STOCKHOLM SCHOOL OF ECONOMICS
Department of Economics
5350 Master's Thesis in Economics
Academic Year 2017–2018

The Dynamics of Sectoral Network Formation

Tadas Gedminas (41061)

Abstract: Recent studies have highlighted the role that linkages between firms and sectors play in impacting shock transmission, R&D development and expansion of trade relationships. However, mechanisms that govern how these linkages form and develop over time are not fully understood. The present work tackles this issue and aims at characterising the international sectoral network formation using a random graph model. The proposed framework incorporates elements of input-output structure, geographical variation of linkages, and path dependence in new link formation. The developed model exhibits properties such as: (i) the distribution of firms and sectors by number of linkages is skewed and fat-tailed; (ii) the average distance of connections increases with the higher number of outward linkages; (iii) the likelihood of new link formation depends on network proximity and geographical distances between sectors. These implications are then tested in an empirical setting using the World Input-Output Database (WIOD). Empirical analysis finds supporting evidence for (i) and (iii), whereas evidence for (ii) is inconclusive.
Keywords: random graph model, WIOD, input-output network
JEL: D85, C67, D57, F14

Supervisor: Mark Sanctuary
Date submitted: 13.05.2018
Date examined: 30.05.2018
Discussant: Julia Baumann and Petra Somnell
Examiner: Maria Perrotta Berlin

Acknowledgements

I would like to thank my supervisors Mark Sanctuary of Stockholm School of Economics and professor Guido Cozzi of University of St. Gallen, for their guidance and insightful feedback. I would also like to thank program director, Anna Dreber of Stockholm School of Economics, for always providing help and assistance when in need. Finally, I am grateful to my family and friends who encouraged and supported me throughout the writing process. Thank you.

Contents

List of Figures

List of Tables

1 Introduction

In the past, the structure of relationships between firms and sectors was not considered to be relevant for studying aggregate phenomena. While this may have been due to lack of necessary modelling tools or access to micro-level data, conceptual arguments were also dismissive (Dupor, 1999). However, recent literature has contested this position and argued that the microstructure of the economy is important. For example, seminal work by Acemoglu et al. (2012) highlighted how an unbalanced sectoral network may lead to higher macroeconomic volatility. Other studies have shown how these structures impact R&D development (Acemoglu et al., 2016a), the transmission of monetary policy (Ozdagli and Weber, 2017) and the synchronisation of business cycle co-movements across countries (di Giovanni et al., 2018).

Although evidence in favour of the importance of sectoral networks has grown in recent years, a complete framework which would characterise how these structures emerge and change over time has not been developed. This has been a challenging task given that existing approaches have been unable to capture empirical facts such as that the distribution of linkages is skewed and fat-tailed (Bernard and Moxnes, 2018), or that the geographical expansion of linkages depends on existing relationships (Chaney, 2014; Morales et al., 2017).

Thus, the contribution of this work is to develop a framework which accounts for these empirical features and also informs about the dynamics of new linkage formation. The proposed model incorporates input-output relationships between firms and includes a geographic dimension. This allows applying the framework to an international sectoral network setting. The resulting model produces a distribution of linkages which qualitatively matches empirical observations and also exhibits path dependence in how new relationships form. In particular, the derived model predicts that as more outward linkages are established, they become more geographically spread out. Some of these elements have been captured in previous studies (Chaney, 2014). However, a contribution of the present work is that the results are derived in an input-output setting. This allows extending implications of the model from firm to sectoral level. Furthermore, the framework highlights how the likelihood of new linkage formation is dependent on both network proximity and geographical distances. Conceptually, the two mechanisms that drive these results are that firms are more likely to find compatible production inputs if their existing suppliers use them and that firms use their existing set of relationships to establish new connections in different locations.

After deriving key results from the model, we bring these aspects to a novel empirical setting using the World Input-Output Database (WIOD). Given that a framework which would allow for input-output relationships and geographic variation has been lacking in the literature, applications of the WIOD for studying network formation have been limited. In the empirical part of the study we first show that, in line with theoretical predictions and previously studied contexts, the WIOD features a skewed and fat-tailed distribution of linkages. Furthermore, the geographic variation of these linkages is also in line with predictions of the model. Thereafter, we study the likelihood of new linkage formation by estimating a probability model and show that the likelihood of new linkage formation increases with closer network proximity, shorter geographic distances and the higher number

of existing connections. On the other hand, results relating to the theoretical prediction that geographic distances become easier to overcome as sectors establish more connections are mixed and inconclusive.

The present study relates most closely to the works of Chaney (2014) and Carvalho and Voigtländer (2014). Chaney (2014) introduced a network-based approach in explaining how firms search for export destinations. The proposed model distinguished local and remote search processes and explained the geographic distribution of French firm export destinations. In line with the present work, the model captures the fact that geographic expansion is path dependent. On the other hand, Carvalho and Voigtländer (2014) related the emergence of network structure to firms' choice of intermediate inputs. Specifically, a pre-existing input-output relationship between sectors could determine the search of new potentially useful production inputs. The model predicts that network proximity between sectors impacts further linkage formation. These studies make important contributions in explaining formation of linkages between firms and sectors, however, the scope of both papers limits their application to an international sectoral network setting. In Chaney (2014) the network formation is limited to the search of customers and is silent on the use of inputs, whereas Carvalho and Voigtländer (2014) are focused only on national input-output relationships. Both of these concerns are addressed in the present study.

The remainder of this work is structured in the following way. Section 2 discusses literature relevant for this study and highlights the existing research gaps. Section 3 describes and develops the theoretical framework. Section 4 presents the descriptive features of the WIOD and describes the empirical strategy for testing predictions of how new linkages between sectors form. Section 5 presents the empirical results and discusses limitations of the study. Section 6 concludes the study.

2 Literature Review

Early research that studied the impact of microstructure focused on whether micro-level shocks could explain macroeconomic fluctuations. The prevailing sentiment from this body of work was that the effect is negligible or can generally be accounted for. For instance, Lucas (1977) suggested that it is possible to apply the law of large numbers argument given the high number of firms in the economy. This would imply that, in expectation, idiosyncratic firm level shocks average out and studying these shocks would not help in explaining business cycle fluctuations. From a different point of view, Hulten (1978) laid out theoretical justifications for attributing micro-level shock significance to the direct share of output. Similarly, Dupor (1999) derived theorems and conditions under which the input-output structure of the economy would be irrelevant for the transmission of sectoral level shocks to aggregate fluctuations.

Recent work, however, has begun to challenge these ideas. Seminal work by Gabaix (2011) proposed a "granular hypothesis" for aggregate fluctuations. The author showed that if the firm size distribution is fat-tailed, micro-level shocks originating in large firms could not be absorbed by the rest of the economy and hence could translate to aggregate fluctuations. Alternatively, Acemoglu et al. (2012) showed that, among other things, output volatility may be impacted by the network structure of input linkages between sectors. More recently, Baqaee and Farhi (2017) suggested that first-order approximations, which were essential for Hulten's (1978) results, are not appropriate if non-linearities exist. Instead, aspects such as structural elasticities of substitution, network linkages, structural returns to scale, and degree of factor reallocation are missed when using first-order approximations. In an international setting, di Giovanni et al. (2018) showed how linkages between firms can explain business cycle co-movements between countries. Given that large firms tend to be more engaged in cross-border trade, shocks originating abroad can still reach the rest of the economy via indirect linkages. While these studies contribute to increasing evidence that microstructure is important, mechanisms behind how these structures emerge remain understudied.

The remainder of the literature review is structured into four parts. The first part covers evidence of microstructure having an impact on the macroeconomy, with the focus on input-output linkages. The second part discusses the role of microstructure in an international setting. The third part highlights key approaches in existing modelling frameworks. The final part discusses empirical methods used in analysing input-output linkages and previous works that have used the WIOD.

2.1 Importance of Microstructure

Initial attempts to account for granular features of the economy were multi-sector general equilibrium models. These models were mainly used to explain co-movements across sectors. For example, Long and Plosser (1983) developed a multi-sector real business cycle model, where shocks originating in one sector are allowed to propagate to other sectors via production input relationships. Horvath (1998) extended the model and showed that the degree

to which sectoral shocks translate to aggregate volatility is not due to the total number of sectors (or firms), but rather to the way they are related. Crucially, if sectors are not connected, the shocks may have a much more sizeable impact.

Empirical work measuring and verifying these effects has been more sparse. Among the studies that have explored these ideas, Conley and Dupor (2003) estimated whether economic distances, measured by input-output relationships, could explain the comovements in productivity between sectors. The authors found significant and positive covariance between sectoral total factor productivity growth and the effect was stronger for sectors that had similar sets of inputs. Also, work by Foerster et al. (2011) showed that accounting for sectoral linkages can explain half of the variation in industrial production during the Great Moderation, whereas variation in large sectors is not as significant. Similarly, Carvalho and Gabaix (2013) studied whether sectoral volatility could help explain aggregate output volatility. The authors calculated a measure referred to as "fundamental volatility" which weights sectoral volatility by share of total output. They found that this measure has high explanatory power and could also provide an explanation for the Great Moderation.

In contrast to these works, recent research has argued that the scope of these effects can be traced back to individual firm behaviour. For example, Gabaix's (2011) work showed that due to fat-tailed firm size distributions, shocks to large firms could have a non-negligible impact on aggregate output. In support of this hypothesis, Carvalho and Grassi (2015) extended the framework. The model derived by the authors incorporates empirically observed features such as aggregate output and productivity persistence, and that volatility is time-varying. The primary driver of these effects are shocks originating in large firms, in line with Gabaix (2011). On the other hand, Grassi (2017) highlighted a conceptual issue with the "granular hypothesis". Specifically, the hypothesis states that large firms can contribute to aggregate fluctuations, but at the same time are unable to impact equilibrium prices and quantities. To account for this, Grassi (2017) introduced a framework where firms engage in oligopolistic competition and respond to productivity shocks strategically, depending on their market power.

It has been argued, however, that the importance of the microstructure is not limited to firm size distributions and instead the structure of linkages between firms may play a more important role (Gabaix, 2016). For example, di Giovanni et al. (2014) analysed whether aggregate fluctuations can be attributed to firm level shocks. The authors were able to contrast the impact of fat-tailed firm size distribution and linkage effect and found that linkages are approximately three times as important in driving aggregate fluctuations.

An overview of the literature behind the role that production networks play in impacting the rest of the economy is summarised in Carvalho (2014). In one of the most influential works within this literature, Acemoglu et al. (2012) developed a formal model to characterise the relevance of input-output linkages in propagating micro-level shocks. The authors suggest that it is not the sparseness of the sectoral networks, but rather the asymmetry in the role that sectors play that results in sectoral shocks leading to higher macroeconomic variation. Alternatively, Acemoglu et al. (2016a) documented the presence of network effect in R&D

development. By studying patent citation network, the authors found that technological progress by the upstream firms is a strong predictor of downstream innovation. On the other hand, Ozdagli and Weber (2017) linked the network structure of firm relationships to the transmission of monetary policy shocks. The authors found evidence that network linkages may have a significant second-order effect on individual firms stocks depending on how firms are linked to each other.

Within growth literature, there has also been an acknowledgement of the role that microstructure may play in explaining differences between countries. Ciccone (2002) showed that the type of industrial technologies that countries adopt could result in sizeable differences in productivity levels and aggregate income. The primary driver of this result was whether the adopted industrial technologies were intermediate input intensive. In line with this argument, Jones (2011) used intermediate input production chains as a potential explanation for income differences across countries. The differences emerge from the insufficient use of intermediate inputs or single sector inefficiencies that propagate through the input-output network and that drag the rest of the economy.

While the above described body of work highlights the relevance of structure of linkages, an important issue that has not been sufficiently addressed in the literature is the process behind the formation of new linkages (Acemoglu et al., 2012; Carvalho, 2014). Existing studies have mostly taken the network structure as given and the impact was studied by changing the structure exogenously. Some notable exceptions that have tried to address this issue are Carvalho and Voigtländer (2014), Lim (2017), and Oberfield (2018). In particular, Carvalho and Voigtländer (2014) applied a random graph model, closely related to Jackson and Rogers (2007), to study the network structure of input-output linkages. The proposed model differentiated between essential and variety inputs and the main theoretical predictions of the model were tested using U.S. sectoral data. The study showed that network proximity between sectors is relevant for predicting new link formation.

2.2 Microstructure in International Trade

Independently from developments in macroeconomics, international trade literature has also turned focus to microstructure. This is most evident in recent research which highlights firm heterogeneity in explaining trade flows. For example, Bernard et al. (2003) developed an extended version of Ricardian trade model to account for qualitative facts of U.S. firm trade. The model allows for many countries, geographic barriers and imperfect competition, and it captures aspects such as higher productivity among exporters, the small fraction of firms which export and that among exporters, the share of exports to total output is not large. A significant contribution of this work is that it introduced firm level trade models. Helpman et al. (2008) extended Melitz (2003) model to allow decomposing the impact of trade on the intensive and extensive margins. The authors argued that since previous estimations of gravity trade models ignored countries that do not trade, it may have led to systematically biased estimation. Among the findings in the paper, the authors showed that the growth in trade between 1970 and 1990 was mostly in the intensive margin. On the other hand, Eaton

et al. (2011) documented empirical regularities which were not accounted for by traditional models. The paper highlighted the importance of firm efficiency as the key attribute in explaining variation across firms in market entry. Armenter and Koren (2015) showed that a basic Melitz trade model does not quantitatively match empirical data regarding export size and frequency. The authors suggested that a different source of firm heterogeneity needs to be included in the model to match observed data.

One potential source of firm heterogeneity that has been proposed to explain these empirical observations is informational frictions. Allen (2014) used data on regional agricultural prices in the Philippines and documented the role of information frictions in impacting price dispersion. The study found that approximately half of the price dispersion was due to information frictions. Study by Albornoz et al. (2012) focused on Argentinian firms, which when searching for new trade location gave up exporting very shortly, even in the presence of significant entry costs. However, other exporters increased their foreign sales and expanded to new destinations in relation to the original entry market. The paper proposed a mechanism where, for an individual exporter, profitability in a given location is initially uncertain. On the other hand, if firms do decide to enter the market, they may learn more precisely whether their expansion is profitable and then choose to either extend their trade in that location or expand to similar destinations. This leads to what the authors refer to as "sequential exporting", where the possibility of profitable expansion makes the initial entry cost worthwhile, even in the presence of high failure rates. The authors verified that the results were not driven by firm heterogeneity, country-specific shocks, credit constrains or impact of rivals.

