Correlation Between Upstreamness & Downstreamness in Random Economies

by

Dylan Terry-Doyle

Individual Research Project CS01

Department of Mathematics
King's College London
United Kingdom

Candidate Number: AD07044
Supervisor: Pierpaolo Vivo
14th September, 2023

Correlation Between Upstreamness & Downstreamness in Random Economies

Abstract

Antràs and Chor (2018) put forth a "puzzling" finding in which the country level upstreamness had a strong positive correlation with downstreamness over the period 1995-2011, with a slope of approximately +1. Where upstreamness and downstreamness respectively measure the average distance from final consumption and from primary inputs of an industry or country in global value chains, which can be computed using global Input/Output tables, e.g. from the World Input Output Database (WIOD). This paper presents a sparse random model of Input/Output tables, where each element of the table is defined as the product of a Bernoulli random variable and a Pareto random variable, further extending the model developed in Bartolucci et al. (2023). It is shown that even for cases of high sparsity in the random model, the correlation slope approaches +1 as the number of industries (or countries) increases. Further validating the results in Bartolucci et al. (2023) that the perfectly positive correlation between upstreamness and downstreamness is rather an inevitable outcome due to the structural constraints of non-negative entries and substochastic rows of the matrices used in the definition of upstreamness and downstreamness, which are derived from Input/Output tables, than a result from the specific structural relationship of Input/Output tables.

Contents

1 Introduction

Global value chains (GVCs) are an integral process of the global economy, undergoing dramatic structural change in the past 40 years and even more so recently as the shocks to GVCs caused by the pandemic are reconciled by the global economy. Hence, it is imperative to understand the role GVCs have in determining the flows of exchange in both goods and money at different levels of granularity in the global economy, from industrial sectors to countries. Input/Output analysis, developed by W. Leontief (Leontief, 1936, 1986), is often used when faced with the challenge of analysing GVCs at both an intranational and international basis. Furthermore, the increasing availability of detailed data collected to form Input/Output (I-O) tables of the global economy at the industrial sector level, such as the World Input Output Database (WIOD) (Timmer et al., 2015), has facilitated the recent development of measures used to capture key elements from the complexity of GVCs.

In particular, Antràs and Chor (2012), Miller and Temurshoev (2015), and Fally (2012) developed measures of upstreamness, which measures the position of an industrial sector or country along the output supply chain[1], and downstreamness, which measures the distance of an industrial sector or country from primary factors of production (value-added, i.e. raw materials) along the input supply chain[2]. In more recent work by Antràs and Chor (2018), the author's posited a "puzzling" finding in which there is a strong positive correlation between upstreamness and downstreamness[3] approximately equal to +1 and remains constant over time. Such a relation is "puzzling" due to the theory of comparative advantage. Particularly as represented by that of the Heckscher-Ohlin model, in which the theory suggests that internationally traded goods are indirectly related to the natural endowment of country specific factors of production, such as land, labour, and capital. Therefore, under such a model, international trade can be viewed as indirect factor arbitrage (Leamer, 1995). Thus, as proposed by Antràs and Chor (2018), given this theory of international trade, one would expect that countries with a comparative advantage in natural resources would have increasingly specialised in early stage production, which would place them more downstream than upstream, and vice versa for countries with a comparative advantage in late stage production processes that are closer to final demand. Therefore, one would expect a weakly positive or negative correlation between upstreamness and downstreamness that would become less correlated over time. However, as shown by Antràs and Chor (2018), the actual correlation between upstreamness and downstreamness for countries is closer to +1 and remains relatively constant over time.

Two explanations of the correlation observed between upstreamness and downstreamness are given in Antràs and Chor (2018). The first is that increasing trade costs will increase the positive correlation between upstreamness and downstreamness, because in closed economies

[1] A relatively upstream sector or country sells a small share of its output as final demand on the output supply chain.

[2] A relatively downstream sector or country uses little value-added in its production process relative to intermediate inputs.

[3] In Antràs and Chor (2018) the author's often use a simpler definition of upstreamness and downstreamness of F/GO and VA/GO, respectively, which represent the share of global output going to either final demand (F) or value-added (VA).

value-added is equivalent to final consumption, therefore there would be no difference between upstreamness and downstreamness. The second is the industrial composition, a larger service sector share would increase the positive correlation between upstreamness and downstreamness because services tend to have smaller production chains. However, Antràs and Chor (2018) demonstrate that trade costs have fallen between the years 1995-2011 which suggest this is not a reason for the strong correlation. Although, they show that the service industries have indeed risen in output share from 59.5% in 1995 to 65.6% in 2011, which could provide an explanation for the positive correlation. They further develop a structural model of the international trade flows, in which they found that the trade costs would have reduced the correlation from 0.868 (1995) to 0.666 (2011) and the change in services share would have increased the correlation to 0.889 (2011), a marginal increase and still below the actual value of 0.912 in 2011. However, the use of a Cobb-Douglas production function (CDPF) with constant shares is cause for concern, as shown in Shaikh (1974) when the shares of production are constant the CDPF will necessarily fit a broad class of production data due to an algebraic relationship not a production relationship between inputs. Hence, a constant elasticity of substitution or Leontief production function would be better suited. Moreover, the model does not provide an explanation for the existing strong correlation, only factors that may have changed the correlation over time.

Recent work by Bartolucci et al. (2023) have demonstrated that such a strongly positive correlation coefficient may exist purely based on the structure of the matrices used to define upstreamness and downstreamness, which are derived from I-O tables, that they have non-negative elements and are row-substochastic. Bartolucci et al. (2023) prove that for a random I-O table of a closed economy where the elements are independently and identically distributed (i.i.d.) exponential random variables, and the elements of the final demand vector are also i.i.d. exponential random variables, then the correlation coefficient between the rank-1 estimate of upstreamness and downstreamness has a slope of $+1$ for any N, where no assumptions of the structure of the I-O table have been made. Bartolucci et al. (2023) further demonstrate that the findings from the analytic model are confirmed by numerical simulation for a range of distributions of the I-O matrix and final demand vector. However, they propose extensions to this work such as analysing the effect of sparsity of I-O tables on the correlation between upstreamness and downstreamness, and to investigate heterogeneity on random world I-O tables made of country blocks with different parameter values.

A key assumption in Bartolucci et al. (2023) is the approximation of upstreamness and downstreamness with the simpler rank-1 estimations. Bartolucci et al. (2020b) show that the resolvent of a matrix A (i.e. the Leontief inverse) and the corresponding influence of the ith node is approximated by the rank-1 estimations (Eq. (21)), this has been shown to hold for a variety of matrices and can then be directly applied to the analysis of upstreamness and downstreamness because they share a similar functional form. Bartolucci et al. (2020a) used such analysis on I-O tables in which their universality theorem established a formula for the average output multipliers for a class of row-substochastic matrices A, those with spectral radius $\rho(A) < 1$, which solely depends on the averages of each row. Hence, this theorem provides an approximation for the output multipliers corresponding to an economic sector if

the intermediate demand is known, without the requirement of any matrix inversion. The rank-1 model was tested against empirical data and showed significant accuracy in predicting the value for the upstreamness multipliers.

This paper intends to extend the analysis in Bartolucci et al. (2023) to that of a sparse model and for Pareto distributed elements of the I-O table and final demand vector. Making use of analytic techniques developed in Bartolucci et al. (2023) and using similar numerical simulations to confirm the results. Moreover, numerical simulation of a heterogeneous block world I-O table is presented to further corroborate the findings to the applicability at the country level analysis which is usually seen in the literature.

The structure of the paper is as follows. Section 2 provides an overview of input output analysis and the key concepts. Section 3 gives a detailed derivation and definition of upstreamness and downstreamness, the rank-1 estimations of each measure, and the correlation from the data of the WIOD. Section 4 provides a definition of the covariance and slope, empirical analysis of the distribution of the I-O matrix A and final demand vector $\boldsymbol{F}$, and a definition of the random model probability density functions. Section 5 derives the analytic model of the covariance and slope, using numerical approximations to determine the behavior for large N as the model does not have a closed form solution. Sections 6 displays numerical simulations of the model to validate the results in Section 5. Section 7 provides a discussion of the paper, interpreting the results and the consequence in the wider literature, as well as extensions that could be made to this research. Finally, Section 8 gives conclusive remarks of the paper.

2 Overview of Input-Output Analysis

Input-output analysis is method of economic analysis developed by W. Leontief in which the economy can be viewed as a system of interconnected balance sheets. This method was devised to formalise François Quesnay's *Tableau Économique* with empirical data, where Leontief (1936) created the first I-O table for the United States using data from the 1919 census. Of particular interest is the revenue accounts of economic agents, due to the nature of economic transactions such information can be represented by a matrix which captures the flow of goods from one economic agent to another. Hence, I-O analysis makes possible the reconstruction of an economy's complex network. Figure 1 displays a schematic overview of a world input output table (WIOT), which shows a global economic system of M countries and N sectors. This can be represented as two matrices for intermediary use (left) and final demand (right). Hence, the size of the intermediary use matrix is $MN \times MN$ and the final demand matrix is $M \times MN$.

Assuming an I-O table for a single country with N industries, the I-O coefficient matrix A can be defined by dividing each entry for the intermediary input by the total output of the sector, given by the vector $\boldsymbol{Y}$. Hence, the entry of matrix A is defined as $a_{ij} = d_{ij}/Y_i$, where d_{ij} is the dollar (or unit) amount of intermediary input from sector j to sector i. Thus, the entry a_{ij} of matrix A represents the amount of sector i's product that sector j must use to produce a single product. Then, given the vector of final demand $\boldsymbol{F}$, the expression of the economic

			Use by country-industries							Final use by countries			Total use
			Country 1			…	Country M			Country 1	…	Country M	
			Industry 1	…	Industry N	…	Industry 1	…	Industry N		…		
Supply from country-industries	Country 1	Industry 1											
		…											
		Industry N											
	……												
	Country M	Industry 1											
		…											
		Industry N											
Value added by labour and capital													
Gross output													

Figure 1: Schematic Outline of a World Input–Output Table (WIOT) with M countries and N industries (Timmer et al., 2015).

systems becomes:

$$\boldsymbol{Y} = A\boldsymbol{Y} + \boldsymbol{F}. \tag{1}$$

Hence, the solution to this system of linear equations is given by:

$$\boldsymbol{Y} = [\mathbb{I}_N - A]^{-1}\,\boldsymbol{F}, \tag{2}$$

where $\mathbb{I}_N$ is the identity matrix of size $N \times N$ and $[\mathbb{I}_N - A]^{-1}$ is called the Leontief inverse of matrix A. If the spectral radius of matrix A is less than one, $\rho(A) < 1$, then the Leontief inverse can be expanded as an infinite power series of A (Bartolucci et al., 2020a):

$$[\mathbb{I}_N - A]^{-1} = \mathbb{I}_N + A + A^2 + \cdots = \sum_{n=0}^{\infty} A^n. \tag{3}$$

Considering a positive shock to final demand, the first contribution to the Leontief inverse represents the direct increase in final output of all sectors needed to meet the increase in demand. Then, the second contribution represents the amount of extra output needed to meet the increase in inputs from the first contribution, and so on. In the kth contribution, this represents the kth increase in output needed to meet the chain of all prior increments. Therefore, the Leontief inverse captures the technology interdependence of the economy (Bartolucci et al., 2020a).

3 Upstreamness & Downstreamness

3.1 Definition

This subsection will follow closely the definition of upstreamness and downstreamness as presented in Bartolucci et al. (2023). The definition of upstreamness and downstreamness as given in Antràs and Chor (2012) starts by considering a closed economy composed of N industrial sectors with no inventories. Then, for each sector $i \in \{1, 2, ..., N\}$, gross output Y_i is defined by the sum of its use as a final good F_i and intermediary input Z_i, where Z_i is the sum of each

sector j's use of industry i's product (a_{ij}):

$$Y_i = F_i + Z_i = F_i + \sum_{j=1}^{N} a_{ij}, \tag{4}$$

where $a_{ij} = d_{ij}Y_j$, for which d_{ij} is the dollar amount of sector i's product needed to produce one dollars worth of sector j's output. Then by substituting the expression of a_{ij} into Eq. (4) and iterating this identity, an infinite series is obtained that reflects the use of sector i's product in the value chain of this hypothetical economy:

$$Y_i = F_i + \sum_{j=1}^{N} d_{ij}F_j + \sum_{j=1}^{N}\sum_{k=1}^{N} d_{ij}d_{ik}F_j + \ldots, \tag{5}$$

where the terms in the first sum is the direct use of i as an input and contributions beyond the first sum are the indirect uses of i as an input (Antràs and Chor, 2012, pp. 2160). Defining $(D)_{ij} = d_{ij}$ as the matrix of dollar values and using the infinite power series representation of the matrix inverse:

$$\sum_{n=0}^{\infty} D^n = \left[\mathbb{I}_N - D\right]^{-1}, \tag{6}$$

where $\mathbb{I}_N$ is once again the $N \times N$ identity matrix, then Eq. (5) in matrix form is:

$$\boldsymbol{Y} = \left[\mathbb{I}_N - D\right]^{-1}\boldsymbol{F}, \tag{7}$$

where $\boldsymbol{Y}$ is the output vector and $\boldsymbol{F}$ is the final demand column vector. Hence, this is the same as Eq. (2) defined in Section 2.

