A New Interpretation of the Economic Complexity Index*

Penny Mealy† J. Doyne Farmer‡ Alexander Teytelboym§

Abstract

The Economic Complexity Index (ECI) introduced by Hidalgo and Hausmann (2009) has been successful in explaining differences in GDP/capita and economic growth across countries. There has been confusion, however, about what it means and why it works. The ECI was originally motivated as an enhancement of diversity which is defined as the number of products a country is competitive in. However, the ECI is orthogonal to diversity. Nor is the ECI an eigenvalue centrality measure – in fact, the standard eigenvalue centrality measure applied to the export similarity matrix is equivalent to diversity. Instead we show that the ECI can be understood in terms of spectral clustering. It corresponds to an approximate solution to the problem of partitioning a graph into two parts in order to minimize the connections between the parts. It can also be viewed as an optimal one-dimensional ordering that clusters countries with similar exports together and minimizes the distance between countries. We present two empirical examples that involve regional employment in occupations and industries rather than exports, in which diversity fails to be a distinguishing feature of the data. These particular regional settings illustrate how the ECI can be useful even when diversity is not.

---

*We would like to thank Simon Angus, R. Maria Del Rio Chanona, Ricardo Hausmann, Eric Kemp-Benedict, Neave O'Cleary, Devavrat Shah, and Muhammad Yildrim for useful conversations and assistance in the development of this paper.

†Institute for New Economic Thinking at the Oxford Martin School,Smith School for Enterprise and the Environment, and St Edmund Hall. E-mail: `penny.mealy@inet.ox.ac.uk`.

‡Institute for New Economic Thinking at the Oxford Martin School, Department of Mathematics, and Christ Church College. E-mail: `doyne.farmer@maths.ox.ac.uk`.

§ St Catherine's College, Department of Economics, and Institute for New Economic Thinking at the Oxford Martin School. Email: `alexander.teytelboym@economics.ox.ac.uk`.

1 Introduction

The Economic Complexity Index (ECI) has proven to be useful in predicting GDP growth and giving insights into economic development.[1] Hidalgo and Hausmann (2009) define the ECI for a country through a recursive, self-referential procedure, in which a country is more complex if it exports more complex products, and similarly a product is more complex if it is exported by more complex countries. By defining a matrix $\widetilde{M}$, whose rows and columns correspond to countries, and whose entries describe the similarities in the countries' exports, the ECI of a country can be defined as the eigenvector associated with its second-largest eigenvalue (see Section 2).[2]

We show that the ECI is equivalent to a spectral clustering method introduced by Shi and Malik (2000) for the problem of partitioning an undirected weighted graph with similarity matrix S into two components.[3] The goal is to minimize the sum of edge weights cutting across the partition while making the size of the two components reasonably similar. Using spectral clustering, Shi and Malik (2000) provide an approximate solution to this problem based on minimizing the *normalized cut (Ncut) criterion*. Their approximate solution is a simple transformation of the *normalized Fiedler vector*, which is the eigenvector corresponding to the second-smallest eigenvalue of the normalized Laplacian of S.[4] We show here that the ECI is equivalent to their approximate solution.

When interpreted as a clustering algorithm the ECI sorts countries into two clusters. It does this by assigning each country a real number on an interval with both positive and negative values, such that countries with similar ECI have similar exports. Countries with positive ECI are in one cluster and countries with negative ECI are in the other cluster, and the absolute value of a country's ECI measures the distance of any given country to the boundary between the clusters.

[1] An alternative measure of economic complexity called *Fitness* was introduced by Tacchella et al. (2012). It is beyond the scope of this paper to study this measure here.

[2] Hidalgo and Hausmann (2009) and Hausmann et al. (2014) standardize the ECI by subtracting the mean and dividing by the variance in order to make comparisons across different years. We use the unstandardized version throughout this paper in order to make the connection to spectral clustering.

[3] Spectral clustering was originally developed for image recognition and is now used for dimensionality reduction, community detection and data clustering, with applications to a variety of problems including high performance computing, web page ranking, information retrieval and RNA motif classification.

[4] The normalized Laplacian is a stochastic matrix, so the smallest eigenvalue is automatically zero and so the second-smallest eigenvalue is the leading eigenvalue of interest. The word "normalized" applies to the Fiedler vector because it is the eigenvalue of a normalized Laplacian, rather than due to any normalization of the vector.

For the economic applications presented here ECI does not yield a clear separation into disjoint clusters, but it nonetheless provides a useful rank ordering that places countries with similar exports near each other. There are potentially $C!$ ways to order a set of C countries. Appropriating another result of Shi and Malik (2000), the ECI is the unique way to assign a real number to each country in order to minimize the sum of the squared distances between countries, where the distances are weighted according to the similarity matrix S. In contrast to its application to graph partitioning, the ECI is an *exact* solution to this problem. It thus provides a reduction of the high dimensional space of countries and their exports onto a single dimension that proves to be very useful for problems such as understanding the relationship between exports and GDP.

A visual representation can be seen in Fig. 1. Drawing on country trade data for 2013, we show the S graph on the right hand side, where countries are represented as nodes and weighted links are given by $S_{cc'}$.[5] Here countries are coloured by their ECI, with darker shades of green representing higher complexity and darker shades of pink representing lower complexity. On the left hand side of the plot we show ECI values for each country plotted in ascending order. Countries with an ECI value greater than zero are in the green cluster and countries with a negative ECI are in the pink cluster. The ECI value (above or below zero) provides an indication of the distance from the cut. In this case the clusters are not well-defined—one cluster blends into the other—but the ECI nonetheless provides a very useful rank ordering.

[5] We are only visualising links with a weight larger than a given threshold. In this case we plot all links with a weight > 3.

