BANCA D'ITALIA
EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

A probabilistic method for reconstructing the foreign direct investments network in search of ultimate host economies

by Nadia Accoto, Valerio Astuti and Costanza Catalano

April 2023

Number 760

BANCA D'ITALIA
EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

A probabilistic method for reconstructing the foreign direct investments network in search of ultimate host economies

by Nadia Accoto, Valerio Astuti and Costanza Catalano

Number 760 – April 2023

The series Occasional Papers *presents studies and documents on issues pertaining to the institutional tasks of the Bank of Italy and the Eurosystem. The* Occasional Papers *appear alongside the* Working Papers *series which are specifically aimed at providing original contributions to economic research.*

The Occasional Papers *include studies conducted within the Bank of Italy, sometimes in cooperation with the Eurosystem or other institutions. The views expressed in the studies are those of the authors and do not involve the responsibility of the institutions to which they belong.*

The series is available online at www.bancaditalia.it .

ISSN 1972-6643 (online)

Designed by the Printing and Publishing Division of the Bank of Italy

A PROBABILISTIC METHOD FOR RECONSTRUCTING THE FOREIGN DIRECT INVESTMENT NETWORK IN SEARCH OF ULTIMATE HOST ECONOMIES

by Nadia Accoto*, Valerio Astuti* and Costanza Catalano*

Abstract

The Ultimate Host Economies (UHEs) of a given country are defined as the ultimate destinations of Foreign Direct Investment (FDI) originating in that country. Bilateral FDI statistics struggle to identify them due to the non-negligible presence of conduit jurisdictions, which provide attractive intermediate destinations for pass-through investments due to favorable tax regimes. At the same time, determining UHEs is crucial for understanding the actual paths followed by FDI among increasingly interdependent economies. In this paper, we first reconstruct the global FDI network through mirroring and clustering techniques, starting from data collected by the International Monetary Fund. Then we provide a method for computing an (approximate) distribution of the UHEs of a country by using a probabilistic approach to this network, based on Markov chains. More specifically, we analyze the Italian case.

JEL Classification: C51, C60, F23, G15.
Keywords: foreign direct investment, ultimate host economies, network reconstruction, clustering, absorbing Markov chain.
DOI: 10.32057/0.QEF.2023.0760

Contents

* Bank of Italy, Statistical Analysis Directorate.

1 Introduction[1]

Foreign Direct Investment (FDI) is a category of financial cross-border investments in which an investor (direct investor) of one economy makes an investment in an enterprise (direct investment enterprise) of another economy that allows having control or a significant degree of influence on the management of that enterprise.[2] FDI statistics, which provide information on investments between immediate counterparts in two different countries, are key indicators of countries' participation in the global economy; they are usually differentiated into *inward* FDI (investments *received* by an economy) and *outward* FDI (investments *made* by an economy).

However, in a world that is more and more interconnected, such statistics are not sufficient to reconstruct the investment chains, due to the increasing presence of multinational enterprises and countries that act as tax havens[3] or investment hubs[4]. Indeed, investments can pass through the so-called Special Purpose Entities (SPEs), i.e. enterprises commonly created and registered in tax havens and investments hubs that allow tax optimization by channelling investments through economies, before arriving to the final investment recipient country. Therefore, a large portion of FDI transits in and out of some countries before reaching their final destination, producing no real economic value in the crossing country.

This led international organizations and national compilers to consider the development of *extended* experimental statistics, such as inward FDI by Ultimate Investing Economies (UIE), i.e. the starting economies of the investment chains where the first investments are originated, and outward FDI by Ultimate Host Economies (UHE), i.e. the final recipient economies of the investment chains. Compiling FDI statistics by UIE and UHE, other than by immediate counterpart country, in fact, would make the FDI statistics more complete and useful from a macroeconomic point of view, highlighting who ultimately controls the investments, the ultimate destination, and the financial connections between economies.

The fourth edition of the OECD's Benchmark Definition of Foreign Direct Investment [12] recommends to compile inward investment positions according by UIE; as of today, only few countries (Italy included[5]) provide such additional information as experimental statistics. FDI statistics by UHE belong to the

[1]We thank Marta Crispino, Valerio Della Corte, Andrea Del Monaco, Silvia Fabiani, Stefano Federico, Fadi Hassan, Giacomo Oddo, Giacomo Romanini, Alfonso Rosolia, Silvia Sabatini and Simonetta Zappa for useful discussions and suggestions. The views expressed in this paper are those of the authors and do not involve the responsibility of the Bank of Italy and/or the Eurosystem.

[2]By definition, the direct investor owns 10 percent or more of the voting power in the direct investment enterprise [12].

[3]Countries whose economies are (almost) entirely dedicated to the provision of offshore services. In this paper we consider as tax havens the countries listed in [4].

[4]Jurisdictions with real economies but that also facilitate transit of investments due to favourable tax and investment conditions.

[5]Information on the residency of ultimate investor is collected directly from reporting firms in the annual FDI survey and then cross-validated with commercial databases (Orbis).

research agenda of international organizations as the IMF ([7] §1.43) and the OECD ([12], §672). A specific guidance and methodology concerning the UHE concept is being developed in the scope of the international statistical manuals' revision process ([8]), which is still ongoing. Recently, experimental methods have been developed on that by the FDI statistical community [2, 3], although no countries have yet published comprehensive statistics on UHE.

In this framework, we propose a model to estimate, for each country, the distribution of its outward FDI by UHE. The mathematical model is based on absorbing Markov chains, and it has as input the reconstructed global outward FDI network.

The data on outward FDI are taken from the Coordinated Direct Investment Survey (CDIS) database, where the IMF yearly collects, on a voluntary basis, information about FDI stocks from all the world's countries. Even though the process of compiling the FDI statistics can have its own characteristics in each reporting country, our model is not affected by these differences, given that we exploit only proportions of outward investments from each country and not the absolute declared values. Since the reporting countries in the CDIS database are only 90 over 246 world's countries, and some of the disseminated data are kept confidential by the reporting country itself, some imputation methods are necessary to reconstruct the full network. In particular, we make use of mirroring and clustering techniques. The Markov chain model (inspired by [1]) is then set up on this reconstructed network, providing, for each country, an estimate of the percentage of its outward FDI towards each other country as final recipient.

The paper is structured as follows: Section 2 contains the description of how the outward FDI network is fully reconstructed, addressing the problem of missing data, while Section 3 describes the mathematical model used to estimate the FDI distribution by UHE for every country. Finally, Section 4 reports the results of the model in the Italian case and Section 5 presents some conclusions and future work.