To define the measure of upstreamness for the ith industry U_{1i} Antràs and Chor (2012) multiply Eq. (5) by the distance from final use and normalise by the output of sector i:

$$\begin{aligned} U_{1i} =& 1 \times \frac{F_i}{Y_i} + 2 \times \frac{\sum_{j=1}^{N} d_{ij}F_j}{Y_i} + 3 \times \frac{\sum_{j=1}^{N}\sum_{k=1}^{N} d_{ij}d_{ik}F_j}{Y_i} + \ldots \\ =& \frac{\left(\left[\mathbb{I}_N - D\right]^{-2}\boldsymbol{F}\right)_i}{Y_i}. \end{aligned} \tag{8}$$

Then, using the definition of $\boldsymbol{Y}$ in Eq. (7) and substituting this into the definition of upstreamness, one obtains the upstreamness vector as:

$$\boldsymbol{U}_1 = \left[\mathbb{I}_N - A_U\right]^{-1}\mathbf{1}, \tag{9}$$

where $\mathbf{1}$ is the column vector of ones and

$$A_U = Y^{-1}A = \begin{pmatrix} \frac{a_{11}}{Y_1} & \cdots & \frac{a_{1N}}{Y_1} \\ \vdots & \ddots & \vdots \\ \frac{a_{N1}}{Y_N} & \cdots & \frac{a_{NN}}{Y_N} \end{pmatrix}, \tag{10}$$

where $Y = \text{diag}(\boldsymbol{Y})$ is the diagonal matrix of the output vector. Substituting the definition of

a_{ij} into A_U gives $(A_U)_{ij} = d_{ij}Y_j/Y_i$, which corresponds to the share of sector i's gross output purchased as an input to sector j. Thus, the matrix A_U has non-negative elements and is row-substochastic, meaning

$$\sum_{j=1}^{N}(A_U)_{ij} \leq 1, \qquad \forall\, i \in \{1, \ldots, N\}. \tag{11}$$

From this definition of upstreamness, the further a sector's position up the value chain the greater the weight it has, the sector's output goes through many production stages before being used as final demand. Hence, a sector that sells primarily to final demand, will have a smaller measure of upstreamness compared to a sector that sells its output primarily as an input to other sectors. Therefore, by construction, the upstreamness of sector i $(\boldsymbol{U}_1)_i \geq 1$, when it is exactly equal to one, then no output of industry i is purchased as an input to any other industry, hence, it's output is only used in final demand.

Fally (2012) independently proposed a measure of distance from final demand, where a recursive definition of upstreamness was given:

$$U_{2i} = 1 + \sum_{j=1}^{N} \frac{d_{ij}Y_j}{Y_i} U_{2j}. \tag{12}$$

This recursive relationship implies that sectors which sell a large share of their output as inputs to relatively upstream sectors, will also make that sector relatively upstream. As shown in Antràs et al. (2012), this definition is equivalent to the one in Antràs and Chor (2012). Using the fact that $d_{ij}Y_j = a_{ij}$, the $\boldsymbol{U}_2$ measure of upstreamness is also given by:

$$\boldsymbol{U}_2 = \left[\mathbb{I}_N - A_U\right]\mathbf{1}. \tag{13}$$

Miller and Temurshoev (2015) were first to introduce the measure of downstreamness, which defines the average distance of sector i from its providers of primary inputs. By using the accounting identity from the input demand side, that the output of sector i Y_i must be equal to the value of its primary inputs V_i (value-added) and its intermediary input purchases from all other sectors, thus:

$$Y_i = V_i + Z_i = V_i + \sum_{j=1}^{N} a_{ij} = V_i + \sum_{j=1}^{N} d_{ij}Y_j, \tag{14}$$

which in matrix form can be represented as:

$$\boldsymbol{Y} = \left[\mathbb{I}_N - D^T\right]^{-1}\boldsymbol{V}. \tag{15}$$

Hence, the downstreamness measure for sector i is defined as:

$$\begin{aligned} D_{1i} =& 1 \times \frac{V_i}{Y_i} + 2 \times \frac{\sum_{j=1}^{N} d_{ji} V_j}{Y_i} + 3 \times \frac{\sum_{j=1}^{N} \sum_{k=1}^{N} d_{ji} d_{ki} V_j}{Y_i} + \dots \\ =& \frac{\left(\left[\mathbb{I}_N - D^T \right]^{-2} \boldsymbol{V} \right)_i}{Y_i}. \end{aligned} \tag{16}$$

Then, using Eq. (15), the downstreamness vector can be written in matrix form as:

$$\boldsymbol{D}_1 = \left[\mathbb{I}_N - A_D \right] \mathbf{1}, \tag{17}$$

where the matrix A_D is given by:

$$A_D = (AY^{-1})^T = \begin{pmatrix} \frac{a_{11}}{Y_1} & \cdots & \frac{a_{N1}}{Y_1} \\ \vdots & \ddots & \vdots \\ \frac{a_{1N}}{Y_N} & \cdots & \frac{a_{NN}}{Y_N} \end{pmatrix}, \tag{18}$$

where Y is again the same diagonal matrix of the output vector, $Y = \text{diag}(\boldsymbol{Y})$. Hence, as with matrix A_U, matrix A_D has non-negative elements and is row-substochastic, such that:

$$\sum_{j=1}^{N} (A_D)_{ij} \leq 1, \qquad \forall\, i \in \{1, \dots, N\}. \tag{19}$$

Furthermore, it is evident from the construction of A_U and A_D that they share the same diagonal elements $(A_U)_{ii} = a_{ii}/Y_i = (A_D)_{ii}$.

As with the measure of upstreamness, Fally (2012) introduced an equivalent recursive definition of downstreamness:

$$D_{2i} = 1 + \sum_{j=1}^{N} d_{ji} D_{2j}, \tag{20}$$

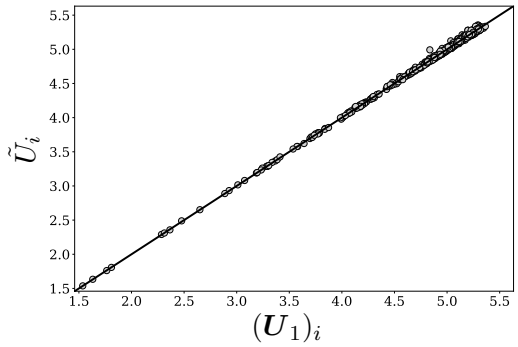
which can be expression as Eq. (17) by using the fact that $a_{ji} = d_{ji} Y_i$.

3.2 Rank-1 Estimation

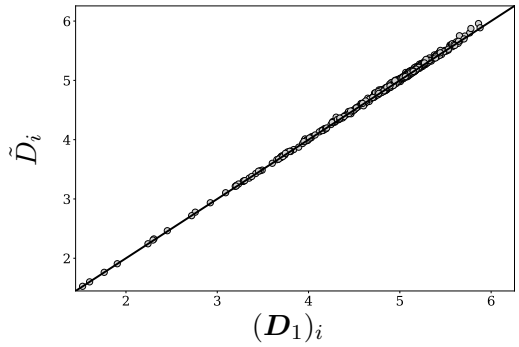
To perform the analysis in Section 4 it is first necessary to define simpler measures for upstreamness and downstreamness, which currently involve a matrix inversion which makes it a non-trivial task to evaluate the random model. By using the results from Bartolucci et al. (2020b), which show that the rank-1 estimation of the full interaction matrix accurately approximates the measures of centrality on the network, for which Bartolucci et al. (2020a) further demonstrate that this result is applicable to the measures of upstreamness and downstreamness. One can define the simpler rank-1 estimators for upstreamness ($\tilde{U}_i$) and downstreamness ($\tilde{D}_i$) as:

$$\begin{aligned} \tilde{U}_i &= 1 + \frac{r_i}{1 - \frac{1}{N} \sum_{j=1}^{N} r_j} \\ \tilde{D}_i &= 1 + \frac{r'_i}{1 - \frac{1}{N} \sum_{j=1}^{N} r'_j}, \end{aligned} \tag{21}$$

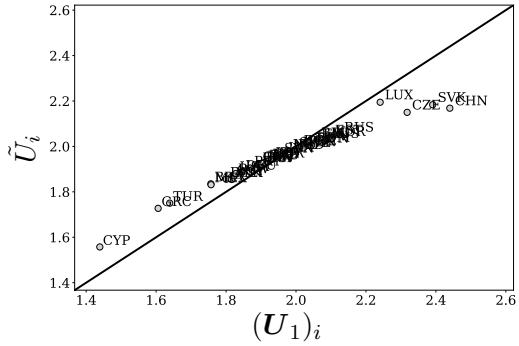
where $r_i = \sum_{j=1}^{N} (A_U)_{ij}$ is the row sum of the matrix A_U and $r'_i = \sum_{j=1}^{N} (A_D)_{ij}$ is the row sum of the matrix A_D.



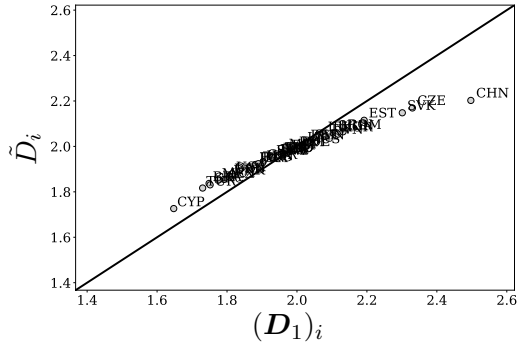
(a) Upstreamness vs. Rank-1 Estimation (Random).



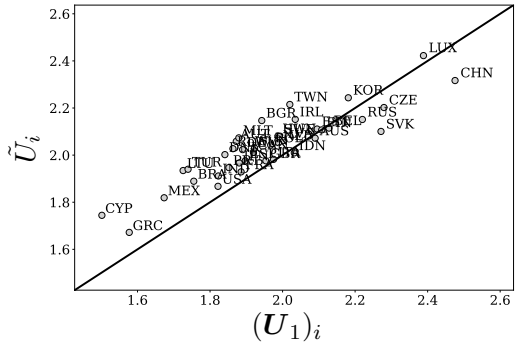
(b) Downstreamness vs. Rank-1 Estimation (Random).



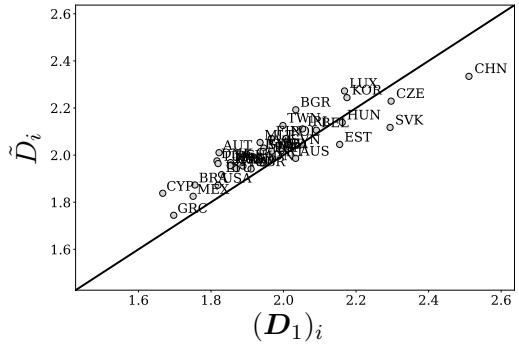
(c) Upstreamness vs. Rank-1 Estimation (1995).



(d) Downstreamness vs. Rank-1 Estimation (1995).



(e) Upstreamness vs. Rank-1 Estimation (2011).



(f) Downstreamness vs. Rank-1 Estimation (2011).

Figure 2: Correlation of upstreamness $(\boldsymbol{U}_1)_i$ and downstreamness $(\boldsymbol{D_1})i$ vs. the rank-1 estimations $\tilde{U}_i$ and $\tilde{D}_i$ respectively. In panels (a) and (b), the data was generated from a Pareto distributed random model with parameters $N = 500$ (number of points), $p = 1$, $\alpha = 2.5$, $\alpha_F = 1.5$, $m = 1$, and $m_F = 100$, see Section 4 for details of the model. Whereas, in panels (c) and (d), and (e) and (f) data from the WIOD was used for the years 1995 and 2011 respectively, using country level measures of upstreamness, downstreamness, and the rank-1 estimations.

From Figure 2 it is evident that the rank-1 estimations of upstreamness and downstreamness map with sufficient accuracy to the original definition of both measures. In Sub-Figures 2a and 2b the correlation between the original measures of upstreamness and downstreamness vs.

the rank-1 estimations is shown. Where the data has been modelled with Pareto distributed disorder of the random matrices A_U and A_D used to define upstreamness and downstreamness (see Section 4 for more details of the random model) which clearly demonstrates a strong correlation to the rank-1 estimations. Furthermore, Sub-Figures 2c-2f show the correlation of the rank-1 estimations for country level measures of upstreamness and downstreamness for the years 1995 and 2011, which also show a strong positive correlation.

Furthermore, in Table 1 it is evident that for each year from 1995-2011 the rank-1 estimations of upstreamness and downstreamness have a strong positive correlation with the original measures, all above 0.9 and statistically significant at the 1% level. Therefore, we can confidently use this rank-1 approximations as a substitute for the more complex original definitions of upstreamness and downstreamness. Although, as proposed by Bartolucci et al. (2020b), this correlation to the original matrix inversion measure may break down in the presence of high sparsity of the interaction matrix, which will need further investigation on our model.

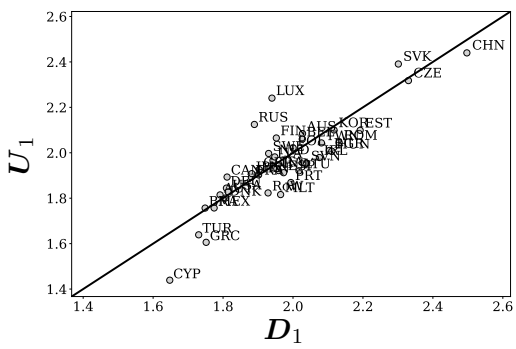
Correlation:	$\text{Corr}((\boldsymbol{U}_1)_i, \tilde{U}_i)$	$\text{Corr}((\boldsymbol{D}_1)_i, \tilde{D}_i)$
1995	0.982***	0.985***
1996	0.981***	0.985***
1997	0.983***	0.984***
1998	0.982***	0.982***
1999	0.984***	0.983***
2000	0.986***	0.984***
2001	0.986***	0.984***
2002	0.985***	0.984***
2003	0.984***	0.982***
2004	0.985***	0.983***
2005	0.983***	0.981***
2006	0.980***	0.978***
2007	0.977***	0.975***
2008	0.955***	0.950***
2009	0.952***	0.945***
2010	0.955***	0.948***
2011	0.958***	0.950***

Table 1: Pearson correlation coefficients of the rank-1 estimations of upstreamness and downstreamness over time, rounded to 3 decimal points. Where ***, **, and * denote significance at the 1%, 5%, and 10% levels respectively.

