 4 in the SDGs-GENEPY S_c, and South Africa (ZAF), moves from 110 in degree to 49 in the SDGs-GENEPY score. Instead, Chile (CHL) moves from the 28-th position in degree, to the 51-th in the SDGs-GENEPY score and Cuba (CUB), which downgrades

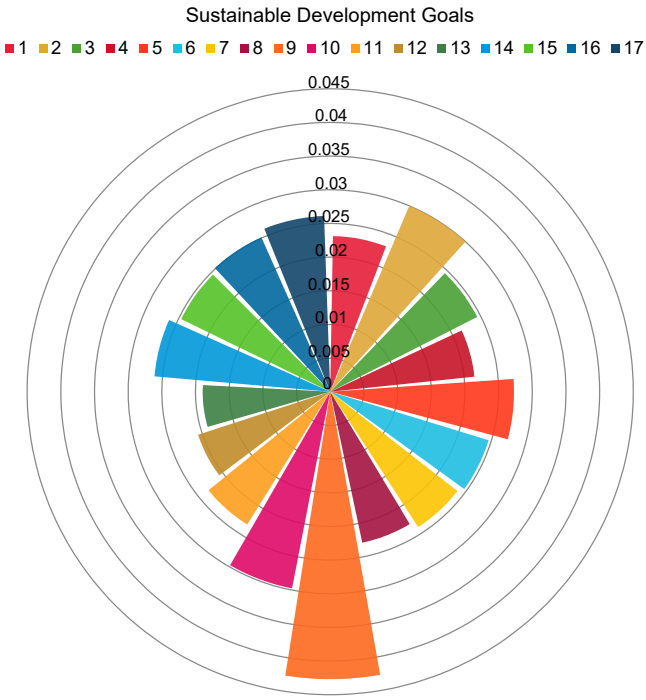

Figure 4.7: The SDGs-GENEPY weights of the Sustainable Development Goals. The radial bar chart plots the SDGs-GENEPY weights Y_g/k'_g for all Goals, see Eqs. (4.4)

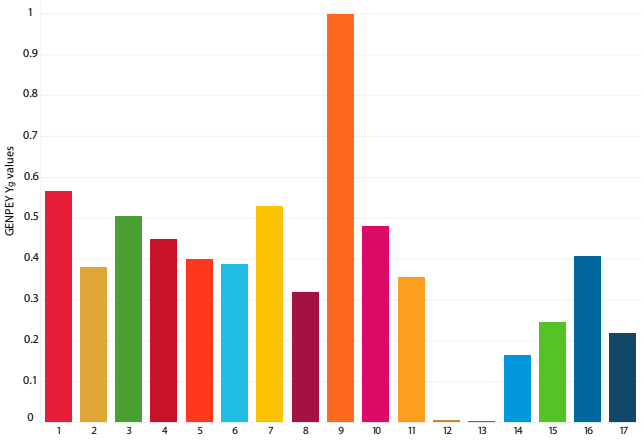

Figure 4.8: The SDGs-GENEPY Y_g values of the Sustainable Development Goals, see Eqs. (4.4). The values are normalised between 0 and 1.

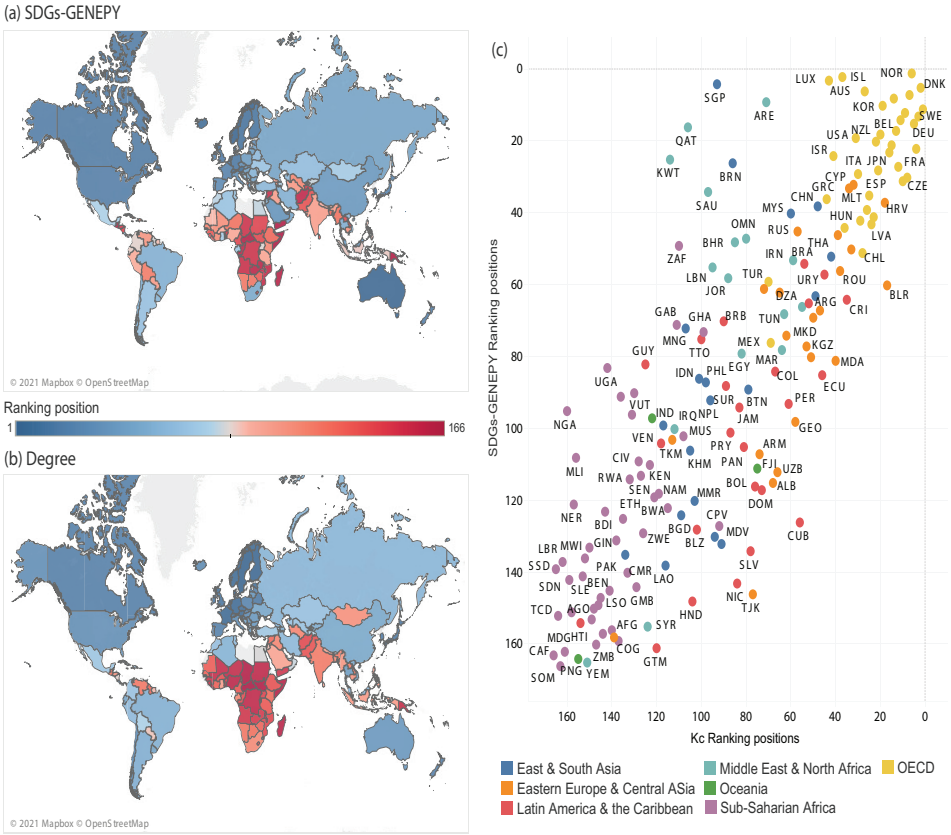

Figure 4.9: Countries rankings according to the degree and SDGs-GENEPY values. In panel (a), countries are colored according to the ranking position computed by the SDGs-GENEPY index. In panel (b) shows the ranking position computed by the degree or, equivalently, the SDG Index [82]. In both maps, ranking position is defined according to descending score (1 = best performer, 166 = worst performer). In panel (c) we scatter the values of the two rankings: on the x-axis is the degree ranking, on the y-axis, the SDGs-GENEPY one. Countries are color-coded according to their Region as specified in the legend, in accordance with the region division in the 2020 Dashboard [82].

from the 56-th position in degree to the 126-th in SDGs-GENEPY S_c.

To explain the reasons behind these variations, we refer to Norway as a relevant example: Norway is among the top absolute performers within SDG 9 (having largest weighting value Y_g/k'_g), together with South Korea and Singapore. Most countries perform poorly within this Goal – only 50% of the value is above the 40% of Goal achievement –, as also represented in Figure 4.2. As a consequence, the SDGs-GENEPY framework assigns a higher weight to countries which are better performers in this Goal. Also, Norway figures as the best absolute performer in Goal 10, and reaches good performances in Goal 2, thus explaining the upgrading of the North-European country from the sixth to the first position in the SDGs-GENEPY S_c ranking. Another relevant example is represented by the case of Singapore, a nation that due to its outstanding performances in more knowledge-intensive SDGs, has reached the third position in SDGs-GENEPY. In contrast, Norway and Singapore are among the worst relative performers in SDGs 13 and 12, respectively (see Figure 4.2), but their low performances in these SDGs are comparatively less relevant within the SDGs-GENEPY framework, due to the lower weight values assigned to these two Goals.

4.4 A Note on the Use of the SDG Index and Dashboards

Notwithstanding the call for efforts toward the standardization in the data collection by all National Statistical Systems, NSSs, launched by the Cape Town Global Action Plan in 2017 [176], the data accessible at the UN Statistics Division (available at `https://unstats.un.org/sdgs/indicators/database/`) clearly show that work is still needed to have a comprehensive, homogeneous, and extensive database covering all countries and years under the Agenda 2030 and beyond. For this reason, the input data we are using are taken from the 2020 SDG Index and Dashboard [82], which represent a commendable step forward in data collection, homogenization and assessment of countries progresses in sustainable development. The aim of the Dashboard is to provide yearly rankings of UN countries based on an aggregated score of all Goals' performances. The score is intended to be readable as a percentage of achievement of all the Goals, ranging from 0 to 100; therefore, countries close to 100 are approaching the complete fulfilling of the Agenda's Goals according to the indicators used to compute the score [90]. The score is constructed upon a number of indicators providing

quantitative information about countries performances. All listed indicators are normalized according to an optimum and a minimum value of indicator performance to ensure comparability and aggregation of measurements (we refer the reader to [82, 90] for further details). Listed indicators are updated every year, accounting for advances in monitoring and research. In order to provide statistical-sound results, we only refer to 2020 data, thus not inferring any possible missing data back in other years' Dashboards. The 2020 data-set constitutes of 115 indicators across the Goals, 30 of which are specifically defined for the members of the Organization for Economic Co-operation and Development (OECD). The Dashboard only includes countries covering at least 85% of the indicators, totalling 166 out of 193 UN countries. To have OECD-specific indicators entails that, with respect to the same Goal g, the term N_{cg} (from which, in Eq (4.1), the value of performance P_{cg} is obtained) differs between OECD and other countries. The Dashboard also introduces Regional scores, assigning countries to 7 different Regions around the world, namely: Sub-Saharan Africa, Middle East and North Africa – MENA –, East and South Asia, Eastern Europe and Central Asia, Latin America and the Caribbean – LAC –, Oceania and OECD group, which we use to color-code countries in Figure 4.9. In line with the methodology exemplified with the SDG Index, we replaces countries' missing data with the Regional score in that same Goal [90].