2 The outward FDI network

We consider the bilateral data on outward stocks of FDI by counterpart country on year 2019, taken from the CDIS database [9]. The data in the CDIS database are presented according to the Extended Directional Principle ([7], section 6.42): a direct investment reported by a country A can be *inward*, representing the investments made in A by foreign countries, or *outward*, representing the investments made by A abroad. Direct investments abroad cover assets and liabilities between resident direct investors and their non-resident direct investment enterprises[6]. Direct investment in the reporting economy includes all liabilities and assets between resident direct investment enterprises and their non-resident direct investors. Assets and liabilities between resident and non-resident fellow

[6]Here we use *economy* as synonym of *country* and *resident/non-resident* to indicate economic entities that belong/do not belong to the reporting economy. For more details see [7]

enterprises are classified as outward (inward) investment if the ultimate controlling parent is resident (non-resident). The data are broken down by financial instruments (equity and debt instruments)[7]. The CDIS reporting economies are primarily requested to provide data on inward FDI, but they can complement them also providing data on outward FDI.

Our aim is to reconstruct the weighted, directed network of bilateral outward FDI between the world's countries and to use it to gain information about the ultimate hosting economies of each country (in particular, of Italy).

A weighted, directed network is formally defined as a triple (V, E, W) where V is a set of vertices (nodes), $E \subseteq V \times V$ is a set of links (edges) and $W = \{w_{ij} : (i, j) \in E\}$ is a set of real values that identifies the weight of each link (see also [15] for reference). In our case, we want to reconstruct the network having as nodes all the world's countries (as listed in the CDIS database), a link from i to j if country i has some outward FDI towards country j, with weight w_{ij} equal to the total outward FDI stocks of i towards j.

This network will be the starting point of the subsequent Markov chain model that will provide approximated FDI statistics by UHE. The main difficulties in reconstructing the network using CDIS data can be summarized as follows:

1. not all countries provide data on their outward FDI;
2. some reported amounts are negative,[8] while we need positive weights for the Markov chain model;
3. reporting countries can decide to flag some of their data as *confidential*, thus masking the value of the investment.[9]

In particular, the reporting countries are only 90 over the 246 possible ones and the confidential data are 12.1% of the available data (25.0% of the nonzero data available), spread over 39 reporting countries. The percentage of confidential data varies between countries: it goes from a maximum of 91.8% of confidential data for Hong Kong to a non-null minimum of 0.8% of confidential data for Guatemala, and no confidential data for 51 countries (see also Figure 1).

We handled item 2. by considering the absolute value of the negative amounts, as setting these negative values to zero would have been in contrast with the information that such link (investment) exists.[10] In the next sections we describe the solutions adopted to handle items 1. and 3.

[7] Debt instruments refer to debt positions between affiliated enterprises, see §6.26 [7]

[8] This could happen, for example, when the debt positions from the subsidiary to the parent company exceed the value of the investment of the parent in the subsidiary.

[9] This expedient is often used to avoid that individual information are indirectly disclosed by deduction from reported data. In this case, we know that a link (investment) between two countries exists but we do not know its weight.

[10] We tried to consider only the *equity* part of the outward FDI stocks, thus excluding the debt instruments component. However, by doing so we observe that we lose information on the FDI network (around 200 records less), while the percentage of negative amounts slightly decrease (from 1.8% to 1.5%). Therefore we preferred to keep the highest amount of information possible, while "paying" the fee of few more negative amounts.

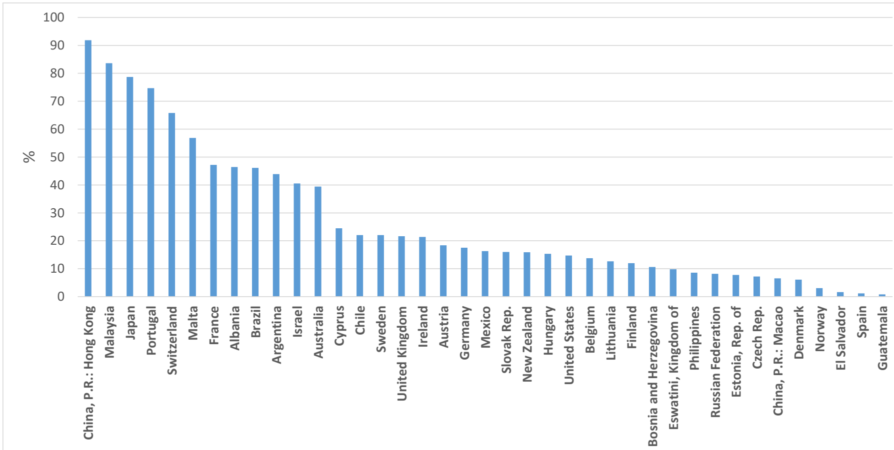

Figure 1: Percentage of confidential data in the outward FDI database by country. Only countries with at least one confidential data are displayed.

2.1 Missing data: mirroring the inward FDI

The CDIS database provides both the bilateral inward and outward FDI statistics supplied by the reporting countries.[11] The inward and outward database, in theory, should be symmetric, as for every countries i and j it should hold:

$$\text{outward FDI of } i \text{ towards } j \ = \text{inward FDI of } j \text{ from } i. \tag{1}$$

In real data, these two values might not coincide as the one on the left hand side of (1) is reported by country i, while the one on the right hand side is reported by country j. Property (1) can nonetheless be of help in (partially) solving items 1. and 3. listed at the beginning of the section. Indeed, in the inward database the reporting countries result to be 122 (in the outward database reporting countries are 90), thus providing many additional data with which we can enrich the network. We complement the outward database with data taken from the inward database and not vice versa, because in the model we want to exploit the outward direction of the investment when looking for the UHEs. In particular, we use the inward database to:

- impute the confidential data. If the outward FDI of country i to country j is confidential, we check the inward FDI reported by country j from i: if it exists and it is different from zero, we substitute the confidential data with this amount;
- add missing links. If the outward FDI of country i to country j is equal to zero or it is not reported, we check the inward FDI reported by country j

[11]All the reporting countries of the outward database are also reporting countries in the inward database but not vice versa.

from i: if it exists and it is different from zero, we add a link from i to j with this amount as weight.[12]

After these operations, we again consider the absolute values of the negative amounts that might have been substituted.