In studying how Chinese firms expanded to new export destinations after removal of trade restrictions, Defever et al. (2015) also found presence of path dependence in how firm chose new destinations. The spatial correlation is higher than what would be expected from a standard gravity trade model and further provides evidence that trade destination choices are not independent. Morales et al. (2017) proposed a model to combine these mechanisms with conventional gravity trade models. The authors refer to the mechanism as "extended gravity", where further expansion is dependent on existing destination. The implications of the model were tested empirically, where the extended gravity effect was attributed to four factors: whether new export destinations with regards to existing ones shared a border, were in the same continent, spoke the same language and had similar income per capita levels. The authors found that entry sunk costs are significantly lower in the presence of at least one of these factors. While recent empirical work has emerged to highlight limitations of existing heterogeneous firm trade models, information frictions arguments have been raised in the past . For example, Rauch (1999) found evidence that for trade in differentiated products, proximity and historical ties can help explain observed trade patterns. The author proposed that network-search type models may be required to give theoretical justification for these observations.

More explicitly, business and social networks have been suggested as potential mechanisms that could help firms overcome informational frictions (Rauch, 2001; Chaney, 2016).

Garmendia et al. (2012) study these aspects in explaining intranational trade in Spain. The paper empirically showed the existence of regional home bias in trade, which disappears after controlling for social and business networks. In one of the most relevant papers for the present work, Chaney (2014) developed and an illustrated application of random graph model in international trade setting. The starting point of the modelling framework was the idea that firms face information barriers and thus are unable to identify all of the potentially profitable markets. Instead, firms learn about new locations to trade by performing a local and remote search. Remote search is modelled as preferential network search, where firms use their connections in locations abroad to conduct search of new markets. After developing the model, the study showed that firm connections and geographic distributions of these connections implied by the model matched data of French firms.

With regards to previously discussed research in macroeconomics, there is also recent literature on elements of the microstructure amplifying cross-border spillovers. Burstein et al. (2008) studied and documented the synchronisation of business cycles between countries that are engaged in trade along supply chains. The key mechanism behind the extent of this synchronisation was the elasticity of substitution when choosing intermediate inputs. Alternatively, di Giovanni et al. (2018) empirically analysed whether French firm international trade linkages can explain the correlation between business cycle co-movements. The authors observed that at least two-thirds of these co-movements could be attributed to direct and indirect linkages. The direct linkage effect primarily came from larger firms, which were more likely to have international connections and accounted for at least half of France's value added.

2.3 Network Structure Modelling

First approaches that captured the structure of sectoral linkages were previously discussed multi-sector general equilibrium models. In this regard, a key foundational paper which incorporated this approach is Long and Plosser (1983). This paper introduced sectoral level granularity in a real business cycle framework. In turn, other works continued to use this framework as the basis for further study of sectoral linkages with the most notable example being Acemoglu et al. (2012). Similarly, Atalay (2017) developed a multi-sectoral general equilibrium model, which highlighted the role of elasticity of input substitution as the principal mechanism behind the extent of sectoral shock propagation. A recent example by Pasten et al. (2018) developed a multi-sectoral model which allows for heterogeneity in sectors by price stickiness. This model, which is more in line with prevailing New Keynesian approaches in macroeconomics, shows that the degree of idiosyncratic shock propagation is much more related to price stickiness rather than sector size effects or asymmetries in the input-output structure. While these approaches are most closely related to traditional macroeconomic models, they are silent on how the structure of linkages between sectors emerges and changes over time.

Alternative approaches, on the other hand, have tried capturing granularity by modelling individual firm behaviour. This has been more common in international trade literature,

whereas applications in macroeconomics have been more limited. Some recent examples include Chaney (2018) who provided a microfounded explanation behind the gravity trade equation. The study focused on the distance coefficient and proposed a model where assuming Pareto firm size distribution and that larger firms are more likely to engage in exports of further distances could explain the empirical estimates of the distance coefficient. An exception in macroeconomics, however, is di Giovanni et al. (2014). Instead of basing the model on previously mentioned general equilibrium models, the paper used approaches more commonly applied in international trade (Melitz, 2003; Eaton et al., 2011). This gave a more micro-founded explanation behind aggregate fluctuations and allowed to decompose firm level effects, sectoral level shocks and the importance of linkages between firms. However, while in contrast to multi-sectoral models these approaches capture more granular elements, they still miss out how the structure of linkages between firms and sectors forms.

More rigorous treatment of endogenising production network formation has been tackled by Oberfield (2018) and Lim (2017). Oberfield (2018) developed a model where entrepreneurs have an option to produce a good in many ways and hence make the choice of input decision by cost-optimisation. This feature endogenously produces outcomes where "star" suppliers emerge - highly productive input producers that are widely adopted. Lim (2017) developed a model which relaxed Oberfield's (2018) assumption that production recipes are available to entrepreneurs exogenously. However, the model still features rich structure where firms are allowed to differ in the way they are connected to other firms, the role they play in the supply chain, and they are dynamically looking and changing their partners. To the best of my knowledge, these two works develop the richest frameworks for studying firm network formation. However, the richness of these models comes with the trade-off and introduces challenges for testing implications of these models empirically.

An alternative approach that has been proposed in the study of the dynamics of firm and sector network formation is to incorporate models from the broader field of network literature. One of the first works to integrate these models in production network analysis was Atalay et al. (2011). The paper proposed a preferential attachment network formation model, where new links are more likely to form between nodes (sectors) that are already highly connected. The authors showed that this framework captured the distributions of sector linkages more accurately, especially at the tails of the distribution. Alternatively, Carvalho and Voigtländer (2014) incorporated models from Jackson and Rogers (2007) in the study of production network formation. By distinguishing essential inputs and variety inputs for production, the authors showed how implications for the dynamics of network formation could be extended from firm to sectoral level. Afterwards, the authors present supporting empirical evidence for the proposed mechanism. In international trade literature, Chaney's (2014) work is closest to this approach. By incorporating preferential attachment in the model and by distinguishing between a local and remote search of customers, the paper develops a model for studying the geographic variation of customers. While these types of models are appealing due to their tractability and ability to match qualitative features of distributions of linkages, the mechanisms in these models are more mechanical and miss out

on optimising and equilibrium behaviour.

2.4 Empirical Study of Network Structure

Empirical approaches that have been used to study network structure both in macroeconomic and international trade literature often depend on the modelling framework that has been used. In particular, highly structured multi-sectoral general equilibrium models are matched empirically by calibration and are then used to study how much variation in the data can be attributed to various shocks and elements of the microstructure. A notable exception to this is Acemoglu et al. (2016b). In this paper, the authors contrasted the impact of supply and demand side shocks along the supply chain. In line with theoretical predictions, the authors found that demand-side shocks propagated upstream, and supply-side shocks downstream of the supply chain.

A more recent and novel approach to the empirical study of the importance of input-output linkages has been to exploit exogenous variation in supply chains. Barrot and Sauvagnat (2016), captured variation in natural disasters in the U.S. to measure how idiosyncratic shocks propagate through production networks. Similarly, Carvalho et al. (2016) studied the aftermath of the Great East Japan Earthquake of 2011. In their study, the authors used exogenous variation in the spatial dimension and studied how shocks originating in earthquake struck regions propagated upstream and downstream through the rest of supply chain. The authors found that the shock propagation effect could account for a 1.2 percentage point decline in Japan's gross output.

Focusing on the literature that is relevant to the present study and that has studied which factors influence the likelihood of new link formation, Carvalho and Voigtländer (2014) is a notable example. The authors constructed and estimated a probability model for predicting input adoptions. The study found that closer network proximity in input-output relationships positively impacts the likelihood of input adoption. In addition, the authors further studied whether closer proximity could influence the time it takes for an input to be adopted and found supporting evidence of the effect. In international trade literature, Chaney (2014) used a probability model to estimate the likelihood of entering new export destinations. The empirical model was used for motivating the modelling framework and the estimation technique was similar to Carvalho and Voigtländer (2014).

Concerning the data used in present work, most common application of the WIOD have been on the extent of country engagement in global value chains. For example, Timmer et al. (2014) highlighted a number of qualitative facts about global value chains. Namely, there has been a tendency of increasing fragmentation of production, with growing share of foreign value added being part of the production. Furthermore, there is a growing trend of use of capital and high-skilled labour in production, especially in developed economies. This aspect has also been documented in Los et al. (2015), highlighting that regional fragmentation has not increased as much as global fragmentation. In an attempt to provide an example case for application of the WIOD, Timmer et al. (2015) presented how the automotive production has chained over time. The focus of these and related studies have mostly been on the intensive

margin of trade. In contrast, in the present work, we focus on the extensive margin of trade - binary relationships between sectors.

In concluding the literature review, we focus this study in exploring an existing gap within literature and studying how international sectoral linkages form. Thus, the research questions that we aim to answer is: *What determines the likelihood of new network linkage formation between sectors?* To answer this question we first develop a model that incorporates input-output relationships and includes a geographic dimension. After developing the framework and showing key implications of the model, we test them in an empirical setting using the WIOD.

3 Theoretical Framework

The theoretical framework is based on the works of Jackson and Rogers (2007), Chaney (2014) and Carvalho and Voigtländer (2014). From Carvalho and Voigtländer (2014) we implement ideas of distinguishing different types of inputs. The spatial dimension is included following Chaney (2014). Finally, derivations of the key properties of the model are related to Jackson and Rogers (2007). The aim of the theoretical framework is to capture empirical features that characterise the sectoral network structure in an international input-output setting found in previous works and the WIOD dataset. In the first part of the modelling framework, we highlight implications of the model at the firm level and then show which of the properties are preserved when aggregating to sectoral level.

3.1 Initial Setup

The model is constructed in a discrete time setting. We define the set I_t as the set of all firms (varieties) that exist in period t. Individual elements i of the set I_t correspond to individual firms. We assume that the set is finite, discrete, and consists of n_t number of firms. Each period the number of firms grows at an exogenous rate γ. Firms are differentiated by their choice of input sets. When a new firm is born it draws a subset of firms $K_i \subset I_t$ that become essential input providers. This set is drawn randomly and uniformly from the set of all firms, I_t. Denote the size of this set m_K. Essential inputs can be thoughts as production inputs without which the firm would not be able to operate.

Alternatively, firm's search for variety inputs. These inputs can be thought as being used for differentiating the product. Instead of searching for variety inputs randomly, the firm considers its network neighbourhood of essential input providers. Specifically, candidates for variety inputs are firms that source at least one of the firm in the set K_i as input. This feature aims at capturing the fact that when a firm searches for non-essential inputs, it is more likely to find technologically compatible input if it is used by one of its existing suppliers. From this set, the firm chooses variety input set N_i, which has size m_N[1].

Next, we introduce the spatial dimension to the model. We define location space by set C, which consist of discrete elements c representing individual locations. Intuitively, the set C can be thought as representing the set of countries. Note that within each location the same firm set I_t exists. In other words, if we take an individual firm i within the set and if it is defined by some essential input set K_i, an identical firm exists in all other locations. While this is a strict symmetry assumption, firms that are defined by the same set of inputs will differ across locations in the number of customers they have (firms that use them as inputs). Alternatively, this may lead to situations where a K_i type firm becomes dominant within a location, while in other locations the same type of firm has only a few customers. Another interpretation of this setup would be to think of I_t as the set of production recipes which are

[1] A key difference from present work and Jackson and Rogers (2007) is that we abstract from modelling probability of adoption as we do not endogenize this process. In Carvalho and Voigtländer (2014) for example, the probability of adopting an essential input was assumed to be equal to 1 and probability of adopting a variety input was endogenized to be between 0 and 1.

uniformly available to entrepreneurs across all locations. However, over time the utilisation of these recipes may produce asymmetries across locations. In addition, given that the set I_t grows at an exogenous rate γ it is also assumed to be symmetric across all locations

Having defined firm and location sets separately, we next describe how they link within the model. First of all, when a firm draws its set of essential inputs, it independently draws the location of the input provider. For example, if a firm is born in location c_0, when it draws one of the essential input providers from set K_i, it also draws the location of the input provider c. For this we introduce a probability function defined over the whole location set C, which corresponds to the likelihood of a search originating in location c_0 drawing a match in c: $g(c_0, c)$. In this setting we further assume that the location set C is equal to the integer set $C = \mathbb{Z}^2$. This way a distance measure can be defined as the absolute difference between the two locations: $|c_0 - c|$. Hence, the probability function can be expressed as: $g(|c_0 - c|)$.

Note that the probability function is also symmetrical $g(c_0, c) = g(c, c_0)$, since $|c_0 - c| = |c - c_0|$. For the remainder of the framework we impose two additional assumptions regarding properties of the spatial matching probability function $g(.,.)$: (i) probability of a match is decreasing in distance, and (ii) the distribution has a finite second moment. As it will be shown later, these are sufficient assumptions to derive key properties related to the geographical distribution of linkages between firms.