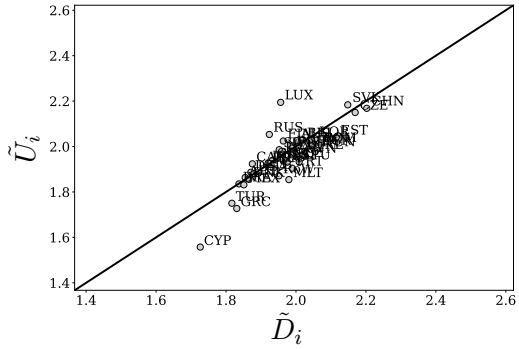
3.3 Correlation

To demonstrate this "puzzling" correlation between upstreamness and downstreamness, data from the WIOD 2013 Release (Timmer et al., 2015) has been used to analyse this relationship for the years 1995 and 2011 as in Antràs and Chor (2018). In Figure 3 it is clear that there exists over time an almost perfect correlation between upstreamness and downstreamness, where points for each country in the $(\boldsymbol{U}_1, \boldsymbol{D}_1)$ plane are clustered around the line with slope +1 in

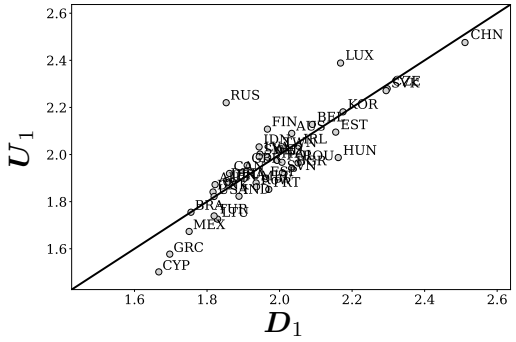
both 1995 and 2011. Additionally, the rank-1 estimations display a similar correlation, which is expected due to the strong correlation with the original measures. Hence, the findings suggest that this relationship is robust over time.



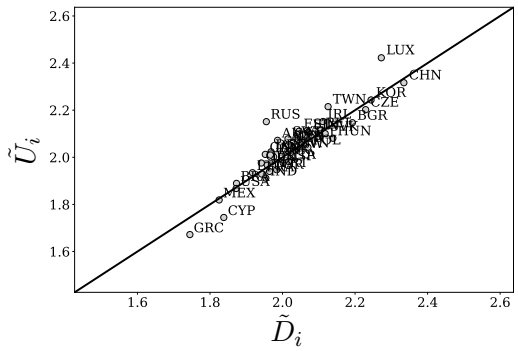
(a) Correlation in 1995.



(b) Correlation of rank-1 estimates in 1995.



(c) Correlation in 2011.



(d) Correlation of rank-1 estimates in 2011.

Figure 3: Country level correlation of upstreamness ($\boldsymbol{U}_1$) and downstreamness ($\boldsymbol{D}_1$) from WIOD. Each point is a country's position in the ($\boldsymbol{U}_1$, $\boldsymbol{D}_1$) or ($\tilde{U}_i$, $\tilde{D}_i$) plane, where the black line has a slope of +1. Panels (a) and (b) show the correlation in 1995 of both the original measures of upstreamness and downstreamness and the rank-1 estimations, respectively. Panels (c) and (c) show the correlation in 2011 of both the original measures of upstreamness and downstreamness and the rank-1 estimations, respectively.

To further analyse the correlation of upstreamness and downstreamness over time, the Pearson correlation coefficient for the years 1995-2011 have been computed in Table 2, in which the correlation has remained strongly positive over this period and in each year is significant at the 1% level for both the original measures and the rank-1 estimations. This suggests that the correlation may not be due to economic factors, but rather the structural constraints of the I-O table as proposed by Bartolucci et al. (2023), due to the insignificant change in the correlation over time and the persistence of this relationship.

Correlation:	$\mathrm{Corr}(\boldsymbol{D}_1, \boldsymbol{U}_1)$	$\mathrm{Corr}(\tilde{D}_i, \tilde{U}_i)$
1995	0.864***	0.825***
1996	0.874***	0.843***
1997	0.895***	0.876***
1998	0.895***	0.879***
1999	0.867***	0.854***
2000	0.844***	0.841***
2001	0.869***	0.868***
2002	0.859***	0.856***
2003	0.870***	0.871***
2004	0.876***	0.882***
2005	0.883***	0.888***
2006	0.889***	0.891***
2007	0.899***	0.902***
2008	0.920***	0.939***
2009	0.935***	0.952***
2010	0.929***	0.951***
2011	0.928***	0.951***

Table 2: Pearson correlation coefficients of GVC measures over time rounded to 3 decimal points. Where ***, **, and * denote significance at the 1%, 5%, and 10% levels respectively.

4 The Random Model

As in Bartolucci et al. (2023) the random model considered is that of a closed economy with N sectors. The $N \times N$ matrices A_U and A_D used to define upstreamness and downstreamness (Eqs. (10) and (18), respectively) are assumed to be generated from a random $N \times N$ Input/Output (interaction) matrix A between sectors and the final demand vector $\boldsymbol{F}$ is also drawn from a random model. Thus, no underlying assumptions regarding the structural relationship between sectors has been made, the only conditions on A_U and A_D is that their elements are non-negative and that they are row-substochastic.

4.1 Covariance & Slope

Given the assumption that the underlying dynamics of the I-O matrix A and the final demand vector $\boldsymbol{F}$ are modelled randomly, one can define the covariance between upstreamness and downstreamness of the ith industry as:

$$\mathrm{Cov}((\boldsymbol{U}_1)_i, (\boldsymbol{D}_1)_i) = \mathbb{E}[(\boldsymbol{U}_1)_i(\boldsymbol{D}_1)_i] - \mathbb{E}[(\boldsymbol{U}_1)_i]\mathbb{E}[(\boldsymbol{D}_1)_i], \tag{22}$$

where the expectation $\mathbb{E}[\cdot]$ is taken with respect to the joint probability density function (PDF) of the entries of A. Because upstreamness and downstreamness are defined by a complicated matrix inversion (Eqs. (9) and (17), respectively), computing the covariance in Eq. (22) is a

non-trivial task. Therefore, making use of the rank-1 estimations from Bartolucci et al. (2020a) and Bartolucci et al. (2020b), as described in Section 3.2 in Eq. (21), which are correlated with the more complicated actual definitions of upstreamness and downstreamness. Thus, it is sufficient to compute the covariance between the simpler rank-1 estimations of upstreamness and downstreamness:

$$\mathrm{Cov}(\tilde{U}_i, \tilde{D}_i) = \mathbb{E}[\tilde{U}_i, \tilde{D}_i] - \mathbb{E}[\tilde{U}_i]\mathbb{E}[\tilde{D}_i]. \tag{23}$$

Moreover, as in Bartolucci et al. (2023) we make the further simplification that due to the Law of Large Numbers (LLN), the values of $(1/N)\sum_{i=j}^{N} r_j$ and $(1/N)\sum_{i=j}^{N} r'_j$ from the Eq. (21) will quickly converge to their non-fluctuating averages, $\mathbb{E}[r]$ and $\mathbb{E}[r']$ respectively. Additionally, by removing the ith dependence of r and r', because every sector is statistically equivalent in the random model, thus, they can be considered as the sum of the first row of A_U and A_D respectively. Therefore, the covariance of interest is given by:

$$\mathcal{C}_N = \frac{\mathbb{E}[rr'] - \mathbb{E}[r]\mathbb{E}[r']}{(1 - \mathbb{E}[r])(1 - \mathbb{E}[r'])}. \tag{24}$$

Figure 2 in Section 3.2 shows that there is an almost perfect correlation between the upstreamness and downstreamness and the rank-1 estimations for numerical simulations of the random model.

The slope between $\tilde{D}_i$ and $\tilde{U}_i$ can be determined using Eq. (23) by assuming that there is a linear relationship with no intercept and a slope of S between $\tilde{D}_i$ and $\tilde{U}_i$:

$$\tilde{U}_i = S\tilde{D}_i. \tag{25}$$

Then, substituting Eq. (25) into the form for the covariance from Eq. (23), we get:

$$\mathrm{Cov}(\tilde{U}_i, \tilde{D}_i) = S\left(\mathbb{E}[\tilde{U}_i^2] - \mathbb{E}[\tilde{U}_i]^2\right). \tag{26}$$

Finally, solving for the slope S gives the expression:

$$S = \frac{\mathrm{Cov}(\tilde{U}_i, \tilde{D}_i)}{\mathrm{Var}(\tilde{U}_i)}, \tag{27}$$

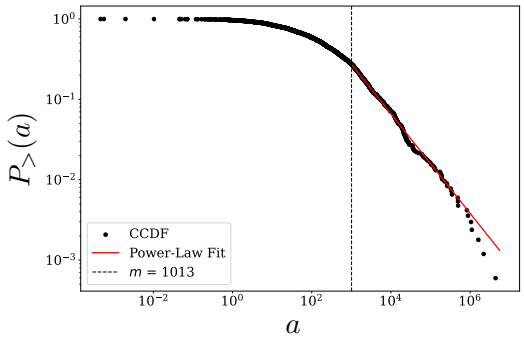
where $\mathrm{Var}(\tilde{U}_i) = \mathbb{E}[\tilde{U}_i^2] - \mathbb{E}[\tilde{U}_i]^2$ is the variance of $\tilde{U}_i$. Again, from the LLN, this can be further simplified by making the assumption that $(1/N)\sum_{j=1}^{N} r_j$ can be replaced by the non-fluctuating average $\mathbb{E}[r]$. Hence, the expression for the slope S can be further simplified in terms of the covariance in Eq. (24) of the non-fluctuating averages of $\mathbb{E}[r]$ and $\mathbb{E}[r']$:

$$S = \frac{\mathcal{C}_N(1 - \mathbb{E}[r])^2}{\mathbb{E}[r^2] - \mathbb{E}[r]^2}. \tag{28}$$

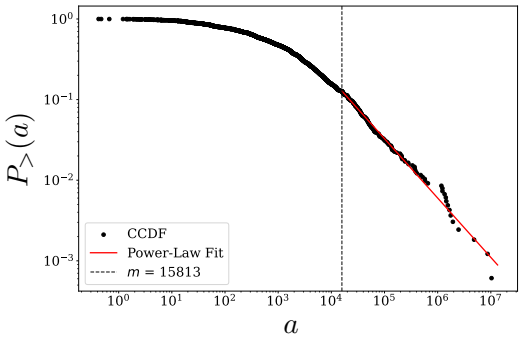
Therefore, the slope will have a value of $+1$ if $\mathbb{E}[rr']$ is equal to $\mathbb{E}[r^2]$ and $\mathbb{E}[r']$ is equal to $\mathbb{E}[r]$.

4.2 Empirical Analysis

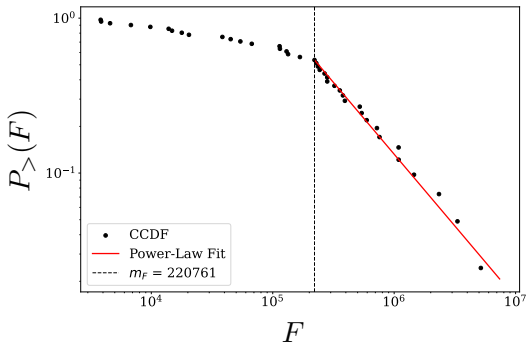
To determine the probability distribution of the random model, empirical analysis has been performed on the data from the WIOD (Timmer et al., 2015) for the country level I-O matrix and final demand vector for the years 1995 and 2011. The complementary cumulative distribution function (CCDF) of both the I-O matrix A and final demand vector $\boldsymbol{F}$ have been analysed in Figure 4, where CCDFs have been plotted on a log-log scale, the optimal values of the Pareto distribution are shown by the solid red line and the dashed line is the minimum cut-off value of the distribution. From Figure 4 the Pareto distribution fits the tails of the data in each year 1995 and 2011. However, the CCDF of the matrix A has a less pronounced fit than the final demand vector $\boldsymbol{F}$, particularly for small a (the elements of A), in which the CCDF may be better represented by a log-normal or exponential distribution. However, due to the large tail behaviour this will be best captured using the Pareto distribution because the tail values will dominate the behaviour of the distribution which will be particular critical when using the rank-1 approximation of the Leontief inverse matrix for upstreamness and downstreamness.



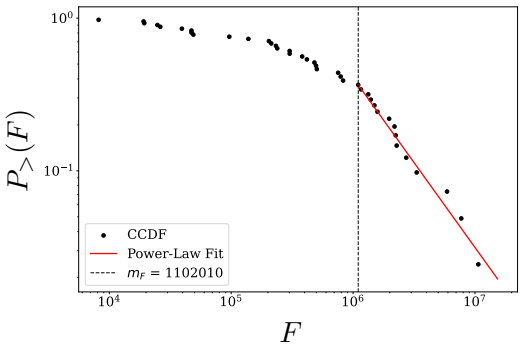
(a) CCDF of A in 1995.



(b) CCDF of A in 2011.



(c) CCDF of $\boldsymbol{F}$ in 1995.



(d) CCDF of $\boldsymbol{F}$ in 2011.

Figure 4: Country level analysis of complementary cumulative distribution function.

Therefore, the random model that will be developed will use a Pareto distribution for both the matrix A and vector $\boldsymbol{F}$, and will be further extended to include a sparsity parameter to control the sparsity of A as a Pareto distributed random variable is greater than zero by definition.

4.3 Model Definition

We first consider a simple random model in which the entries of the interaction matrix $(A)_{ij} = a_{ij}$ are independently and identically distributed (i.i.d.) from a Pareto distribution with PDF:

$$p(a_{ij}) = \begin{cases} ca_{ij}^{-(\alpha+1)} & \text{if } a_{ij} \geq m \\ 0 & \text{if } a_{ij} < m, \end{cases} \tag{29}$$

where $m = a_{\min} > 0$ is the minimum value of a_{ij} and $c = \alpha m^{\alpha}$ is the normalisation constant. Hence, the entries of A are strictly positive and contain no underlying economic structure. Moreover, the entries of the final demand vector $\boldsymbol{F}$ are also assumed to be i.i.d. according to a Pareto distribution with PDF:

$$p_F(F_i) = \begin{cases} c_F F_i^{-(\alpha_F+1)} & \text{if } F_j \geq m_F \\ 0 & \text{if } F_i < m_F, \end{cases} \tag{30}$$

where $m_F = F_{\min} > 0$ is the minimum value of F_i and $c_F = \alpha_F m_F^{\alpha_F}$ is the normalisation constant.