Concluding Remarks

The problem of defining aggregated scores in sustainable development is a recurrent one, also required to track the path toward the achievement of the Goals within the Agenda 2030 and many strategies can be pursued for their computation (see, e.g., [81, 82]). Nevertheless, the complexity of the Agenda 2030 should not be neglected when defining aggregated scores (the complexity is related to the presence of trade-offs and synergies among the Sustainable Development Goals and the heterogeneity of countries' challenges and responses [79, 81]). In light of this complexity, in this work we have introduced a novel perspective on sustainable development in which we addressed, within a network science framework, the need for ranking countries for their status with respect to what set by the Agenda. In particular, we show that the countries-SDGs system can be structured as a bipartite network and that, by using the centrality tools, different weighting approaches naturally emerge for the computation of aggregated scores to rank countries accordingly.

Thanks to this network representation of the system, we show that the SDG Index identified by Sachs *et al.* [82] – which, in line with the Agenda's principles, considers equal weights for all Goals – corresponds to measure the degree of countries. In network science, the degree centrality measures the local behavior of the node and it does not account for the complex interconnections of the system. A first step toward the use of global metrics to account for the structure of the network is the use of the eigenvector centrality. However, we have demonstrated that in the countries – SDGs system, the information about the degree of countries recurs even when evolving to the eigen-centrality. Besides the formal reasoning about the spectral gap, the strong correlation between the two centrality metrics is due to the fact that countries' performances in SDGs are highly correlated, especially if they have similar degree (see Figure 4.10). This fact highlights that countries set in similar development conditions [177] tend to emulate each other performances [178] and explains why, when ranked for their degree, nearby positioned countries show similar behavioral patterns (see Fig. 4.2). Nevertheless, the relative spectra shown in Fig. 4.2 also show the heterogeneity of all countries' performances around the world also beyond their average value (or equivalently, the degree). This suggests the need for more subtle metrics able to unravel the complexity of the system. We address this change of weighting perspective through the GENeralized Economic complexity framework (SDGs-GENEPY) [120].

The SDGs-GENEPY approach we propose for the creation of one aggregated score brings two main positive advancements. Firstly, the weights $w_g = Y_g/k'_g$ are self-emerging from the data, and they account for the relative performances of countries as measured by term k'_g. Secondly, the division of the SDGs-GENEPY S_c values for k_c – intrinsic of the computation of the index – removes the undesired degree-bias which is known to affect eigenvector-based centrality measures [170], thus providing useful insights about the countries' status in sustainable development. These characteristics of the SDGs-GENEPY framework can be interpreted in light of further considerations about the Agenda 2030. Countries whose relative performance value P_{cg}/k_c in Goal g is greater than that in other SDGs, give a higher contribution to the term k'_g. Its inverse $1/k'_g$ possibly diagnoses structural limitations in achieving the Goal: higher values of $1/k'_g$ are obtained for those Goals in which only few countries have positive relative performances. Therefore, we can assume that heavier (in sense of weights) Goals are also those ones that some countries favor with national strategies, to the detriment of other Goals. This is witnessed by the fact of having found Climate

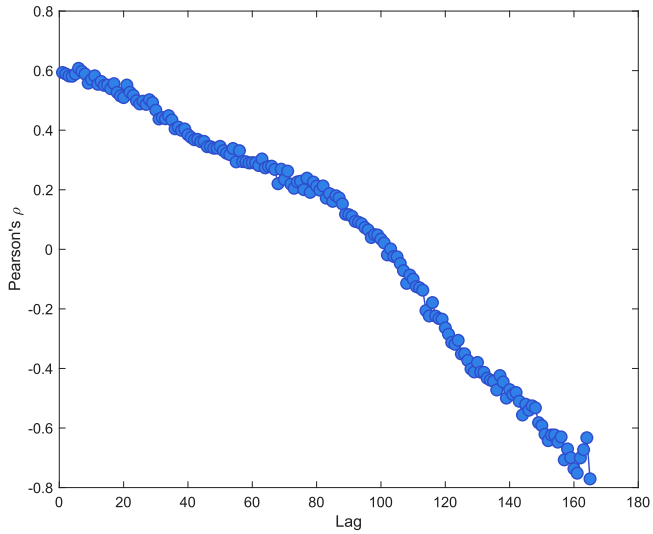

Figure 4.10: **Pearson's correlation values of countries' performances.** We consider the correlation between the rows of the matrix of performances **P**. The rows of the matrix, and so countries, are ordered – top to bottom – according to decreasing value of countries' degree, k_c. This allows one to order countries according to the similarities in their development conditions. For sake of representation, we summarize correlation through the computation of lagged correlations: each point on the plot represents the average value of all correlations between countries positioned at a defined lag step. Therefore, the last point of the plot defines the correlation between the first and last rows (i.e., countries), for which a negative correlation value is obtained.

Action and Innovation as, respectively, the lowest and greatest weighted Goals, whose $w_g = Y_g/k'_g$ values are mainly determined by the relative performances of high-income and sustainable-outperforming countries, such as Norway (see Figure 4.2). Evidence of the validity of such analysis can be found in Norway's strategies of development, among the most relevant example in this study. Norway is currently diversifying its industrial sector by enhancing investments in the Research and Development area, so to face the reduction in prices of crude oil [179, 180] (see, also, the Climate Action Tracker, `https://climateactiontracker.org/countries/norway/`). In fact, Norway is one of the world-wide leader exporter of crude oil [140], a fact that puts under the spotlight Norway's shared responsibility in Climate Action and the permanent presence of trade-offs between economic and environmental issues at the world level [181]. Therefore, in the SDGs-GENEPY indexing approach, the heterogeneity of countries and contrasting policy implementations are naturally embedded through the data and brought up

by the algorithm, determining the weights of SDGs. This hierarchy testifies the shared global responsibility in sustainable development and the intrinsic compromise among political willingness, opportunities and capacities to move toward sustainable development [82, 174, 175]. This compromise is even more evident in countries with more favorable conditions to fulfill the Agenda, resulting in higher 'knowledge' (i.e., policy and intervention designs and implementations; awareness and preparedness to face the challenges [62, 75, 82, 171–173]).

In light of these considerations, we can interpret the SDGs-GENEPY ranking of countries as a picture of shared responsibilities, where it emerges the possibility for nations to act like role-models and promote the achievement of global sustainable development. In light of the emulation phenomena among countries [178], we argue that to identify role-model countries is rather relevant and in can pave the way to a new strategy for boosting sustainable development in the next decade. In particular, our ranking can be used as an '*ex post*' and complementary tool to the Rapid Integrated Assessment – RIA – analysis [61] which the United Nations conduct to monitor the willingness of countries in integrating the Goals within their national strategies. In this sense, our analysis would effectively provide insights about the implementation of such plans, also providing a tool for comparing the efforts across countries. Moreover, such approach can be suitably adapted to sub-national level by using regional data on sustainability performance, thus revealing crucial features of countries' regional efficiency in sustainable development.

In conclusion, the burgeoning literature in the field of SDGs assessments suggests the presence of different ideologies about how to properly measure the status of countries for their sustainability path. In light of the complexity of the system defined within the Agenda 2030, we realize that the understanding of such paths should not be shrunk to a single indicator analysis. Therefore, to fully understand countries' paths toward sustainable development, we suggest the use of different and complementary mathematical approaches, as, e.g., the computation of both the degree and SDGs-GENEPY ranking. Such parallel analysis would provide a bird's-eye view of the conditions of countries to achieve sustainable development while providing a list of change-making places and actions that can help meeting the 2030 deadline.

5

Final Remarks

The present thesis contributes to interdisciplinary research literature in complexity science, economics, innovation and sustainable development. The central theme of this contribution is represented by network science, which has been recognized to be a powerful mathematical tool to disentangle the complexity of systems, and its applications range in many fields. Nevertheless, the heterogeneity of some network science applications, especially the ones concerning with the use of centrality measures, possibly mine the use of this theory in providing novel insights about the systems under study.

Therefore, as first contribution, we began with addressing the general problem of identifying central nodes (i.e., important) in a network. In the first part of this work, in Chapter 2, we recast the question *"what does it mean to be central in a network ?"* within a statistical framework. Within such framework, different centrality measures are used as explaining variables in an estimation exercise of the adjacency matrix of the network and their performances are compared according to their ability to provide better estimates. Under this statistical perspective, the frequently used one-dimensional centrality metrics for undirected (i.e., degree, eigenvector and Katz centrality) and directed networks (i.e., degree and hub-authority centrality), provide the same information and perform poorly in reconstructing the network. As a solution, we introduced multi-dimensional centrality metrics to improve the estimation results, also providing with the useful feature of ranking the nodes in the network thanks to the use of the statistical concept of *unique contribution*.