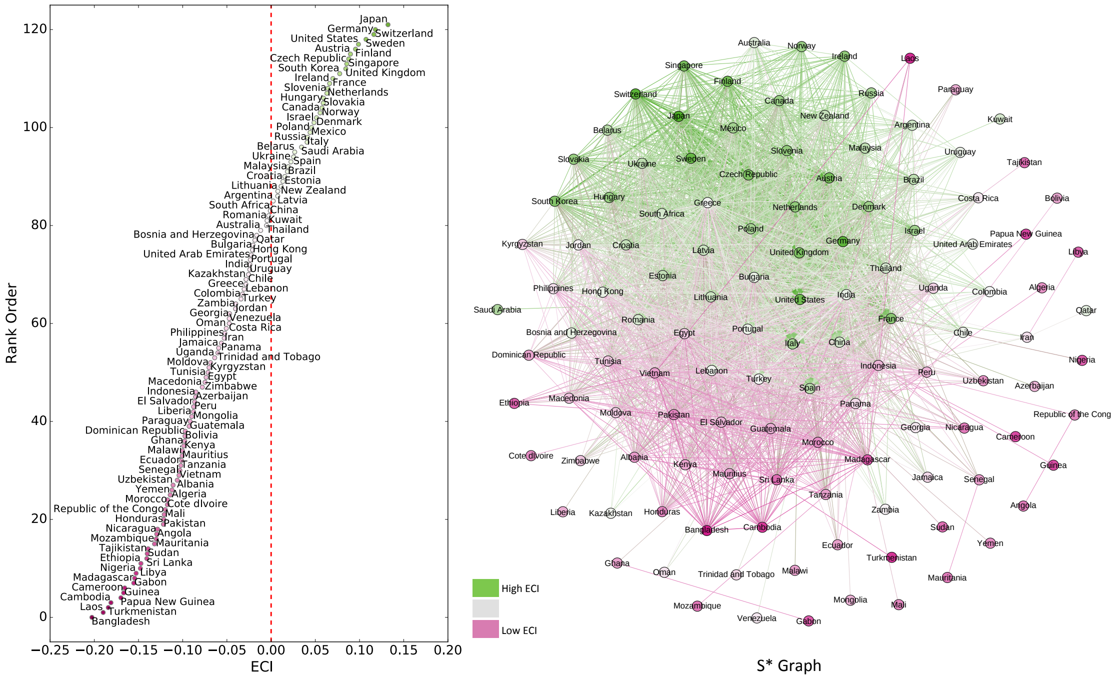

Figure 1: Visual representation of how the ECI vector partitions the S graph. The S graph (on the right of the plot) shows country nodes coloured by their ECI value, with shades of green representing more positive ECI and shades of pink representing more negative ECI. The ECI rank ordering of countries (on the left of the plot) shows how countries having positive ECI fall on one side of the partition and countries having negative ECI fall on the other side of the partition.

Much of the conventional wisdom about the ECI is based on discussion of *diversity* and *ubiquity*. The diversity of a country is the number of products for which a country's exports are competitive and ubiquity is the number of countries that are competitive in a product. The diversity is the row (or column) sum of the export similarity matrix S i.e. it is the degree centrality of the network that this matrix represents. As shown by Kemp-Benedict (2014), the diversity vector is *orthogonal* to the ECI, meaning the dot product of the ECI and the diversity vector is 0. Hence the ECI captures a more subtle property of the network that is distinct from diversity.[6]

While the ECI might appear to resemble a measure of eigenvector centrality (Morrison et al., 2017), we show that this resemblance is superficial. The matrix $\widetilde{M}$

[6] The Fitness measure of Tacchella et al. (2012) is strongly correlated with diversity. Mariani et al. (2015) show that Spearman Rank correlations for Fitness and diversity tend to be between 0.94 and 0.97 (see also Morrison et al. (2017)).

defines a directed network. As Newman (2010) argues, the eigenvector centrality for a directed network can be defined in terms of the left eigenvector corresponding to the leading eigenvalue of $\widetilde{M}$ (which is 1 because $\widetilde{M}$ is row-stochastic). We show that the eigenvector centrality of $\widetilde{M}$ is proportional to the degree centrality of S, i.e. it is proportional to the diversity. The eigenvector centrality of S is also highly correlated with diversity. In contrast the ECI provides an approximation to the cut of the network into two disjoint sets such that the sum of link weights crossing the boundary is minimized.

In the context of trade data we argue that both diversity and the ECI capture important—and *different*—information relevant to the process of economic development. The pattern of diversification that countries experience over the development path is well established in the economics literature. Countries tend to diversify early in their developmental phase, and begin specializing at much higher levels of per capita income (Imbs and Wacziarg, 2003). In contrast, the ECI both groups together countries with similar exports and separates countries with dissimilar exports. It therefore sheds light on the type of production capabilities that separate high- and low-income countries and gives empirical insights into the differences in countries' productive capabilities at different developmental stages.

In some settings the ECI becomes much more useful than diversity in explaining variation in economically-relevant variables. We present a brief preview of forthcoming work for two examples.[7] One of these examples is based on the Integrated Public Use Microseries (IPUMS) data from the US Census (Ruggles et al., 2017), which surveys occupations and industries in each state, and the other is the UK Business Register and Employment Survey (BRES), which details industrial employment for local regions within the UK. For these two examples states and regions are not well-distinguished based on their diversity,[8] whereas they are well-distinguished by ECI. In both cases the ECI correlates strongly with economic variables such as income and earnings per capita.

This paper is organized as follows. In Section 2, we review the definition of the ECI. In Section 3 we examine the relationships between the ECI, diversity, and eigenvector centrality. In Section 4 we describe the normalized cut criterion and explain its formal relationship to the ECI. In Section 5 we discuss these results in relation to development and show two examples where the ECI is more useful than diversity for explaining differences in regional earnings and income per capita. Section 6 concludes.

[7] We only present a brief preview in these two cases because the analysis has been developed with other co-authors and will be presented more fully elsewhere.

[8] This may be because trade costs are lower within countries than they are between countries, so that possessing a diverse range of productive capabilities is less important.

2 The Economic Complexity Index

Hidalgo and Hausmann (2009) define the Economic Complexity Index (ECI) using an algorithm that operates on a binary country-product matrix M with elements M_{cp}, indexed by country c and product p. $M_{cp} = 1$ if country c has a *revealed comparative advantage* (RCA) > 1 in product p, where RCA is calculated using the Balassa (1965) index, given by

$$RCA_{cp} = \frac{x_{cp}/\sum_p x_{cp}}{\sum_c x_{cp}/\sum_c \sum_p x_{cp}}, \tag{1}$$

where x_{cp} is country c's exports of product p, and $M_{cp} = 1$ if $RCA_{cp} > 1$ and $M_{cp} = 0$ otherwise.

Summing across the rows and columns of M gives a country's *diversity* (denoted $k_c^{(0)}$) and product *ubiquity* (denoted $k_p^{(0)}$), defined as

$$k_c^{(0)} = \sum_p M_{cp}, \tag{2}$$

$$k_p^{(0)} = \sum_c M_{cp}. \tag{3}$$

In the original "Method of Reflections", Hidalgo and Hausmann (2009) measure of country diversity and product ubiquity is iterated to calculate the average values associated with country and product nodes' neighbours from the previous iteration step as shown in Eqs. (4) and (5), yielding

$$k_c^{(N)} = \frac{1}{k_c^{(0)}} \sum_p M_{cp} k_p^{(N-1)}, \tag{4}$$

$$k_p^{(N)} = \frac{1}{k_p^{(0)}} \sum_c M_{cp} k_c^{(N-1)}. \tag{5}$$

In the limit as $N \to \infty$ these converge to constant vectors, e.g. $k_c^{(\infty)} = k$ independent of c. In their original procedure Hidalgo and Hausmann (2009) used a large but not too large value of N, e.g. $N = 18$. This produced useful deviations from the constant vector, with a strong positive correlation between $k_c^{(N)}$ and log GDP per capita.