This procedure let us impute 682 values over the 1954 starting confidential data, and it let us add 7771 links (969 links that were reported as zeros in the outward FDI database plus 5927 extra links referring to non-reporting countries in the outward FDI database). The final network has 246 nodes (the total number of CDIS countries) and 14712 links, 3356 of which are still confidential (22.8% of the total data different from zero)[13].

The network results to be strongly connected, i.e. there is a path of outward FDI from any country to any other country, and sparse (sparsity score=0.76, i.e. only 24.3% of links over all the possible ones are present[14]). The next section describes how we impute the values of the remaining confidential data.

2.2 Missing data: clustering techniques

We propose in this section a second step of imputation, based on the proximity of countries in the "*outward investment space*". More specifically, each country i defines a real-valued vector $\vec{v}_i$ in a 246-dimensional space, where each j-th component v_i^j of the vector denotes the amount invested by country i in country j. All of these components have non-negative values, such that the vectors span only a very small wedge of the embedding space. In addition, countries are obviously very different in terms of the total amount invested - which is linked to the magnitude of these vectors - but useful information on the investment pattern is contained in the *ratios* between the components. The underlying assumption of our imputation strategy is that countries that are close in this space have a similar investment pattern, such that, if country i has a confidential value for investments towards a given country j, the average proportion of investment towards country j from a set of countries which are close to i will function as a good estimator for the missing value.

For this reason we define an algorithm to study the similarity of the investment vectors (for a review of proximity measures used for clustering algorithms, see [6]). We consider the scalar product between the normalized vectors after having subtracted their means:

$$\rho_{ij} = \vec{\tilde{v}}_i \cdot \vec{\tilde{v}}_j = \sum_{k=1}^{246} \tilde{v}_i^k \, \tilde{v}_j^k \tag{2}$$

[12]Note: in this case, we also keep the inward FDI data that are confidential, as they signal the presence of a link that was not reported in the outward FDI database.

[13]The percentage of confidential data is increased because we have complemented the outward FDI database also with the confidential data from the inward FDI database.

[14]The sparsity score of a directed network (V, E) is defined as $(1- \mid E \mid / \mid V \mid^2)$, with $\mid \cdot \mid$ the cardinality of the corresponding set.

where $\tilde{v}_i^k = \frac{(\vec{v}_i - \overline{v}_i)}{\|\vec{v}_i - \overline{v}_i\|}$ and $\tilde{v}_j^k = \frac{(\vec{v}_j - \overline{v}_j)}{\|\vec{v}_j - \overline{v}_j\|}$ are the standardized vectors, $\overline{v}_i$ and $\overline{v}_j$ are the mean vectors, and for any vector $\vec{v}$ we define its norm as $\|\vec{v}\| = \left[\sum_k v^k \cdot v^k\right]^{\frac{1}{2}}$. The vectors $\vec{\tilde{v}}$ belong to the sphere of radius 1 in the 246-dimensional space, so that $\sigma_{ij} = \cos\alpha$, where α is the angle formed by the vectors $\vec{\tilde{v}}_i$ and $\vec{\tilde{v}}_j$. Therefore it holds that $-1 \leqslant \cos\alpha = \rho_{ij} \leqslant 1$.[15]

Clearly, the existence of missing values will have an influence on the evaluation of the proximity score, in that we ignore the portion of the investment pattern that is hidden in confidential data. If for a given country the number of confidential entries is too large, we will probably have a distorted representation of the behaviour of the country. In the evaluation of the proximity score, neglecting the confidential values altogether would be equivalent to assuming that no investment is present in these cases. This assumption seems however unrealistic, in that one would think there are more reasons to have confidentiality on something existing, than on an absent investment. For this reason, to represent confidential data we choose a small, positive number - for definiteness fixed at 1.[16]

The imputation procedure works as follows. Whenever we encounter a confidential value for the investment from country i to country j, we select the N_{close} countries with the highest values of the proximity scores with i having a non-confidential value for the investments towards j (such that the list of N_{close} countries can be different for any imputed confidential value). In addition we put a minimum threshold on the proximity score for the country to be considered as "close" to the one with the confidential value.[17] From this set of neighbour countries, we compute the average avg_{ij} of the proportion of the investments towards country j. If the set of countries selected is empty (for example, if every country with the proximity score above the chosen threshold has also a missing value for country j), we impute the confidential value with a zero value. Otherwise, let S_i be the sum of the known values of the outward FDI from country i (which may come both from the CDIS database and the mirroring technique), $\text{conf}(i)$ be the list of countries for which i has confidential values and

$$M_i = \sum_{j \in \text{conf}(i)} \text{avg}_{ij} \quad .$$

Then:

- if i is a reporting country in the outward CDIS database, we retrieve from it the total outward FDI of country i towards the rest of the world

[15] Other proximity measures were tried, as for example the scalar product of the normalized investment vectors $\sigma_{ij} = (\vec{v}_i \cdot \vec{v}_j)/\|\vec{v}_i\|\,\|\vec{v}_j\|$, but the results turned to be pretty similar.

[16] The exact value of this placeholder does not play any role in the clustering, and is exploited only for bookkeeping reasons.

[17] The proximity score ρ_{ij} can assume negative values, so we impose the constraint $\rho_{ij} > 0$, i.e. a maximum spanned angle of $\pi/2$.

$(world_i)$.[18] Then we impute the value

$$\frac{(world_i - S_i)\,\mathrm{avg}_{ij}}{M_i}, \tag{3}$$

in this way, the sum of the total outward FDI of country i remains equal to the reported value $world_i$;

- if i is not a reporting country in the outward CDIS database, we impute the value
$$\frac{S_i\,\mathrm{avg}_{ij}}{(1-M_i)}, \tag{4}$$
in this way, the imputed value is indeed the $\mathrm{avg}_{ij}\%$ of the final total outward sum.[19]

Due to the mirroring, we underline that it may happens that $S_i > world_i$. In this case we use formula (4) instead of formula (3) for the imputation. Table 1 provides some examples of the three closest countries in term of proximity index for four selected countries. These four countries are the ones presenting the maximum percentage of confidential data.

Table 1: Clustering examples with proximity indices.