Taking the multi-dimensional metrics of nodes' centrality as the starting point, Chapter 3 contributes to the literature in the Economic Complexity field. This new economic theory faces the shrinking of information that affects most economic models and indicators (as the Gross Domestic Products) by introducing a data-driven analysis of the innovation and growth potential of countries at the world level. The theory is network based, as it

introduces the structuring of the export data as a bipartite system of countries and traded commodities. In order to analyse the innovation potential of countries, special centrality metrics have been introduced to deal with such system. However, also in this field the contrasting theories on how to measure the importance of the nodes – in this particular case, innovation – undermine their potential. By exploiting the results on multi-dimensional metrics of Chapter 2 and the linear algebra upon which these are constructed, we reconcile the most notorious metrics of Economic Complexity in one single value, the GENeralized Economic comPlexitY index (GENEPY). We have shown that the GENEPY is able to track the trajectories of growth of countries without the need to add exogenous information about countries' economies. Among the many other advantages of the GENEPY index that we discussed (its linearity and neatness), the GENEPY formulation paves the way to the micro-foundation of the economic complexity. In fact, the GENEPY index has the same structure as the microeconomics-based EXPY index and this is not an *a priori* construction. Therefore, the GENEPY framework can actually shed new light about the microeconomics capabilities of countries to boost their performances at the macroeconomic level.

The final contribution of this thesis deals with sustainable development and, in particular, the Agenda 2030. The global call for actions that the United Nations have introduced in 2015 targets sustainable development of all countries through the definition of the 17 Sustainable Development Goals. These Goals promote a more equal, just and sustainable future in which socio-economic and environmental development are in balance. The system of countries and Goals within the Agenda 2030 is a complex one, due to the demonstrated synergies and trade-off among the Goals and, nonetheless, due to the heterogeneity of countries. While network theory has been used to unveil the interconnections among the Goals, no substantial literature has been found that concerns with its use to rank countries according to their status in sustainable development; although, the strategy of indexing in the development field is a recurrent one. Therefore, we have introduced the bipartite network representation of the countries – Goals system that allowed us to analyse the definition of rankings of countries through centrality metrics. Against the need to embed the network complexity of the interconnections between countries and Goals, we introduced the use of the GENEPY framework to this system to shed new light about country's efficiency in sustainable development. In fact, we suggest that the adoption of the economic complexity theory to the field of sustainable development can possibly identify role-model countries to address change making actions

and sheds further lights about the trade-offs between environmental and economic issues.

In a nutshell, in this work we have presented unique and novel network science approaches to track and address global performances of innovation and sustainability.

5.1 Limitations, Future Works and Perspectives

The results of this thesis are mainly based on statistical analyses and data-driven approaches. Each of our contributions presents room for improvements and contexts of application.

Firstly, the statistical perspective about network centrality has only been defined for a subset of centrality metrics, i.e., the degree and eigenvector-based centrality. While some centrality metrics based on paths measuring, as the closeness, can be recast in this statistical perspective through specifically-defined matrix, this might not be true for other metrics. Against many efforts, an exact solution of this statistical perspective for the PageRank centrality has not been defined. The PageRank is a peculiar centrality metrics which provides a random walk interpretation of the network; and its popularity is increasing. Differently from the other measures set in directed networks, the Page Rank centrality neglects one of the two properties of the nodes – namely, the outgoing edges – and it only measures the in-centrality of the nodes. This asymmetric characteristic of the metrics already biases the use of the Page Rank as an estimator of the original network, which is intrinsically characterized by two dimensions as defined by the directions of the edges. Therefore, we address future work toward matrix manipulation that can allow one to recover the Page Rank centrality within the proposed statistical framework. Also, we address exploration of the results concerning different errors, estimator functions definitions or statistical estimation method (as using the maximum likelihood, perhaps). Eventually, further work could be dedicated to considering the Shapley values instead of the Unique Contribution for valuing the estimation power of the centrality scores.

Secondly, the GENEPY framework defines centrality metrics for bipartite systems. Although we have proven its efficacy in shedding new light on the system of sustainable development, the application of the GENEPY framework cannot be generalized to all bipartite systems due to its mathematical linear nature. In fact, we have discussed the role of non-linearity in unveiling the nestedness feature of some systems, in particular the ecological

ones, where entities remain trapped in some peculiar states, not having the possibility to explore all the possible states in finite (or even infinite) observation time. Instead, in the economic complexity and the sustainable development field we expect all the countries to be able – sooner or later – to improve the sophistication of their export baskets and to complete the fulfillment of the Goals, respectively. The systems that are able to explore all their possible states in finite times are called *ergodic*. We speculate that the GENEPY framework, due to its linear nature, could be applied to more ergodic systems (i.e., systems where the ergodicity condition is reached on time scales not to far from the characteristic time scales of the system); whereas, non-ergodic systems, such as that of plant-pollinators, may require other algorithms suitable for their intrinsic non-linear nature. Therefore, we plan to address future work in the understanding of the role of linearity in these non-ergodic systems. Moreover, as concerns the field of economic complexity, we address future work toward the micro-foundation of the field due to the similarities with the EXPY framework and the understanding of the possible predicting power of the GENEPY index about economic growth.

Thirdly, the application of network science to the Agenda 2030 is a powerful tool to address efforts toward the fulfillment of the Goals. Nevertheless, these directions hardly depend on the input data used to run the GENEPY analysis. For more extensive and comprehensive studies of countries' efficacy, we address future work toward the collection and homogenization of countries' performances within the indicators of sustainable development, especially those ones in line with the Agenda's definitions.

In conclusion, this thesis confirms the importance of network science and data-driven approaches in providing new perspectives about socio-economics systems, perspectives that should accompany more standard doctrines in defining change making actions.

Appendices

A

The Implications of Becoming a Relevant Exporter for Water Resources

Food demand has an environmental impact on the water resources [182]. In fact, food production, including crops and animal products, requires water to be processed. This consumption of water resources for production of goods (or even services) is defined as Water Footprint (WF) [183] and it shows a global variability of its own (it depends on climate factors, land characteristics, type of cultivation, irrigation system, among the many other factors) [184]. For agricultural products, the most water consumption happens during the process of growing the crop or tree, i.e., from seeding to harvest. Other steps of the supply chain generally involve a smaller amount of water [183]. The amount of water which is embedded in a good at the end of the production stages is defined as Virtual Water Content (VWC) [184].

By considering VWC, the international trade of agricultural products also entails trading the water resources of the countries in which production occurs [185]. Therefore, the impact of trade on national local resources varies according to the water requirements of the product traded and to the water availability of the exporting country [182]. Nevertheless, the analyses we here present about the relationship between market competition dynamics and water footprint shows that trade choices of countries sometimes produce the over-exploitation of water resources.

We consider two different crop products, maize and vanilla, which are set in two different market dynamics [186]. On the one hand, maize is considered to be a *staple* product, i.e., it is set in a relatively more competitive market and its production is ubiquitous. In determining its pricing values, maize producers also tend to account for its water footprint on the national water resources [186]. On the other hand, the so-called *cash* crops, such as vanilla, are often produced in situations of oligopoly, in which large producing and

trading countries mainly determine prices without accounting for the water footprint [186]. In economic complexity terms, maize is less complex than vanilla, and the latter constitute a niche product.

The law of comparative advantage in economics states that, under free trade conditions, countries tend to produce more of a good for which they can economically gain from trade, thus assigning production to export rather than to local consumption [187]. As stated in Chapter 3, the relevance of this advantage can be computed through the Revealed Comparative Advantage method [121], which accounts for the global export of the product and the share of the countries within the global market.

From the 'water footprint' point of view, such advantage may become a main driver of water-resources depletion, also providing poor gain in comparative advantage. To provide an example, in Table A.1 we report the top three worldwide producers of maize (USA, China and Brazil) and vanilla (Madagascar, Indonesia and China), detailing the volumes of their production and export, the water footprint of their production and their RCA values in that product.

As we show in Table A.1, the Revealed Comparative Advantage of Madagascar in the export of vanilla is relevantly greater than 1. The export of vanilla from Madagascar is almost two third of the worldwide trade of vanilla. In terms of water cost, the virtual water export through vanilla is almost 1 billion cubic meters. Brazil generates a water footprint of 24.8 billion cubic meters on its water resources to gain 13.62 points of RCA in maize export. Roughly the same water footprint is generated in USA for the export of maize, however determining a lower RCA score for the country. However, not all the top three worldwide producers of maize and vanilla relevantly gain in trade from the over-exploitation of their water resources. This is particularly true for China, which generates a big water footprint for the generation of the export of maize and vanilla that provide a very small Revealed Comparative Advantage, 0.0017 and 0.009, respectively. In terms of Economic Complexity, these values are clearly meaningless, especially if compared to the other countries' RCA values in the same products and to the standard threshold value of $RCA = 1$. Nevertheless, such exports are clearly relevant for the water resources of China. A small export contribution of 0.13 % of the Chinese vanilla produces a water footprint of 2.7 million cubic meter.