3.2 Dynamics of Customer Acquisition

To proceed further, we consider how an individual firm's customers change over time. We define the total number of customers that a firm i has at time t as $d_i(t)$. In network literature, this would correspond to the outdegree of a firm. We further define $f_{i,t}(c)$, which is the number of costumers that a firm i has in location c at time t. The relationship between the two variables is:

$$\sum_{c \in C} f_{i,t}(c) = d_i(t) \tag{3.2.1}$$

The key process of interest in this framework is how a firm acquires customers over time in some location c. Formally, this implies characterising $f_{i,t+1}(c) - f_{i,t}(c)$ process. In this setup, firms acquire new customers by new firms using them as inputs. More specifically, a new firm may acquire firm i as an essential input or a variety input. First, focusing on the acquisition as an essential input, this happens when firm i and location c_0 are selected. Probability of drawing location c_0 has been described above and depends on probability function $g(c_0, c)$. For drawing firm i as essential input, note that essential inputs are chosen by a uniform search, hence the probability that the firm will be selected is $\frac{1}{n_t}$. Given that each firm draws m_K essential inputs and that in each period the number of new firms is γn_t, the expected increase in customers in some location c from essential input adoption is described by equation:

[2] In Chaney (2014) the main conclusions are derived using the integer set for defining location. However, the author further shows that conclusions also hold for different location sets. For the present work we maintain integer set for location set throughout the rest of the modelling framework.

$$\gamma n_t \left(\frac{m_K}{n_t} \right) \cdot g(c_0, c) = \gamma m_K \cdot g(c_0, c) \tag{3.2.2}$$

This implies that the number of customers that become essential input users in location c increases with the exogenous growth rate, γ, and with larger number of inputs chosen as essential inputs, m_K. Furthermore, the expected number decreases if the firm's location, c_0, is further away from c. Next, the firm may also become a variety input via network neighbourhood search. For this to occur, first of all, the new born firm in location c needs to identify one of the firm i's existing customers as an essential input, such that it becomes part of the set of potential variety inputs, N_i. Moreover, instead of considering the origins of the firm, the starting point of the search process is from location of its customers c_d. Finally, it needs to be weighted by the number of customers that the firm has in this location, i.e. $f_{i,t}(c_d)$ and the fact that $\frac{m_K}{n_t}$ of these customers would be selected in expectation.

Next, consider the size of the network effect. The expected number of total linkages is equal to $m = m_K + m_N$, which corresponds to the total number of inputs that any firm uses. Given that in total m_K inputs are taken as essential inputs, the total size of the network neighbourhood from which a firm might choose variety inputs is $m_K \cdot m = m_K \cdot (m_K + m_N)$. Since the total number of firms that will be selected as variety inputs is m_N, the corresponding network search probability is $\frac{m_N}{m_K \cdot m}$. As before, we need to multiply this expression by the total number of new firms born: γn_t. Combining all of these components, the expected increase in number of customers from variety input adoption is:

$$\gamma n_t \cdot \left(\frac{m_K}{n_t} \right) \cdot \left(\frac{m_N}{m_K \cdot (m_K + m_N)} \right) \cdot \left[\sum_{c_d \in C} f_{i,t}(c_d) \cdot g(c_d, c) \right] = \frac{\gamma m_N}{m} \cdot \left[\sum_{c_d \in C} f_{i,t}(c_d) \cdot g(c_d, c) \right] \tag{3.2.3}$$

Similar as with essential inputs, the likelihood of becoming adopted as a variety input increases if the growth rate of firms is higher and more variety inputs are adopted. An important element which introduces path dependence in the model, is that now the location matching happens with respect to where existing customers are located, not the origins of the firm. Hence, the likelihood is higher if the location c is closer where the firm has a larger number of customers $f_{i,t}(c_d)$. Combining increases from essential and variety input adoption, the process of new customer acquisition in location c is described by:

$$f_{i,t+1}(c) - f_{i,t}(c) = \gamma m_K \cdot g(c_d, c) + \frac{\gamma m_N}{m} \sum_{c_d \in C} f_{i,t}(c_d) \cdot g(c_d, c) \tag{3.2.4}$$

While this characterized how a firm acquires customers in a particular location it can be shown that the expected total number of customers that a firm has is independent of the location matching. We direct the steps of the solution to Appendix A.1. The resulting equation characterises how the total number of firm's customers evolves over time:

$$d_i(t+1) - d_i(t) = \gamma m_K + d_i(t) \cdot \frac{\gamma m_N}{m} \tag{3.2.5}$$

From Equation (3.2.5) we can directly observe the path dependence in the customer acquisition process which is captured by the fact that the increase depends on $d_i(t)$, i.e. the total number of customers at time t. In the next part we solve the difference equation and show how it relates to the distribution of firms by customers.

3.3 Characterising Firm Distribution by Customers

In this form, Equation (3.2.5) is a linear first-order autonomous difference equation. Before solving the equation we assume that when a firm is born at time t_0, no other firm uses it as input, i.e. $d_i(t_0) = 0$. We further define $r = \frac{m_K}{m_N}$, as the measure of relative importance of essential inputs to variety inputs. Solution to the difference equation is given in Equation (3.3.1) whereas the steps are reported in Appendix A.2.

$$d_i(t) = rm\left[\left(1 + \frac{\gamma}{1+r}\right)^{t-t_0} - 1\right] \tag{3.3.1}$$

Using the solution to the difference equation, we can obtain the distribution of firms by number of customers. Define $F_t(d)$ as the cumulative distribution function, which is the fraction of firms with number of customers smaller than d. Conversely, $1 - F_t(d)$ is the fraction of firms with number of customers larger than d. Then it follows:

Proposition 3.1. *If the customer acquisition is characterised by Equation (3.3.1), the cumulative distribution of firms by number of customers, $F_t(d)$ follows:*

$$F_t(d) = 1 - \left(\frac{rm}{d+rm}\right)^{\log(1+\gamma)\cdot\left(\log(1+\frac{\gamma}{1+r})\right)^{-1}} \tag{3.3.2}$$

The proof of Proposition 3.1 is described below. First, note than in Equation (3.3.1), if we fix t and d, the equation has a unique solution for t_0. In other words, at a given time t, in expectation, firms which have the number of customers d, are those that were born in period t_0. Denote this period as $t^\star$. In this case, the expression $1 - F_t(d)$ then corresponds to the fraction of firms that are older than $t^\star$, which leads to the expression: $1 - F_t(d) = \frac{n_{t^\star}}{n_t}$. Next, we have that $n_t = (1+\gamma)n_{t-1} \Rightarrow \frac{n_{t-1}}{n_t} = (1+\gamma)^{-1}$. Moreover, $t^\star = t^\star + t - t = t - (t - t^\star)$. Hence the expression $\frac{n_{t^\star}}{n_t} = (1+\gamma)^{-(t-t^\star)} = 1 - F_t(d)$. Using Equation (3.3.1) we can solve for $t - t^\star$ for a given d:

$$d = rm\left[\left(1 + \frac{\gamma}{1+r}\right)^{t-t^\star} - 1\right] \tag{3.3.3}$$

$$\frac{d+rm}{rm} = \left(1 + \frac{\gamma}{1+r}\right)^{t-t^\star} \tag{3.3.4}$$

$$\log\left(\frac{d+rm}{rm}\right) = (t-t^\star)\log\left(1 + \frac{\gamma}{1+r}\right) \tag{3.3.5}$$

$$t - t^\star = \log\left(\frac{d+rm}{rm}\right)\cdot\left(\log(1+\frac{\gamma}{1+r})\right)^{-1} \tag{3.3.6}$$

Using the expression for the cumulative distribution we get:

$$1 - F_t(d) = (1+\gamma)^{-(t-t^\star)} \tag{3.3.7}$$

$$1 - F_t(d) = (1+\gamma)^{-\log(\frac{d+rm}{rm})\cdot\left(\log(1+\frac{\gamma}{1+r})\right)^{-1}} \tag{3.3.8}$$

Next, using the logarithm property $a^{\log b} = b^{\log a}$, Equation (3.3.8) can be rearranged to obtain:

$$1 - F_t(d) = \left(\frac{d+rm}{rm}\right)^{-\log(1+\gamma)\cdot\left(\log(1+\frac{\gamma}{1+r})\right)^{-1}} \tag{3.3.9}$$

$$= \left(\frac{rm}{d+rm}\right)^{\log(1+\gamma)\cdot\left(\log(1+\frac{\gamma}{1+r})\right)^{-1}} \tag{3.3.10}$$

$$F_t(d) = 1 - \left(\frac{rm}{d+rm}\right)^{\log(1+\gamma)\cdot\left(\log(1+\frac{\gamma}{1+r})\right)^{-1}} \tag{3.3.11}$$

Which corresponds to Proposition 3.1. As it will be shown below, it is useful to consider the firm distribution by customers in a log-log scale. We plot this distribution in Figure 3.1, whereas we derive the analytical features of the distribution in Appendix A.3, which show that the distribution in log-log scale is concave in the left-tail and linear in the right tail.

As discussed in Jackson and Rogers (2007) the difference between the left and right tail has an intuitive explanation. When the firm is young it has a low number of other firms using it as an input (d small), hence the majority of new connections come from other firms choosing the firm via random uniform search (being selected as an essential input). On the other hand, as the number of customers grows, eventually the network search mechanism becomes dominant which generates fat-tailed distribution, which is approximately linear in log-log scale. This implies that the network search speeds up as the number of existing customers increases, which results in strong path dependence and in the emergence of dominant firms that have a large number of customers.

3.4 Geographical Distribution of Customers

Next, we derive results related to the geographic distribution of customers. For this, consider again Equation (3.2.4), which describes how a firm acquires new customers in location c. To simplify notation, we suppress the subscript i and assume that the origins of the firm is $c_0 = 0$. This simplifies to:

$$f_{t+1}(c) - f_t(c) = \gamma m_K \cdot g(|c|) + \frac{\gamma m_N}{m} \cdot \left(\sum_{c_d \in C} f_t(c_d) \cdot g(|c_d - c|)\right) \tag{3.4.1}$$

In order to proceed further, first focus on the term in the brackets: $\sum_{c_d \in C} f_t(c_d) \cdot g(|c_d - c|)$. Note that $f_t(.)$ can be thought as a random variable defined across location set and

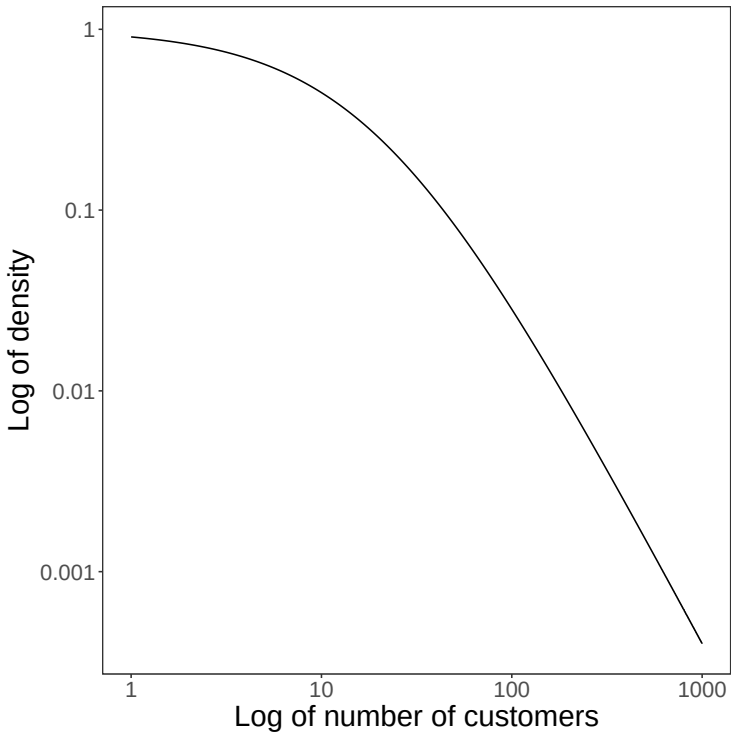

Figure 3.1: **Distribution of Firms By Number of Customers**. Counter-cumulative distribution of firms by customers obtained in Equation (3.3.11). Both the density and number of customers are expressed in log scale. The parameters were set to: $\gamma = 0.02$, $r = 1$ and $m = 20$.

$g(|.|)$ is a probability distribution function defined across the same set. This expression then corresponds to the product of the two functions across their domain, or the discrete convolution product of the two functions:

$$f_{t+1}(c) - f_t(c) = \gamma m_K \cdot g(|c|) + \frac{\gamma m_N}{m} \cdot \left(\sum_{c_d \in C} f_t(c_d) \cdot g(|c_d - c|) \right) \tag{3.4.2}$$

$$= \gamma m_K \cdot g(|c|) + \frac{\gamma m_N}{m} \cdot \underbrace{f_t(c) * g(|c|)}_{\text{convolution product}} \tag{3.4.3}$$

In this form we can apply the convolution theorem, which says that the convolution product is equal to the product of Fourier transformed functions. More explicitly: $f(c) * g(|c|) = \hat{f}(\omega) \cdot \hat{g(\omega)}$, where $\hat{f}(\omega) = \sum_{c \in \mathbb{Z}} f(c) e^{-i\omega c}$, and $\hat{g}(\omega) = \sum_{c \in \mathbb{Z}} g(|c|) e^{-i\omega c}$. Using the Fourier transformations, we can rewrite Equation (3.4.3) as:

$$\hat{f}_{t+1}(\omega) - \hat{f}_t(\omega) = \gamma m_K \cdot \hat{g}(\omega) + \frac{\gamma m_N}{m} \hat{f}_t(\omega) \cdot \hat{g}(\omega) \tag{3.4.4}$$

As in the previous section, we have a linear first-order difference equation with $\hat{f}_{t=t_0} = 0$. The equation permits solution to:

$$\hat{f}_t(\omega) = rm \cdot \left[\left(1 + \frac{\gamma \hat{g}(\omega)}{1+r}\right)^{t-t_0} - 1 \right] \tag{3.4.5}$$

The steps of the solution are presented in Appendix A.4. Keeping the result in mind, note that $g(|.|)$ defines the likelihood of a single location match. Alternatively, we can define the expected geographic density of customers for a firm which already has d number of customers. This would take the form $g_d(c) = \frac{f_d(c)}{d}$. Furthermore, as shown in the proof of Proposition 3.1, we can transform these functions from referring to time, to number of customers. Hence, we can define a corresponding function $g_t(c)$, which represents how geographic density changes over time, $g_t(c) = \frac{f_t(c)}{d_t}$. In this form, we can apply a Fourier transformation to $g_t(c)$:

$$\hat{g}_t(\omega) = \frac{\hat{f}_t(\omega)}{d(t)} \tag{3.4.6}$$

Since we have derived the terms $\hat{f}_t(\omega)$ and $d(t)$, this leads to the expression:

$$\hat{g}_t(\omega) = \frac{\hat{f}_t(\omega)}{d(t)} = \frac{rm \cdot \left[(1 + \frac{\gamma \hat{g}(\omega)}{1+r})^{t-t_0} - 1 \right]}{rm \cdot \left[(1 + \frac{\gamma}{1+r})^{t-t_0} - 1 \right]} = \frac{\left[(1 + \frac{\gamma \hat{g}(\omega)}{1+r})^{t-t_0} - 1 \right]}{\left[(1 + \frac{\gamma}{1+r})^{t-t_0} - 1 \right]} \tag{3.4.7}$$

To proceed further, we formally define a random variable C_t for location realizations. A key property is the second moment of random variable C_t, since it corresponds to the average squared distance of customers. We formally define Δ_t as the average squared distance of customers of a firm of age $(t - t_0)$:

$$\Delta_t \equiv \sum_{c \in C} c^2 g_t(|c|) = E[C_t^2] \tag{3.4.8}$$

Alternatively, as before we can define Δ_d as the average squared distance of customers, of a firm that already has d number of customers. In this form it is possible to obtain how the average distance of customers changes over time and, more importantly, conditional on the number of existing customers.