We then consider a more complicated random model (which will be the basis of analysis in this paper) that has a parameter $p \in (0, 1]$ which can increase ($p \to 1$) or decrease ($p \to 0$) the sparsity of the interaction matrix A. Let $X \sim \text{Pareto}(\alpha)$ and $Y \sim \text{Bernoulli}(p)$ be i.i.d. random matrices of size $N \times N$, then define the interaction matrix as the Hadamard product (element-wise product) $A = Y \circ X$, therefore the entries of A are given by $a_{ij} = y_{ij}x_{ij}$. Thus, the PDF of a_{ij} is given by:

$$p(a_{ij}) = (1-p)\delta(a_{ij}) + pca_{ij}^{-(\alpha+1)}\theta(a_{ij} - m), \tag{31}$$

where $\delta(a_{ij})$ is the Dirac delta function, $\theta(a_{ij} - m)$ is the Heaviside step function, $m = x_{\min} > 0$ is the minimum value of x_{ij} and $c = \alpha m^{\alpha}$ (from normalisation). Therefore, setting $p = 1$ recovers the first simple random model for the PDF of element a_{ij} in Eq. 29. Furthermore, the PDF of the elements of the final demand function remains the same as in the simple model, given in Eq. (30).

The interaction matrix A is used to define the matrices A_U and A_D, hence, from Eqs. (10) and (18), then using the definition of Y_i in Eq. (4) and the accounting identity $\sum_j a_{ij} + F_i = \sum_j a_{ij} + V_i$, the elements of A_U and A_D are given by:

$$(A_U)_{ij} = \frac{a_{ij}}{\sum_j a_{ij} + F_i}, \tag{32}$$

$$(A_D)_{ij} = \frac{a_{ji}}{\sum_j a_{ij} + F_i}. \tag{33}$$

Which, for appropriately chosen parameter values, particularly small α_F and large m_F relative to α and m, the elements of A_U and A_D will be non-negative and their rows will be substochastic as desired.

From Eqs. (24) and (28) the expressions of interest is to compute the covariance and slope relative to the rank-1 estimations of the first row of A_U and A_D, r and r' respectively. Thus, by definition, they are given by:

$$r = \sum_j (A_U)_{1j} = \frac{\sum_j a_{1j}}{\sum_j a_{1j} + F_1}, \tag{34}$$

$$r' = \sum_j (A_D)_{1j} = \frac{\sum_j a_{j1}}{\sum_j a_{1j} + F_1}. \tag{35}$$

Therefore, this is sufficient to compute the pairwise covariance and slope that is of interest.

5 Results

The results for the sparse random model described in Section 4 are presented below. A closed form solution for the model did not exist, therefore the analysis depends on numerical approximations of the equations, which is further shown to fit numerical simulations of the random model.

Given the definition of the row-substochastic sums of A_U and A_D, r and r' in Eqs. (34) and (35) respectively. To compute the covariance and slope functions of the rank-1 estimations, Eqs. (24) and (28) respectively, it is necessary to compute the expectations $\mathbb{E}[r]$, $\mathbb{E}[r']$, $\mathbb{E}[rr']$ and $\mathbb{E}[r^2]$. Denoting $a_{1j} = a_j$ and $F_1 = F$, then using the PDFs for a_j and F described in Eqs. (31) and (30) respectively, the expectation of r is given by:

$$\mathbb{E}[r] = \int_0^\infty \prod_{i=1}^N da_i p(a_i) \int_{m_F}^\infty dF p_F(F) \frac{\sum_k a_k}{\sum_k a_k + F}. \tag{36}$$

Using the identity:

$$\frac{1}{\xi} = \int_0^\infty ds\, e^{-\xi s} \tag{37}$$

for $\xi > 0$, then:

$$\begin{aligned} \mathbb{E}[r] = \int_0^\infty ds \int_0^\infty \prod_{i=1}^N da_i \int_{m_F}^\infty dF \left[(1-p)\delta(a_i) + pca_i^{-(\alpha+1)}\theta(a_i - m)\right] c_F F^{-(\alpha_F+1)} \\ \times e^{-s(\sum_k a_k + F)} \sum_k a_k. \end{aligned} \tag{38}$$

Each a_i is i.i.d., hence:

$$\begin{aligned}\mathbb{E}[r] =& c_F N \int_0^\infty ds \left(\int_0^\infty da \left[(1-p) e^{-sa} \delta(a) + pca^{-(\alpha+1)} e^{-sa} \theta(a-m) \right] \right)^{N-1} \\ & \times \int_0^\infty dy \left[(1-p) y e^{-sy} \delta(y) + pcy^{-\alpha} e^{-sy} \theta(y-m) \right] \int_{m_F}^\infty dF \, F^{-(\alpha_F+1)} e^{-sF} \\ =& c_F N \int_0^\infty ds \left((1-p) \int_0^\infty da \, e^{-sa} \delta(a) + pc \int_m^\infty da \, a^{-(\alpha+1)} e^{-sa} \right)^{N-1} \\ & \times \left((1-p) \int_0^\infty dy \, y e^{-sy} \delta(y) + pc \int_m^\infty dy \, y^{-\alpha} e^{-sy} \right) \int_{m_F}^\infty dF \, F^{-(\alpha_F+1)} e^{-sF} \\ =& pcc_F N \int_0^\infty ds \left[(1-p) + pcs^\alpha \Gamma(-\alpha, ms) \right]^{N-1} s^{\alpha_F+\alpha-1} \Gamma(-\alpha_F, m_F s) \Gamma(1-\alpha, ms),\end{aligned} \tag{39}$$

where $\Gamma(z,x) = \int_x^\infty t^{z-1} e^{-t} dt$ is the upper incomplete gamma function.

Similarly, the expectation of r' is given by:

$$\mathbb{E}[r'] = \int_0^\infty \prod_{i=1}^N da_i p(a_i) \prod_{j=2}^N db_j p(b_j) \int_{m_F}^\infty dF p_F(F) \frac{a_1 + \sum_{k \geq 2} b_k}{\sum_k a_k + F}, \tag{40}$$

where $b_j = a_{j1}$ for $j = 2, ..., N$. Again, using the identity in Eq. (37), the expression for the expectation becomes:

$$\begin{aligned}\mathbb{E}[r'] =& \int_0^\infty ds \int_0^\infty \prod_{i=1}^N da_i \prod_{j=2}^N db_j \int_{m_F}^\infty dF \left[(1-p)\delta(a_i) + pca_i^{-(\alpha+1)} \theta(a_i - m) \right] \\ & \times \left[(1-p)\delta(b_j) + pcb_j^{-(\alpha+1)} \theta(b_j - m) \right] c_F F^{-(\alpha_F+1)} e^{-s(\sum_k a_k + F)} \\ & \times \left(a_1 + \sum_{k \geq 2} b_k \right).\end{aligned} \tag{41}$$

Each a_i and b_j is i.i.d., hence:

$$\begin{aligned}\mathbb{E}[r'] =& c_F \left(\int_0^\infty db \left[(1-p)\delta(b) + pcb^{-(\alpha+1)} \theta(b-m) \right] \right)^{N-1} \\ & \times \int_o^\infty ds \left(\int_0^\infty da \left[(1-p) e^{-sa} \delta(a) + pca^{-(\alpha+1)} e^{-sa} \theta(a-m) \right] \right)^{N-1} \\ & \times \int_0^\infty dy \left[(1-p) y e^{-sy} \delta(y) + pcy^{-\alpha} e^{-sy} \theta(y-m) \right] \int_{m_F}^\infty dF \, F^{-(\alpha_F+1)} e^{-sF} \\ & + c_F (N-1) \left(\int_0^\infty db \left[(1-p)\delta(b) + pcb^{-(\alpha+1)} \theta(b-m) \right] \right)^{N-2} \\ & \times \int_o^\infty ds \left(\int_0^\infty da \left[(1-p) e^{-sa} \delta(a) + pca^{-(\alpha+1)} e^{-sa} \theta(a-m) \right] \right)^{N} \\ & \times \int_0^\infty dz \left[(1-p) z \delta(z) + pcz^{-\alpha} \theta(z-m) \right] \int_{m_F}^\infty dF \, F^{-(\alpha_F+1)} e^{-sF}.\end{aligned} \tag{42}$$

By definition: $\int_0^\infty db\left[(1-p)\delta(b)+pcb^{-(\alpha+1)}\theta(b-m)\right]=\int_0^\infty dbp(b)=1$, therefore:

$$\begin{aligned}\mathbb{E}[r'] =& c_F\int_o^\infty ds\left((1-p)\int_0^\infty da\, e^{-sa}\delta(a)+pc\int_m^\infty da\, a^{-(\alpha+1)}e^{-sa}\right)^{N-1}\\ &\times\left((1-p)\int_0^\infty dy\, ye^{-sy}\delta(y)+pc\int_m^\infty dy\, y^{-\alpha}e^{-sy}\right)\int_{m_F}^\infty dF\, F^{-(\alpha_F+1)}e^{-sF}\\ &+c_F(N-1)\left((1-p)\int_0^\infty dz\, z\delta(z)+pc\int_m^\infty dz\, z^{-\alpha}\right)\\ &\times\int_o^\infty ds\left((1-p)\int_0^\infty da\, e^{-sa}\delta(a)+pc\int_m^\infty da\, a^{-(\alpha+1)}e^{-sa}\right)^{N}\\ &\times\int_{m_F}^\infty dF\, F^{-(\alpha_F+1)}e^{-sF}.\\ =& pcc_F\Bigg[\int_0^\infty ds\left[(1-p)+pcs^\alpha\Gamma(-\alpha,ms)\right]^{N-1}s^{\alpha_F+\alpha-1}\Gamma(-\alpha_F,m_Fs)\Gamma(1-\alpha,ms)\\ &+\frac{m^{1-\alpha}}{\alpha-1}(N-1)\int_0^\infty ds\left[(1-p)+pcs^\alpha\Gamma(-\alpha,ms)\right]^{N}s^{\alpha_F}\Gamma(-\alpha_F,m_Fs)\Bigg].\end{aligned}\tag{43}$$

The expectation of rr' is given by:

$$\begin{aligned}\mathbb{E}[rr'] &= \int_0^\infty\prod_{i=1}^N da_ip(a_i)\prod_{j=2}^N db_jp(b_j)\int_{m_F}^\infty dFp_F(F)\frac{\sum_k a_k}{\sum_k a_k+F}\frac{a_1+\sum_{k\geq 2}b_k}{\sum_k a_k+F}\\ &=\int_0^\infty\prod_{i=1}^N da_ip(a_i)\prod_{j=2}^N db_jp(b_j)\int_{m_F}^\infty dFp_F(F)\\ &\quad\times\frac{a_1^2+a_1\sum_{k\geq 2}b_k+a_1\sum_{k\neq 1}a_k+\sum_{k\neq 1}a_k\sum_{k\geq 2}b_k}{\left(\sum_k a_k+F\right)^2}.\end{aligned}\tag{44}$$

Using the identity:

$$\frac{1}{\xi^2}=\int_0^\infty ds\, se^{-\xi s}\tag{45}$$

for $\xi>0$, then:

$$\begin{aligned}\mathbb{E}[rr'] =& \int_0^\infty ds\int_0^\infty\prod_{i=1}^N da_i\prod_{j=2}^N db_j\int_{m_F}^\infty dF\left[(1-p)\delta(a_i)+pca_i^{-(\alpha+1)}\theta(a_i-m)\right]\\ &\times\left[(1-p)\delta(b_j)+pcb_j^{-(\alpha+1)}\theta(b_j-m)\right]c_FF^{-(\alpha_F+1)}se^{-s(\sum_k a_k+F)}\\ &\times\left(a_1^2+a_1\sum_{k\geq 2}b_k+a_1\sum_{k\neq 1}a_k+\sum_{k\neq 1}a_k\sum_{k\geq 2}b_k\right).\end{aligned}\tag{46}$$