Table A.1: For the top three producers of maize and vanilla worldwide, we detail their water footprint of production and export and the economic gain (detailed as price per unit, share of export worldwide, volume of dollars gained from the export and the Revealed Comparative Advantage). The data on the water footprints are taken from the CWASI database [188]; data on export volumes in dollars are taken and processed from the CEPII-BACI dataset [140]. Data on prices of production are taken from Falsetti *et al.* [186]. All data refer to 2015.

Variables	**Maize**			**Vanilla**		
Producers/ Exporter	**USA**	**CHN**	**BRA**	**MDG**	**IDN**	**CHN**
Water footprint of production (m^3/ton)	479	881	723	339 868	90 783	64 478
Volume of production (ton)	34.5 M	26.5 M	8.5 M	292	200	57
Price of production per unit (USD/ton)	130	602	227	3 412	1 708	16 912
Export share (% global amount of export)	30.6 %	17.5 %	0.03 %	62.2 %	5.19 %	0.13 %
Volume of water export (m^3)	21.3 B	8.4 M	24.8 B	0.96 B	34.1 M	2.7 M
Volume of export (USD)	9.7 B	7.93 M	5.18 B	290 M	24.2 M	626 K
Revealed Comparative Advantage (relevant exporter for $RCA > 1$)	3.42	0.0017	13.62	3 641	4.6	0.009

B

The GENEPY Ranking of Countries in Economic Complexity

Results refer to 2017. Further details are provided in Chapter 3.

Table B.1: Ranking positions of countries for their economic complexity, computed according to ascending values of GENEPY. Results refer to year 2017. For further details refer to Chapter 3.

Position	iso-3 code	Country	Position	iso-3 code	Country	Position	iso-3 code	Country
1	JPN	Japan	60	VGB	British Virgin Islands	119	CUQ	Curaçao
2	KOR	Korea, Rep. of Korea	61	COL	Colombia	120	SLE	Sierra Leone
3	DEU	Germany	62	LBN	Lebanon	121	RWA	Rwanda
4	CHE	Switzerland-Liechtenstein	63	PRK	Korea, Dem. People's Rep. of	122	TGO	Togo
5	USA	United States of America	64	PAK	Pakistan	123	NIC	Nicaragua
6	CHN	China	65	URY	Uruguay	124	ETH	Ethiopia
7	CZE	Czech Republic	66	LKA	Sri Lanka	125	MOZ	Mozambique
8	GBR	United Kingdom	67	ARG	Argentina	126	SAU	Saudi Arabia
9	HKG	Hong Kong (SARC)	68	BRB	Barbados	127	TJK	Tajikistan
10	SWE	Sweden	69	KHM	Cambodia	128	BDI	Burundi
11	SGP	Singapore	70	MKD	The former Yugoslav Rep. of Macedonia	129	BTN	Bhutan
12	AUT	Austria	71	MDA	Moldova, Rep.of	130	KWT	Kuwait
13	ITA	Italy	72	KGZ	Kyrgyzstan	131	BEN	Benin
14	MYS	Malaysia	73	JOR	Jordan	132	MWI	Malawi
15	BEL	Belgium-Luxembourg	74	UZB	Uzbekistan	133	GHA	Ghana
16	ISR	Israel	75	SYR	Syrian Arab Republic	134	CMR	Cameroon
17	THA	Thailand	76	KAZ	Kazakstan	135	AZE	Azerbaijan

Table B.1 continued from previous page

Position	iso-3 code	Country	Position	iso-3 code	Country	Position	iso-3 code	Country
18	FRA	France	77	MUS	Mauritius	136	CUB	Cuba
19	NLD	Netherlands	78	CRI	Costa Rica	137	CIV	Côte d'Ivoire
20	FIN	Finland	79	AUS	Australia	138	TTO	Trinidad and Tobago
21	SVN	Slovenia	80	MNE	Montenegro	139	GIB	Gibraltar
22	SVK	Slovakia	81	CHL	Chile	140	BHS	Bahamas
23	HUN	Hungary	82	DOM	Dominican Republic	141	COD	Democratic Republic of the Congo
24	POL	Poland	83	PER	Peru	142	SDN	Sudan (2011)
25	PHL	Philippines	84	ARE	United Arab Emirates	143	DZA	Algeria
26	IRL	Ireland	85	ARM	Armenia	144	TKM	Turkmenistan
27	IND	India	86	ALB	Albania	145	MHL	Marshall Islands
28	ESP	Spain	87	ABW	Aruba	146	NER	Niger
29	DNK	Denmark	88	MMR	Myanmar	147	GMB	Gambia
30	ROU	Roumania	89	MAR	Morocco	148	GUY	Guyana
31	MEX	Mexico	90	SYC	Seychelles	149	NCL	New Caledonia
32	HRV	Croatia	91	BGD	Bangladesh	150	PNG	Papua New Guinea
33	EST	Estonia	92	FJI	Fiji	151	BRN	Brunei Darussalam
34	IDN	Indonesia	93	PRY	Paraguay	152	VEN	Venezuela
35	VNM	Viet Nam	94	ISL	Iceland	153	SUR	Suriname
36	LTU	Lithuania	95	KEN	Kenya	154	NGA	Nigeria
37	PRT	Portugal	96	GEO	Georgia	155	BFA	Burkina Faso
38	TUR	Turkey	97	BHR	Bahrain	156	SOM	Somalia
39	BLR	Belarus	98	ATG	Antigua and Barbuda	157	ERI	Eritrea

Table B.1 continued from previous page

Position	iso-3 code	Country	Position	iso-3 code	Country	Position	iso-3 code	Country
40	BRA	Brazil	99	MDG	Madagascar	158	COG	Congo
41	CYP	Cyprus	100	JAM	Jamaica	159	YEM	Yemen
42	BGR	Bulgaria	101	LAO	Lao People's Democratic Republic	160	GIN	Guinea
43	LVA	Latvia	102	GTM	Guatemala	161	QAT	Qatar
44	CAN	Canada	103	HND	Honduras	162	LBR	Liberia
45	NOR	Norway	104	UGA	Uganda	163	VUT	Vanuatu
46	UKR	Ukraine	105	OMN	Oman	164	SLB	Solomon Islands
47	TUN	Tunisia	106	SEN	Senegal	165	GAB	Gabon
48	RUS	Russian Federation	107	IRN	Islamic Republic of Iran	166	MRT	Mauritania
49	SRB	Serbia	108	BLZ	Belize	167	FLK	Falkland Islands (Malvinas)
50	NPL	Nepal	109	ZMB	Zambia	168	GRL	Greenland
51	MAC	Macau	110	MNG	Mongolia	169	TCD	Chad
52	SLV	El Salvador	111	HTI	Haiti	170	CYM	Cayman Islands
53	ZAF	South Africa	112	BOL	Bolivia	171	MDV	Maldives
54	MLT	Malta	113	VCT	Saint Vincent and the Grenadines	172	LBY	Libyan Arab Jamahiriya
55	BIH	Bosnia and Herzegovina	114	ZWE	Zimbabwe	173	AGO	Angola
56	EGY	Egypt	115	AFG	Afghanistan	174	GNB	Guinea-Bissau
57	GRC	Greece	116	ECU	Ecuador	175	SSD	South Sudan
58	NZL	New Zealand	117	TZA	Tanzania, United Rep. of	176	IRQ	Iraq

Table B.1 continued from previous page

Position	iso-3 code	Country	Position	iso-3 code	Country	Position	iso-3 code	Country
59	PAN	Panama	118	MLI	Mali	177	GNQ	Equatorial Guinea

C

The GENEPY Ranking of Countries in Sustainable Development

Results refer to 2020. Further details are provided in Chapter 4.

Table C.1: Ranking positions of countries for their sustainable development, computed according to ascending values of GENEPY. Results refer to year 2020. For further details refer to Chapter 4.