In Hausmann et al. (2014), the "Method of Reflections" is reframed as an eigenvalue problem, where they show that inserting Eq. (5) into Eq. (4) and rewriting gives

$$
\begin{aligned}
k_c^{(N)} &= \frac{1}{k_c^{(0)}} \sum_p M_{cp} \frac{1}{k_p^{(0)}} \sum_{c'} M_{c'p} k_{c'}^{(N-2)}, \\
&= \sum_{c'} k_{c'}^{(N-2)} \sum_p \frac{M_{cp} M_{c'p}}{k_c^{(0)} k_p^{(0)}}, \\
&= \sum_{c'} \widetilde{M}_{cc'} k_{c'}^{(N-2)},
\end{aligned} \tag{6}
$$

where

$$
\widetilde{M}_{cc'} \equiv \sum_p \frac{M_{cp} M_{c'p}}{k_c^{(0)} k_p^{(0)}} = \frac{1}{k_c^{(0)}} \sum_p \frac{M_{cp} M_{c'p}}{k_p^{(0)}}. \tag{7}
$$

The Economic Complexity Index (ECI) is the eigenvector χ corresponding to the second-largest eigenvalue of $\widetilde{M}$ (Hausmann et al., 2014).

Letting C be the number of countries, $\widetilde{M}$ is an asymmetric, $C \times C$, row-stochastic matrix. This means that each of the rows of $\widetilde{M}$ sum to one, so that its entries can be regarded as conditional probabilities in a Markov transition matrix. This interpretation of $\widetilde{M}$ as a Markov transition matrix was discussed in Hidalgo and Hausmann (2009) and further developed by Kemp-Benedict (2014). When applied to country trade data one can think of $\widetilde{M}$ as a weighted similarity matrix, reflecting how similar two countries' export baskets are. Each element represents the products that country c has in common with country c', weighted by each product's scarcity (inverse of the product's ubiquity) and normalized by the diversity of country c. So if two countries both export a fairly ubiquitous product (such as nails), $\widetilde{M}_{cc'}$ would be lower than if they both exported a fairly rare product (such as lasers). Moreover, if country c exports diverse products, $\widetilde{M}_{cc'}$ would be lower than if country c only exported a few products.

Note that with the earlier recursive definition, while diversity and ubiquity are given as initial conditions, they become irrelevant in the limit as $N \to \infty$. Eqs. (4) and (5) define a linear dynamical system with a stable fixed point attractor, in which the solution is independent of the initial condition. Diversity and ubiquity are relevant only because they are incorporated into the definition of the dynamical system itself. The fact that $\widetilde{M}$ is row-stochastic means that the leading eigenvalue is one and the leading eigenvector is constant. The iterative method gives essentially equivalent results due to the fact that the second eigenvector corresponds to the direction in which the system converges most slowly onto the first eigenvector.

3 ECI, diversity, and eigenvector centrality

3.1 Notation

In this paper, we denote the ECI vector by χ and the ECI of country c is denoted χ_c. We also denote the diversity by d where $d_c = k_c^{(0)}$ is the diversity of country c.

3.2 ECI and diversity

The ECI measure is conceptually cast in Eqs. (4) and (5) in terms of measuring country diversity and product ubiquity, and then iteratively "correcting" a country's diversity by the ubiquity of its products. This follows from the proposed hypothesis that prosperous countries have capabilities that allow them to export a diverse range of products, many of which other countries are not able to produce (Hidalgo and Hausmann, 2009, Hausmann et al., 2014). However, as Kemp-Benedict (2014) showed, *the ECI is orthogonal to diversity* i.e. the dot product of the diversity and the ECI vectors is zero,

$$d \cdot \chi = \sum_c d_c \chi_c = 0.$$

That said, for the export data ECI and diversity are nevertheless positively correlated, as shown in Fig. 2.

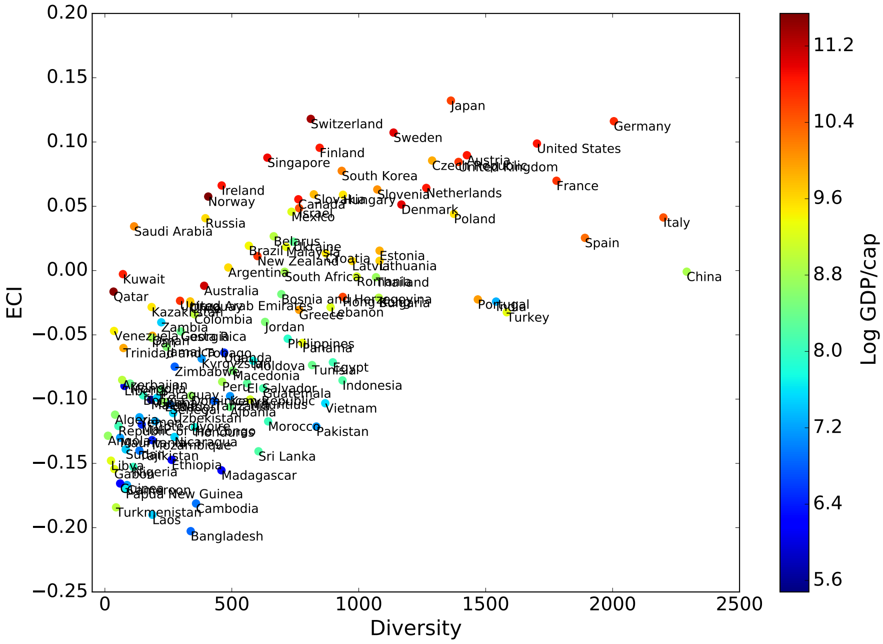

Figure 2: Diversity vs. economic complexity of countries compared to GDP per capita. Data is based on HS6 COMTRADE data for the year 2013. GDP per capita data is from the World Bank.

The fact that these vectors are orthogonal in the vector space of nodes (in this case countries) does not necessarily imply that there is no relationship between them, but the orthogonality of these two vectors illustrates that they are capturing very different aspects of the similarity of countries based on their exports.[9] The difference between them is in a sense obvious: Diversity captures “How many exports?” whereas the ECI captures “What type of exports?”.