Hence, the expectation can broken into four separate integrals:

$$\mathbb{E}[rr']=c_F\left[K_1+K_2+K_3+K_4\right],\tag{47}$$

where

$$
\begin{aligned}
K_1 =& \int_0^\infty ds\, s \int_0^\infty \prod_{i=1}^N da_i \prod_{j=2}^N db_j \int_{m_F}^\infty dF \left[(1-p)\delta(b_j) + pcb_j^{-(\alpha+1)}\theta(b_j - m)\right] \\
& \times \left[(1-p)e^{-sa_i}\delta(a_i) + pca_i^{-(\alpha+1)}e^{-sa_i}\theta(a_i - m)\right] F^{-(-\alpha_F+1)}e^{-sF}a_1^2 \\
=& \left(\int_0^\infty dbp(b)\right)^{N-1} \int_0^\infty ds\, s \left((1-p)\int_0^\infty da\, e^{-sa}\delta(a) + pc\int_m^\infty da\, a^{-(\alpha+1)}e^{-sa}\right)^{N-1} \\
& \times \left((1-p)\int_0^\infty dy\, y^2 e^{-sy}\delta(y) + pc\int_m^\infty dy\, y^{1-\alpha}e^{-sy}\right)\int_{m_F}^\infty dF\, F^{-(\alpha_F+1)}e^{-sF} \\
=& pc\int_0^\infty ds\, [(1-p) + pcs^\alpha\Gamma(-\alpha, ms)]^{N-1}\, s^{\alpha_F+\alpha-1}\Gamma(-\alpha_F, m_F s)\Gamma(2-\alpha, ms),
\end{aligned} \tag{48}
$$

$$
\begin{aligned}
K_2 =& \int_0^\infty ds\, s \int_0^\infty \prod_{i=1}^N da_i \prod_{j=2}^N db_j \int_{m_F}^\infty dF \left[(1-p)\delta(b_j) + pcb_j^{-(\alpha+1)}\theta(b_j - m)\right] \\
& \times \left[(1-p)e^{-sa_i}\delta(a_i) + pca_i^{-(\alpha+1)}e^{-sa_i}\theta(a_i - m)\right] F^{-(-\alpha_F+1)}e^{-sF}a_1 \sum_{k\geq 2} b_k \\
=& (N-1)\left(\int_0^\infty dbp(b)\right)^{N-2} \left((1-p)\int_0^\infty dz\, z\delta(z) + pc\int_m^\infty dz\, z^{-\alpha}\right) \\
& \times \int_0^\infty ds\, s \left((1-p)\int_0^\infty da\, e^{-sa}\delta(a) + pc\int_m^\infty da\, a^{-(\alpha+1)}e^{-sa}\right)^{N-1} \\
& \times \left((1-p)\int_0^\infty dy\, ye^{-sy}\delta(y) + pc\int_m^\infty dy\, y^{-\alpha}e^{-sy}\right)\int_{m_F}^\infty dF\, F^{-(\alpha_F+1)}e^{-sF} \\
=& p^2c^2\frac{m^{1-\alpha}}{\alpha-1}(N-1)\int_0^\infty ds\, [(1-p) + pcs^\alpha\Gamma(-\alpha, ms)]^{N-1}\, s^{\alpha_F+\alpha} \\
& \times \Gamma(-\alpha_F, m_F s)\Gamma(1-\alpha, ms),
\end{aligned} \tag{49}
$$

$$
\begin{aligned}
K_3 =& \int_0^\infty ds\, s \int_0^\infty \prod_{i=1}^N da_i \prod_{j=2}^N db_j \int_{m_F}^\infty dF \left[(1-p)\delta(b_j) + pcb_j^{-(\alpha+1)}\theta(b_j - m)\right] \\
& \times \left[(1-p)e^{-sa_i}\delta(a_i) + pca_i^{-(\alpha+1)}e^{-sa_i}\theta(a_i - m)\right] F^{-(-\alpha_F+1)}e^{-sF}a_1 \sum_{k\neq 1} a_k \\
=& (N-1)\left(\int_0^\infty dbp(b)\right)^{N-1} \\
& \times \int_0^\infty ds\, s \left((1-p)\int_0^\infty da\, e^{-sa}\delta(a) + pc\int_m^\infty da\, a^{-(\alpha+1)}e^{-sa}\right)^{N-2} \\
& \times \left((1-p)\int_0^\infty dy\, ye^{-sy}\delta(y) + pc\int_m^\infty dy\, y^{-\alpha}e^{-sy}\right)^2 \int_{m_F}^\infty dF\, F^{-(\alpha_F+1)}e^{-sF} \\
=& p^2c^2(N-1)\int_0^\infty ds\, [(1-p) + pcs^\alpha\Gamma(-\alpha, ms)]^{N-1}\, s^{\alpha_F+2\alpha-1}\Gamma(-\alpha_F, m_F s) \\
& \times \Gamma(1-\alpha, ms)^2,
\end{aligned} \tag{50}
$$

$$
\begin{aligned}
K_4 =& \int_0^\infty ds\, s \int_0^\infty \prod_{i=1}^{N} da_i \prod_{j=2}^{N} db_j \int_{m_F}^\infty dF \left[(1-p)\delta(b_j) + pcb_j^{-(\alpha+1)}\theta(b_j - m)\right] \\
&\times \left[(1-p)e^{-sa_i}\delta(a_i) + pca_i^{-(\alpha+1)}e^{-sa_i}\theta(a_i - m)\right] F^{-(-\alpha_F+1)}e^{-sF} \sum_{k\neq 1} a_k \sum_{k\geq 2} b_k \\
=& (N-1)^2 \left(\int_0^\infty dbp(b)\right)^{N-2} \left((1-p)\int_0^\infty dz\, z\delta(z) + pc\int_m^\infty dz\, z^{-\alpha}\right) \\
&\times \int_0^\infty ds\, s \left((1-p)\int_0^\infty da\, e^{-sa}\delta(a) + pc\int_m^\infty da\, a^{-(\alpha+1)}e^{-sa}\right)^{N-1} \\
&\times \left((1-p)\int_0^\infty dy\, ye^{-sy}\delta(y) + pc\int_m^\infty dy\, y^{-\alpha}e^{-sy}\right)\int_{m_F}^\infty dF\, F^{-(\alpha_F+1)}e^{-sF} \\
=& p^2c^2\frac{m^{1-\alpha}}{\alpha-1}(N-1)^2\int_0^\infty ds\left[(1-p) + pcs^\alpha\Gamma(-\alpha, ms)\right]^{N-1} s^{\alpha_F+\alpha} \\
&\times \Gamma(-\alpha_F, m_F s)\Gamma(1-\alpha, ms).
\end{aligned} \tag{51}
$$

Therefore, the total expression is given by:

$$
\begin{aligned}
\mathbb{E}[rr'] =& pcc_F\Bigg[\int_0^\infty ds\left[(1-p) + pcs^\alpha\Gamma(-\alpha, ms)\right]^{N-1} s^{\alpha_F+\alpha-1}\Gamma(-\alpha_F, m_F s)\Gamma(2-\alpha, ms) \\
&+ pc\frac{m^{1-\alpha}}{\alpha-1}(N-1)\int_0^\infty ds\left[(1-p) + pcs^\alpha\Gamma(-\alpha, ms)\right]^{N-1} s^{\alpha_F+\alpha} \\
&\quad\times \Gamma(-\alpha_F, m_F s)\Gamma(1-\alpha, ms) \\
&+ pc(N-1)\int_0^\infty ds\left[(1-p) + pcs^\alpha\Gamma(-\alpha, ms)\right]^{N-1} s^{\alpha_F+2\alpha-1} \\
&\quad\times \Gamma(-\alpha_F, m_F s)\Gamma(1-\alpha, ms)^2 \\
&+ pc\frac{m^{1-\alpha}}{\alpha-1}(N-1)^2\int_0^\infty ds\left[(1-p) + pcs^\alpha\Gamma(-\alpha, ms)\right]^{N-1} s^{\alpha_F+\alpha} \\
&\quad\times \Gamma(-\alpha_F, m_F s)\Gamma(1-\alpha, ms)\Bigg]
\end{aligned} \tag{52}
$$

Finally, to compute the slope, the expectation of r^2 is required, given by:

$$
\mathbb{E}[r^2] = \int_0^\infty \prod_{i=1}^{N} da_i p(a_i) \int_{m_F}^\infty dF p_F(F)\left(\frac{\sum_k a_k}{\sum_k a_k + F}\right)^2. \tag{53}
$$

Again, using the identity is Eq. (45):

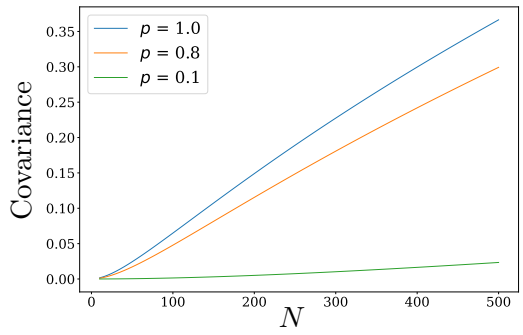
$$
\begin{aligned}
\mathbb{E}[r^2] = &\int_0^\infty ds \int_0^\infty \prod_{i=1}^{N} da_i \int_{m_F}^\infty dF \left[(1-p)\delta(a_i) + pca_i^{-(\alpha+1)}\theta(a_i - m)\right] c_F F^{-(\alpha_F+1)} \\
&\times s e^{-(\sum_k a_k + F)} \left(\sum_k a_k\right)^2 \\
= &c_F \Bigg[N \int_0^\infty ds\, s \left((1-p)\int_0^\infty da\, e^{-sa}\delta(a) + pc\int_m^\infty da\, a^{-(a+1)} e^{-sa}\right)^{N-1} \\
&\times \left((1-p)\int_0^\infty dy\, y^2 e^{-sy}\delta(y) + pc\int_m^\infty dy\, y^{1-\alpha} e^{-sy}\right) \int_{m_F}^\infty dF\, F^{-(\alpha_F+1)} e^{-sF} \\
&+ (N^2 - N)\int_0^\infty ds\, s \left((1-p)\int_0^\infty da\, e^{-sa}\delta(a) + pc\int_m^\infty da\, a^{-(a+1)} e^{-sa}\right)^{N-2} \\
&\times \left((1-p)\int_0^\infty dy\, y e^{-sy}\delta(y) + pc\int_m^\infty dy\, y^{-\alpha} e^{-sy}\right)^2 \int_{m_F}^\infty dF\, F^{-(\alpha_F+1)} e^{-sF} \Bigg] \\
= &pcc_F N \Bigg[\int_0^\infty ds \left[(1-p) + pcs^\alpha \Gamma(-\alpha, ms)\right]^{N-1} s^{\alpha_F+\alpha-1} \Gamma(-\alpha_F, m_F s) \Gamma(2-\alpha, ms) \\
&+ pc(N-1)\int_0^\infty ds \left[(1-p) + pcs^\alpha \Gamma(-\alpha, ms)\right]^{N-2} s^{\alpha_F+2\alpha-1} \Gamma(-\alpha_F, m_F s) \\
&\times \Gamma(1-\alpha, ms)^2 \Bigg].
\end{aligned}
\tag{54}
$$

To investigate the relationship of the model parameters on the covariance and slope functions, the above integrals have been computed numerically. Figure 5 displays the numerical approximations of the covariance and slope functions for different parameter values. In Sub-Figures 5a and 5b is it evident that changing the level of sparsity in the model, by changing the parameter p, has a significant impact on the value of both the covariance and slope functions. In which, the more sparse the model the weaker the relationship between N and the covariance and slope. Hence, the closer p is to 1 (no sparsity) the faster the slope approaches $+1$ for increasing N, and for p closer to 0 ($p = 0.1$) the slower the slope approaches $+1$ for increasing N. Thus, the slope will approximate $+1$ for large enough N even when the sparsity of the system is quite high.

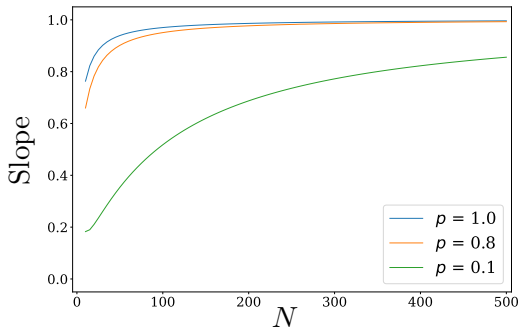
Furthermore, in Sub-Figures 5c and 5d the relationship between the covariance and slope functions is shown for different values of α. In particular, for higher α the faster the slope approaches $+1$ for increasing N, which is consistent with the findings in Bartolucci et al. (2023), because when α is higher the Pareto probability distribution will behave similarly to an exponential distribution as the tail becomes smaller. Additionally, for $\alpha < 1$, the relationship between N and the slope breaks down, which is expected because the Pareto distribution no longer has a finite mean when $\alpha < 1$, hence, the dominant contributions from the final demand vector are no longer significant.

Conversely, in Sub-Figures 5e and 5f the effect of different values of m_F has little impact on the slope values for different N. Although, the effect on the covariance is larger than the other

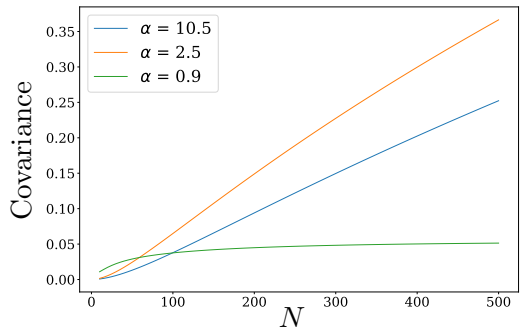
parameters, for lower m_F the higher the covariance as N increases.



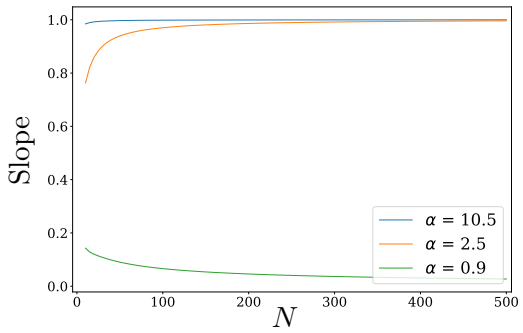
(a) Numerical solution of the covariance for different values of the sparsity parameter p.



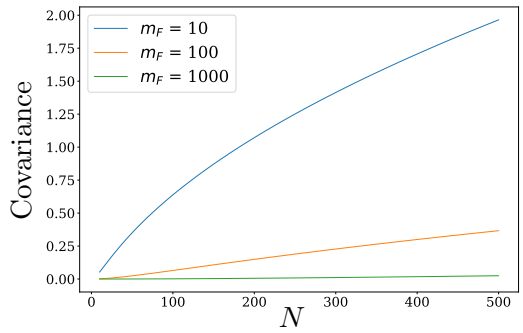
(b) Numerical solution of the slope for different values of the sparsity parameter p.



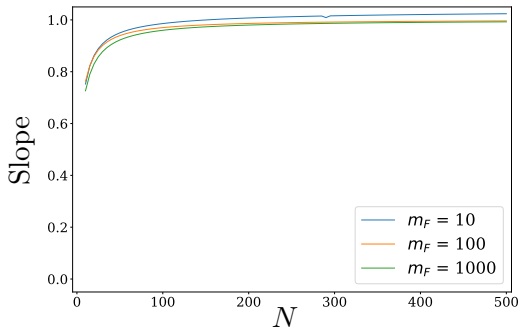
(c) Numerical solution of covariance for different values of the power-law parameter α.



(d) Numerical solution of slope for different values of the power-law parameter α.



(e) Numerical solution of covariance for different values of the power-law parameter m_F.



(f) Numerical solution of slope for different values of the power-law parameter m_F.