Position	iso-3 code	Country	Position	iso-3 code	Country	Position	iso-3 code	Country
1	NOR	Norway	57	URY	Uruguay	113	SEN	Senegal
2	AUS	Australia	58	JOR	Jordan	114	RWA	Rwanda
3	LUX	Luxembourg	59	TUR	Turkey	115	ALB	Albania
4	SGP	Singapore	60	BLR	Belarus	116	BOL	Bolivia
5	DNK	Denmark	61	MNE	Montenegro	117	DOM	Dominican Republic
6	ISL	Iceland	62	KAZ	Kazakhstan	118	NAM	Namibia
7	AUT	Austria	63	VNM	Vietnam	119	BWA	Botswana
8	CHE	Switzerland	64	CRI	Costa Rica	120	MMR	Myanmar
9	ARE	United Arab Emirates	65	ARG	Argentina	121	NER	Niger
10	KOR	Korea, Rep.	66	DZA	Algeria	122	STP	Sao Tome and Principe
11	SWE	Sweden	67	UKR	Ukraine	123	BDI	Burundi
12	NLD	Netherlands	68	TUN	Tunisia	124	BGD	Bangladesh
13	FIN	Finland	69	BIH	Bosnia and Herzegovina	125	ETH	Ethiopia
14	BEL	Belgium	70	BRB	Barbados	126	CUB	Cuba
15	DEU	Germany	71	GAB	Gabon	127	CPV	Cabo Verde
16	QAT	Qatar	72	MNG	Mongolia	128	BLZ	Belize
17	GBR	United Kingdom	73	GHA	Ghana	129	ZWE	Zimbabwe
18	CAN	Canada	74	MKD	North Macedonia	130	LKA	Sri Lanka
19	USA	United States	75	TTO	Trinidad and Tobago	131	DJI	Djibouti
20	NZL	New Zealand	76	MEX	Mexico	132	MDV	Maldives

Table C.1 continued from previous page

Position	iso-3 code	Country	Position	iso-3 code	Country	Position	iso-3 code	Country
21	IRL	Ireland	77	AZE	Azerbaijan	133	GIN	Guinea
22	FRA	France	78	MAR	Morocco	134	SLV	El Salvador
23	JPN	Japan	79	EGY	Egypt, Arab Rep.	135	PAK	Pakistan
24	ISR	Israel	80	KGZ	Kyrgyz Republic	136	MWI	Malawi
25	KWT	Kuwait	81	MDA	Moldova	137	LBR	Liberia
26	BRN	Brunei Darussalam	82	GUY	Guyana	138	LAO	Lao PDR
27	SVN	Slovenia	83	UGA	Uganda	139	SSD	South Sudan
28	ESP	Spain	84	COL	Colombia	140	CMR	Cameroon
29	ITA	Italy	85	ECU	Ecuador	141	SLE	Sierra Leone
30	CZE	Czech Republic	86	IDN	Indonesia	142	SDN	Sudan
31	EST	Estonia	87	PHL	Philippines	143	NIC	Nicaragua
32	MLT	Malta	88	SUR	Suriname	144	GMB	Gambia, The
33	CYP	Cyprus	89	BTN	Bhutan	145	LSO	Lesotho
34	SAU	Saudi Arabia	90	MRT	Mauritania	146	TJK	Tajikistan
35	PRT	Portugal	91	BFA	Burkina Faso	147	BEN	Benin
36	GRC	Greece	92	NPL	Nepal	148	HND	Honduras
37	HRV	Croatia	93	PER	Peru	149	COM	Comoros
38	CHN	China	94	JAM	Jamaica	150	TGO	Togo
39	SVK	Slovak Republic	95	NGA	Nigeria	151	COD	Congo, Dem. Rep.
40	MYS	Malaysia	96	TZA	Tanzania	152	TCD	Chad
41	POL	Poland	97	VUT	Vanuatu	153	AGO	Angola
42	HUN	Hungary	98	GEO	Georgia	154	HTI	Haiti
43	LVA	Latvia	99	IND	India	155	SYR	Syrian Arab Republic

Table C.1 continued from previous page

Position	iso-3 code	Country	Position	iso-3 code	Country	Position	iso-3 code	Country
44	LTU	Lithuania	100	IRQ	Iraq	156	MOZ	Mozambique
45	RUS	Russian Federation	101	PRY	Paraguay	157	SWZ	Eswatini
46	BGR	Bulgaria	102	MUS	Mauritius	158	AFG	Afghanistan
47	OMN	Oman	103	TKM	Turkmenistan	159	COG	Congo, Rep.
48	BHR	Bahrain	104	VEN	Venezuela, RB	160	ZMB	Zambia
49	ZAF	South Africa	105	PAN	Panama	161	GTM	Guatemala
50	SRB	Serbia	106	KHM	Cambodia	162	MDG	Madagascar
51	CHL	Chile	107	ARM	Armenia	163	CAF	Central African Republic
52	THA	Thailand	108	MLI	Mali	164	PNG	Papua New Guinea
53	IRN	Iran, Islamic Rep.	109	CIV	Cote d'Ivoire	165	YEM	Yemen, Rep.
54	BRA	Brazil	110	KEN	Kenya	166	SOM	Somalia
55	LBN	Lebanon	111	FJI	Fiji			
56	ROU	Romania	112	UZB	Uzbekistan			

Nomenclature

Abbreviation

EC	Economic Complexity
ECI	Economic Complexity Index
FC	Fitness and Complexity algorithm
GDP	Gross Domestic Product
GDP PPP	Gross Domestic Product at Power Purchasing Parity
GENEPY	GENeralized Economic comPlexitY
HS	Harmonized System
MR	Method of Reflections
PCI	Product Complexity Index
RCA	Relative Comparative Advantage
SDG	Sustainable Development Goal
SDGs	Sustainable Development Goals
SE	Sum of squared Errors
SVD	Singular Value Decomposition
TSS	Total Sum of Squares
UC	Unique Contribution
UN	United Nations

Recurrent symbols

λ	Eigenvalues of a generic matrix
σ	Singular values of a generic matrix
k	Degree of a node

List of Figures

List of Tables

References

1. Capra, F. *The Hidden Connections: A Science for Sustainable Living* (Anchor, 2004).
2. Domínguez-García, V. & Munoz, M. A. Ranking Species in Mutualistic Networks. *Scientific Reports* **5,** 8182 (2015).
3. Borgatti, S. & Everett, M. Models of Core/Periphery Structures. *Social Networks* **21,** 375–395 (2000).
4. Hausmann, R., Hwang, J. & Rodrik, D. What You Export Matters. *Journal of Economic Growth* **12,** 1–25 (2007).
5. Giustolisi, O., Ridolfi, L. & Simone, A. Embedding the Intrinsic Relevance of Vertices in Network Analysis: The Case of Centrality Metrics. *Scientific Reports* **10,** 1–11 (2020).
6. Newman, M. E., Barabási, A. & Watts, D. J. *The Structure and Dynamics of Networks.* (Princeton University Press, 2006).
7. Barabási, A. *et al. Network Science* (Cambridge University Press, 2016).
8. Ladyman, J., Lambert, J. & Wiesner, K. What Is a Complex System? *European Journal for Philosophy of Science* **3,** 33–67 (2013).
9. Newman, M. E. *Network - An Introduction* (Oxford University Press, 2010).
10. Boguslawski, P. *Modelling and Analysing 3D Building Interiors with the Dual Half-Edge Data Structure* PhD thesis (University of Glamorgan Pontypridd, Wales, UK, 2011).
11. Caldarelli, G. *Scale-Free Networks: Complex Webs in Nature and Technology* (Oxford University Press, 2007).
12. Bavelas, A. Communication Patterns in Task-Oriented Groups. *Journal of the Acoustical Society of America* **22,** 725–730 (1950).
13. Padgett, J. F. & Ansell, C. K. Robust Action and the Rise of the Medici, 1400-1434. *American Journal of Sociology* **98,** 1259–1319 (1993).
14. Caplow, T. A Theory of Coalitions in the Triad. *American Sociological Review* **21,** 489–493 (1956).

15. Bonacich, P. Power and Centrality: A Family of Measures. *American Journal of Sociology* **92,** 1170–1182 (1987).

16. Colizza, V., Barrat, A., Barthélemy, M. & Vespignani, A. The Role of the Airline Transportation Network in the Prediction and Predictability of Global Epidemics. *Proceedings of the National Academy of Sciences* **103,** 2015–2020 (2006).

17. Christakis, N. A. & Fowler, J. H. Social Network Sensors for Early Detection of Contagious Putbreaks. *PloS One* **5** (2010).

18. Pastor-Satorras, R., Castellano, C., Van Mieghem, P. & Vespignani, A. Epidemic Processes in Complex Networks. *Reviews of Modern Physics* **87** (2015).

19. Lemaitre, J. *et al.* Rainfall as a Driver of Epidemic Cholera: Comparative Model Assessments of the Effect of Intra-Seasonal Precipitation Events. *Acta Tropica* **190,** 235–243 (2019).

20. Guimera, R., Mossa, S., Turtschi, A. & Amaral, L. N. The Worldwide Air Transportation Network: Anomalous Centrality, Community Structure, and Cities' Global Roles. *Proceedings of the National Academy of Sciences* **102,** 7794–7799 (2005).

21. Schweitzer, F. *et al.* Economic Networks: The New Challenges. *Science* **325,** 422–425 (2009).

22. Hidalgo, C. A. & Hausmann, R. The Building Blocks of Economic Complexity. *Proceedings of the National Academy of Sciences* **106,** 10570–10575 (2009).

23. Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A. & Pietronero, L. A New Metrics for Countries' Fitness and Products' Complexity. *Scientific Reports* **2** (2012).

24. Borgatti, S. P., Mehra, A., Brass, D. J. & Labianca, G. Network Analysis in the Social Sciences. *Science* **323,** 892–895 (2009).

25. Rinaldo, A., Banavar, J. R. & Maritan, A. Trees, Networks, and Hydrology. *Water Resources Research* **42** (2006).

26. Porta, S. *et al.* Street Centrality and Densities of Retail and Services in Bologna, Italy. *Environment and Planning B: Planning and Design* **36,** 450–465 (2009).