Proposition 3.2. *The average squared distance of customers, Δ_d, conditional on the number of existing customers, d, is increasing in the number of existing customers d and takes the form:*

$$\Delta_d = A \cdot \ln\left(1 + \frac{d}{rm}\right) \cdot \left(1 + \frac{rm}{d}\right) \tag{3.4.9}$$

$$A = \left[\sum_{c \in C} c^2 g(|c|) \right] \cdot \frac{\gamma}{1 + r + \gamma} \cdot \left[\ln\left(1 + \frac{\gamma}{1+r}\right) \right]^{-1} \tag{3.4.10}$$

The proof of the proposition is provided in Appendix A.5. The proof relies on the fact that $g_t(|c|)$ is a probability function, hence the Fourier transformation can be related to the characteristic function of the distribution. Specifically, it can then be used to obtain moments of $g_t(|c|)$ by taking derivatives of $\hat{g}_t$ and evaluating them at $\omega = 0$. After expressing how squared distance of customers changes over time, the final step derives the expression for the squared distance of customers depending on the number of customers that a firm has (Δ_d).

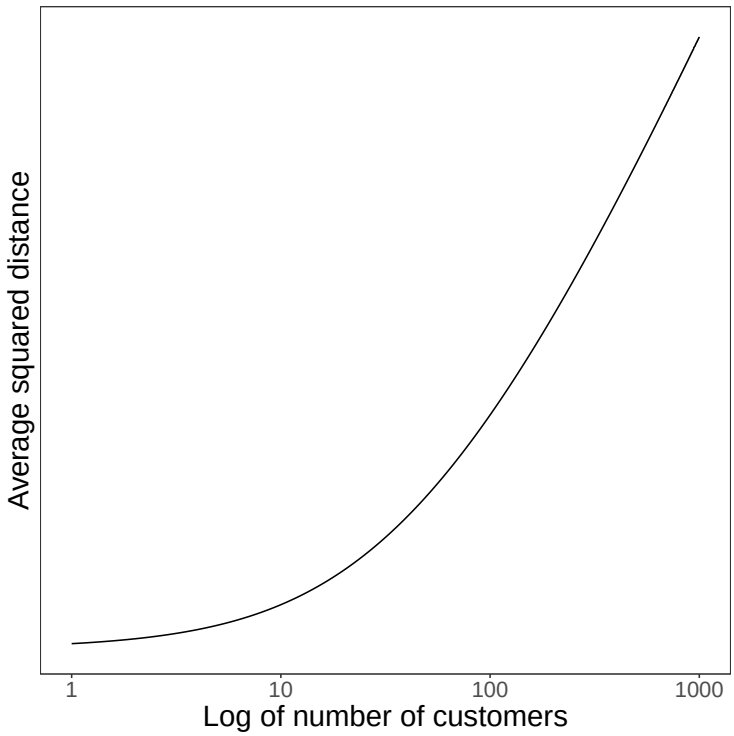

Figure 3.2: **Average Distance of Customers**. Relationship between number of customers (d) and the average squared distance of customers Δ_d, (Equation (3.4.9)). The x-axis is plotted in log scale. The parameters were set to: $\gamma = 0.02$, $r = 1$ and $m = 20$.

Note, that the part $\ln\left(1+\frac{d}{rm}\right)\cdot\left(1+\frac{rm}{d}\right)$ implies that Δ_d is increasing in d. To see it more clearly, Figure 3.2 shows the relationship explicitly, using the same assumptions as in Figure 3.1. Note that to maintain consistency with results in this paper and with comparability with previous work, the x-axis (number of customers), is re-scaled in log scale.

The key result in these derivations is that the average squared distance of firm's customers increases as the total number of customers increases. This mechanism suggests that as the firm acquires new customers they become more spread out. To understand what is driving this result, we can contrast this with the distribution obtained when aggregating over all locations (Figure 3.1). As it was shown and argued, the process of customer acquisition is dominated by different elements depending on how many customers a firm has. When the firm has a low number (small d), customer acquisition is mainly driven by new firms using it as an essential input. If we consider the customer acquisition Equation (3.4.1), the likelihood of this occurring is associated with local search (originating in c_0). Hence, when this mechanism is dominant, the average distance of customers is small (see left-tail in Figure 3.2). However, over time, when further customer acquisition becomes dominated by network based search, the distribution of customers across location increases much more substantially and depends on the geographic location of existing customers. This is also the result obtained in Chaney (2014) and which has been documented empirically.

Another important note about the results with regards to location is that the two assumptions about $g(|.|)$ were that the likelihood of location decreases with distance and that

$g(|.|)$ has a finite second moment. Hence, the result does not rely on strong assumptions about the underlying distribution of location matching.

3.5 Sectoral Level Implications

An advantageous aspect of modelling input adoption by distinguishing essential and variety inputs is that it is then possible to extend implications of the model from firm to sectoral level. This can be achieved by assuming that firms are classified into sectors based on their set of essential inputs. Before proceeding, note that the set of firms I_t can also be represented by a $n_t \times n_t$ matrix of directed relationships between firms, where an element of the matrix $b_{ij} = 1$ would indicate that firm i is using firm j as an input. As it will be discussed later, it is also useful to define a binary vector μ_{K_i} of length n_t for each firm, where elements of the vector correspond to other firms and are equal to 1 if the firm is an essential input provider for firm i.

Similarly, a sector classification s_j is defined by a binary vector μ_{s_j} of length n_t, which indicates whether sourcing the corresponding firm is necessary to be classified into a sector. We assume that sectors are symmetric in the number of inputs that are used to define a sector (number of elements equal to 1 in μ_{s_j}), which is equal to x. Furthermore, we assume that there is a finite number of sectors J. Formally, a firm is classified into sector s_j if it has the largest overlap between the vector that is used to define its set of essential inputs (μ_{K_i}) and vector that is used to define sector s_j (μ_{s_j})[3]. Given that in expectation the overlap between the two vectors would not be perfect, we define k_{s_j} as the expected number of elements that are equal to 1 in both μ_{K_i} and μ_{s_j}. Since sectors are assumed to be symmetric in the number of inputs that define them, the expected value of k_{s_j} is the same for all sectors.

Similarly as with firms, we can define an input-output sector matrix $J \times J$, where each element a_{ij} corresponds to a binary variable which indicates if there are firms classified to sector s_i that source inputs from s_j. An important aspect that is different from similar aggregation described in Carvalho and Voigtländer (2014) is that sectors s_i and s_j are allowed to be in different locations, denoted c_i and c_j respectively. Next, if a relationship between two sectors does not exist, we define a network proximity measure $\eta_{s_i,s_j}(c) = (\mu_{s_j})'\nu_{s_i}(c)$, where $\nu_{s_i}(c)$ is the number of firms that are used to classify into sector s_j and are located in c, that use one of s_i firms, located in c_i as input. Remember that when a firm is born in location c_j, it may source its essential inputs from location other than c_j. This introduces an additional complexity to the framework, hence we need to consider network proximity measure in all possible location. As it will be shown later on, this setup implies that the likelihood of a new relationship forming between sectors s_j and s_i depends jointly on the network proximity $\eta_{s_i,s_j}(c)$ and location matching function $g(.,.)$.

Before proceeding further to sectoral level implications, we introduce an additional variable. By considering a firm i located in c_i, we define $i_{s_j}(c)$ as the number of firms in location c, that would be classified into sector s_j, that source the firm i as an input. We can relate

[3]Distance between two binary vectors can be measured as the Hamming distance, where the distance is greater the larger the number of non-matching elements exists. In this context a firm is classified into sector s_j if it has the minimum Hamming distance with respect to μ_{K_i}.

this value to the sector proximity measure, if we consider all firms i that define sector s_i. Specifically, the network proximity measure is equal to:

$$\eta_{s_i,s_j}(c) = \sum_{i \in s_i} i_{s_j}(c) \tag{3.5.1}$$

The next proposition describes on which factors the likelihood of a new connection forming depends on:

Proposition 3.3. *If we take two sectors s_j and s'_j, both located respectively in c_j and c'_j, which do not use any inputs from sector s_i located in c_i, s_j is said to be more likely to adopt sector s_i as input than s'_j if:*

$$\sum_{c \in C} g(c_j, c) \cdot \eta_{s_i,s_j}(c) > \sum_{c \in C} g(c'_j, c) \cdot \eta_{s_i,s'_j}(c) \tag{3.5.2}$$

The proof of the proposition is in Appendix A.6. The result suggests that the likelihood of adoption is dependent on the local network proximity between sectors weighted by the distance between the location and where sector s_j is. Furthermore, this likelihood is not dependent on location of the sector s_i directly. Instead, since sectoral adoption only happen due to network neighbourhood search, we only need to consider where sector s_i contacts are located. On the other hand, if we consider the expected expansion of connections, sector s_i is more likely to have more connections near c_i.

3.6 Summary of Theoretical Framework

Using a random graph model with differentiated types of inputs and location we showed that the following properties can be derived: (i) from Section 3.3, the distribution of firms by number of customers (in log-scale) is concave in the left-tail and is linear in the right-tail, which reflects a fat-tailed distribution; (ii) from Section 3.4, the average distance of customers increases with total number of customers and is not dependent on the underlying probability function of location matching; (iii) from Section 3.5, the likelihood of input adoption increases if network proximity, weighted by geographic matching function, is larger. In contrast to previous works, (i) and (ii) have been derived before (Chaney, 2014). However, the contribution of the work is that the results of (i) and (ii) were obtained allowing for input-output relationships between firms. Furthermore, implications of (iii) with regards to the joint importance of network proximity and geographic distances are novel.

From the derived results and propositions of the framework, we highlight three hypothesis which we will consider in the empirical part of the study when analysing the WIOD. First, given the derived properties of the distribution by number of outward linkages in Section 3.3, the first hypothesis states that:

Hypothesis 1 (H1)**.** *The distribution of sectors by number of outward linkages is skewed and exhibits a fat right tail.*

Next, considering theoretical prediction regarding increasing average distance of customers from Proposition 3.2 the hypothesis states:

Hypothesis 2 (H2)**.** *The average squared distance of customers increases as sectors accumulate more outward linkages.*

Finally, the last hypothesis focused on the likelihood of new network formation and is based on Proposition 3.3:

Hypothesis 3 (H3)**.** *The likelihood of new linkage formation increases with shorter network proximity and shorter geographical distances.*

Hypothesis H3 is considered using an econometric estimation, whereas Hypotheses H1 and H2 are primarily analysed qualitatively.

4 Empirical Analysis of Sectoral Network Formation

4.1 Data

The primary source of data for the present empirical study is the World Input-Output Database (WIOD) (Timmer et al., 2015), November 2016 release. The dataset covers inter-country industry-by-industry trade flows, for 43 countries and 54 sectors[4]. The considered data release covers the period between 2000 and 2014 and reports yearly trade flows in current prices, denominated in millions of US$. Using this data, it is possible to track how much a given sector in a country imports from and exports to every other sector in every other country. The data also tracks total output, final demand, and changes in inventory and gross capital for each sector in the sample. All unmatched flows are gathered together under "Rest-of-the-World" classification. Given that the estimation will rely on geographic matching, we omit this category.

In addition to information regarding trade flows, together with the 2016 WIOD trade data release, a socio-economic data companion was released in February 2018. This data contains additional information about each of the sectors that are part of the 2016 release: number of employees, total hours worked, labour and capital compensation and nominal capital stock. Furthermore, the data contains industry level deflators for value added, total output and intermediate inputs.

Since theoretical implications are tied to geographic variation of linkages between sectors, we also collect data on geographic distances between countries. We follow Chaney (2014) and use distance measures calculated and reported by CEPII. In particular, we used a measure defined as the distance between country centroids which correspond to population weighted coordinates of major cities within the country.

Since in this study we focus on the extensive margin of trade - binary relationships between sectors - and due to potentially noisy trade data, we impose a cut-off point. Following Carvalho (2010), we set that a relationship between two sectors exists, if, from the point of view of importing sector, the inflow is at least 1% of the total sectoral intermediate input use. For robustness, we tested the sensitivity of econometric results based on less and more restrictive thresholds. In the next section we characterize the main features of WIOD.

4.2 Qualitative features of WIOD

In this section, we focus on the key descriptive features of WIOD that relate to the theoretical predictions. Specifically, the focus is on whether the distribution of sectors by outward linkages, as proposed in Hypothesis H1, is skewed and exhibits a fat right tail. Furthermore, in relation to Hypothesis H2, we check whether the average distance of linkages increases the more connections a sector has. First of all, we visualise the international sectoral network in Figure 4.1, where sectors by country are highlighted.