Figure 5: Numerical solutions of the covariance and slope functions for increasing N for different values of the model parameters. Parameter values, panels (a) and (b) $p = \{1, 0.8, 0.1\}$, $\alpha = 2.5$, $\alpha_F = 1.5$, $m = 1$, and $m_F = 100$; panels (c) and (d) $p = 1$, $\alpha = \{10.5, 2.5, 0.9\}$, $\alpha_F = 1.5$, $m = 1$, and $m_F = 100$; panels (e) and (f) $p = 1$, $\alpha = 2.5$, $\alpha_F = 1.5$, $m = 1$, and $m_F = \{10, 100, 1000\}$.

5.1 Limit of Large N: p of $\mathcal{O}(\frac{1}{N})$

To investigate the behaviour of these functions for increasing sparsity with the system size N, we take the limiting case as $N \to \infty$ for p of $\mathcal{O}(1/N)$. Hence, we define p as:

$$p := \frac{\nu}{N}, \tag{55}$$

where $0 < \nu \leq N$, thus keeping $p \in (0, 1]$. Then, substituting this definition of p and taking the limit as $N \to \infty$ of the covariance function (which in this case can be broken into the limits of each expectation, because in the limit of large N none of expectations depend on N). Hence, the expectation of r becomes:

$$\begin{aligned}
\lim_{N\to\infty} \mathbb{E}[r] = &\lim_{N\to\infty} \frac{\nu}{N} cc_F N \int_0^\infty ds \left[1 - \frac{\nu}{N} + \frac{\nu c s^\alpha \Gamma(-\alpha, ms)}{N}\right]^{N-1} s^{\alpha_F+\alpha-1} \\
&\Gamma(-\alpha_F, m_F s)\Gamma(1-\alpha, ms) \\
=&\nu cc_F \int_0^\infty ds \lim_{N\to\infty} \left[1 + \frac{\nu c s^\alpha \Gamma(-\alpha, ms) - \nu}{N}\right]^{N-1} s^{\alpha_F+\alpha-1} \\
&\Gamma(-\alpha_F, m_F s)\Gamma(1-\alpha, ms) \\
=&\nu cc_F \int_0^\infty ds \exp\left[\nu c s^\alpha \Gamma(-\alpha, ms) - \nu\right] s^{\alpha_F+\alpha-1} \Gamma(-\alpha_F, m_F s)\Gamma(1-\alpha, ms).
\end{aligned} \tag{56}$$

Similarly:

$$\begin{aligned}
\lim_{N\to\infty} \mathbb{E}[r'] = &\lim_{N\to\infty} \left(\frac{\nu}{N} cc_F \int_0^\infty \left[1 + \frac{\nu c s^\alpha \Gamma(-\alpha, ms) - \nu}{N}\right]^{N-1} s^{\alpha_F+\alpha-1} \Gamma(-\alpha_F, m_F s) \right. \\
&\quad \times \Gamma(1-\alpha, ms) \\
&\left. + \frac{\nu}{N} cc_F \frac{m^{1-\alpha}}{\alpha - 1}(N-1) \int_0^\infty \left[1 + \frac{\nu c s^\alpha \Gamma(-\alpha, ms) - \nu}{N}\right]^{N} s^{\alpha_F} \Gamma(-\alpha_F, m_F s) \right) \\
=&\nu cc_F \frac{m^{1-\alpha}}{\alpha - 1} \int_0^\infty ds \exp\left[\nu c s^\alpha \Gamma(-\alpha, ms) - \nu\right] s^{\alpha_F} \Gamma(-\alpha_F, m_F s),
\end{aligned} \tag{57}$$

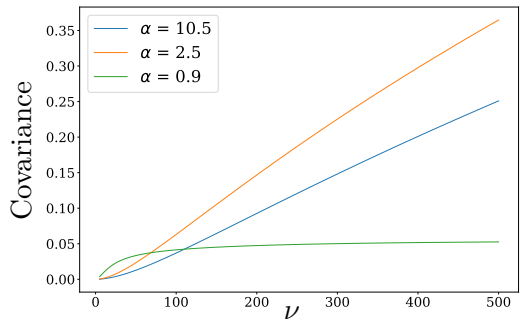
$$
\begin{aligned}
\lim_{N\to\infty}\mathbb{E}[rr'] = \lim_{N\to\infty}\Bigg(&\frac{\nu}{N}cc_F\int_0^\infty ds\left[1+\frac{\nu cs^\alpha\Gamma(-\alpha,ms)-\nu}{N}\right]^{N-1}s^{\alpha_F+\alpha-1}\\
&\times\Gamma(-\alpha_F,m_Fs)\Gamma(2-\alpha,ms)\\
&+\frac{\nu^2}{N^2}c^2c_F\frac{m^{1-\alpha}}{\alpha-1}(N-1)\int_0^\infty ds\left[1+\frac{\nu cs^\alpha\Gamma(-\alpha,ms)-\nu}{N}\right]^{N-1}s^{\alpha_F+\alpha}\\
&\times\Gamma(-\alpha_F,m_Fs)\Gamma(1-\alpha,ms)\\
&+\frac{\nu^2}{N^2}c^2c_F(N-1)\int_0^\infty ds\left[1+\frac{\nu cs^\alpha\Gamma(-\alpha,ms)-\nu}{N}\right]^{N-1}s^{\alpha_F+2\alpha-1}\\
&\times\Gamma(-\alpha_F,m_Fs)\Gamma(1-\alpha,ms)^2\\
&+\frac{\nu^2}{N^2}c^2c_F\frac{m^{1-\alpha}}{\alpha-1}(N-1)^2\int_0^\infty ds\left[1+\frac{\nu cs^\alpha\Gamma(-\alpha,ms)-\nu}{N}\right]^{N-1}s^{\alpha_F+\alpha}\\
&\times\Gamma(-\alpha_F,m_Fs)\Gamma(1-\alpha,ms)\Bigg)\\
=&\nu^2c^2c_F\frac{m^{1-\alpha}}{\alpha-1}\int_0^\infty ds\exp\left[\nu cs^\alpha\Gamma(-\alpha,ms)-\nu\right]s^{\alpha_F+\alpha}\Gamma(-\alpha_F,m_Fs)\\
&\times\Gamma(1-\alpha,ms),
\end{aligned}
\tag{58}
$$

$$
\begin{aligned}
\lim_{N\to\infty}\mathbb{E}[r^2] = \lim_{N\to\infty}\Bigg(&\frac{\nu}{N}cc_FN\int_0^\infty ds\left[1+\frac{\nu cs^\alpha\Gamma(-\alpha,ms)-\nu}{N}\right]^{N-1}s^{\alpha_F+\alpha-1}\\
&\times\Gamma(-\alpha_F,m_Fs)\Gamma(2-\alpha,ms)\\
&+\frac{\nu^2}{N^2}c^2c_FN(N-1)\int_0^\infty ds\left[1+\frac{\nu cs^\alpha\Gamma(-\alpha,ms)-\nu}{N}\right]^{N-2}s^{\alpha_F+2\alpha-1}\\
&\times\Gamma(-\alpha_F,m_Fs)\times\Gamma(1-\alpha,ms)^2\Bigg)\\
=&\nu cc_F\int_0^\infty ds\exp\left[\nu cs^\alpha\Gamma(-\alpha,ms)-\nu\right]s^{\alpha_F+\alpha-1}\Gamma(-\alpha_F,m_Fs)\Gamma(2-\alpha,ms)\\
&+\nu^2c^2c_F\int_0^\infty ds\exp\left[\nu cs^\alpha\Gamma(-\alpha,ms)-\nu\right]s^{\alpha_F+2\alpha-1}\Gamma(-\alpha_F,m_Fs)\\
&\times\Gamma(1-\alpha,ms)^2.
\end{aligned}
\tag{59}
$$

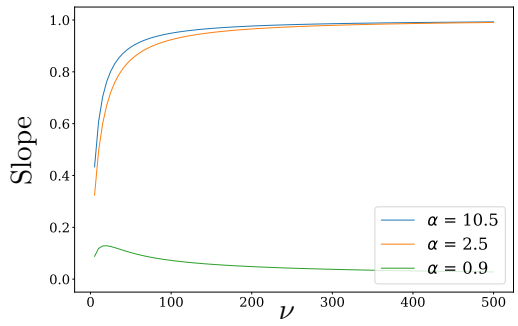
Numerical methods have again been used to evaluate these integrals for different parameter values, because they also do not have a closed form solution. Of particular interest is the effect of the sparsity parameter ν, which reduces the sparsity for larger ν, on the limiting covariance and slope functions for different parameter values. As shown in Figure 6, Sub-Figures 6a and 6b show the relationship for different values of α on the covariance and slope functions. Where a higher value of α, which reduces the size of the tail, corresponds to lower values of the covariance and a slope that approaches $+1$ faster as ν increases. Furthermore, similarly to the base line model, when $\alpha < 1$ the relationship between the covariance and slope in terms of ν breaks down because the contributions from the final demand vector become insignificant.

Furthermore, Sub-Figures 6c and 6d show the relationship between ν and m_F on the covariance and slope. In which, a lower value of m_F corresponds to a higher value of the covariance and a slightly higher value for the slope as ν increases.

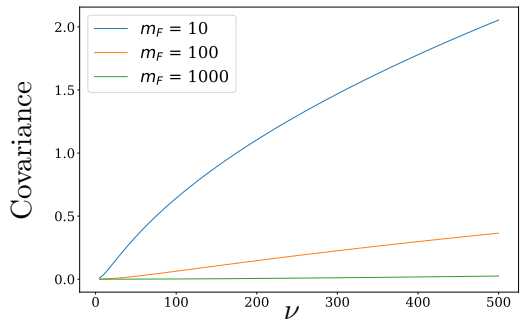
Therefore, for p of $\mathcal{O}(1/N)$ the limiting slope can be said to be robust against the effects of sparsity for a given threshold of ν, approximately $\nu > 100$, and will still be approximately equal $+1$ even when the sparsity increases with the size of the system.



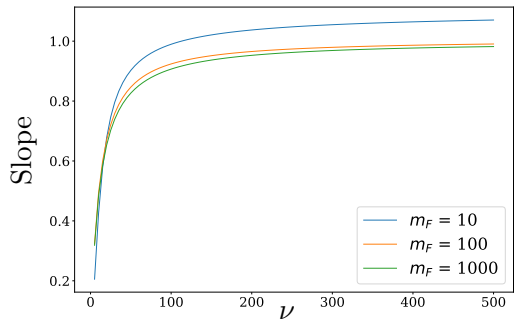
(a) Numerical solution of covariance for different values of the power-law parameter α.



(b) Numerical solution of slope for different values of the power-law parameter α.



(c) Numerical solution of covariance for different values of the power-law parameter m_F.



(d) Numerical solution of slope for different values of the power-law parameter m_F.

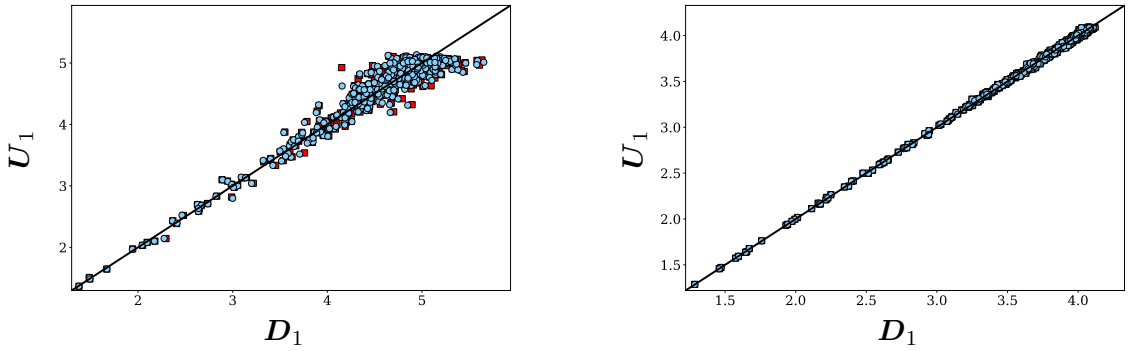
Figure 6: Numerical solutions of the covariance and slope functions for increasing ν for different values of the model parameters. Parameter values, panels (a) and (b) $\alpha = \{10.5, 2.5, 0.9\}$, $\alpha_F = 1.5$, $m = 1$, and $m_F = 100$; panels (c) and (d) $\alpha = 2.5$, $\alpha_F = 1.5$, $m = 1$, and $m_F = \{10, 100, 1000\}$.

6 Numerical Simulations

To further corroborate the findings from the numerical approximations of the analytic solution, we also perform numerical simulations of the random model defined in Section 4 by generating the $N \times N$ random input use matrix A, which is then used to compute the vectors for upstreamness $\boldsymbol{U}_1$ and downstreamness $\boldsymbol{D}_1$, defined in Eqs. (9) and (17) respectively.

Initially we generate the simple random model where there is no sparsity (p=1), such that each entry of A is generated according to a Pareto distribution with parameters α and m, where α is the slope of the power law and m is the cut-off. Furthermore, the entries of the final demand vector $\boldsymbol{F}$ are also generated according to a Pareto distribution with parameters α_F and m_F, where $\alpha > \alpha_F$ and $m < m_F$. Figure 7 shows results of numerical simulations for different values of α. In Sub-Figure 7a $\alpha = 2.5$, in which the points are clustered around the line with slope $+1$ and is similar to that of the country level scatter plot in Figure 3. In Sub-Figure 7b the

points of the scatter plot are almost perfectly on the line with slope +1, this is congruent with the results of the numerical approximations of the random model which predicted a slope closer to +1 for higher α. In both cases, the rank-1 estimations map closed to the original values for upstreamness and downstreamness, with particular accuracy in the high α case.