27. Bullmore, E. & Sporns, O. Complex Brain Networks: Graph Theoretical Analysis of Structural and Functional Systems. *Nature Reviews Neuroscience* **10,** 186–198 (2009).

28. Rubinov, M. & Sporns, O. Complex Network Measures of Brain Connectivity: Uses and Interpretations. *Neuroimage* **52,** 1059–1069 (2010).
29. Chen, J. *et al.* COVID-19 Infection: The China and Italy perspectives. *Cell Death & Disease* **11,** 1–17 (2020).
30. Gössling, S., Scott, D. & Hall, C. M. Pandemics, Tourism and Global Change: A Rapid Assessment of COVID-19. *Journal of Sustainable Tourism,* 1–20 (2020).
31. WHO. *Coronavirus Disease 2019 (COVID-19) Situation Report – 88* tech. rep. (World Health Organization, 2020).
32. Davis, A., Gardner, B. B. & Gardner, M. R. *Deep South: A Social Anthropological Study of Caste and Class* (University of South Carolina Press, 2009).
33. Hidalgo, C. A., Klinger, B., Barabási, A. & Hausmann, R. The Product Space Conditions the Development of Nations. *Science* **317,** 482–487 (2007).
34. Leavitt, H. J. Some Effects of Communication Patterns on Group Performance. *Journal of Abnormal and Social Psychology* **46** (1951).
35. Shaw, M. Group Structure and the Behavior of Individuals in Small Groups. *Journal of Psychology* **38,** 139–149 (1 1954).
36. Freeman, L. Centrality in Social Networks, Conceptual Clarification. *Social Networks* **1,** 215–239 (1979).
37. Katz, L. A New Status Index Derived from Sociometric Analysis. *Psychometrika* **18** (1953).
38. Bonacich, P. Factoring and Weighting Approaches to Status Scores and Clique Identification. *Journal of Mathematical Sociology* **2,** 113–120 (1 1972).
39. Newman, M. E. A Measure of Betweenness Centrality Based on Random Walks. *Social Networks* **27,** 39–54 (2005).
40. Brin, S. & Page, L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks* **30,** 101–117 (1998).
41. Estrada, E. & Rodríguez-Velázquez, J. Subgraph Centrality in Complex Networks. *Physical Review E* **71** (5 2005).
42. Benzi, M. & Klymko, C. Total Communicability as a Centrality Measure. *Journal of Complex Networks* **1,** 124–149 (2013).

43. Golub, G. H. & Van Loan, C. F. *Matrix Computations* (JHU Press, 2012).

44. Brandes, U. *Network Analysis: Methodological Foundations* (Springer Science & Business Media, 2005).

45. Koschützki, D. *et al.* in *Network Analysis* 16–61 (Springer, 2005).

46. Liao, H., Mariani, M., Medo, M., Zhang, Y. & Zhou, M.-Y. Ranking in Evolving Complex Networks. *Physics Reports* **689,** 1–54 (2017).

47. Morrison, G. *et al.* On Economic Complexity and the Fitness of Nations. *Scientific Reports* **7,** 15332 (2017).

48. Hausmann, R. *et al. The Atlas of Economic Complexity: Mapping Paths to Prosperity* (MIT Press, 2014).

49. Tacchella, A., Mazzilli, D. & Pietronero, L. A Dynamical Systems Approach to Gross Domestic Product Forecasting. *Nature Physics* **14,** 861 (2018).

50. Schumpeter, J. *The Theory of Economic Development* (Transaction Publishers, 1934).

51. Aghion, P. & Howitt, P. *A Model of Growth through Creative Destruction* tech. rep. (National Bureau of Economic Research, 1990).

52. Romer, P. M. Endogenous Technological Change. *Journal of Political Economy* **98,** S71–S102 (1990).

53. Rodrik, D. Goodbye Washington Consensus, Hello Washington Confusion? A Review of the World Bank's Economic Growth in the 1990s: Learning from a Decade of Reform. *Journal of Economic Literature* **44,** 973–987 (2006).

54. Rodrik, D. *One Economics, Many Recipes: Globalization, Institutions, and Economic Growth* (Princeton University Press, 2008).

55. Nelson, R. R. *An Evolutionary Theory of Economic Change* (Harvard University Press, 2009).

56. Easterly, W. & Levine, R. What Have we Learned From a Decade of Empirical Research on Growth? It's Not Factor Accumulation: Stylized Facts and Growth Models. *The World Bank Economic Review* **15,** 177–219 (2001).

57. Hulten, C. R. in *New Developments in Productivity Analysis* 1–54 (University of Chicago Press, 2001).

58. Inoua, S. A Simple Measure of Economic Complexity. *arXiv preprint arXiv:1601.05012* (2016).

59. Mariani, M. S., Vidmer, A. & Medo Matsúšand Zhang, Y.-C. Measuring Economic Complexity of Countries and Products: Which Metric to Use? *The European Physical Journal B* **88,** 293 (2015).

60. UN General Assembly. *Transforming our World: The 2030 Agenda for Sustainable Development* tech. rep. (United Nations, 2015).

61. Abud, M., Molina, G., Pacheco, A., Pizarro, G. & et al. *A Multi-Dimensional Focus for the Agenda 2030* tech. rep. (United Nations Development Program, 2017).

62. Griggs, D. *et al.* Sustainable Development Goals for People and Planet. *Nature* **495,** 305–307 (2013).

63. Deaton, A. *The Great Escape: Health, Wealth, and the Origins of Inequality* (Princeton University Press, 2013).

64. UN General Assembly. *The Sustainable Development Goals Report* tech. rep. (United Nations, 2019).

65. UN General Assembly. *The Sustainable Development Goals Report* tech. rep. (United Nations, 2020).

66. UN General Assembly. *Shared Responsibility, Global Solidarity: Responding to the Socio-Economic Impacts of Covid-19* tech. rep. (United Nations, 2020).

67. Barbier, E. B. & Burgess, J. C. Sustainability and Development after COVID-19. *World Development* **135,** 105082 (2020).

68. UN General Assembly. *Work of the Statistical Commission pertaining to the 2030 Agenda for Sustainable Development* tech. rep. A/RES/71/313 (2020).

69. Griggs, D., Nilsson, M., Stevance, A., McCollum, D., *et al. A Guide to SDG Interactions: From Science to Implementation* (International Council for Science, Paris, 2017).

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

71. Nerini, F. F. *et al.* Connecting Climate Action with Other Sustainable Development Goals. *Nature Sustainability* **2,** 674–680 (2019).

72. Nilsson, M. *et al.* Mapping Interactions between the Sustainable Development Goals: Lessons Learned and Ways Forward. *Sustainability Science* **13,** 1489–1503 (2018).

73. Van Soest, H. L. *et al.* Analysing Interactions among Sustainable Development Goals with Integrated Assessment Models. *Global Transitions* **1,** 210–225 (2019).

74. Sachs, J. D. *et al.* Six Transformations to Achieve the Sustainable Development Goals. *Nature Sustainability* **2,** 805–814 (2019).

75. Guerrero, O. A. & Castañeda Ramos, G. Policy Priority Inference: A Computational Method for the Analysis of Sustainable Development. *Available at SSRN* (2020).

76. Requejo-Castro, D., Giné-Garriga, R. & Pérez-Foguet, A. Data-Driven Bayesian Network Modelling to Explore the Relationships between SDG 6 and the 2030 Agenda. *Science of the Total Environment* **710,** 136014 (2020).

77. Tremblay, D., Fortier, F., Boucher, J.-F., Riffon, O. & Villeneuve, C. Sustainable Development Goal Interactions: An Analysis Based on the Five Pillars of the 2030 Agenda. *Sustainable Development* **28,** 1584–1596 (2020).

78. Gentili, P. L. Why Is Complexity Science Valuable for Reaching the Goals of the UN 2030 Agenda? *Rendiconti Lincei. Scienze Fisiche e Naturali,* 1–18 (2021).

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

80. *Leaders* in The Economist. *The 169 Commandments* (The Economist, 2015).

81. Biggeri, M., Clark, D. A., Ferrannini, A. & Mauro, V. Tracking the SDGs in an 'Integrated' Manner: A Proposal for a New Index to Capture Synergies and Trade-Offs between and within goals. *World Development* **122,** 628–647 (2019).

82. Sachs, J. *et al. The Sustainable Development Goals and COVID-19. Sustainable Development Report 2020* tech. rep. (Bertelsmann Stiftung and Sustainable Development Solutions Network (SDSN), 2020).

83. Cooley, A. & Snyder, J. *Ranking the World* (Cambridge University Press, 2015).

84. Anand, S. & Sen, A. *Human Development Index: Methodology and Measurement* tech. rep. (Human Development Report Office, 1994).

85. Alkire, S., Roche, J. M., Santos, M. E. & Seth, S. *Multidimensional Poverty Index 2011: Brief Methodological Note* (Oxford University Press, 2011).

86. Nardo, M., Saisana, M., Saltelli, A. & Tarantola, S. Tools for Composite Indicators Building. *European Commission, Ispra* **15,** 19–20 (2005).