3.3 ECI and eigenvector centrality

Morrison et al. (2017) have asserted that the ECI is a “standard eigenvalue centrality algorithm”. Eigenvector centrality is defined as the eigenvector corresponding to the largest eigenvalue of an adjacency matrix and this definition is standard for undirected networks. In the case of directed networks, the natural definition is to

[9] The correlation subtracts the product of the means and normalizes by the square root of the variances. Thus unless the mean of one of the variables is zero, zero dot product does not imply zero correlation. Neither diversity nor the unstandardized ECI have mean zero. Standardization of the ECI means that its dot product with diversity is no longer zero.

take the left eigenvector corresponding to the leading eigenvalue of the adjacency matrix (Newman, 2010, p. 178).[10]

Therefore, the eigenvector centrality vector of $\widetilde{M}$ is the left (row) eigenvector x corresponding to the largest eigenvalue of the following eigenvalue equation

$$x\widetilde{M} = \lambda x. \tag{8}$$

Since $\widetilde{M}$ is row-stochastic, its largest eigenvalue is 1. Hence we are in interested in solutions to

$$x\widetilde{M} = x. \tag{9}$$

As the rows $\widetilde{M}$ have been normalized by diversity, it is easy to check that any vector proportional to d is a solution to Eq. (9). Since *eigenvector centrality of $\widetilde{M}$ is proportional to diversity*, it does not add anything to what we already know about $\widetilde{M}$.

For many purposes it is more convenient to work with the export similarity matrix

$$S = \sum_p \frac{M_{cp} M_{c'p}}{k_p^{(0)}}. \tag{10}$$

By construction S is symmetric and satisfies $S_{cc'} \geq 0$. If we define D as a diagonal matrix with diagonal entries $D_{cc} = d'_c = \sum_c W_{cc'}$ and $D_{i \neq j} = 0$, we can re-write S as

$$S = D\widetilde{M} \tag{11}$$

Since S is symmetric matrix, it represents an undirected network and it is more convenient for studying eigenvector centrality. In Figure 3 we show that the *eigenvector centrality of S is highly correlated with diversity*. This is unsurprising. Diversity is the degree centrality of S and degree centrality is correlated with eigenvector centrality in many networks.

The fact that the ECI is defined as the eigenvector associated with the *second*-largest eigenvalue, whereas eigenvalue centrality is associated with the leading eigenvalue, makes it clear that *ECI is not eigenvector centrality*. In fact, in the export data, eigenvector centrality is much more closely correlated to diversity than to the ECI (see Figure 3).

[10] In Newman (2010)'s exposition the adjacency matrix is transposed (i.e. $\widetilde{M}_{cc'}$ is an edge from c' to c), this is equivalent to finding the right eigenvector of the transpose of $\widetilde{M}$. The right leading eigenvector of $\widetilde{M}$ is a constant vector since $\widetilde{M}$ is row-stochastic.

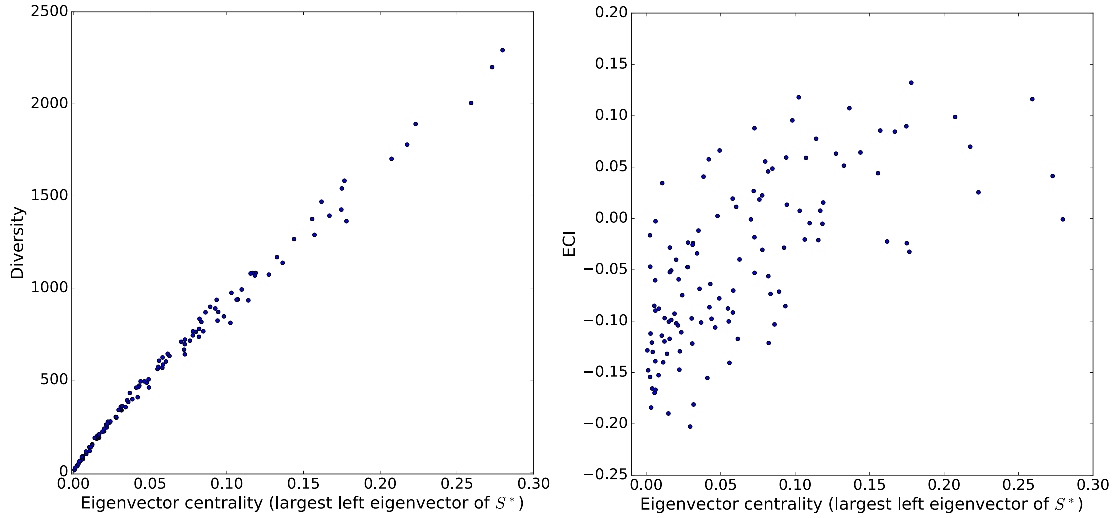

Figure 3: Eigenvector centrality of the export similarity matrix S vs country diversity (d) and eigenvector centrality of S vs the ECI. Data is based on H6 COMTRADE data for the year 2013.

4 ECI and spectral clustering

In this section we show that the ECI emerges naturally from a commonly-used spectral clustering technique.

4.1 Spectral clustering and the $Ncut$ criterion

Spectral clustering refers to a set of methods for partitioning a network into groups of nodes that are similar within clusters and dissimilar across clusters. Spectral clustering has various applications, such as image processing, where it is used to segment images into homogeneous components and community detection in social networks. Spectral clustering methods are also related to the conductance of graphs and the mixing rates of Markov chains.

Consider an undirected graph $G = (V, E)$ with vertices V and edges E. We allow the graph G to be weighted, with non-negative weights so the adjacency matrix entries are $S_{ij} \geq 0$ where $S_{ij} = S_{ji}$. (Though the export matrix is one possible example, we now allow S to be any matrix with these properties). The degree of vertex i is defined as

$$d_i = \sum_{j \in V} S_{ij}, \tag{12}$$

and the size or "volume" of a set of vertices $A \subseteq V$ can be measured as

$$vol(A) = \sum_{i \in A} d_i. \tag{13}$$

For our purposes here we only need to consider the problem of partitioning a graph into two disjoint sets. One way to do this is by solving the *cut* problem. The objective is to find a partition of V into complementary sets A and $\bar{A}$ that minimize the number of links between the two sets. The *cut* problem is to find the minimum of

$$cut(A, \bar{A}) = \sum_{i \in A, j \in \bar{A}} S_{ij}. \tag{14}$$

This objective function has the undesirable property that its solution often partitions a single node from the rest of the graph. To avoid this problem, Shi and Malik (2000) proposed the *normalized cut* ($Ncut$) *criterion*, which penalizes solutions that are not properly balanced. The objective is to partition the graph in such a way that each cluster contains a reasonable number of vertices. In particular they propose minimizing the objective function

$$Ncut(A, \overline{A}) = (\frac{1}{vol(A)} + \frac{1}{vol(\overline{A})}) \sum_{i \in A, j \in \bar{A}} S_{ij}. \tag{15}$$

Let D be the diagonal degree matrix with $D_{ii} = d_i$ and $D_{i \neq j} = 0$. Shi and Malik (2000) showed that finding the minimum value of $Ncut$ is equivalent to solving the optimization problem

$$\min_A Ncut(A, \bar{A}) = \min_y \frac{y^T (D - S) y}{y^T D y}, \tag{16}$$

subject to $y_i \in \{1, -vol(A)/vol(\bar{A})\}$ and $y^T D\mathbf{1} = 0$.