Country	**Clustering** (first 3 countries)	**Proximity score**
China, P.R.: Hong Kong	China, P.R.: Macao	0.808
	Samoa	0.806
	Western Sahara	0.806
Malaysia	Myanmar	0.814
	Grenada	0.800
	Nepal	0.798
Japan	Canada	0.914
	Mexico	0.903
	Pitcairn Islands	0.868
Portugal	Uruguay	0.826
	Venezuela, Rep. Bolivariana de	0.760
	Andorra, Principality of	0.687

To test the performance of the algorithm and in order to choose N_{close}, we consider two out-of-sample tests: first, we select a pair of countries and hide the value of the investment from the first to the second. We impute the hidden value, and evaluate the difference between the imputed value and the real one.

[18] In the CDIS database, countries also report data by counterpart macroareas, including *World*. This value never appears as confidential.

[19] In some rare cases it may happen that $M_i > 1$. We then use in formula (4) the coefficients $\hat{\mathrm{avg}}_{ij} = \gamma_i\,\mathrm{avg}_{ij}$, with $\gamma_i = (\%$ of confidential data of country $i)/M_i$. This solution fails in the even more rare case where also all the nonzero data of country i are confidential. We then impute the values $(\delta\,\mathrm{avg}_{ij})/M_i$, where δ is an arbitrary constant, normally set to 100. The choice of δ will never play a role, as in the Markov chain model we will make use of ratios.

As a second test, we hide 10 outward investments of a single country and impute them. In this way, for this second test we can evaluate the correlation between the series of true values and the series of imputed ones.[20]

With $N_{close} = 3$, the percentage of zero values imputed, i.e. when the algorithm fails to find a value to impute, is close to 5%. We report some summary statistics of the error distribution of the imputation process in Table 2. In particular, the pairwise error is simply the difference between the imputed proportion of investments toward a given country and the true value, and refers to Test 1; we excluded the cases in which the imputed value is equal to zero. To have a comparative measure of these values, the average proportion, corresponding to the inverse of the number of countries in the network, is $246^{-1} \approx 4.1 \times 10^{-3}$. The correlation statistics between the series of imputed and true values refer to Test 2, where we excluded the cases in which the series of true values is composed of zeros only.
Table 2 also reports some summary statistics of the proximity score between each country and its selected neighbours.

Table 2: Summary statistics of the imputation process results

	Mean	25-th percentile	Median	75-th percentile
Proximity	0.50	0.28	0.53	0.70
Pairwise error (10^{-3})	-0.39	-0.49	0.00	1.20
Correlation	0.53	0	0.69	0.99

The value $N_{close} = 3$ is the one chosen for reconstructing the final outward FDI network. Selecting a higher value for this parameter does not significantly improve the estimation error of either the pairwise error or the correlation measure.

3 In search of ultimate host economies

Once we have reconstructed the network, we can proceed to build the model to estimate, for each country, the distribution of its FDI by ultimate host economies. We here provide a description on how such model works, that will be formally presented in the next section.

A multinational enterprise investing from home country i in host country j could establish an intermediate step through a third country k. This intermediate step is merely financial, as in country k no real "productive" investment takes place, and it is generally qualified as conduit investment. In the model we allow for a conduit component in each country, representing the percentage of FDI received passing through the country.

We now want to simulate the investment process on the FDI network following the investment from the investor to the final recipient. Starting from

[20]For the first test we simulate the imputation of all 246^2 investment values on the network. For the second test we hide a total of 10^6 values.

country i (the origin of the investment), we know from bilateral data the distribution of the investments of i towards its immediate counterpart countries (the out-neighbours of i in the network). If w_{ij} represents the magnitude of the outward FDI of country i towards j, and $s_i = \sum_j w_{ij}$ is the total outward FDI of country i, we can say that an investment made by i has a probability of $\pi_{ij} = w_{ij}/s_i$ to be invested in country j. Suppose now that the investment has been indeed made towards country j: it either stays there (country j behaves as a non-conduit jurisdiction) and the investments chain stops making j the UHE of i, or country j behaves as a conduit and the investment passes through towards another country. The investment will pass from country j to a country k in the network according to the probability π_{jk} computed from country j's FDI bilateral data.[21] Again, either the investment stops there, making k the UHE of i, or country k behaves as a conduit and the investment passes through towards another country. Eventually, the investment will stop in some final recipient economy, that is the UHE.

What we have just described is the behaviour of a random walk on (a modified version of) the FDI network. In the next section we provide the rigorous mathematical framework of this model.

3.1 The Markov chain model

Consider the FDI network and suppose for each country/node i to be split into its conduit component i_c and its non-conduit component i_{nc}, thus producing a total of $246 \cdot 2 = 492$ nodes. To each country j we associate a probability $cond(j)$ to act as a conduit: how we estimate this parameter for every country will be addressed in the next section. Here we underline that we are making the strong assumption that the percentage of passing-through investments of a country j *does not depend* on the country from which it receives the investment: this assumption comes from the scarcity of the available data, as will be explained in the next section.[22] We now proceed to define a (discrete-time) Markov chain process on such augmented network.

A Markov chain is identified by a set of *states* $\{s_1, \dots, s_n\}$ and a set of probabilities $\{p_{ij}\}$ called *transition probabilities*. A Markov chain is a process that starts in one the states and moves successively from one state to another: if the chain is in state s_i, it will move at the next step to state s_j with probability p_{ij}. An initial probability distribution η on the states specifies the starting state[23]. For more formal definitions in the framework of Markov processes we refer the reader to [11].

In our case, the states of the Markov chain are the nodes i_λ, $\lambda \in \{c, nc\}$, of the augmented network. We now need to define the probability $p_{i_\lambda j_\mu}$ that the

[21] In principle it is not guaranteed that conduit investments will follow the same distribution π_{jk} as non-conduit ones. Being unable to estimate the particular distribution of conduit investments, we make an uninformative assumption and conflate the two (in principle different) distributions.

[22] A generalized model where such probabilities depend also on the investing country will be object of future work.

[23] I.e. η is an entry-wise nonnegative vector of lenght n whose entries sum up to 1.

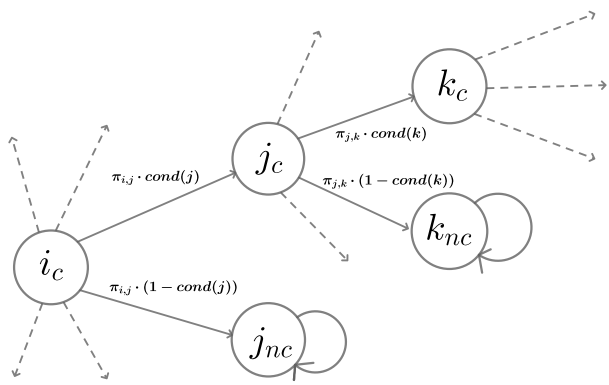

Figure 2: Visual representation of the Markov chain.