87. European Commission, J. & OECD. *Handbook on Constructing Composite Indicators: Methodology and User Guide* (OECD publishing, 2008).

88. Booysen, F. An Overview and Evaluation of Composite Indices of Development. *Social Indicators Research* **59,** 115–151 (2002).

89. Schmidt-Traub, G., Kroll, C., Teksoz, K., Durand-Delacre, D. & Sachs, J. D. National Baselines for the Sustainable Development Goals Assessed in the SDG Index and Dashboards. *Nature Geoscience* **10,** 547–555 (2017).

90. Lafortune, G., Fuller, G., Moreno, J., Schmidt-Traub, G. & Kroll, C. *SDG Index and Dashboards. Detailed Methodological Paper* tech. rep. (Bertelsmann Stiftung and Sustainable Development Solutions Network (SDSN), 2018).

91. Sciarra, C., Chiarotti, G., Laio, F. & Ridolfi, L. A Change of Perspective in Network Centrality. *Scientific Reports* **8** (2018).

92. Rothenberg, R. B. *et al.* Choosing a Centrality Measure: Epidemiologic Correlates in the Colorado Springs Study of Social Networks. *Social Networks* **17,** 273–297 (1995).

93. Kiss, C. & Bichler, M. Identification of Influencers -— Measuring Influence in Customer Networks. *Decision Support Systems* **46,** 233–253 (2008).

94. Pietronero, L. *et al.* Economic Complexity: "Buttarla in caciara" vs a Constructive Approach. *arXiv preprint arXiv:1709.05272* (2017).

95. Benzi, M. & Klymko, C. A Matrix Analysis of Different Centrality Measures. *arXiv preprint arXiv:1312.6722* (2014).

96. Schoch, D., Valente, T. W. & Brandes, U. Correlations among Centrality Indices and a Class of Uniquely Ranked Graphs. *Social Networks* **50,** 46–54 (2017).

97. Valente, T. W., Coronges, K., Lakon, C. & Costenbader, E. How Correlated Are Network Centrality Measures? *Connections (Toronto, Ont.)* **28,** 16 (2008).

98. Perra, N. & Fortunato, S. Spectral Centrality Measures in Complex Networks. *Physical Review E* **78** (2008).

99. Meghanathan, N. in *Intelligent Systems in Cybernetics and Automation Theory* 11–20 (Springer, 2015).

100. Li, C., Li, Q., Van Mieghem, P., Stanley, H. E. & Wang, H. Correlation Between Centrality Metrics and their Application to the Opinion Model. *The European Physical Journal B* **88** (2015).

101. Zachary, W. W. An Information Flow Model for Conflict and Fission in Small Groups. *Journal of Anthropological Research* **33,** 452–473 (1977).

102. Newton, R. & Spurrell, D. A Development of Multiple Regression for the Analysis of Routine Data. *Applied Statistics,* 51–64 (1967).

103. Nimon, K. Regression Commonality Analysis: Demonstration of an SPSS Solution. *Multiple Linear Regression Viewpoints* **36,** 10–17 (2010).

104. Nathans, L. L., Oswald, F. L. & Nimon, K. Interpreting Multiple Linear Regression: A Guidebook of Variable Importance. *Practical Assessment, Research & Evaluation* **17** (2012).

105. Everett, M. & Borgatti, S. Induced, Endogenous and Exogenous Centrality. *Social Networks* **32,** 339–344 (2010).

106. Bianconi, G. & Barabási, A.-L. Competition and Multiscaling in Evolving Networks. *Europhysics Letters* **54,** 436 (2001).

107. Caldarelli, G., Capocci, A., De Los Rios, P. & Munoz, M. A. Scale-Free Networks from Varying Vertex Intrinsic Fitness. *Physical Review Letters* **89,** 258702 (2002).

108. Rencher, A. *Methods of Multivariate Analysis* (John Wiley & Sons, Inc., 2002).

109. Eckart, C. & Young, G. The Approximation of One Matrix by Another of Lower Rank. *Psychometrika* **1,** 211–218 (1936).

110. Skillicorn, D. *Understanding Complex Datasets: Data Mining with Matrix Decompositions* (CRC press, 2007).

111. Iacobucci, D., McBride, R. & Popovich, D. L. Eigenvector Centrality: Illustrations Supporting the Utility of Extracting More Than One Eigenvector to Obtain Additional Insights into Networks and Interdependent Structures. *Journal of Social Structure* (2017).

112. Newman, M. E. Finding Community Structure in Networks Using the Eigenvectors of Matrices. *Physical Review E* **74** (2006).

113. Davis, T. A. & Hu, Y. The University of Florida Sparse Matrix Collection. *ACM Transactions on Mathematical Software (TOMS)* **38,** 1 (2011).

114. Kleinberg, J. M. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM-SIAM Symposium on Discrete Algorithms* **46,** 604–632 (1999).

115. Kaplan, W. & Lewis, D. *Calculus and Linear Algebra: Vector spaces, Many-Variable Calculus, and Differential Equations* (John Wiley & Sons, 1970).

116. Everett, M. & Borgatti, S. The Dual-Projection Approach for Two-Mode Networks. *Social Networks* **35,** 204–210 (2013).

117. Cimini, G., Mastrandrea, R. & Squartini, T. Reconstructing Networks. *arXiv preprint arXiv:2012.02677* (2020).

118. Pugliese, E. *et al.* Unfolding the Innovation System for the Development of Countries: Coevolution of Science, Technology and Production. *Scientific Reports* **9,** 1–12 (2019).

119. Borgatti, S. & Everett, M. A Graph-Theoretic Perspective on Centrality. *Social Networks* **28,** 466–484 (2006).

120. Sciarra, C., Chiarotti, G., Ridolfi, L. & Laio, F. Reconciling Contrasting Views on Economic Complexity. *Nature Communications* **11,** 1–10 (2020).

121. Balassa, B. Trade Liberalisation and "Revealed" Comparative Advantage. *The Manchester School* **33,** 99–123 (1965).

122. Caldarelli, G. *et al.* A Network Analysis of Countries' Export Flows: Firm Grounds for the Building Blocks of the Economy. *PloS One* **7** (2012).

123. Cristelli, M., Gabrielli, A., Tacchella, A., Caldarelli, G. & Pietronero, L. Measuring the Intangibles: A Metrics for the Economic Complexity of Countries and Products. *PloS One* **8,** e70726 (2013).

124. Kemp-Benedict, E. An Interpretation and Critique of the Method of Reflections. *Munich Personal RePEc Archive* (2014).

125. Gabrielli, A. *et al.* Why We Like the ECI+ Algorithm. *arXiv preprint arXiv:1708.01161* (2017).

126. Pugliese, E., Zaccaria, A. & Pietronero, L. On the Convergence of the Fitness-Complexity Algorithm. *The European Physical Journal Special Topics* **225,** 1893–1911 (2016).

127. Lin, J.-H., Tessone, C. & Mariani, M. Nestedness Maximization in Complex Networks through the Fitness - Complexity Algorithm. *Entropy* **20,** 768 (2018).

128. Felipe, J., Kumar, U., Abdon, A. & Bacate, M. Product Complexity and Economic Development. *Structural Change and Economic Dynamics* **23,** 36–68 (2012).

129. Poncet, S. & de Waldemar, F. S. Export Upgrading and Growth: The Prerequisite of Domestic Embeddedness. *World Development* **51,** 104–118 (2013).

130. Alsén, A. *et al. National Strategy for Sweden: From Wealth to Well-Being* tech. rep. (Boston Consulting Group, 2013).

131. Cristelli, M., Tacchella, A., Cader, M., Roster, K. & Pietronero, L. *On the Predictability of Growth* tech. rep. (The World Bank, 2017).

132. Brito, S., Magud, N. E. & Sosa, S. *IMF Working Paper: Real Exchange Rates, Economic Complexity, and Investment* tech. rep. (International Monetary Fund, 2018).

133. Bruno, M., Saracco, F., Garlaschelli, D., Tessone, C. J. & Caldarelli, G. The Ambiguity of Nestedness Under Soft and Hard Constraints. *Scientific Reports* **10,** 1–13 (2020).

134. Atmar, W. & Patterson, B. D. The Measure of Order and Disorder in the Distribution of Species in Fragmented Habitat. *Oecologia* **96,** 373–382 (1993).

135. Smith, R. L., Smith, T. M., Hickman, G. C. & Hickman, S. M. *Elements of Ecology* (Benjamin Cummings Menlo Parie, CA, 1998).

136. Frankel, J. A. & Romer, D. H. Does Trade Cause Growth? *American Economic Review* **89,** 379–399 (1999).

137. Baudena, M. *et al.* Revealing Patterns of Local Species Richness along Environmental Gradients with a Novel Network Tool. *Scientific Reports* **5,** 11561 (2015).

138. Tu, C., Carr, J. & Suweis, S. A Data Driven Network Approach to Rank Countries Production Diversity and Food Specialization. *PloS One* **11,** e0165941 (2016).