Due to the fact that y_i is restricted to one of two possible values, this is not a simple linear algebra problem, and finding the true minimum of the $Ncut$ criterion has been shown to be NP-hard. An approximate solution can found by relaxing this restriction and letting y_i take on any real value. An approximate solution is obtained by finding the eigenvector $y^{[2]}(S)$ corresponding to the second-smallest eigenvalue of the generalized eigenvalue equation

$$(D - S)y = \lambda D y. \tag{17}$$

$L_S = D - S$ is the *Laplacian* matrix of S. By making the substitution

$$y = D^{-1/2} z, \tag{18}$$

this can be rewritten as a standard eigenvalue equation

$$D^{-\frac{1}{2}}(D - S)D^{-\frac{1}{2}} z = \hat{L}_S z = \lambda z, \tag{19}$$

where $\hat{L}_S = D^{-\frac{1}{2}}(D - S)D^{-\frac{1}{2}}$ is the *normalized Laplacian* of S. Because the normalized Laplacian is a stochastic matrix its smallest eigenvalue is zero. As Shi and Malik (2000) show, the eigenvector $z^{[2]}(S)$ associated with the second-smallest eigenvalue of $\hat{L}_S$, which is called the *normalized Fiedler vector*, yields a useful approximate minimum to the $Ncut$ criterion.[11] Transforming back to y using Eq. (18) to solve the original problem gives the solution

$$y^{[2]}(S) = D^{-1/2} z^{[2]}(S). \tag{20}$$

The solution $y^{[2]}$ provides a useful approximate solution that minimizes the normalized cut criterion, that is equal to a simple transformation of the normalized Fiedler vector.

4.2 Relationship between ECI and the $Ncut$ criterion

Recall that $\widetilde{M}$ is the matrix whose second largest eigenvalue is the ECI. To see the relation between spectral clustering and the ECI, note that the similarity matrix $S = D\widetilde{M}$ characterizing the exports of countries is in the form used to minimize the normalized cut criterion. Multiplying both sides of Eq. (19) by $D^{-\frac{1}{2}}$ and re-arranging terms gives

$$D^{-1} S D^{-\frac{1}{2}} z = (1 - \lambda) D^{-\frac{1}{2}} z. \tag{21}$$

Substituting $\widetilde{M} = D^{-1} S$ gives

$$\widetilde{M} D^{-\frac{1}{2}} z = (1 - \lambda) D^{-\frac{1}{2}} z. \tag{22}$$

The eigenvalue equation for $\widetilde{M}$ is

$$\widetilde{M} y = \lambda_{\widetilde{M}} y. \tag{23}$$

[11] Here we are only considering partitioning into two clusters. An approximate solution to the problem of partitioning into k clusters is given by the k eigenvectors corresponding to the k smallest eigenvalues. In fact image recognition problems are typically solved by recursively partitioning into two clusters.

Comparing the last two equations makes it clear that the eigenvalues and eigenvectors of $\widetilde{M}$ are related to those of S by

$$\lambda_{\widetilde{M}} = 1 - \lambda, \tag{24}$$
$$y = D^{-\frac{1}{2}} z. \tag{25}$$

Thus the second smallest eigenvalue of the Laplacian of S corresponds to the second largest eigenvalue of $\widetilde{M}$, and comparison to Eq. (20) makes it clear that the *ECI is equivalent to the spectral clustering solution of the normalized cut criterion*, i.e.

$$\chi = y^{[2]}(S) = D^{-\frac{1}{2}} z^{[2]}(S). \tag{26}$$

To summarize, the ECI is related to the normalized Fiedler vector by a simple transformation, in precisely the same way that the solution to the normalized cut criterion problem is related to the normalized Fiedler vector.

ECI can alternatively be defined as the average of the Product Complexity Index (PCI) (Hausmann et al., 2014). We should emphasize that our spectral clustering interpretation can also be applied to the PCI. Hence, PCI is the spectral clustering solution of the normalized cut criterion applied to the product-by-product matrix $\widetilde{M}_{pp}$ (which produced by swapping the c and p indices in Eq. (7)) after normalizing it by diagonal ubiquity matrix. From our perspective, PCI separates the product-by-product matrix into two clusters of products which are similar within clusters and different across clusters.

We can also use spectral clustering to highlight the difference between eigenvector centrality and the ECI. This difference becomes apparent if one interprets these results in terms of a random walk. Because $\widetilde{M}$ is row-stochastic, its entries can be interpreted as transition probabilities, as emphasized by Kemp-Benedict (2014). One can then imagine a random walk through the network with these transition probabilities. Eigenvector centrality measures the asymptotic likelihood of visiting any given node, which in this case corresponds to the diversity of a given country. When applied to the similarity network S, eigenvector centrality produces a ranking of the country such that a country's rank is equal to the sum of the ranks of countries whose exports are similar. The ECI, in contrast, it corresponds to a partition of the network that gives the least flux between the two components of the export similarity matrix.

4.3 Minimal distance interpretation of ECI

The ECI can also be interpreted in terms of a minimum distance criterion. To motivate why this is useful, suppose we hypothesize that there exists a relationship between the exports of a country and some quantity of interest, such as GDP per capita. To reduce the dimensionality of the problem it would be useful to find a rank ordering that places countries with similar exports close to each other. However, if there are C countries, there are $C!$ possible rank orderings. For $C \approx 100$ this number is intractably large, and searching for all possible rank orderings would be impossible. It turns out that the ECI provides a unique ranking that minimizes the distance between countries, where the distances are weighted according to the similarity matrix S.