Markov chain moves from state i_λ to state j_μ, in other words the probability that the investment passes from node i_λ to node j_μ. We remind that π_{ij} represents the percentage of outward FDI from country i to country j. For each pair of nodes $(i_\lambda j_\mu)$, $\lambda, \mu \in \{c, nc\}$, we set:

- $p_{i_{nc}i_{nc}} = 1$ and hence $p_{i_{nc}j_\mu} = 0$ for all $j_\mu \neq i_{nc}$.
  This represents the fact that country i act as a non-conduit, so the investment remains in such country and does not pass through to any other country.

- $p_{i_c j_c} = \pi_{ij} \cdot cond(j)$.
  This represents the fact that i is acting as a conduit letting the investment passing to j, and j will also let the investment pass through (acts as a conduit).

- $p_{i_c j_{nc}} = \pi_{ij} \cdot (1 - cond(j))$.
  This represents the fact that i is acting as a conduit letting the investment passing to j, while j does not let the investment pass through (acts as a non-conduit), i.e. it is the final recipient.

See also Figure 2 for a graphical representation of the Markov chain. The non-conduit states i_{nc} in the literature are called *absorbing* states, because once the investment enters that state it cannot leave (no outgoing links but a selfloop). All the states that are not absorbing, the states i_c in our case, are called *transient*. We can arrange all the above probabilities in a matrix $P = [p_{i_\lambda j_\mu}]$ of dimension 492×492 where both the row and the column indices are ordered first by the absorbing states and then by the transient ones. The matrix P will then have the following structure:

$$P = \begin{array}{c} \\ \text{abs} \\ \text{trans} \end{array} \begin{array}{c} \begin{array}{cc} \text{abs} & \text{trans} \end{array} \\ \left[\begin{array}{cc} I & 0 \\ R & Q \end{array} \right] \end{array}, \tag{5}$$

where I is the 246×246 identity matrix, R represents the transition probabilities $p_{i_c j_{nc}}$ from transient states to absorbing states, and Q represents the transition probabilities $p_{i_c j_c}$ between transient states.

A Markov chain with at least one absorbing state that is reachable from any other state is called an *absorbing* Markov chain: the Markov chain defined by P in (5) is indeed absorbing. If we want to model the investment chain originating from a given country i, the starting distribution η will have all zero entries but $\eta_{i_c} = 1$. This initial condition on transient states is a mere modeling expedient to ensure that the investment moves towards another country in the first step (as otherwise it would never leave an initial absorbing state i_{nc}). This is crucial in the model as we are considering *foreign* direct investments, which by definition are directed towards countries different from the country that originated them.

An absorbing Markov chain is characterized by the fact that, independently on the starting state, it will eventually end up in one of the absorbing states (and remain there forever) with probability 1. The following theorem provides the long-run distribution of transition probabilities on an absorbing Markov chain.

Theorem 1 (Theorems 11.3-11.5 in [5])**.** *The limiting distribution of an absorbing Markov chain with transition matrix as in (5) is given by:*

$$P^* = \lim_{n \to \infty} P^n = \begin{array}{c} \\ \mathit{abs} \\ \mathit{trans} \end{array} \begin{array}{c} \begin{array}{cc} \mathit{abs} & \mathit{trans} \end{array} \\ \left[\begin{array}{cc} I & 0 \\ (I-Q)^{-1}R & 0 \end{array} \right] \end{array}. \tag{6}$$

Moreover, if the chain starts from the transient state i_c, the $i-$th component of the vector $\mathbf{t} = (I-Q)^{-1}\mathbf{1}$ represents the expected number of steps before the chain enters an absorbing state, where $\mathbf{1}$ is the all-ones column vector. We call it the expected time of absorption*.*

The interpretation of the limiting matrix P^* is the following: for any country that acts as source of the investment (the rows of the matrix), the process will end up, after a sufficiently large number of investment steps, with probability 1, in one of the UHE, modeled as absorbing states (the columns of the matrix). In particular, if we want to retrieve the FDI distribution by UHE of a country i, we just need to take the $i-$th row of the fundamental matrix $(I-Q)^{-1}R$: this is the FDI distribution by UHE over the (non-conduit version of the) world's countries. In other words, the share of FDI from country i that ends up in country j as final recipient is $P^*_{i_c, j_{nc}}$, i.e. the entry of P^* on row i_c and column j_{nc}; this is equivalent to perform the vector-matrix multiplication $\eta^\top P^*$, with $\eta_{i_c} = 1$ and all its other entries equal to zero. Consequently, the i-th entry of the vector $\mathbf{t}$ defined in Theorem 1 represents the expected number of countries the investment originated in country i crosses before arriving in its ultimate host economy.

3.2 Estimation of the conduit parameters

We here describe, for each country i, how we estimate the parameter $cond(i)$ i.e. the probability that an investment arriving in i would pass through it heading to another country.

As we have already mentioned in the introduction, most conduit investments in the world take place through a limited set of jurisdictions that act as global hubs. These hubs can be divided into two groups:

(a) *tax havens*: small jurisdictions whose economy is entirely, or almost entirely, dedicated to the provision of offshore financial services;

(b) *investment hubs*: countries that have a substantial real economic activity but also act as conduit jurisdictions due to their favourable tax and investment regimes.

We consider as tax havens 38 countries,[24] as listed by the European Commission in the Balance of Payments vademecum [4], plus 4 other countries[25] identified by Casella [1]. Note that we consider Hong Kong and Singapore belonging to group (b) instead of group (a), despite appearing in the Balance of Payments vademecum, due to their relevant size in term of population, comparable to other investments hubs such as Luxembourg. Since tax havens act fully as conduit jurisdictions, we associate to them a probability $cond(i) = 1$.