139. Albeaik, S., Kaltenberg, M., Mansour, A. & Hidalgo, C. Improving the Economic Complexity Index. *arXiv preprint arXiv:1707.05826* (2017).

140. Gaulier, G. & Zignago, S. *BACI: International Trade Database at the Product-Level* Working Papers 2010-23 (CEPII, Oct. 2010). `http://www.cepii.fr/CEPII/fr/publications/wp/abstract.asp?%09NoDoc=2726`.

141. Soete, L., Schneegans, S., Eröcal, D., Angathevar, B. & Rasiah, R. *UNESCO Science Report: Towards 2030* tech. rep. (United Nations Educational, Scientific and C, 2015).

142. Randall, J. E. & Ironside, R. G. Communities on the Edge: An Economic Geography of Resource-Dependent Communities in Canada. *Canadian Geographer/Le Géographe Canadien* **40,** 17–35 (1996).

143. Battersby, B., Ewing, R., *et al. International Trade Performance: The Gravity of Australia's Remoteness* (Treasury, 2005).

144. Guttmann, S. & Richards, A. Trade Openness: an Australian Perspective. *Australian Economic Papers* **45,** 188–203 (2006).

145. Mealy, P., Farmer, J. D. & Teytelboym, A. Interpreting Economic Complexity. *Science Advances* **5** (2019).

146. Newman, M. E. Spectral Methods for Community Detection and Graph Partitioning. *Physical Review E* **88,** 042822 (2013).

147. Lucińska, M. & Wierzchoń, S. T. Clustering Based on Eigenvectors of the Adjacency Matrix. *International Journal of Applied Mathematics and Computer Science* **28,** 771–786 (2018).

148. Dobbs, R. *et al. Urban World: Cities and the Rise of the Consuming Class* tech. rep. (McKinsey Global Institute, 2012).

149. Allen, F., Qian, J. & Qian, M. Law, Finance, and Economic Growth in China. *Journal of Financial Economics* **77,** 57–116 (2005).

150. Cetorelli, N. & Goldberg, L. S. Global Banks and International Shock Transmission: Evidence from the Crisis. *IMF Economic Review* **59,** 41–76 (2011).

151. Bendini, R. *Exceptional Measures: The Shanghai Stock Market Crash and the Future of the Chinese Economy* tech. rep. (Policy Department, Directorate-General for External Policies, European Union, 2015).

152. Pritchett, L. & Summers, L. H. *Asiaphoria Meets Regression to the Mean* tech. rep. (National Bureau of Economic Research, 2014).

153. UNDESA. *World Population Prospects: The 2012 Revision, Volume II, Demographic Profiles (ST/ESA/SER.A/345)* tech. rep. (United Nations, Department of Economic and Social Affairs, Population Division, 2013).

154. Saraceno, L. *Economic Complexity: Exploration of Ranking Metrics and Investigation of Causality Between Complexity and Gross Domestic Product per Capita.* MA thesis (Politecnico di Torino, 2020).

155. Pugliese, E., Chiarotti, G. L., Zaccaria, A. & Pietronero, L. Complex Economies Have a Lateral Escape from the Poverty Trap. *PloS One* **12,** e0168540 (2017).

156. Liao, H. & Vidmer, A. A Comparative Analysis of the Predictive Abilities of Economic Complexity Metrics using International Trade Network. *Complexity* **2018** (2018).

157. Brummitt, C. D., Gomez-Lievano, A., Hausmann, R. & Bonds, M. H. Machine-Learned Patterns Suggest that Diversification Drives Economic Development. *arXiv preprint arXiv:1812.03534* (2018).

158. Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A. & Pietronero, L. Economic Complexity: Conceptual Grounding of a New Metrics for Global Competitiveness. *Journal of Economic Dynamics and Control* **37,** 1683–1691 (2013).

159. Sciarra, C., Chiarotti, G., Ridolfi, L. & Laio, F. Network-Driven Rankings of Countries' Status in the Sustainable Development Goals. *Submitted to Scientific Reports, preprint available at essoar.10506441.1* (2021).

160. Brundtland, G. H., Khalid, M., Agnelli, S., Al-Athel, S. & Chidzero, B. Our Common Future. *New York* **8** (1987).

161. Way, C. *The Millennium Development Goals Report 2015* tech. rep. (United Nations, 2015).

162. Servaes, J. *Sustainable Development Goals in the Asian Context* (Springer, 2017).

163. Brown, L. R. *Plan B: Rescuing a Planet Under Stress and a Civilization in Trouble* (WW Norton & Company, 2003).

164. Capra, F. & Luisi, P. L. *The Systems View of Life: A Unifying Vision* (Cambridge University Press, 2014).

165. Allen, C., Metternicht, G. & Wiedmann, T. Initial Progress in Implementing the Sustainable Development Goals (SDGs): A Review of Evidence From Countries. *Sustainability Science* **13,** 1453–1467 (2018).

166. Jacob, A. Mind the Gap: Analyzing the Impact of Data Gap in Millennium Development Goals'(MDGs) Indicators on the Progress toward MDGs. *World Development* **93,** 260–278 (2017).

167. Cho, J., Isgut, A., Tateno, Y., *et al. An Analytical Framework for Identifying Optimal Pathways towards Sustainable Development* tech. rep. (United Nations Economic, Social Commission for Asia, and the Pacific (ESCAP), 2016).

168. Benzi, M. & Klymko, C. On the Limiting Behavior of Parameter-Dependent Network Centrality Measures. *SIAM Journal on Matrix Analysis and Applications* **36,** 686–706 (2015).

169. Benzi, M. & Klymko, C. A Matrix Analysis of Different Centrality Measures. *arXiv preprint arXiv:1312.6722* (2014).

170. Benzi, M., Estrada, E. & Klymko, C. Ranking Hubs and Authorities Using Matrix Functions. *Linear Algebra and its Applications* **438,** 2447–2474 (2013).

171. Volkery, A., Swanson, D., Jacob, K., Bregha, F. & Pintér, L. Coordination, challenges, and innovations in 19 national sustainable development strategies. *World Development* **34,** 2047–2063 (2006).

172. Kroll, C. *Sustainable Development Goals: Are the Rich Countries Ready?* (Citeseer, 2015).

173. Lopez-Calva, L.-F. *et al. World Development Report 2017: Governance and the Law* tech. rep. (The World Bank, 2017).

174. Garmer, L. *SDG Accelerator and Bottleneck Assessment* tech. rep. (Technical report, United Nations Development Programme, New York, NY, 2017).

175. Ashford, N. A. in *Innovation-Oriented Environmental Regulation* 67–107 (Springer, 2000).

176. StatCom. *Cape Town Global Action Plan for Sustainable Development Data (prepared by the high-level group for partnership, coordination and capacity-building for statistics for the 2030 Agenda for Sustainable Development)* tech. rep. (United Nations Statistical Commission et al., 2017).

177. Baldwin, E., Carley, S. & Nicholson-Crotty, S. Why Do Countries Emulate Each Others' Policies? A Global Study of Renewable Energy Policy Diffusion. *World Development* **120,** 29–45 (2019).

178. Reinert, E. *Emulation vs. Comparative Advantage: Competing and Complementary Principles in the History of Economic Policy* 2009.

179. Hou, Z., Keane, J., Kennan, J. & te Velde, D. W. *The Oil Price Shock of 2014* tech. rep. (Working Paper. Overseas Development Institute, London, UK, 2015).

180. OECD. *OECD Science, Technology and Innovation Outlook 2018* (OECD publishing, 2008).

181. Beg, N. *et al.* Linkages between Climate Change and Sustainable Development. *Climate Policy* **2,** 129–144 (2002).

182. D'Odorico, P. *et al.* Global Virtual Water Trade and the Hydrological Cycle: Patterns, Drivers, and Socio-Environmental Impacts. *Environmental Research Letters* **14,** 053001 (2019).

183. Hoekstra, A. Y. & Mekonnen, M. M. The Water Footprint of Humanity. *Proceedings of the National Academy of Sciences* **109,** 3232–3237 (2012).

184. Tuninetti, M., Tamea, S., D'Odorico, P., Laio, F. & Ridolfi, L. Global Sensitivity of High-Resolution Estimates of Crop Water Footprint. *Water Resources Research* **51,** 8257–8272 (2015).

185. Tuninetti, M., Tamea, S. & Dalin, C. Water Debt Indicator Reveals where Agricultural Water Use Exceeds Sustainable Levels. *Water Resources Research* **55,** 2464–2477 (2019).

186. Falsetti, B., Vallino, E., Ridolfi, L. & Laio, F. Is Water Consumption Embedded in Crop Prices? A Global Data-Driven Analysis. *Environmental Research Letters* **15,** 104016 (2020).

187. Morales Meoqui, J. Smith's and Ricardo's Common Logic of Trade. *Munich Personal RePEc Archive* (2010).

188. Tamea, S., Tuninetti, M., Soligno, I. & Laio, F. Virtual Water Trade and Water Footprint of Agricultural Goods: The 1961–2016 CWASI Database. *Earth System Science Data Discussions,* 1–23 (2020).