To see how this is true, we draw on Shi and Malik (2000), who pointed out that while the second smallest eigenvector $y^{[2]}(S)$ only approximates the normalized cut criterion, it exactly minimizes

$$\frac{\sum_{i,j} (y_i - y_j)^2 S_{ij}}{\sum_i y_i^2 d_i}, \tag{27}$$

subject to the constraint

$$\sum_i y_i d_i = 0. \tag{28}$$

We have written this in coordinate notation for clarity. Given the similarity matrix S, the problem is to find real numbers y_i for each node i that minimize Eq. (27). To understand how this problem differs from that of finding the normalized cut, compare to Eq. (16). The denominators are the same and the constraints are the same, but the numerator in Eq. (16) is replaced here by the sum of the squared distances between nodes weighted according to the similarity matrix S.

The numerator is suggestive of principal components analysis or ordinary least squares, but the problem being solved is different. Rather than being given a set of data points and finding a vector or vectors that minimize the distance from the data, here we are given a matrix S that characterizes the similarity of the nodes and we need to assign each node i a real number y_i. The solution $\chi_i = y_i^{[2]}(S)$ has the special property that it minimizes the sum of the squared distance between nodes, weighted by the similarity matrix S.[12] This provides a precise justification for the intuition of Hausmann et al. (2014, p. 24), who stated "[The ECI] is the eigenvector that captures the largest amount of variance in the system and is our measure of economic complexity." This also justifies Kemp-Benedict's (2014) interpretation of the ECI as ranking countries according to how similar their exports are.

[12] The constraint of Eq. (28) is the orthogonality condition between diversity and the ECI, which guarantees that the interval spanned from lowest ECI to highest ECI contains zero, and more specifically that the sum of the diversity of the countries in one cluster equals the sum of the diversities of the countries in the other cluster.

5 What might this mean in the development context?

In the context of trade data, previous work has tended to conceptualise the ECI and diversity as being related measures. As we show in Fig. 2 and 4, ECI and diversity are positively correlated, and also both positively correlated to per capita GDP. However, since ECI is orthogonal to diversity, these measures capture different information. We now discuss each of these in turn.

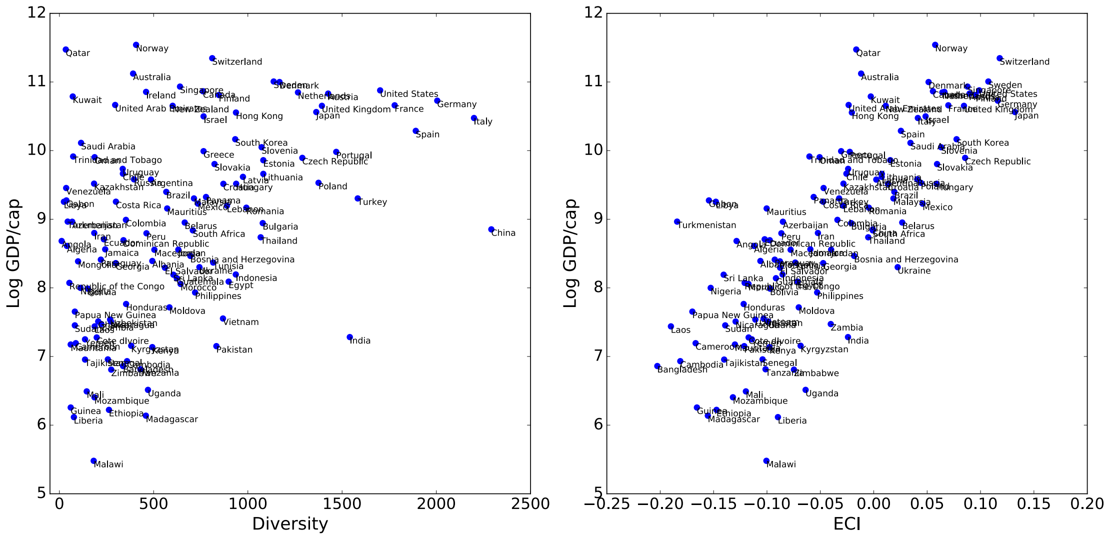

Figure 4: Country diversity vs log GDP/capita and ECI vs log GDP/capita in based on the HS6 COMTRADE export data for 2013.

5.1 Diversity, ECI and Economic Development

The relationship between diversification and development is well established in the economics literature. The general finding is that countries tend to follow a U-shaped pattern, where they first diversify and then begin specialising relatively late in the development process (Imbs and Wacziarg, 2003). This pattern aligns with other empirical studies that have described a positive association between export diversification and economic growth, which tends to be stronger for less developed countries (Al-Marhubi, 2000, Herzer and Nowak-Lehnmann D, 2006, Hesse, 2009).[13]

13 For example, the positive relationship between per capita income and diversity is can be generated by: (1) dampened impacts of sector-specific shocks (Acemoglu and Zilibotti, 1997),

In contrast to diversity, the ECI provides additional information relevant for economic development. As we illustrate in Fig. 1, the ECI provides a rank ordering of countries in terms of how similar their exports are to each other. The fact that this ordering is useful in explaining variation in per capita GDP and predicting growth suggests that different types of exports (and by extension, productive capabilities) are associated with different growth and development outcomes.

There is some evidence in the literature to support this notion. Hausmann et al. (2007) show that specializing in some products will bring higher growth than specializing in others. Manufacturing products in particular show strong unconditional convergence in labour productivity (Rodrik, 2012). Within developing countries technologically sophisticated products are more strongly associated with export and income growth (Lall, 2000).

5.2 The ECI can be more useful than diversity in regional settings

While the ECI has traditionally been applied to country trade data, it is also interesting to consider how it performs in different contexts, such as regional employment data.

5.2.1 UK Local Authorities and Industries

Using data from the UK Business Register and Employment Survey (BRES), we construct a binary region-industry matrix W. This matrix is similar to the country-product matrix M, but defined for UK local authorities and industries. While M_{cp} binary values are calculated using the Balassa index for country exports (see Eq. (1)), we construct W on the basis of a region r's *Location Quotient* (LQ) in industry i.

$$LQ_{ri} = \frac{e_{ri}/\sum_p e_{ri}}{\sum_r e_{ri}/\sum_r \sum_i e_{ri}} \tag{29}$$

where e_{ri} is the number of people employed in industry i ind region r.

and (2) preferences for diversity which can be better met with local industries where trading costs are high (Murphy et al., 1989). The tendency to specialise later in the development process is cast in terms of the efficiency benefits emphasised in traditional Ricardian trade theory as well as theories of geographical agglomeration in "New Trade Theory" and economic geography (Krugman, 1991).

As shown in Figs. 5 and 6, applying the ECI methodology to data on regions and industries provides an ordering of local authorities that is strongly positively correlated with log earnings per capita (Pearson $R = 0.76$, p-value $= 3.9 \times 10^{-61}$). Interestingly, we find that diversity is negatively correlated to ECI (Pearson $R = -0.42$, p-value $= 1.3 \times 10^{-14}$) and log earnings per capita (Pearson $R = -0.34$, p-value $= 9.6 \times 10^{-10}$).