Regarding the investments hubs belonging to group (b), some countries report their FDIs made by Special Purpose Entities (SPEs) on an annual basis [13]. These data can be used to compute the conduit probability of country i as the ratio between the inward FDI through SPEs of i and its total inward FDI:

$$cond(i) = \frac{\text{Inward FDI of } i \text{ in resident SPEs}}{\text{Total Inward FDI of } i}. \tag{7}$$

In particular, as conduit probability we take the average of the above ratio on the years 2017-2019.[26] In (7) we consider *inward* FDIs beacuse we are trying to estimate the percentage of investments made to country i that will pass through it towards other countries, relative to the total investments received. The computed conduit probabilities of these self-reporting countries are reported in Table 3. To them we add four countries, namely United Kingdom, Ireland, Hong Kong and Singapore, where we consider as conduit probability the values estimated by Casella [1], through a regression method based on GDP[27]. The

[24]Andorra, Anguilla, Antigua and Barbuda, Aruba, Bahamas, Bahrain, Barbados, Belize, Bermuda, Cayman Islands, Cook Islands, Curaçao, Dominica, Gibraltar, Grenada, Guernsey, Isle of Man, Jersey, Lebanon, Liberia, Liechtenstein, Marshall Islands, Mauritius, Montserrat, Nauru, Niue, Panama, Philippines, Saint Lucia, Saint Vincent and the Grenadines, Samoa, Seychelles, Sint Maarten, St Kitts and Nevis, Turks and Caicos Islands, Vanuatu, British Virgin Islands, U.S. Virgin Islands.

[25]Malta, Monaco, Netherlands Antilles, San Marino.

[26]If some annual data are missing, we consider the average on the years with available data.

[27]Since Casella uses the outward FDI to compute such probabilities, for each country i we correct the values by the coefficient (Total Outward FDI of i/Total Inward FDI of i). Singapore does not report its outward FDIs, so we used the sum of the imputed values.

countries listed in Table 3 are the countries that we consider belonging to group (b).

Table 3: Conduit probability of self-reporting SPEs and estimated hubs (%).

Country	$cond(i)$
Luxembourg	94.08
Netherlands	68.79
Hungary	57.22
Switzerland	20.39
Austria	19.86
Denmark	18.90
Spain	7.66
Sweden	7.30
Belgium	6.85
Portugal	5.83
Norway	4.95
United States	4.35
Iceland	4.17
Lithuania	3.78
Finland	3.19
Estonia	2.77
Poland	0.92
South Korea	0.41
Hong Kong	78.94
Ireland	60.31
Singapore	24.72
United Kingdom	20.16

To all the countries that belong neither to group (a) nor to group (b) we assign a conduit probability of 0, thus implying that they always act as final recipient.

4 The Italian case

4.1 Results

The weighted FDI network reconstructed in Section 2 and the conduit probabilities computed in Section 3.2 are all the ingredients we need to run our Markov chain model (Section 3.1). We here focus on the Italian case although we remind that our model provide results on the FDI distribution by UHE for *every* country.

Figure 3 and 4 compare the Italian bilateral data on outward FDI from the CDIS database (presented as percentage over the total outward FDI) with the results we obtain from the model in terms of UHE (percentage of outward FDI towards the final recipient countries).

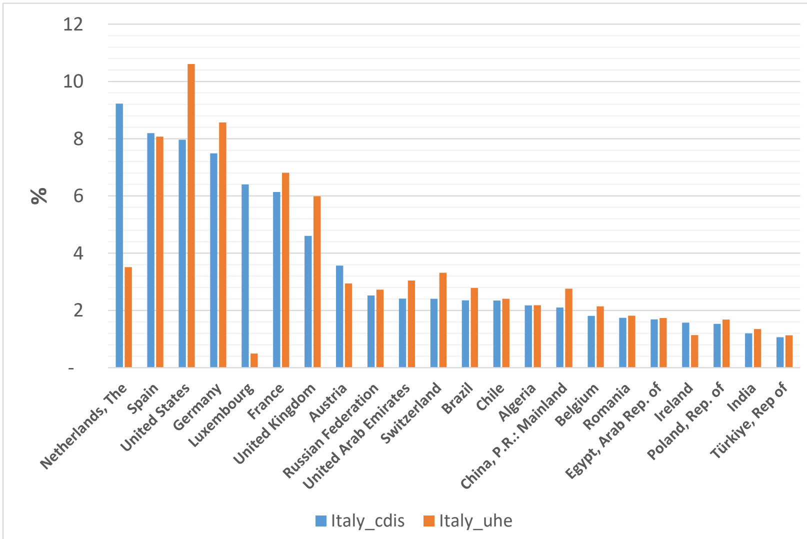

Figure 3: Comparison between the Italian bilateral FDI from CDIS and the estimated distribution of Italian FDI by ultimate host economies, sorted in decreasing order by bilateral data. Only values over 1% are displayed.

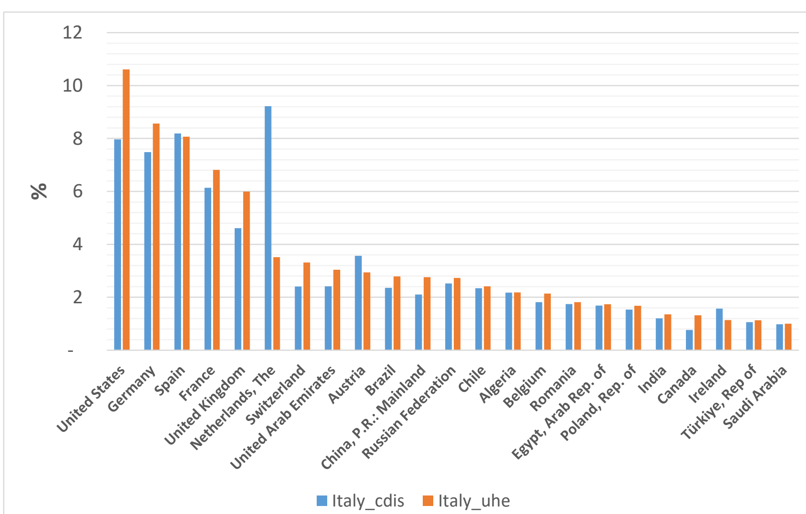

Figure 4: Comparison between the Italian bilateral FDI from CDIS and the estimated distribution of Italian FDI by ultimate host economies, sorted in decreasing order by ultimate host economies. Only values over 1% are displayed.