[4]In Appendix A.7 we report the list of industries and countries that were part of the sample

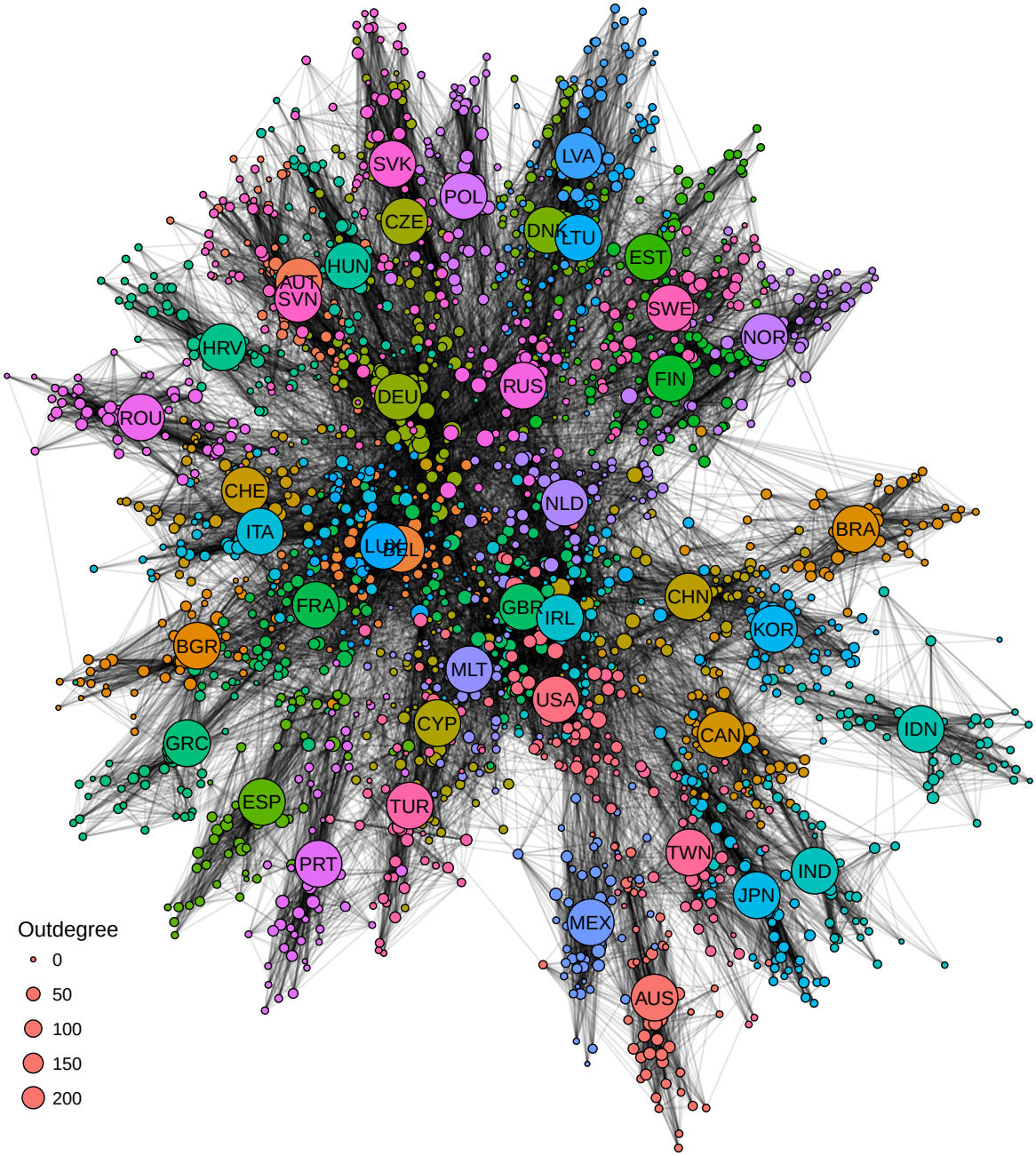

Figure 4.1: **WIOD Network**. Sectoral network structure by country in 2014. A link between two sectors indicates that one of the sector constitutes at least 1% of intermediate input use of the other sector. Large country labelled nodes correspond to average location of country sectors within the network. Country abbreviations and names are matched in Appendix A.7. Data source: author's rendering of WIOD (Timmer et al., 2015).

To visualise this network we used Fruchterman and Reingold (1991) plotting algorithm. The algorithm plots the network nodes (in this cases sectors) such that more connected nodes are grouped. Furthermore, to highlight how each country is positioned within the

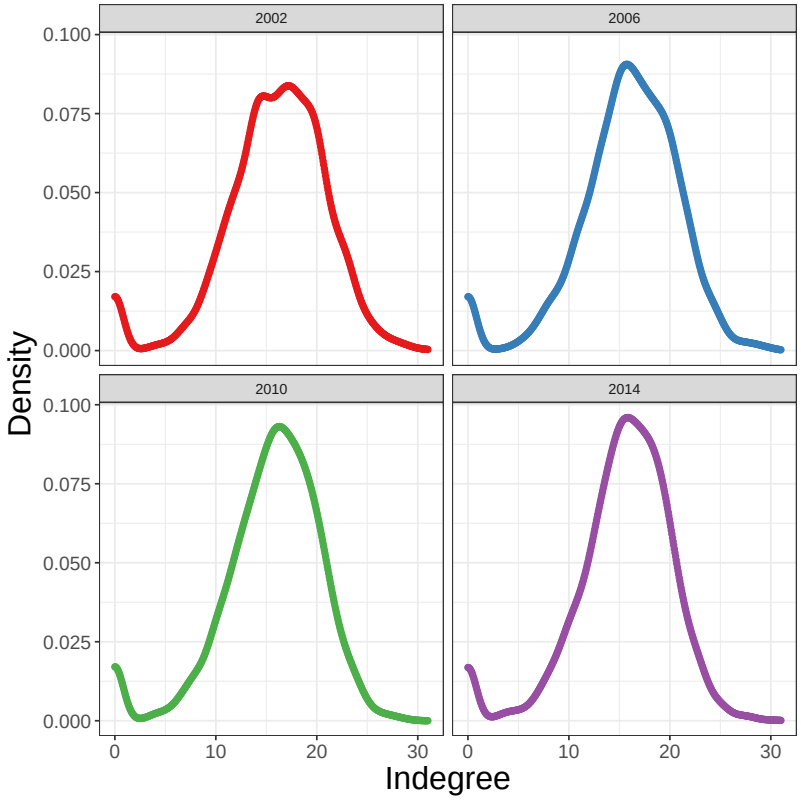

Figure 4.2: **Sectoral Indegree Distribution in WIOD**. The plots show the empirical density of sectors in WIOD by number of sectors that they import inputs from (indegree). Panels are for different sample years. Data source: author's rendering of WIOD (Timmer et al., 2015).

network, we include country labelled centroids which correspond to the average location of country-specific sector nodes in the plot.

First, note that countries which have sectors that are most connected are within the centre of the network. These countries are Germany, Russia, the Netherlands, the U.K. and Belgium. On the other hand, if we focus on countries that are at the periphery of the network we can observe the emergence of geographic regions. For example, the top right corner consists of Scandinavian countries and the Baltic States, the top left corner consist of Central European countries and the bottom left consists of countries in the Mediterranean region. This indirectly matches the prediction of the theoretical framework and supports Hypothesis H2, given that less connected sectors tend to be more local in their connections, hence we can observe geographic regions by a pure network proximity visualisation. Conversely, sectors that are highly connected tend to have more spread out relationships and hence are found in the centre of the network.

Having highlighted some of the features from an aggregate network overview, we consider the distributions of sectors by the number of linkages. Specifically, this is relevant for checking whether the WIOD supports Hypothesis H1 which claims that the distribution of outward linkages should be skewed and have a fat-right tail. The distribution of linkages has been one of the key motivations for applying network-based models in studying firm and

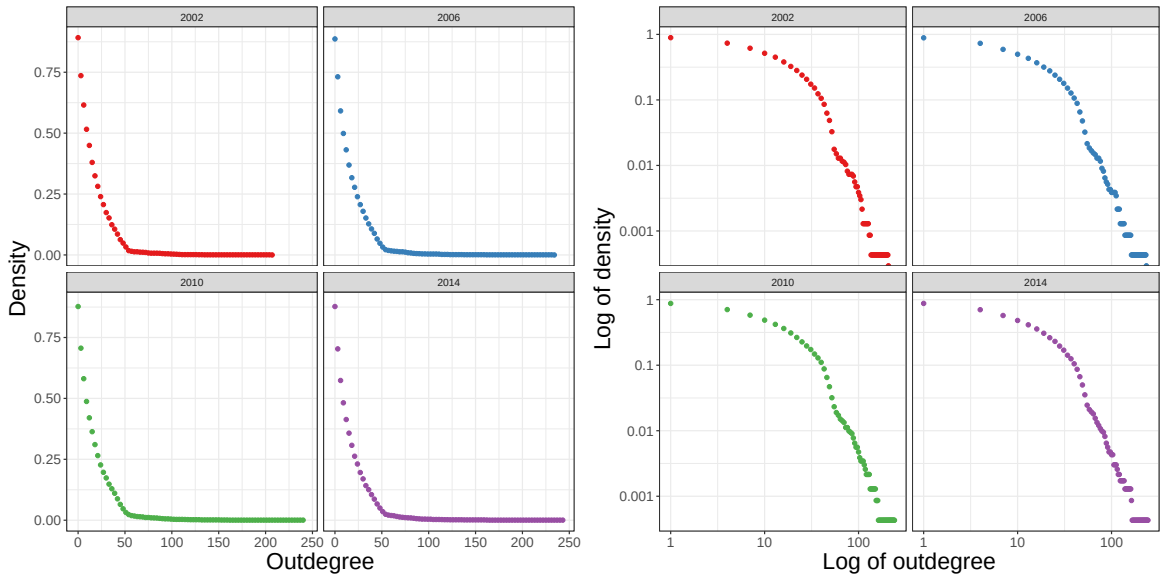

Figure 4.3: **Sectoral Outdegree Distribution in WIOD**. Left panel shows empirical density of sectors in WIOD by number of sectors that they export to (outdegree), plotted in non-scaled values. The right panel shows the same distribution, but with density and number of sectors adjusted in log of base 10 scale. Each plot within each panel is for different sample years. Data source: author's rendering of WIOD (Timmer et al., 2015).

sectoral linkages due to the emergence of power-law type distributions (Jackson and Rogers, 2007; Gabaix, 2016). Moreover, a sharp asymmetry between the indegree (number of inputs that a sector uses) and the outdegree (number of sectors that use it as an input) has been documented empirically in other international contexts (Bernard and Moxnes, 2018), hence it is relevant to check whether the WIOD network exhibits these features as well.

We first consider the empirical density of indegree which is presented in Figure 4.2, where each panel is for different years (2002, 2006, 2010, 2014). We can see that the distribution in most years is centred and single peaked. A similar characterisation of sector linkages has been observed in the case of the U.S. (Carvalho, 2010). This supports the modelling assumption that the number of inputs that are used is stable and is not time varying.

Next, we consider the distribution by outdegree, which is presented in Figure 4.3. First, we focus on the counter-cumulative distribution of sectors by outdegree in non-scaled values (left panel). In support of Hypothesis H1, it can be seen that the distribution is skewed and exhibits a right fat-tail which is characterised by a disproportionate number of sectors that have a very high outdegree. To relate it more closely to the underlying distribution derived in the theoretical model, we plot in the right panel the distribution in log-log scale, where both the outdegree and the density are expressed in logs of base 10. Once again, the pattern is similar to the one observed in previous studies (Carvalho, 2010; Chaney, 2014) and qualitatively matches the theoretical distribution that is obtained in the previous section (see Figure 3.1 and Appendix A.3). Specifically, the outdegree distribution is concave in the left tail and is approximately linear in the right tail. As discussed in the theoretical framework, this may reflect that the behaviour in the tails is different. For a sector that is initially

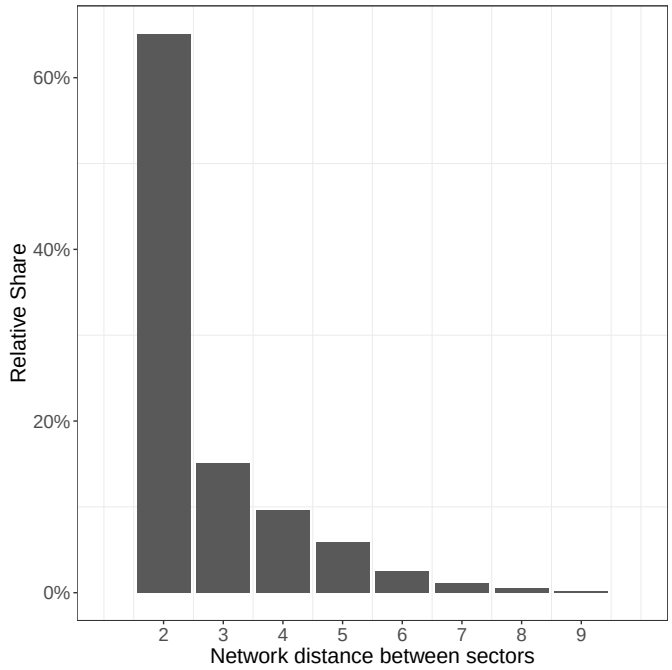

Figure 4.4: **Input Adoption Events by Network Proximity**. Network distance between sectors between which an adoption event occurred. Network distance between sectors is measured by the shortest network path.

not well connected, most of the new connections will come from uniform random search. However, once the sector becomes more connected, the new connections will primarily come from network neighbourhood search.

In addition to focusing on the overall network, we also consider adoption events separately. While the precise definition of what is meant by an adoption event will be introduced in later sections, intuitively it represents events, when a connection between an input providing and an input receiving sector did not exist in 2000 but was eventually established at some point. First, we consider the network proximity distance, measured by the shortest network path between sectors in Figure 4.4. We can see a clear trend that the majority of new connections that were formed were between sectors that had close proximity to each other. This result qualitatively supports Hypothesis H3, which states that the likelihood of adoption would be increasing with closer network proximity.

Alternatively to network proximity, we also consider the average squared distance of new connections from input sending sector's point of view. Figure 4.5 reports in log-scale how the average squared distance of new connections depends on the existing number of outward linkages and number of countries serviced. This relates to the theoretical prediction, that as the sector becomes an input provider to more sectors, the new connections will tend to be geographically further away. While the increase is not uniform and noisy in the right-tail, an upward trend can be observed. This suggests that the higher number of connections that a sector had at reference year 2000, the more likely it was to form connections with sectors that were further away which provides support for Hypothesis H2.