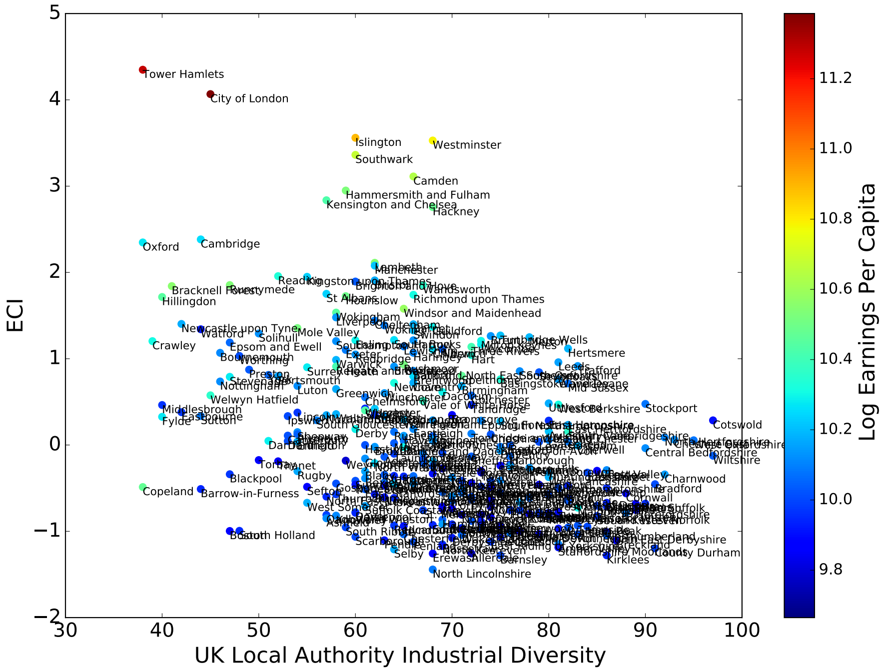

Figure 5: Industrial Diversity vs ECI for UK Local Authorities. Data is based on BRES data for the year 2011. Earnings data is based on 2011 workplace earnings from the UK Office of National Statistics.

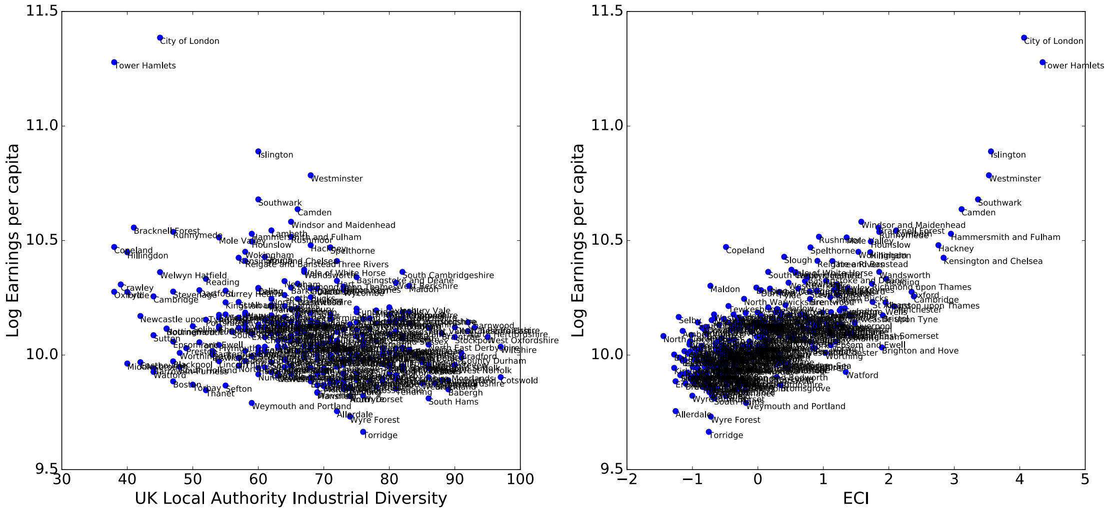

Figure 6: Industrial diversity vs Log earnings per capita and ECI vs Log earnings per capita for data on industrial employment in UK local authorities (2011).

5.2.2 US States and Occupations

We find a similar result when examining employment data for US states and occupations.[14] Here we draw on IPUMS census data for the US and construct W_2 on the basis of a state's location quotient in *occupation* i. In Figs. 7 and 8 we show that once again, ECI is positively correlated to log state per capita GDP (Pearson $R = 0.63$, p-value $= 7.46 \times 10^{-7}$). However, there is no particular relationship between ECI and diversity (Pearson $R = 0.10$, p-value 0.47).

In these regional settings the fact that diversity becomes less important in explaining earnings and income per capita may be due to the fact that trade costs are lower. In these cases the ECI is particularly useful, as we will discuss in detail in other work.

[14]Similar results can also be seen for data on US states and industries.

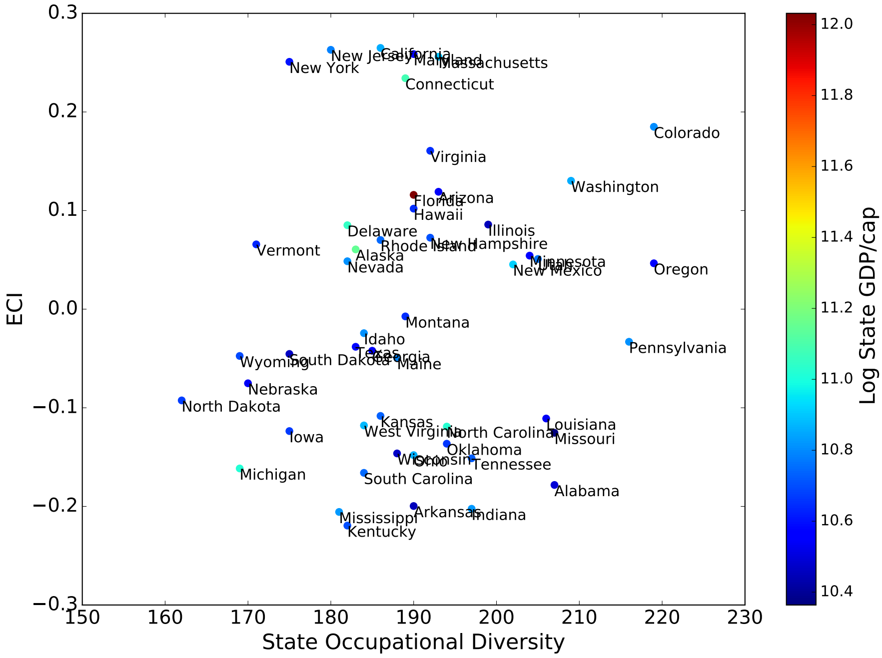

Figure 7: Occupational Diversity and Complexity of US States. Data is based on IPUMS data for the year 2010. State GDP per capita (2010) data is from the Bureau of Economic Analysis.