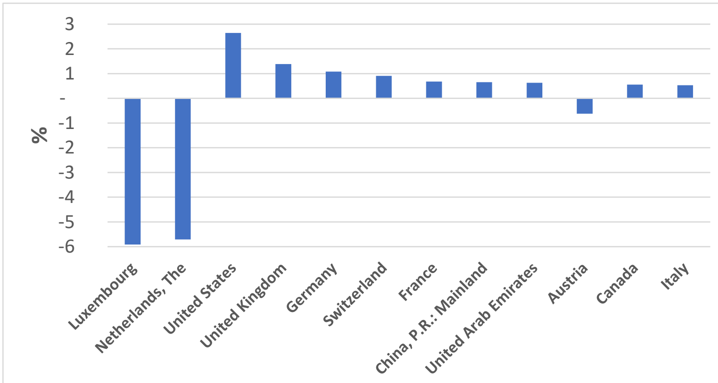

Figure 5: Difference between the estimated distribution by ultimate host economies and the Italian bilateral FDI. Only differences over 0.5 in absolute value are displayed.

We can notice that countries such as The Netherlands and Luxembourg, which are ranking respectively first and fifth according to Italian bilateral data, often let these investments pass through to other destinations: indeed, in the Italian FDI distribution by UHE, their ranks drop respectively to the sixth and 22th position. This result was somehow expected, as both countries are characterized by favorable tax regimes and by the presence of a large number of SPEs in their territory. These countries are also the ones that show the biggest (absolute) percentage difference between the model output by UHE and bilateral data (Figure 5).

Contrarily, countries such as the United States, Germany and France show the opposite behaviour: the volume of Italian investments that they receive as ultimate recipient is larger than what is reported in bilateral data. This means that some investments originated in Italy have been channeled through investments hubs and/or tax havens before ending up in these countries. Figure 4 shows that the main Italian partners in terms of UHE are the United States, Germany, Spain and France (first four positions), while ranking respectively 3rd, 4th, 2nd and 6th when considering bilateral data.

It is also interesting to observe that Italy itself appears as an Italian UHE: the results indeed show that a small percentage of Italian FDI (around 0.5%) returns to Italy, highlighting the presence of round-tripping phenomena.[28] Finally, the results obtained at the country level, in particular for Luxembourg, The Netherlands and USA, are reflected in the aggregation across main areas (Figure 6).

By computing the vector $\mathbf{t}$ in Theorem 1, the expected number of passing-through countries that an Italian investment visits before reaching its final des-

[28]Round-tripping refers to capital that leaves the economy and then goes back to it, see also §6.46 [7].

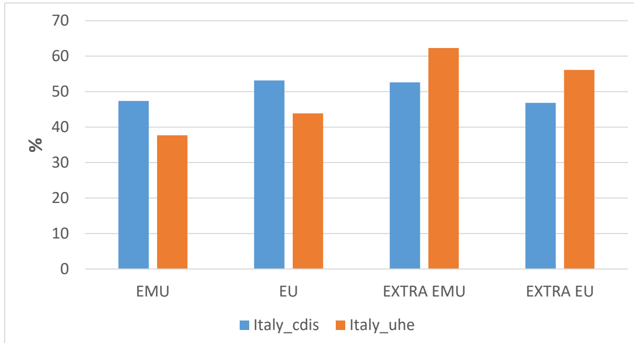

Figure 6: Comparison between the Italian bilateral FDI from CDIS and the estimated distribution of Italian FDI by ultimate host economies for main areas.

tination turns to be around 1.3. In other words, we expect an Italian investment to be channelled through one or (sometimes) two investment hubs before ending up in its ultimate host economy.

For robustness purposes, we also relaxed the hypothesis in Section 3.2 by allowing for a (small) conduit probability to every country that belonged neither to group (a) not to group (b), i.e. to which it was initially assigned a conduit probability $cond(i) = 0$. In this way we take into account both the fact that some countries do not report data on their SPEs, and that some economies might witness some small passing-through investments regardless the presence of SPEs on their territories. In particular, we performed two tests by choosing respectively $cond(i) = 10^{-3}$ and $cond(i) = 10^{-4}$ for such countries. These values were chosen as we wanted them to be positive but being *at most* of the same order of magnitude of the smallest values in Table 3. The model outputs showed that, in the final Italian FDI distribution by UHE, each value was changing at most of, respectivey, 10^{-3} and 10^{-4} as order of magnitude, thus being negligible for the main Italian UHE partners. The same applies to the expected time to absorption: the entry of the vector $\mathbf{t}$ increases at most of 10^{-3} order of magnitude, thus being negligible with respect to its value.

4.2 Results on different years

The results showed in the previous section where making use of the bilateral CDIS data of year 2019. The same exercise can be performed on different years; we here focus on the period 2013-2020 as before 2013 the data used to be compiled according to a different methodology. For the computation of the conduit probabilities we proceed as described in Section 3.2. In particular, for the self-reporting SPEs countries, their conduit probability at year X is computed by taking the average of (7) on the years X, $X-1$ and $X-2$ (moving average). For some countries the SPEs' time series 2013-2020 was not fully available, as some data were missing. In that cases, we either averaged on

Table 4: Conduit probabilities of self-reporting SPEs and estimated hubs for different years (%).

Country	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**
Luxembourg	93.62	94.60	94.46	94.56	94.06	94.82	94.08	85.68
The Netherlands	82.70	80.15	74.86	70.33	68.53	68.69	68.79	58.89
Hungary	58.75	56.46	55.96	59.12	61.69	58.88	57.22	59.61
Switzerland	12.89	12.89	14.60	16.23	18.07	19.92	20.39	18.80
Austria	37.01	35.73	33.88	29.06	26.57	19.86	19.86	19.86
Denmark	11.68	16.11	19.86	23.12	23.14	24.74	18.90	11.91
Spain	1.62	1.58	1.55	3.12	4.83	7.09	7.66	7.54
Sweden	8.33	8.39	8.33	8.00	7.72	7.34	7.30	7.11
Belgium	10.87	10.79	8.91	6.33	3.73	6.24	6.85	7.40
Portugal	14.34	12.54	11.00	9.93	8.76	7.20	5.83	4.24
Norway	1.25	1.02	1.13	1.17	1.59	3.40	4.95	6.17
United States	4.35	4.35	4.35	4.35	4.35	4.35	4.35	4.32
Iceland	37.28	33.03	32.18	28.43	20.08	11.32	4.17	4.12
Lithuania	3.51	3.51	3.51	3.51	2.57	2.24	3.78	9.37
Finland	3.08	3.08	3.08	3.08	3.08	3.30	3.19	2.94
Estonia	1.07	1.07	1.78	2.62	2.75	2.85	2.77	2.48
Poland	3.08	1.53	1.04	1.01	0.92	0.92	0.92	0.92
South Korea	1.10	1.10	0.77	0.58	0.39	0.36	0.41	0.36
Hong Kong	71.39	76.71	76.54	75.65	75.28	74.60	78.94	82.02
Ireland	85.08	96.77	66.57	66.23	60.63	59.50	60.31	58.69
Singapore	29.56	26.56	29.02	32.64	30.69	26.49	24.72	27.73
United Kingdom	23.49	18.21	22.05	21.23	21.22	19.70	20.16	20.19

the available data or, if all the data on years X, $X-1$ and $X-2$ were missing, we imputed the value by taking the one of the closest available year. Table 4 reports the computed conduit probabilities for each year and country.