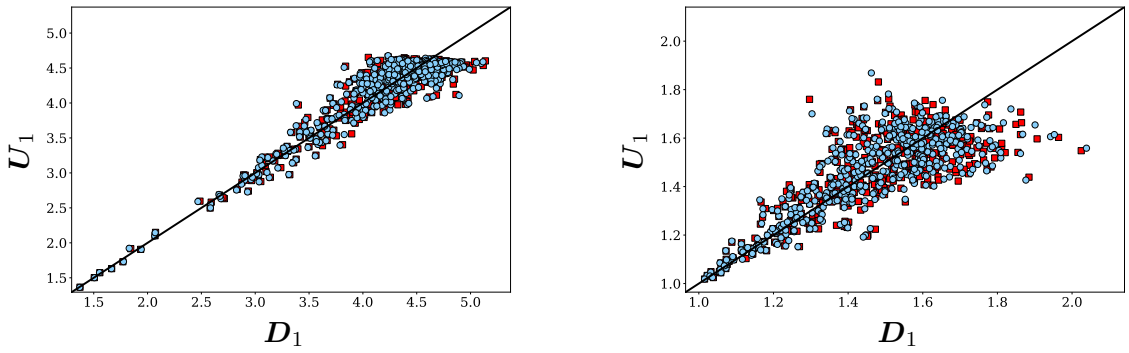


(a) Numerical simulation with low α.

(b) Numerical simulation with high α.

Figure 7: Numerical simulation of the scatter plot (correlation) between upstreamness and downstreamness $((\boldsymbol{U}_1)_i, (\boldsymbol{D}_1)_i)$ (blue circles), and the rank-1 estimations $(\tilde{U}_i, \tilde{D}_i)$ (red squares), for the simple random model with different values of α. In panel (a) $\alpha = 2.5$ and in panel (b) $\alpha = 10.5$, for both panels $N = 500$, $\alpha_F = 1.5$, $m = 1$, $m_F = 100$, and the thick black line has slope +1.

Similarly, we generate the random matrix A and random vector $\boldsymbol{F}$ for the sparse random model for different parameter values of p, when $p \to 1$ the sparsity of random matrix A reduces and when $p \to 0$ the sparsity of the random matrix A increases. Therefore, the probability density function to generate the random matrix A has the form given in Eq. (31), which is derived from the product of a Bernoulli random variable with parameter p (the sparsity) and a Pareto random variable, again with parameters α and m, the distribution of $\boldsymbol{F}$ is kept the same. Figure 8 displays the effect of the model when the sparsity is changed on the correlation between upstreamness and downstreamness. In Sub-Figure 8a where $p = 0.8$, such that the sparsity of the random matrix A is relatively low, does not correspond to a significant change in the models behavior compared to the non sparsity case with the same parameters in Sub-Figure 7a. This result is in line with the numerical approximations of the analytic solution, in which there is an insignificant change to the slope behavior when $p = 0.8$ compared to $p = 1$ as seen in Sub-Figure 5b. However, in Sub-Figure 8b where $p = 0.1$ the correlation of the model is less pronounced and has an increased variance around the line with slope +1, again this behaviour was predicted by the analytic approximations in Sub-Figure 5b, hence, these numerical simulations confirm the theoretic results obtain from the analytic model.

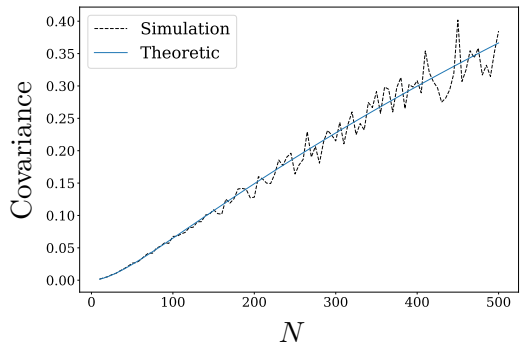


(a) Numerical simulation with low sparsity.

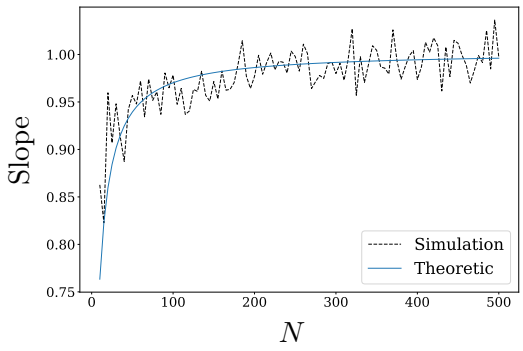
(b) Numerical simulation with high sparsity.

Figure 8: Numerical simulation of the scatter plot (correlation) between upstreamness and downstreamness $((\boldsymbol{U}_1)_i, (\boldsymbol{D}_1)_i)$ (blue circles), and the rank-1 estimations $(\tilde{U}_i, \tilde{D}_i)$ (red squares), for the sparse random model with different values of p. In panel (a) $p = 0.8$ and in panel (b) $p = 0.1$, for both panels $N = 500$, $\alpha = 2.5$, $\alpha_F = 1.5$, $m = 1$, $m_F = 100$, and the thick black line has slope $+1$.

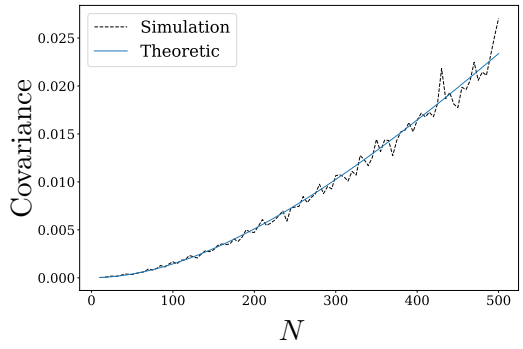
To further analyse the numerical simulations, we generate the rank-1 estimations of the covariance and slope functions, where the expectations in the definition of the rank-1 covariance and slope have been calculated over a batch of $M = 500$ simulations for each N. In Figure 9 it is evident that the theoretic curves are a good fit to the numerical simulations. Sub-Figures 9a and 9b show the theoretic fit to the numerical simulations in the non-sparse regime where $p = 1$, for both the covariance and slope the rank-1 numerical simulations are well approximated by the theoretic curves. Additionally, Sub-Figures 9c and 9d display that even in the sparse regime where $p = 0.1$ the theoretic solutions maintain a good fit to the numerical simulations for the rank-1 covariance and slope functions.



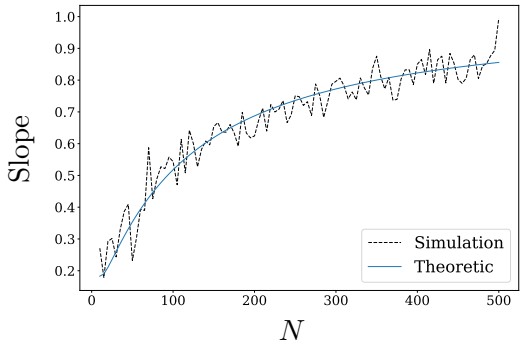
(a) Covariance: theoretic vs rank-1 simulation.



(b) Slope: theoretic vs rank-1 simulation.



(c) Sparse covariance: theoretic vs rank-1 simulation.



(d) Sparse slope: theoretic vs rank-1 simulation.

Figure 9: Numerical simulation of the rank-1 estimations averaged over $M = 500$ runs for the covariance and slope functions compared to the analytic approximations. In Sub-Figures (a) and (b) $p = 1$, in Sub-Figures (c) and (d) $p = 0.1$, in all Sub-Figures $\alpha = 2.5$, $\alpha_F = 1.5$, $m = 1$, and $m_F = 100$.

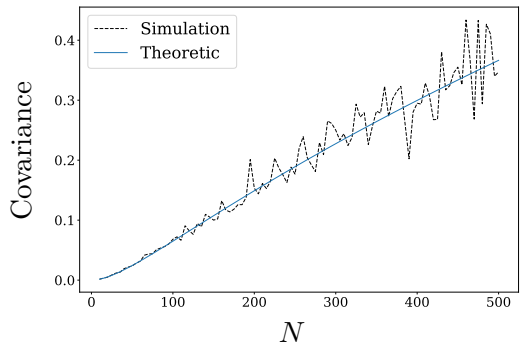
Moreover, similar analysis has been carried out for the actual covariance and slope defined using the 'true' measures of upstreamness and downstreamness that involve the full matrix inversion, as defined in Eqs. (9) and (17) respectively, where the covariance and slope have been calculated for the first element of $\boldsymbol{U}_1$ and $\boldsymbol{D}_1$ ($i = 1$). Hence, finding:

$$\mathrm{Cov}((\boldsymbol{U}_1)_1, (\boldsymbol{D}_1)_1) = \mathbb{E}[(\boldsymbol{U}_1)_1(\boldsymbol{D}_1)_1] - \mathbb{E}[(\boldsymbol{U}_1)_1]\mathbb{E}[(\boldsymbol{D}_1)_1], \tag{60}$$

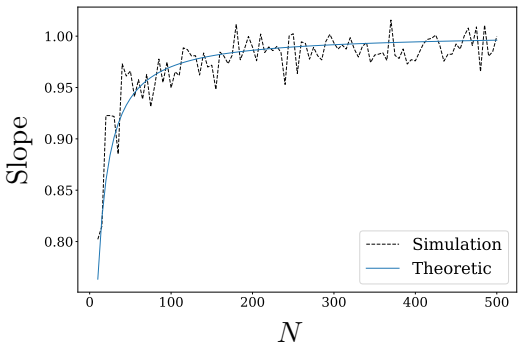
and

$$S = \frac{\mathrm{Cov}((\boldsymbol{U}_1)_1, (\boldsymbol{D}_1)_1)}{\mathrm{Var}[(\boldsymbol{U}_1)_1]}, \tag{61}$$

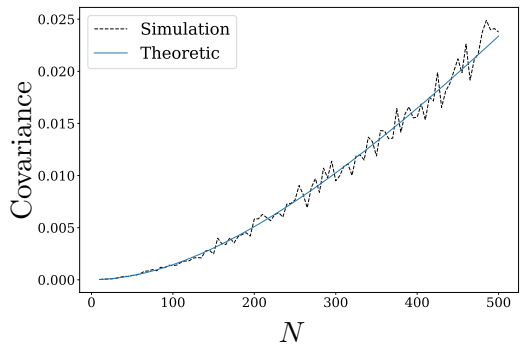
where the expectations were again found by averaging over $M = 500$ runs for each N. Figure 10 shows the results of the simulated actual covariance and slope, which again produces a good fit to the theoretical curves in all cases. In the non-sparse case in Sub-Figures 10a and 10b, where $p = 1$, the theoretic curves show a good fit to the simulated covariance and slope, however, the variance seems to be larger than the simulated rank-1 estimations particularly for the covariance for large N. In the sparse regime in Sub-Figures 10c and 10d, where $p = 0.1$, the theoretic curves are still a good fit to the simulated data, hence the theoretic rank-1 estimations provide an accurate prediction of the covariance and slope even when the matrix A is sparse.



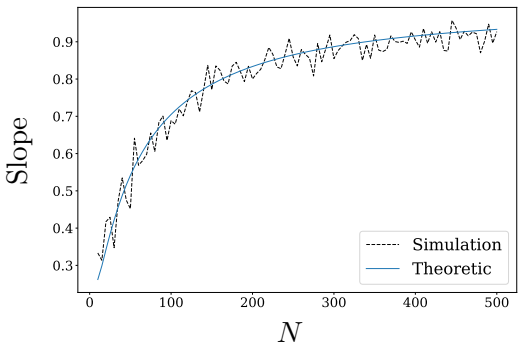
(a) Covariance: theoretic vs actual simulation.



(b) Slope: theoretic vs actual simulation.



(c) Sparse covariance: theoretic vs actual simulation.



(d) Sparse slope: theoretic vs actual simulation.

Figure 10: Numerical simulation of the actual upstreamness and downstreamness for the covariance and slope functions compared to the analytic approximations. Calculated for $i = 1$, such that the covariance and slope is taken with respect to the first element of $\boldsymbol{U}_1$ and $\boldsymbol{D}_1$, such that $\mathrm{Cov}((\boldsymbol{U}_1)_1, (\boldsymbol{D}_1)_1)$. In Sub-Figures (a) and (b) $p = 1$, in Sub-Figures (c) and (d) $p = 0.1$, in all Sub-Figures $\alpha = 2.5$, $\alpha_F = 1.5$, $m = 1$, and $m_F = 100$.

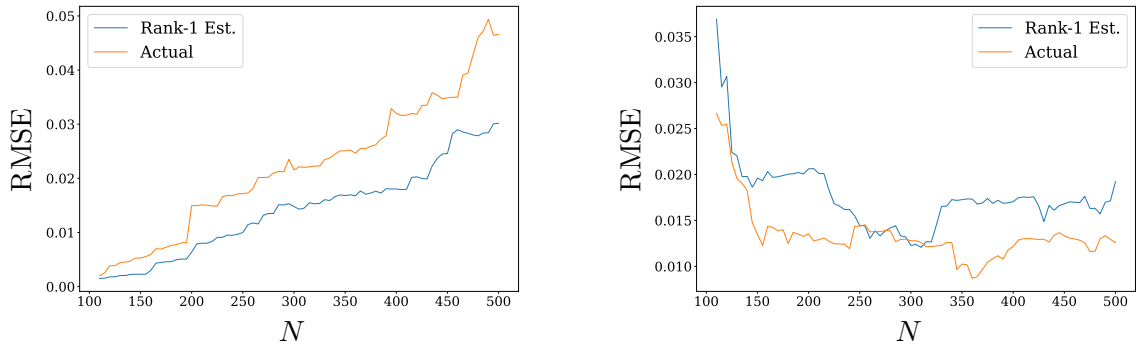
To further highlight these deviations of the simulated scenarios from the theoretic curves, a rolling root mean squared error (RMSE) has been used to give a better understanding of the accuracy of these measures as N increases. The rolling RMSE with a look-back parameter k for a given $n \in \{k, \ldots, N\}$ is defined as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{k} \sum_{i=n-k+1}^{n} (\hat{Y}_i - Y_i)^2}, \tag{62}$$

where $\hat{Y}_i$ is the predicted value of the theoretic covariance or slope at point at $i \in \{1, \ldots, N\}$, and Y_i is the simulated covariance or slope for either actual or rank-1 approximation at point $i \in \{1, \ldots, N\}$. Figure 11 shows the rolling RMSE for both the actual measures of the covariance and slope and the rank-1 estimations, in which they both display similar behaviour for increasing the system size N. In particular, Sub-Figure 11a shows that both the actual and rank-1 measures have an increasing RMSE for the covariance as N increases, thus the dispersion of the simulated covariance around the theoretic covariance increases and becomes less accurate. Moreover, this actual measure has a higher RMSE than the rank-1 estimations, suggesting that the actual measure of the covariance is not as well approximated by the theoretic covariance. Conversely,

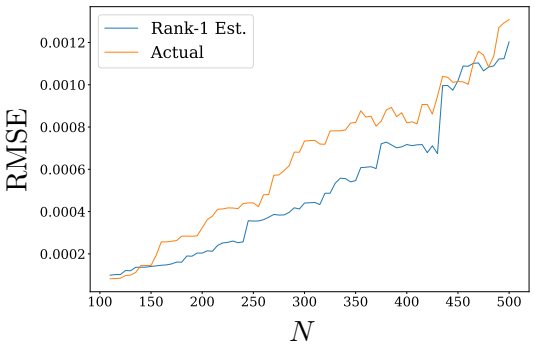
in Sub-Figure 11b for the slope, the actual measure is more accurate throughout the whole range of N than the rank-1 estimation and both become more accurate for larger N.