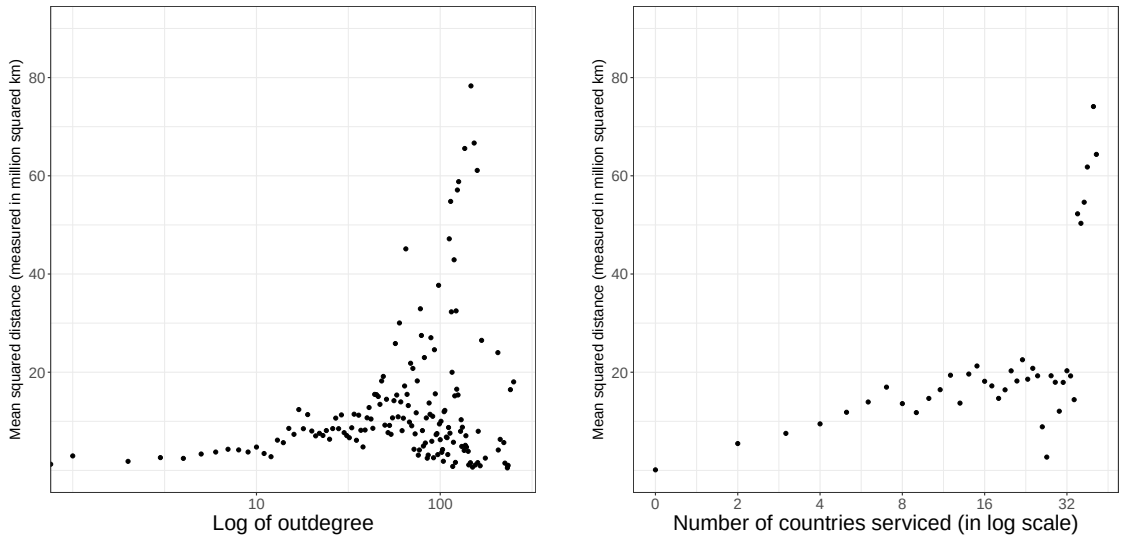

Figure 4.5: **Input Adoption Events by Geographic Distance**. Geographic distance of new connections between sectors. Each point in the plot measures the average squared distance between sectors that formed a new connection, conditional on the number of sectors or countries serviced by the input sending sector prior to the connection forming.

4.3 Econometric Strategy

In this part we focus on the econometric estimation and the definition of key empirical variables. The identification relies on the fact that during the WIOD sample years sectors established new relationships. Specifically, by contrasting the existing network structure at the start of sample years in 2000 with all other years, it is possible to determine out of all possible relationships which ended up forming. Hence the econometric model aims at characterising which variables could predict whether a new connection was established.

As it was mentioned above, due to potential noisiness of trade data a threshold is imposed to constitute an economically meaningful relationship between sectors. Formally, we define a binary variable $Trade^{i}_{s_1,s_2}$, which is equal to 1 if in year i, from the point of view of sector s_1, imports from sector s_2 constituted at least 1% of total intermediate input usage of sector s_1, and is equal to 0 otherwise. Next, given that the key interest of the econometric analysis is to study the likelihood of new network formation, we define a new variable for adoption events. Formally, this variable takes the form:

$$Adopt_{s_1,s_2} = \begin{cases} 1 & \text{, if } Trade^{2000}_{s_1,s_2} = 0 \ \& \ Trade^{i}_{s_1,s_2} = 1, \text{for one of } i = 2001, 2002, ..., 2014 \\ 0 & \text{otherwise} \end{cases} \tag{4.3.1}$$

Based on this definition, the dependent variable corresponds to an event, when for a given set of two sectors, there was no trade relationship at the start of the sample in year 2000, but at least in one of the remaining sample years s_1 sourced inputs from s_2. The empirical model that we estimate is:

$$P(Adopt_{s_1,s_2}|X) = \Phi\Big(\alpha + \beta_1 net\_dist_{s_1,s_2} + \beta_2 D_{c_1=c_2} + \beta_3 g(dist_{c_1,c_2}) + \beta_4 out\_degree_{s_1} + \beta_5 g(dist_{c_1,c_2}) \times out\_degree_{s_1} + \beta_6 avg\_dist_{s_1} + \beta Controls + \epsilon_{s_1,s_2}\Big) \quad (4.3.2)$$

Note that c_1 and c_2 correspond to countries in which sector s_1 and sector s_2 are located respectively. All explanatory variable values are taken as their value in year 2000. Next we explain each of the explanatory variables, their definitions and construction, and their expected impact based on theoretical predictions. We report summary statistics of explanatory variables in Appendix A.8.

First of all, $net\_dist_{s_1,s_2}$ is a measure of network proximity. Imposing the trade cut-off and taking the WIOD sample of 43 countries and 54 industries allows to represent the international input-out table by a 2322×2322 binary matrix, which corresponds to a directed network. Next, given that the theoretical implications are tied to the notion of network proximity, for each possible relationship 2322^2 we calculate the network proximity using Dijkstra's (1959) algorithm, which finds the shortest path between two nodes (country-industry pairs). If Hypothesis H3 is true, the expected sign of the coefficient is negative ($\beta_1 < 0$) since the model predicts that closer proximity should increase the likelihood of new connection forming and the shorter the network path, the closer two sectors are.

An important set of explanatory variables are related to the impact of geographical distances on the likelihood of a new link forming. For the exact choice of variables we follow set of explanatory variables that were used in Chaney (2014), with the crucial difference that the present estimation includes cases where a link may form between sectors that are in the same country.

For this reason, we include a dummy variable $D_{c_1=c_2}$ which is equal to 1 if the two sectors are in the same country. The dummy is included given that there is a theoretical lower limit to the value that the physical distance can take and there may be concerns of non-linearity in the relationship at the limit, which in this case would be captured by a fixed effect.

Furthermore, we define a function $g(dist_{c_1,c_2})$ which takes as input the physical distance between two countries in which the two sectors are located. For the baseline specification we assume that the function $g(.)$ takes the form: $g(dist_{c_1,c_2}) = dist_{c_1,c_2} + dist^2_{c_1,c_2}$. In later sections we consider alternative forms of the $g(.)$ function. Similarly as with the network proximity variable, if Hypothesis H3 is true, we would expect that the impact of distance is negative ($\beta_3 < 0$). In other words the further away the countries where the sectors are located, the lower the likelihood of a link forming between the two sectors.

In the estimation we also include variable $out\_degree_{s_1}$, which corresponds to the number of outward linkages of sector s_1, i.e. number of sectors that used it as an input in year 2000. This variable captures the theoretical feature of the model that sector linkage formation may exhibit path dependence - the more connections a sector has the more likely it is to form new links in the future, hence the expected sign of the estimate is positive ($\beta_4 > 0$). In

addition to including $out\_degree_{s_1}$ as a separate variable, we also include interaction between the $out\_degree_{s_1}$ and the distance function. Finally, related to this argument we include $avg\_dist_{s_1}$ variable, which is the average distance of existing outward links. The last two variables indirectly relate to Hypothesis H2. Namely, the model predicts that as the sector accumulates more outward linkages, the average distance of these linkages increases. This can also be thought as representing the fact that distance barriers are less significant if the sector has more outward linkages. Hence if Hypothesis H2 is true, we would expect the estimate of the interaction term β_5 to have the opposite sign of β_2. Similarly, if outward linkages are more spread out, that should make the network based search of further distances easier, hence we would expect that higher average distance of customers increases the likelihood of an adoption event ($\beta_6 > 0$).

In addition to variables that relate to the theoretical framework, we also include control variables which may impact the likelihood of adoption. For this part we follow Carvalho and Voigtländer (2014). First of all, we include input receiving and input sending country and industry fixed effects. Moreover, we include a measure of geographic isolation, defined by the average distance to all countries in the sample. This variable captures the fact that some countries may be geographically remote and hence sectors in these countries face higher barrier of becoming adopted as inputs in another country. Finally, we include log of number of employees in the sector and log of real value added per employee as a proxy for productivity. The intention of these control variables is to account for industry size effects, i.e. larger sectors more likely to form linkages, and quality of the sector, measured by value added per employee, as it may be the case that more productive sectors are more successful in forming new linkages. Employee number and productivity variables are included both for input sending and receiving sectors.

While the theoretical maximum number of observations is $2322^2 = 5,391,684$, we limit the sample in the following way. We first remove observations where $s_1 = s_2$, i.e. the link between the sector and itself is omitted. Next, we remove sectors that do not link to any other sector and hence a network distance measure cannot be established. Furthermore, we exclude cases when $Trade^{2000}_{s_1,s_2} = 1$, since these sectors already have a link between each other at the reference year 2000.

Due to binary nature of the model, we estimate it as a logit model. Given that the base rate of adoption compared to total number of potential adoption events is low, in estimation results together with logit coefficients we report fully standardised coefficients (Long and Freese, 2014). These coefficients measure how many standard deviations the dependent variable changes by a one standard deviation change in the explanatory variable. Using these coefficients it is possible to comment on the relative importance of explanatory variables. In addition to logit estimation, for robustness we also report results of a linear probability model (LPM) estimated with ordinary least squares. Among other robustness checks we consider alternative cut-off points that constitute an economically significant relationship (5% and 0.1%). Furthermore, we consider a smoothed trade series by applying a 3-year moving average smoothing, to check that results are not driven by potentially noisy yearly

fluctuations.

A common concern within the literature of empirical studies of production networks is that a source of exogenous variation is lacking. Hence there is an important concern regarding omitted variable bias. For this reason, in addition to variables that relate to the theoretical prediction of the model, we include additional specification with control variables that have been found to impact the likelihood of new linkage formation. An additional concern is that the estimation assumes that the impact of the network structure in the year 2000 has a long-lasting effect and hence can influence network formation throughout the whole sample. While this may be a bigger concern for empirical studies of the intensive margin of trade, since the present analysis considers extensive margin trade and economically significant trade, their effect should be long-term. Hence we should expect the effect to be preserved over a longer duration. Other limitations of the estimation are discussed in Section 5.5.

5 Estimation Results

In this section we present results of the econometric estimation described in Section 4.3. First results of the baseline specification are reported, followed by robustness checks where different functional forms for the impact of geographic distance were tested. In addition, the empirical models were re-estimated using smoothed trade series and also allowing for alternative thresholds for defining significant trade relationships. The section is concluded with a discussion of empirical results and limitations of the study.

5.1 Baseline Results

In Table 5.1 results of the baseline specification of the model are reported. Each column corresponds to specifications with different sets of explanatory variables. The first set of columns report results of the estimation using network distance as the only explanatory variable. The next set report results using an extended list of explanatory variables, but without any control variables. The following two columns report results using fixed effect and isolation measure controls, whereas the final set of columns report results using additional sector employment and productivity controls. The separation of the estimation in the last two sets is due to the fact that not for all sectors additional data on employment exists.

Focusing on estimations results that are related to the theoretical predictions of the model, it can be seen that network proximity has a substantial impact on the likelihood of a new network link forming. This result is consistent across all specifications. Furthermore, in comparison with other variables, network proximity has the highest relevance for impacting the likelihood of adoption measured by fully standardised coefficients. In baseline specifications, the standardised network proximity coefficient ranges between -0.84 and -0.49.

Considering the impact of geographic distances the signs and significance of estimates are in line with theoretical expectations. Namely, the larger the distance the lower the likelihood of sector connection forming. It can also be seen that not including controls does not allow capturing the fact that the impact of distance is not linear. Specifically, the negative impact has a diminishing effect, given that the coefficient of the squared distance is positive when including controls. Furthermore, as additional controls are added, the standardised coefficient also increases suggesting that the geographic proximity has a relatively sizeable impact on the likelihood of adoption. Finally, it can also be seen that it is relevant to control for the fact that connecting sectors may be in the same country. Both results relating to network proximity and geographical distances support Hypothesis H3.

Next, considering the impact of the number of existing outward linkages, estimation results suggest that the higher the number of existing connections a sector had in 2000, the higher the likelihood that a new link would form. This result is also in line with theoretical prediction, however, compared to network proximity and distance it is less pronounced. Among results that do not agree with theoretical predictions, the interaction term is negative, suggesting that as the number of outward connections increases it becomes more difficult

for sectors to overcome distance barriers. Furthermore, having outward linkages that are further away also does not help to mitigate the negative impact of distance as indicated by the negative coefficient for the average distance variable. These last two results, in fact, do not support Hypothesis H2. It should be mentioned, however, that in contrast to other key variables these effects are not as relevant as network proximity and geographic distance since standardised coefficients are lower.

For completeness, in Appendix A.9 results of a linear probability model estimation are reported. Given the low base rates, the estimate values are low. However, the sign and significance of coefficients agree with the logit estimation. Moreover, given the high significance of variables, but low standardised coefficient values, suggests that a logit model is more appropriate since due to low base rates the model is capturing probabilities that are near the boundary of the lower limit where non-linearities are of greater concern.

5.2 Changing Definition of Distance

To check that results of the econometric estimation do not depend on assumptions behind the functional form of the distance function, $g(.)$, the econometric model is re-estimated using alternative functional forms that have been used in previous works:

$$g(dist_{c1,c2}) = \begin{cases} dist_{c1,c2} + dist_{c1,c2}^2 & \\ \ln(dist_{c1,c2}) & \text{, Carvalho and Voigtländer (2014)} \\ 1/dist_{c1,c2} & \text{, Chaney (2014)} \end{cases} \tag{5.2.1}$$

All of the definitions of the functional form aim at capturing the potentially non-linear impact of distance on the likelihood of new connection formation. In the end, the expected sign of the effect should be consistent across all measures. It should be noted, however, that the expected sign of the definition according to Chaney (2014) is reversed, since the variable is decreasing as the distance measure increases. The results are presented in Table 5.2.

Similar to the baseline results the impact of distance coefficient is consistently in line with theoretical predictions regardless of the distance definition. Furthermore, controlling for connections forming within the same country is relevant in all cases. Moreover, network proximity is highly significant and relevant for explaining the likelihood of further adoption. Consistently throughout all specifications, the impact of number of outward linkages is also significant and of similar relevance.