Figure 7 shows the model outputs in the Italian case for selected countries among the main Italian partners. We can see that the model behaves quite consistently through out the years, and the sign of the difference between the UHE outputs and the CDIS data is maintained by all the four countries. Overall, the main Italian UHE counterpart countries remain the same during the whole period. The expected time to absorption presents a small but negative trend, starting from 1.39 in 2013 and reaching the lowest value of 1.23 in 2020 (Figure 8).

5 Conclusions and future work

In this paper we first presented a methodology on how to reconstruct the full outward FDI network starting from the (incomplete and with confidentiality issues) CDIS database. Second, we proposed a model to estimate the FDI

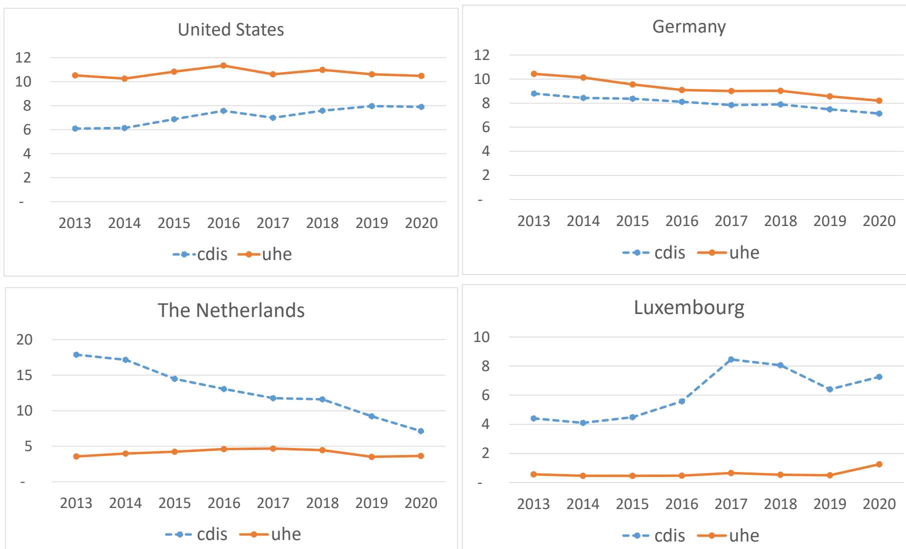

Figure 7: Italian bilateral FDI from CDIS and estimated distribution of Italian FDI by ultimate host economies for selected countries, at different years (% on the total Italian outward FDI).

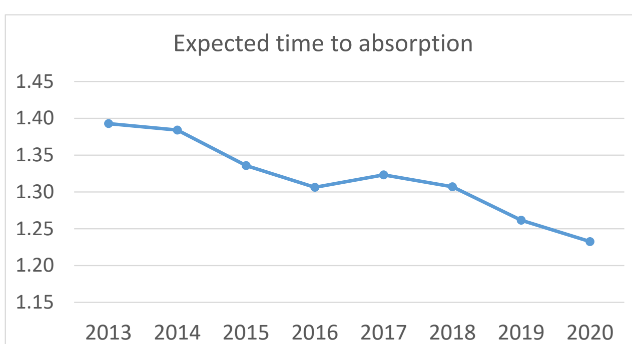

Figure 8: Expected number of passing-through countries that an Italian investment visits before reaching its UHE.

distribution by ultimate host economies for any given country. The results of the model in the Italian case show that some of the main Italian partners in terms of bilateral FDI receive much smaller volumes of investments as final recipients.

Future work would involve further testing the model for the different assumptions that were made by possibly validating it with other experimental FDI statistics by UHE. At the moment, we are not aware of any FDI statistics by UHE publicly available.

Moreover, it would be of interest to refine the model by considering conduit probabilities that depend also on the country making the investment, and not only on the receiving country. Also the search of newer techniques to compute the conduit probabilities would make it possible to extend the list of investments hubs and provide more reliable estimates. Finally, we plan to perform a deeper analysis on the FDI network in terms of connectivity, resistance to shock propagation and centrality measures [10].

References

[1] B. Casella, 'Looking through conduit FDI in search of ultimate investors - a probabilistic approach', *Transnational Corporations*, 1(26), 2019, pp. 109-146.

[2] Deutsche Bundesbank, Direct Investments Statistics, Explanatory notes, 2022.

[3] Deutsche Bundesbank, 'Foreign direct investment stocks at the end of 2020', press release, 6 May 2022.

[4] EUROSTAT, *Balance of Payment vademecum*, Update December 2021 – February 2022.

[5] C. M. Grinstead and J. L. Snell, *Introduction to probability*, American Mathematical Society, 1997.

[6] T. Hastie, R. Tibshirani and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, Springer Series in Statistics, Springer 2009.

[7] IMF, *Balance of Payments and International Investment Position Manual*, Sixth Edition (BPM6), 2009.

[8] IMF, Direct Investments Task Team, *Draft guidance note D6 - Ultimate Investing Economy (UIE)/Ultimate Host Economy (UHE) and Pass-through Funds*, 2022.

[9] IMF, Coordinated Direct Investment Survey, Data by Economy, Inward and Outward Direct Investments Positions.

[10] F. Lima, F. Pinheiro, F. Silva and P. Matos, 'Foreign direct investment – using network analysis to understand the position of Portugal in a global FDI network', *IFC Bulletins chapters*, Bank for International Settlements, 52, 2020.

[11] J. R. Norris, *Markov Chains*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 1997.

[12] OECD, *Benchmark Definition of Foreign Direct Investment*, Fourth edition, 2008.

[13] OECD Statistics, FDI positions Database, main aggregates BMD4, outward directional principle.

[14] J. Stachurski, J. Thomas and T. J. Sargent, 'Economic Networks: Theory and Computation', *QUANTECON*, I, 2022.

[15] R. J. Wilson, *Introduction to Graph Theory*, Longman, 2010.