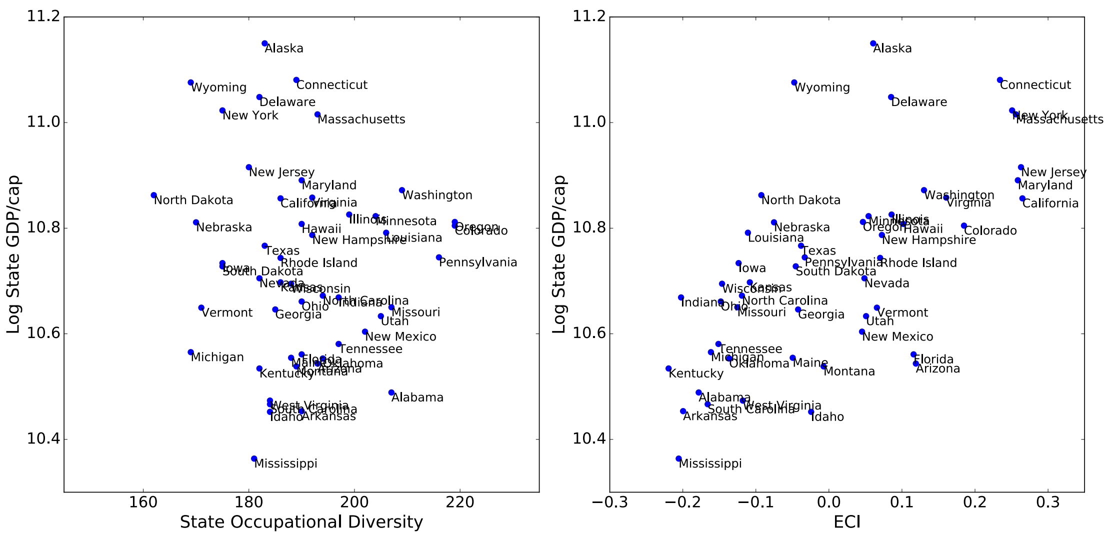

Figure 8: Occupational Diversity vs Log State GDP per capita and ECI vs Log State GDP per capita for US State Occupational Data (2010).

6 Conclusion

In this paper we have shown that, rather than being linked to export diversity or eigenvector centrality, the ECI is equivalent to the spectral clustering solution of normalized cut criterion (Shi and Malik, 2000). The comparison to the clustering problem provides new insights. It shows how the ECI can be thought of as a clustering algorithm that partitions the set of countries into two groups.

Moreover, we show that the ECI can be thought of as a dimension reduction tool. It is the unique way to assign distances to countries such that the sum of their squared distances from each other is minimized, where the distance is measured using a weighted similarity matrix S. The ECI thus takes the complicated high dimensional space of countries and their exports and reduces it to a linear ordering, analogous to the Dewey-Decimal System for classifying books. Simply put, countries with similar ECI export similar products. This new perspective sheds some light on the empirical success of the ECI in explaining variations in GDP and GDP growth across countries—which is distinct from existing findings relating to diversification and development.

There are cases where diversity is less important than ECI. In providing a preview into some forthcoming work, we show that the ECI can be particularly useful in explaining variation in economically-relevant variables across regions that are not well-distinguished on the basis of their diversity.

References

Acemoglu, D. and F. Zilibotti (1997). Was prometheus unbound by chance? risk, diversification, and growth. *Journal of political economy 105*(4), 709–751.

Al-Marhubi, F. (2000). Export diversification and growth: an empirical investigation. *Applied economics letters 7*(9), 559–562.

Balassa, B. (1965). Trade liberalisation and "revealed" comparative advantage. *The Manchester School 33*(2), 99–123.

Hausmann, R., C. A. Hidalgo, S. Bustos, M. Coscia, A. Simoes, and M. A. Yildirim (2014). *The Atlas of Economic Complexity: Mapping paths to prosperity.* MIT Press.

Hausmann, R., J. Hwang, and D. Rodrik (2007). What you export matters. *Journal of Economic Growth 12*(1), 1–25.

Herzer, D. and F. Nowak-Lehnmann D (2006). What does export diversification do for growth? an econometric analysis. *Applied economics 38*(15), 1825–1838.

Hesse, H. (2009). Export diversification and economic growth. *Breaking into new markets: emerging lessons for export diversification*, 55–80.

Hidalgo, C. A. and R. Hausmann (2009). The building blocks of economic complexity. *Proceedings of the National Academy of Sciences 106*(26), 10570–10575.

Imbs, J. and R. Wacziarg (2003). Stages of diversification. *The American Economic Review 93*(1), 63–86.

Kemp-Benedict, E. (2014). An interpretation and critique of the method of reflections. Mimeo.

Krugman, P. R. (1991). *Geography and trade.* MIT Press.

Lall, S. (2000). The technological structure and performance of developing country manufactured exports, 1985-98. *Oxford development studies 28*(3), 337–369.

Mariani, M. S., A. Vidmer, M. Medo, and Y.-C. Zhang (2015). Measuring economic complexity of countries and products: which metric to use? *The European Physical Journal B 88*(11), 293.

Morrison, G., S. V. Buldyrev, M. Imbruno, O. A. Doria Arrieta, A. Rungi, M. Riccaboni, and F. Pammolli (2017). On economic complexity and the fitness of nations. *Scientific Reports 7*(1), 15332.

Murphy, K. M., A. Shleifer, and R. W. Vishny (1989). Industrialization and the big push. *Journal of political economy 97*(5), 1003–1026.

Newman, M. (2010). *Networks: an introduction.* Oxford University Press.

Rodrik, D. (2012). Unconditional convergence in manufacturing. *The Quarterly Journal of Economics 128*(1), 165–204.

Ruggles, S., K. Grenadek, R. Goeken, J. Grover, and M. Sobek (2017). Integrated public use microdata series: Version 7.0.

Shi, J. and J. Malik (2000). Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence 22*(8), 888–905.

Tacchella, A., M. Cristelli, G. Caldarelli, A. Gabrielli, and L. Pietronero (2012). A new metrics for countries' fitness and products' complexity. *Scientific reports 2*, 723.