A notable exception where estimation results do not agree across different distance definitions is with regards to the interaction terms. Note, that since the distance function is re-defined, this implies that the interaction term is now considered together with the new distance function. In contrast to the baseline definition, the interaction term signs are in line with theoretical predictions using Chaney (2014) and Carvalho and Voigtländer (2014) definition and support Hypothesis H2. Namely, the coefficient supports the claim that as the number of outward linkages increases, the negative impact of distance is smaller. Although,

	(1)		(2)		(3)		(4)	
$P(Adopt_{s_1,s_2}\|X)$	β^{logit}	$\hat{\beta}$	β^{logit}	$\hat{\beta}$	β^{logit}	$\hat{\beta}$	β^{logit}	$\hat{\beta}$
$net\_dist_{s_1,s_2}$	-1.4823***	-0.8430	-0.8783***	-0.6292	-0.7365***	-0.4876	-0.7666***	-0.5013
	(0.0137)		(0.0148)		(0.0159)		(0.0180)	
$dist_{c_1,c_2}$			-0.0508***	-0.0847	-0.3224***	-0.4966	-0.3358***	-0.5141
			(0.0140)		(0.0195)		(0.0215)	
$dist^2_{c_1,c_2}$			-0.0001	-0.0033	0.0153***	0.3316	0.0147***	0.3178
			(0.0011)		(0.0013)		(0.0014)	
$out\_degree_{s_1} \times$			-0.0006***	-0.0306	-0.0005***	-0.0260	-0.0003*	-0.0150
$dist_{c_1,c_2}$			(0.0001)		(0.0001)		(0.0001)	
$out\_degree_{s_1}$			0.0129***	0.0858	0.0092***	0.0566	0.0070***	0.0430
			(0.0003)		(0.0005)		(0.0005)	
$avg\_dist_{s_1}$			-0.0001***	-0.0637	-0.0001***	-0.0861	-0.0001***	-0.0412
			(0.0000)		(0.0000)		(0.0000)	
$D_{c_1=c_2}$			2.1010***	0.1096	2.3295***	0.1123	2.3286***	0.1118
			(0.0312)		(0.0414)		(0.0434)	
α	-0.0503		-2.5446***		-6.3960***		-7.1304***	
	(0.0383)		(0.0543)		(0.3594)		(0.4261)	
Fixed Effects	No		No		Yes		Yes	
Additional Controls	No		No		No		Yes	
Adoption Events	14,493		14,493		14,493		13,837	
Observations	3,720,405		3,720,405		3,720,405		3,464,643	
Pseudo R^2	0.222		0.273		0.324		0.336	

Table 5.1: **Baseline Estimation Results**. Column (1) reports estimates of a model using only network proximity to predict adoption. Columns (2) report results using all previously described explanatory variables, except for controls. Column (3) reports results with all explanatory variables, fixed effect and isolation controls, whereas (4) reports results with additional employment and productivity controls. Column β^{logit} reports logit coefficients. Terms in the brackets are robust standard errors. Column $\hat{\beta}$ report fully standardised coefficients. Symbols * indicate statistical significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

the relevance and significance of these estimates limit the strength of these conclusions.

	Baseline		Chaney (2014)		Carvalho and Voigtländer (2014)	
$P(Adopt_{s_1,s_2}\|X)$	β^{logit}	$\hat{\beta}$	β^{logit}	$\hat{\beta}$	β^{logit}	$\hat{\beta}$
$net\_dist_{s_1,s_2}$	-0.7365***	-0.4876	-0.7522***	-0.5188	-0.6919***	-0.4595
	(0.0159)		(0.0156)		(0.0163)	
$dist_{c_1,c_2}$	-0.3224***	-0.4966				
	(0.0195)					
$dist^2_{c_1,c_2}$	0.0153***	0.3316				
	(0.0013)					
$out\_degree_{s_1} \times$	-0.0005***	-0.0260				
$dist_{c_1,c_2}$	(0.0001)					
$out\_degree_{s_1}$	0.0092***	0.0566	0.0100***	0.0644	0.0083***	0.0513
	(0.0005)		(0.0005)		(0.0004)	
$avg\_dist_{s_1}$	-0.0001***	-0.0861	-0.0001***	-0.0760	-0.0001***	-0.0901
	(0.0000)		(0.0000)		(0.0000)	
$D_{c_1=c_2}$	2.3295***	0.1123	3.4115***	0.1713	2.8470***	0.1376
	(0.0414)		(0.0451)		(0.0390)	
$1/dist_{c1,c2}$			0.3909***	0.1218		
			(0.0117)			
$out\_degree_{s_1} \times$			-0.0018***	-0.0163		
$1/dist_{c1,c2}$			(0.0003)			
$\ln(dist_{c1,c2})$					-0.6099***	-0.2391
					(0.0216)	
$out\_degree_{s_1} \times$					0.0008**	0.0103
$\ln(dist_{c1,c2})$					(0.0003)	
α	-6.3960***		-6.3914***		-6.5543***	
	(0.3594)		(0.2949)		(0.3878)	
Fixed Effects	Yes		Yes		Yes	
Additional Controls	No		No		No	
Adoption Events	14,493		14,493		14,493	
Observations	3,720,405		3,720,405		3,720,405	
Pseudo R^2	0.324		0.324		0.326	

Table 5.2: **Changing Definition of Distance Estimation**. The table reports results of re-estimated model using different definitions of distance. In all cases models with all explanatory variables, fixed effects and isolation control were estimated. In Table 5.1 this corresponds to column (3). Column β^{logit} reports logit coefficients. Column $\hat{\beta}$ report fully standardised coefficients. Terms in the brackets are robust standard errors. Symbols * indicate statistical significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

5.3 Changing Threshold Values

The next set of robustness checks consider the stability of results by changing how the definition of an adoption event is specified. In the first step, a smoothing of the WIOD trade series is applied by taking a rolling average window of 3 years for the annual series, and where adoption events are re-calculated using the same cut-off of 1%. The reason for

	(1)		(2)		(3)	
$P(Adopt_{s_1,s_2}\|X)$	β^{logit}	$\hat{\beta}$	β^{logit}	$\hat{\beta}$	β^{logit}	$\hat{\beta}$
$net\_dist_{s_1,s_2}$	-0.9079***	-0.7183	-0.7693***	-0.5873	-0.7478***	-0.5736
	(0.0209)		(0.0217)		(0.0242)	
$dist_{c_1,c_2}$	-0.0423*	-0.0633	-0.3305***	-0.4773	-0.3711***	-0.5412
	(0.0172)		(0.0253)		(0.0286)	
$dist^2_{c_1,c_2}$	0.0001	0.0025	0.0157***	0.3189	0.0161***	0.3312
	(0.0013)		(0.0016)		(0.0018)	
$out\_degree_{s_1} \times$	-0.0007***	-0.0328	-0.0004*	-0.0195	-0.0001	-0.0055
$dist_{c_1,c_2}$	(0.0001)		(0.0002)		(0.0002)	
$out\_degree_{s_1}$	0.0138***	0.0815	0.0101***	0.0578	0.0085***	0.0489
	(0.0004)		(0.0006)		(0.0006)	
$avg\_dist_{s_1}$	-0.0001***	-0.0514	-0.0001***	-0.0613	-0.0000***	-0.0319
	(0.0000)		(0.0000)		(0.0000)	
$D_{c_1=c_2}$	2.1985***	0.1027	2.3317***	0.1051	2.3571***	0.1076
	(0.0382)		(0.0498)		(0.0534)	
α	-2.8817***		-6.5745***		-6.9873***	
	(0.0731)		(0.4369)		(0.5278)	
Fixed Effects	No		Yes		Yes	
Additional Controls	No		No		Yes	
Adoption Events	10,417		10,417		9,988	
Observations	3,721,547		3,721,547		3,466,419	
Pseudo R^2	0.282		0.328		0.338	

Table 5.3: **Rolling Average Estimation**. The table reports results of re-estimated model using trade series smoother with a 3-year window rolling average. Here column (1)-(3) correspond to column (2)-(4) in Table 5.1. Column β^{logit} reports logit coefficients. Terms in the brackets are robust standard errors. Column $\hat{\beta}$ report fully standardised coefficients. Symbols * indicate statistical significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

this procedure is to check whether the results are not driven by one-off noisy increases in trade. In order to maintain that the values are consistent across the years, the trade data was deflated using industry level intermediate input deflators. Estimations results with the new series are reported in Table 5.3.

In comparison with baseline results, even though the number of adoption events is lower, the impact of most variables is preserved, i.e. the sign, significance and relative magnitude of the coefficients is similar. On the other hand, it can also be seen that standardised coefficients are estimated consistently higher in contrast to baseline specification. Altogether these results suggest that the baseline results are not driven by spurious trade relationships.

An additional set of robustness checks related to using different cut-off thresholds. Specifically, adoption events and relationships between sectors were recalculated using more constrained or relax definitions of what constitutes an economically significant relationship. Formally, a relationship between sectors exists if from importing sectors point of view, the input imports are at least 5% (strict conditions) of total intermediate inputs or at least 0.1% (relaxed condition). The result using new definitions are presented in Table 5.4. The comparison models are estimated using fixed effect and isolation controls.

	Baseline		5prc.		0.1prc	
$P(Adopt_{s_1,s_2}\|X)$	β^{logit}	$\hat{\beta}$	β^{logit}	$\hat{\beta}$	β^{logit}	$\hat{\beta}$
$net\_dist_{s_1,s_2}$	-0.7365***	-0.4876	-0.1476***	-0.2629	-1.7435***	-0.5057
	(0.0159)		(0.0170)		(0.0118)	
$dist_{c_1,c_2}$	-0.3224***	-0.4966	-0.5819***	-0.5097	-0.3153***	-0.4751
	(0.0195)		(0.1596)		(0.0052)	
$dist^2_{c_1,c_2}$	0.0153***	0.3316	0.0210	0.1712	0.0133***	0.2793
	(0.0013)		(0.0114)		(0.0003)	
$out\_degree_{s_1} \times$	-0.0005***	-0.0260	-0.0018	-0.0217	0.0000***	0.0047
$g(dist_{c_1,c_2})$	(0.0001)		(0.0013)		(0.0000)	
$out\_degree_{s_1}$	0.0092***	0.0566	0.0281***	0.0576	0.0017***	0.0440
	(0.0005)		(0.0042)		(0.0000)	
$avg\_dist_{s_1}$	-0.0001***	-0.0861	-0.0001***	-0.0618	-0.0006***	-0.1561
	(0.0000)		(0.0000)		(0.0000)	
$D_{c_1=c_2}$	2.3295***	0.1123	3.0612***	0.2971	2.0298***	0.0569
	(0.0414)		(0.2560)		(0.0213)	
α	-6.3960***		-9.7137**		0.1876	
	(0.3594)		(2.6350)		(0.1334)	
Fixed Effects	Yes		Yes		Yes	
Additional Controls	No		No		No	
Adoption Events	14,493		2,244		81,955	
Observations	3,720,405		275,049		4,800,254	
Pseudo R^2	0.324		0.339		0.256	

Table 5.4: **Changing Threshold Estimation**. The table reports results of re-estimated model using different definition of cut-off for significant trade relationship between sectors. In all cases models with all explanatory variables, fixed effects and isolation control were estimated. Column β^{logit} reports logit coefficients. Terms in the brackets are robust standard errors. Column $\hat{\beta}$ report fully standardised coefficients. Symbols * indicate statistical significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Focusing on the contrast between the baseline 1% definition and 5% it can be seen that while the impact of network proximity is preserved, the impact of distance is much less pronounced. This suggests that for economically stronger relationships, network proximity may be more relevant. It should be also mentioned that 5% threshold limits the sample considerably since this leads to a much greater number of sectors that become isolated in the network. By contrasting baseline definition with the 0.1% impact, it can be seen that all measures are significant as in the baseline. However, differences do come from the distance coefficients as the more relaxed cut-off estimates a stronger impact. Overall, these set of robustness checks support the claim that the key results are not driven by the definitions that were used in the baseline specification. On the other hand, the changing importance of network proximity and distance depending on the strength of the relationship may point to a more nuanced effect.

5.4 Summary of Results

Overall, the empirical analysis using the WIOD supports Hypotheses H1 and H3, whereas evidence for H2 is mixed. As was seen in section 4.2 the distribution of sectors by the number of outward linkages is skewed and exhibits a fat-tail in support of Hypothesis H1. Furthermore, by focusing on the econometric specification, the network proximity measure and the impact of distance are significant and in line with model prediction and Hypothesis H3. This result is robust to changes in the way geographic distance enters the empirical model and under alternative characterisations of trade relationships. On the other hand, when it comes to theoretical predictions that suggest that the negative impact of distance is smaller the more customers a sector has or the further away the customers are, the estimation results do not provide supporting evidence. The effects are in line only when considering alternative distance specification, or when the cut-off threshold is relaxed to 0.1%, hence the support for Hypothesis H2 is inconclusive.

5.5 Limitations

While the results of the empirical investigation support some of the theoretical predictions of the model, limitations need to be kept in mind when interpreting these results. First of all, the results of the theoretical framework rely on symmetry assumptions. In particular, the assumption that all type of firms defined by their set of essential inputs simultaneously exist in all locations does not seem plausible. As it was mentioned before, a more plausible interpretation of this idea would be to look at the set I_t as set of production recipes and locations differ in the extent of utilisation of these recipes. However, future works may consider relaxing this assumption and allowing for greater heterogeneity across locations.

In addition to symmetry assumptions, the behaviour behind new connection formation is mechanical rather than driven by economic considerations. This limits the direct application of this model to studying trade flows. On the other hand, it allows focusing on an understudied aspect relating to informational frictions and how they can be overcome. Hence, while this model does not provide explanations on the intensive margin of sectoral trade, it points to a potential way that the extensive margin may be modelled.

Another limitation of the theoretical framework is that while outward linkages change dynamically and firms acquire new customers over time, the set of inputs that the firm selects is static and is determined when the firm is born. This feature does not capture the fact that in the real world it is more likely that firms are searching for new inputs dynamically. This limitation is also a potential area where extensions of the framework may look to introduce optimising behaviour. For example, firms might continue to draw sets N_i and K_i over time, and choose the optimal input set by profit maximization.

With regards to the econometric estimation of the model, there are also limitations that need to be kept in mind. First of all, as it was mentioned above, yearly trade data may be noisy. Hence for defining binary relationships a cut-off is introduced which constitutes that a relationship exists if from the point of view of input receiving sector, the imports constitute at least 1% of total intermediate input use. While this definition is one that has

been used in the literature before (Carvalho, 2010), a more appealing definition would be to take absolute values in trade as in Carvalho and Voigtländer (2014). In this case, however, countries that enter the WIOD sample are not consistent in their sizes, level of development and engagement in international trade. Hence, a relative measure is more appropriate in this setting.