In the sparse case for the covariance in Sub-Figure 11c the difference between the actual and rank-1 estimation of the covariance is less pronounced. Furthermore, the RMSE is lower for both than in the non-sparse case, suggesting that the sparse regime has a more accurate theoretic curve than the non-sparse regime. Although, for the slope in Sub-Figure 11d the trend is less clear than in the non-sparse regime, with some gains in accuracy as N increases. Although, the overall value of the RMSE is higher in the sparse case than the non-sparse case suggesting that the theoretical slope is less accurate when the sparsity of the model is high.

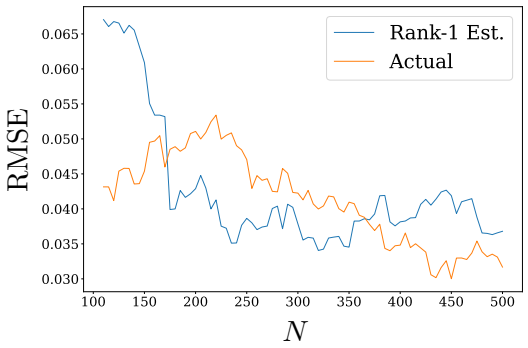


(a) Rolling RMSE of the covariance.

(b) Rolling RMSE of the slope.



(c) Rolling RMSE of the sparse covariance.



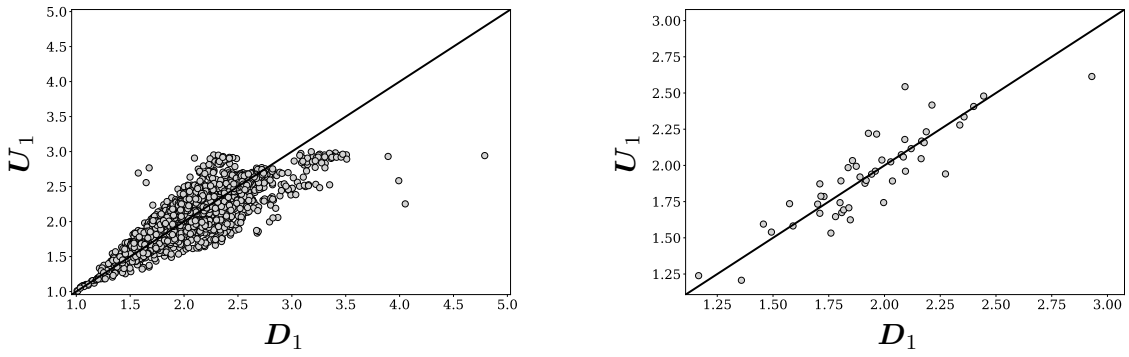
(d) Rolling RMSE of the sparse slope.

Figure 11: Numerical simulation of the actual upstreamness and downstreamness for the covariance and slope functions compared to the analytic approximations. Calculated for $i = 1$, such that the covariance and slope is taken with respect to the first element of $\boldsymbol{U}_1$ and $\boldsymbol{D}_1$, such that $\text{Cov}((\boldsymbol{U}_1)_1, (\boldsymbol{D}_1)_1)$. In Sub-Figures (a) and (b) $p = 1$, in Sub-Figures (c) and (d) $p = 0.1$, in all Sub-Figures $\alpha = 2.5$, $\alpha_F = 1.5$, $m = 1$, and $m_F = 100$.

The above analysis pertains to that of a closed economy framework composed of N different industries, to demonstrate that this random model applies to the wider literature on international trade such as in Antràs and Chor (2018), which first displayed this "puzzling" correlation between upstreamness and downstreamness at the country level. Thus, a block matrix representation of the input use matrix A has been developed. The random matrix A is split into M blocks of size $N \times N$, where each block represents an independent replica of the closed economy model, each with a set of i.i.d. randomly distributed parameter values, hence, the matrix A has

size $MN \times MN$. Each block is distributed according to the PDF in Eq. (31), where the parameters of block $i \in \{1, \ldots, M\}$ are uniformly distributed, $p_i \sim U(p_{\min}, p_{\max})$, $\alpha_i \sim U(\alpha_{\min}, \alpha_{\max})$ and $m_i \sim U(m_{\min}, m_{\max})$. Furthermore, the final demand vector $\boldsymbol{F}$ is composed of M sections of size N, hence the total size is MN, where each country section $i \in \{1, \ldots, M\}$ is Pareto distributed according to the PDF in Eq. (30) with uniformly distributed parameters $\alpha_{Fi} \sim U(\alpha_{F\min}, \alpha_{F\max})$ and $m_{Fi} \sim U(m_{F\min}, m_{F\max})$. The country level I-O matrix A_C is then defined as the block sum of the matrix A, and the country level final demand vector $\boldsymbol{F}_C$ is also defined as the block sum of $\boldsymbol{F}$.

Numerical simulation of this block random model are shown in Figure 12, where the country level upstreamness and downstreamness is defined as the aggregate sum of the industrial sectors for each country block $i \in \{1, \ldots, M\}$ of the input matrix A and final demand vector $\boldsymbol{F}$. It is evident from Figure 12 that the industry level correlation is more dispersed than the country level correlation, which was confirmed using a Pearson correlation coefficient where the industry correlation was given by 0.812 and the country correlation by 0.878 (both with p value < 0.01). Hence, the aggregation to the country level increases the correlation between upstreamness and downstreamness as the effect of outliers is less prominent in the calculation. Therefore, from the simulation in Figure 12 it can be concluded that the country level aggregation of upstreamness and downstreamness from a larger I-O matrix A that includes inter-country trade blocks, has very similar behaviour to that defined in the closed economy framework. This is likely due to the fact that the constraints of the country level matrix A_C are the same as for the single country matrix A, where each element is non-negative and the matrix is row-substochastic.



(a) Industry level correlation.

(b) Country level correlation.

Figure 12: Numerical simulation of the scatter plot (correlation) between upstreamness and downstreamness $((\boldsymbol{U}_1)_i, (\boldsymbol{D}_1)_i)$ for the block random model. Panel (a) shows the industry level correlation and panel (b) shows the country level correlation. The parameters are $M = 50$, $N = 50$, $p_{\min} = 0.01$, $p_{\max} = 0.99$, $\alpha_{\min} = 1.6$, $\alpha_{\max} = 4.6$, $\alpha_{F\min} = 1.2$, $\alpha_{F\max} = 3.2$, $m_{\min} = 0.001$, $m_{\max} = 1$, $m_{F\min} = 100$, $m_{F\max} = 1000$, and the thick black line has slope +1.

7 Discussion

To summarise, we have developed a random model of Input/Output analysis to analyse the "puzzling" correlation between upstreamness and downstreamness as presented in Antràs and

Chor (2018). To accomplish this analysis, we found empirically that the probability distribution of A and $\boldsymbol{F}$ have power-law tails with a relatively low cut-off value. Hence, we used a Pareto distribution to construct the random matrix A and random vector $\boldsymbol{F}$. Furthermore, to extend the analysis in Bartolucci et al. (2023) we further modified the matrix A such that each element was distributed as the product of a Bernoulli random variable with parameter p and a Pareto random variable, this allowed the sparsity of the matrix A to be changed through the parameter p. Although the random model did not have a closed form solution, it was shown by numerical approximation of the integrals, that the slope of the model approached $+1$ as N increased, increasing the number of sectors in the model for a given range of parameter values. In particular, the parameter p had the effect of decreasing the rate at which the slope approached $+1$ for increasing N. Therefore, this indicates that for large N the result that the slope will approach $+1$ when $\alpha > \alpha_F$ and $m < m_F$ is robust even for a very sparse ($p = 0.1$) matrix A.

Furthermore, the limiting case of $N \to \infty$ for p of $\mathcal{O}(1/N)$ was also investigated using numerical approximations, which found that for a very sparse regime (approximately $\nu < 100$) the limiting slope was less than $+1$. Thus, the effect of sparsity on the slope is only significant when $p << 1$ such that the sparsity of the matrix A is very high and increases with system size N. Additionally, the relative difference between the parameters α and α_F had a strong effect on the behavior of the slope, as the difference increased where $\alpha > \alpha_F$ the value of the slope was closer to $+1$ for smaller N. This is in line with the results from Bartolucci et al. (2023) that found for an exponential PDF the slope $S = +1$ for any N, because as α increases the behavior of the Pareto distribution becomes more like an exponential distribution as the tail becomes smaller.

The results of the analytic random model were tested against numerical simulations, which found that the analytic model accurately captured the behaviour of the covariance and slope for increasing N, for both the simulated rank-1 estimations, and the actual measures of upstreamness and downstreamness defined in terms of the more complex matrix inversion. Therefore, the results interpreted from the analytic model can be generalised to the actual measures of the covariance and slope, suggesting that as N increases the correlation between upstreamness and downstreamness will approach $+1$ for p of $\mathcal{O}(1)$ and $\mathcal{O}(1/N)$ for given threshold of sparsity (approximately $\nu > 100$).

Furthermore, for uniformly distributed parameter values, it was found by numerical simulation that in the case of an open economy international trade model (constructed from independent block replicas of the closed economy model on a large input use block matrix A and final demand vector $\boldsymbol{F}$), the aggregation of each country into a country level matrix A_C and final demand vector $\boldsymbol{F}_C$ had very similar behaviour for numerical simulations to that of the closed economy model. This is likely due to the fact that the two constraints on A also applies to A_C, where each element is non-negative and the matrix is row-substochastic. Hence, due to this similarity in structure and behaviour between the closed economy model and the country level model, it infers that the results are applicable to the international trade setting considered in Antràs and Chor (2018) and provides an explanation to this "puzzling" correlation found at the country level between upstreamness and downstreamness.

Further extensions to be made in future studies would be to compute analytically the actual covariance in Eq. (22) of the sparse random model for different probability distributions of the I-O matrix A. Moreover, it would also be interesting to investigate further the country level covariance defined from a larger international I-O matrix A and the effects of different types of heterogeneity in the distribution of the parameters of the model, extending the analysis beyond uniformly distributed parameters.

8 Conclusion

We have developed a sparse Pareto distributed random model to analyse the correlation between upstreamness and downstreamness for a closed economy, in which we found that the slope of the correlation S approaches $+1$ as the number of sectors N increases for a range of parameter values, and was robust for high sparsity when N is large. Furthermore, this analysis showed that, for a sufficiently non-sparse matrix, which increases in sparsity with p of $\mathcal{O}(1/N)$, the slope will also approach $+1$ in the limiting case $N \to \infty$. These results were also confirmed using numerical simulations of the random model, where it was further demonstrated that the results from the closed economy framework can also be applied to the level of international trade between countries.

References

Antràs, P. and Chor, D. (2012) 'Organizing the global value chain', *Econometrica*, 81(6), 2127–2204.

Antràs, P. and Chor, D. (2018) 'On the measurement of upstreamness and downstreamness in global value chains', *World Trade Evolution: Growth, Productivity and Employment*, 5, 126–194.

Antràs, P., Chor, D., Fally, T., and Hillberry, R. (2012) 'Measuring the upstreamness of production and trade flows', *American Economic Review*, 102(3), 412–416.

Bartolucci, S., Caccioli, F., Caravelli, F., and Vivo, P. (2020a) 'Inversion-free leontief inverse: Statistical regularities in input-output analysis from partial information', *arXiv preprint arXiv:2009.06350*.

Bartolucci, S., Caccioli, F., Caravelli, F., and Vivo, P. (2020b) 'Ranking influential nodes in networks from partial information', *arXiv preprint arXiv:2009.06307*.

Bartolucci, S., Caccioli, F., Caravelli, F., and Vivo, P. (2023) 'Correlation between upstreamness and downstreamness in random global value chains', *arXiv preprint arXiv:2303.06603*.

Fally, T. (2012) 'Production staging: measurement and facts', *Boulder, Colorado, University of Colorado Boulder, May*, pp. 155–168.

Leamer, E. E. (1995) 'The heckscher-ohlin model in theory and practice', .

Leontief, W. (1986) *Input-Output Economics*, Oxford University Press.

Leontief, W. W. (1936) 'Quantitative input and output relations in the economic systems of the united states', *The Review of Economic Statistics*, pp. 105–125.

Miller, R. E. and Temurshoev, U. (2015) 'Output upstreamness and input downstreamness of industries/countries in world production', *International regional science review*, 40(5), 443–475.

Shaikh, A. (1974) 'Laws of production and laws of algebra: The humbug production function', *The Review of Economics and Statistics*, pp. 115–120.

Timmer, M. P., Dietzenbacher, E., Los, B., Stehrer, R., and De Vries, G. J. (2015) 'An illustrated user guide to the world input-output database: The case of global automotive production', *Review of International Economics*, 23(3), 575–605.

A Report Code

The code for the project was written in Python, which can be found in the GitHub repository upstreamness-downstreamness by following the link: `https://github.com/DylanTerryDoyle/upstreamness-downstreamness`.