An additional concern that may be raised is that relative to the number of possible relationships, the actual observed base rate of input adoptions is not high. While for the econometric estimation the number of observations is sufficient for obtaining estimates, there still remains a puzzle why the WIOD sample is sparsely connected. This may be a relevant issue to study in the future and also indicates that longer samples may be needed for a more comprehensive empirical study. Specifically, given that network proximity measures may reflect technological compatibility, which corresponds to more secular, supply-side features a longer sample might be necessary to observe sufficient amount of variation. Related to this issue, while the WIOD contains 54 distinct sectors within each country, a larger variation in sectors is desired to capture present heterogeneity in the types of firms and technologies that are observed these days.

Finally, an important aspect which limits the strength of the conclusions is that the proposed mechanisms that underline theoretical framework are defined at the firm level. Hence, empirical validation of these mechanisms will be limited, given that emergent outcomes are studied, but the firm level behaviour is not directly observed. To causally claim that the proposed mechanisms are at play, future studies should consider using firm level data to study input-output relationships and trade destination choices.

6 Conclusion

The study aimed to develop a framework that could help explain international sectoral network formation. The proposed model includes elements of input-output structure and geographic variation of linkages. The developed model exhibits properties such as: (i) the distribution of firms and sectors by number of linkages is skewed and fat-tailed; (ii) the average distance of connections increases with the higher number of outward linkages; (iii) the likelihood of new link formation depends on network proximity and geographical distance between sectors. From these theoretical predictions we formulated hypotheses which were then tested empirically using the WIOD. Empirical analysis showed that the WIOD features skewed and fat-tailed sector distribution of outward linkages. Furthermore, econometric estimation finds supporting evidence that network proximity and geographic distances impact the likelihood of input adoption. These results are consistent and significant under various robustness checks. Finally, the empirical results of the study do not find conclusive evidence for the prediction that as sectors accumulate more outward linkages, new links are connected at further distances.

Concerning previous studies, the present work is in line with existing evidence of how sectoral networks form. However, conceptually one of the primary driving mechanisms of the model were information barriers. In this setup geographic distances acted as a proxy for these barriers. While this simplified the characterisation of the model and correspondence with the data, in reality, information barriers may be associated with a much broader set of factors. Specifically, it may be better captured by the previously discussed "extended gravity" effect, brought forward by Morales et al. (2017), who argue that in addition to geographic distances, accounting for languages and income per capita levels is important. Introducing these elements to the presented framework may be a potential avenue for further research.

In addition, such network-based approaches may be used in extending existing international trade models. While the presented model is limited in focusing only on the extensive margin of trade, it could characterise the mechanism behind how available sets of buyers and suppliers change over time. For example, Bernard and Moxnes (2018) in reviewing existing literature of networks in international trade, point to the fact that firm level asymmetries are directly related to the number of contacts the firms have.

Moreover, defining sectors by their input use may be a useful approach for linking existing frameworks. This could allow uncovering micro-level behaviour of firms by measuring whether predicted emergent properties are observed in sectoral data. For example, the elasticity of input substitution which has been proposed as a key element for the degree of sectoral shock propagation (Atalay, 2017), may, in fact, reflect firm level inefficiencies in searching for new relationships given the presence of informational barriers.

Bibliography

Acemoglu, D., Akcigit, U., and Kerr, W. (2016a). Innovation Network. *Proceedings of the National Academy of Sciences*, 113(41):11483–11488.

Acemoglu, D., Akcigit, U., and Kerr, W. (2016b). Networks and the Macroeconomy: An Empirical Exploration. *NBER Macroeconomics Annual*, 30:273–335.

Acemoglu, D., Carvalho, V. M., Ozdaglar, A., and Tahbaz-Salehi, A. (2012). The Network Origins of Aggregate Fluctuations. *Econometrica*, 80(5):1977–2016.

Albornoz, F., Calvo Pardo, H. F., Corcos, G., and Ornelas, E. (2012). Sequential Exporting. *Journal of International Economics*, 88:17–31.

Allen, T. (2014). Information Frictions in Trade. *Econometrica*, 82(6):2041–2083.

Armenter, R. and Koren, M. (2015). Economies of Scale and the Size of Exporters. *Journal of the European Economic Association*, 13(3):482–511.

Atalay, E. (2017). How Important Are Sectoral Shocks? *American Economic Journal: Macroeconomics*, 9(4):254–280.

Atalay, E., Hortacsu, A., Roberts, J., and Syverson, C. (2011). Network Structure of Production. *Proceedings of the National Academy of Sciences*, 108(13):5199–5202.

Baqaee, D. R. and Farhi, E. (2017). The Macroeconomic Impact of Microeconomic Shocks: Beyond Hulten's Theorem. *NBER Working Paper*, 23145.

Barrot, J.-N. and Sauvagnat, J. (2016). Input Specificity and the Propagation of Idiosyncratic Shocks in Production Networks. *The Quarterly Journal of Economics*, 131(3):1543–1592.

Bernard, A. B., Eaton, J., Jensen, J. B., and Kortum, S. (2003). Plants and Productivity in International Trade. *American Economic Review*, 93(4):1268–1290.

Bernard, A. B. and Moxnes, A. (2018). Networks and Trade. *Annual Review of Economics*, 10:65–85.

Burstein, A., Kurz, C., and Tesar, L. (2008). Trade, Production Sharing, and the International Transmission of Business Cycles. *Journal of Monetary Economics*, 55(4):775–795.

Carvalho, V. and Gabaix, X. (2013). The Great Diversification and Its Undoing. *American Economic Review*, 103(5):1697–1727.

Carvalho, V. M. (2010). Aggregate Fluctuations and the Network Structure of Intersectoral Trade. *Department of Economics and Business, Universitat Pompeu Fabra Economics Working Papers*, 1206.

Carvalho, V. M. (2014). From Micro to Macro via Production Networks. *The Journal of Economic Perspectives*, 28(4):23–47.

Carvalho, V. M. and Grassi, B. (2015). Large Firm Dynamics and the Business Cycle. *CEPR Discussion Paper*, DP10587.

Carvalho, V. M., Nirei, M., Saito, Y. U., and Tahbaz-Salehi, A. (2016). Supply Chain Disruptions: Evidence From the Great East Japan Earthquake. *Cambridge Working Paper Economics*, 1670.

Carvalho, V. M. and Voigtländer, N. (2014). Input Diffusion and the Evolution of Production Networks. *NBER Working Paper*, 20025.

Chaney, T. (2014). The Network Structure of International Trade. *The American Economic Review*, 104(11):3600–3634.

Chaney, T. (2016). Networks in International Trade. In Bramoulle, Y., Galeotti, A., and Rogers, B., editors, *Oxford Handbook of the Economics of Networks*. Oxford University Press, Oxford.

Chaney, T. (2018). Gravity Equation in International Trade: An Explanation. *Journal of Political Economy*, 126(1):150–177.

Ciccone, A. (2002). Input Chains and Industrialization. *The Review of Economic Studies*, 69(3):565–587.

Conley, T. G. and Dupor, B. (2003). A Spatial Analysis of Sectoral Complementarity. *Journal of Political Economy*, 111(2):311–352.

Defever, F., Heid, B., and Larch, M. (2015). Spatial Exporters. *Journal of International Economics*, 95:145–156.

di Giovanni, J., Levchenko, A., and Mejean, I. (2014). Firms, Destinations, and Aggregate Fluctuations. *Econometrica*, 82(4):1303–1340.

di Giovanni, J., Levchenko, A., and Mejean, I. (2018). The Micro Origins of International Business Cycle Comovement. *American Economic Review*, 108(1):82–108.

Dijkstra, E. W. (1959). A Note on Two Problems in Connexion With Graphs. *Numerische Mathematik*, 1:269–271.

Dupor, B. (1999). Aggregation and Irrelevance in Multi-Sector Models. *Journal of Monetary Economics*, 43(2):391–409.

Eaton, J., Kortum, S., and Kramarz, F. (2011). An Anatomy of International Trade: Evidence from French Firms. *Econometrica*, 79(5):1453–1498.

Foerster, A. T., Sarte, P.-D. G., and Watson, M. W. (2011). Sectoral versus Aggregate Shocks: A Structural Factor Analysis of Industrial Production. *Journal of Political Economy*, 119(1):1–38.

Fruchterman, T. M. J. and Reingold, E. M. (1991). Graph Drawing by Force-directed Placement. *Software: Practice and Experience*, 21(11):1129–1164.

Gabaix, X. (2011). The Granular Origins of Aggregate Fluctuations. *Econometrica*, 79(3):733–772.

Gabaix, X. (2016). Power Laws in Economics: An Introduction. *Journal of Economic Perspectives*, 30(1):185–206.

Garmendia, A., Llano, C., Minondo, A., and Requena, F. (2012). Networks and the Disappearance of the Intranational Home Bias. *Economics Letters*, 116(2):178–182.

Grassi, B. (2017). IO in I-O : Size , Industrial Organization and the Input-Output Network Make a Firm Structurally Important. *Unpublished Manuscript*.

Helpman, E., Melitz, M., and Rubinstein, Y. (2008). Estimating Trade Flows: Trading Partners and Trading Volumes. *The Quarterly Journal of Economics*, 123(2):441–487.

Horvath, M. (1998). Cyclicality and Sectoral Linkages: Aggregate Fluctuations from Independent Sectoral Shocks. *Review of Economic Dynamics*, 1(4):781–808.

Hulten, C. R. (1978). Growth Accounting with Intermediate Inputs. *The Review of Economic Studies*, 45(3):511–518.

Jackson, M. O. and Rogers, B. W. (2007). Meeting Strangers and Friends of Friends: How Random Are Social Networks? *American Economic Review*, 97(3):890–915.

Jones, C. I. (2011). Intermediate Goods and Weak Links in the Theory of Economic Development. *American Economic Journal: Macroeconomics*, 3(2):1–28.

Lim, K. (2017). Firm-to-Firm Trade in Sticky Production Networks. *Unpublished Manuscript*.

Long, J. B. J. and Plosser, C. I. (1983). Real Business Cycles. *Journal of Political Economy*, 91(1):39–69.

Long, J. S. and Freese, J. (2014). *Regression Models for Categorical Dependent Variables Using Stata*. Stata Press, Third edition.

Los, B., Timmer, M. P., and de Vries, G. J. (2015). How Global Are Global Value Chains? A New Approach to Measure International Fragmentation. *Journal of Regional Science*, 55(1):66–92.

Lucas, R. E. (1977). Understanding Business Cycles. In *Carnegie-Rochester Conference Series on Public Policy*.

Melitz, M. J. (2003). The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity. *Econometrica*, 71(6):1695–1725.

Morales, E., Sheu, G., and Zahler, A. (2017). Gravity and Extended Gravity: Using Moment Inequalities to Estimate a Model of Export Entry. *NBER Working Paper*, 19916.

Oberfield, E. (2018). A Theory of Input-Output Architecture. *Econometrica*, 86(2):559–589.

Ozdagli, A. K. and Weber, M. (2017). Monetary Policy Through Production Networks: Evidence from the Stock Market. *NBER Working Paper*, 23424.

Pasten, E., Schoenle, R., and Weber, M. (2018). Price Rigidities and the Granular Origins of Aggregate Fluctuations. *NBER Working Paper*, 23750.

Rauch, J. E. (1999). Networks Versus Markets in International Trade. *Journal of International Economics*, 48:7–35.

Rauch, J. E. (2001). Business and Social Networks in International Trade. *Journal of Economic Literature*, 39(4):1177–1203.

Timmer, M. P., Dietzenbacher, E., Los, B., Stehrer, R., and 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.

Timmer, M. P., Erumban, A. A., Los, B., Stehrer, R., and de Vries, G. J. (2014). Slicing Up Global Value Chains. *Journal of Economic Perspectives*, 28(2):99–118.

A Appendix

A.1 Customer Acquisition Process Independence

The starting point of the derivation is the customer acquisition equation:

$$f_{i,t+1}(c) - f_{i,t}(c) = \gamma m_K \cdot g(c_d, c) + \frac{\gamma m_N}{m} \sum_{c_d \in C} f_{i,t}(c_d) \cdot g(c_d, c) \tag{A.1.1}$$

To see that the results do not depend on the location matching probability function $g(.,.)$, consider adding up both the left-hand side and the right-hand side of Equation (A.1.1) across the location set C. First, consider the left-hand side:

$$\sum_{c \in C} \big[f_{i,t+1}(c) - f_{i,t}(c)\big] = \sum_{c \in C} \big[f_{i,t+1}(c)\big] - \sum_{c \in C} \big[f_{i,t}(c)\big] = d_i(t+1) - d_i(t) \tag{A.1.2}$$

Hence, the left-hand side becomes the change in total number of customers over time. Next, consider the first term of the right-hand side. Remember that $g(c_0, c)$ is a distribution function. Furthermore, if we fix location c_0, then sum of $g(c_0, c)$ of all possible values of $c \in C$ is equal to 1. This implies:

$$\sum_{c \in C} \gamma m_K \cdot g(c_0, c) = \gamma m_K \sum_{c \in C} g(c_0, c) = \gamma m_K \tag{A.1.3}$$

The next summation term requires more attention. Note that the expression is:

$$\sum_{c \in C} \big[\frac{\gamma m_N}{m} \sum_{c_d \in C} f_{i,t}(c_d) \cdot g(c_d, c)\big] = \frac{\gamma m_N}{m} \sum_{c \in C} \sum_{c_d \in C} \big[f_{i,t}(c_d) \cdot g(c_d, c)\big] \tag{A.1.4}$$

When adding the first sum with respect to c_d, both $f_{i,t}(.)$ and $g(.,.)$ are varying, which does not directly converge to a useful result. On the other hand, given that both summations are discrete, but infinite ($C = \mathbb{Z}$), and that all the components of the sum are always greater than zero, but finite, we can reverse the order of summation. In which case, when c is varied, but c_d is kept fixed, the term $g(c_d, c)$ reduces to 1, due to the same probability distribution feature. This leads to:

$$\frac{\gamma m_N}{m} \sum_{c_d \in C} \sum_{c \in C} \big[f_{i,t}(c_d) \cdot g(c_d, c)\big] = \frac{\gamma m_N}{m} \sum_{c_d \in C} f_{i,t}(c_d) = \frac{\gamma m_N}{m} \cdot d_i(t) \tag{A.1.5}$$

The second step of the summation directly follows from the definition in Equation (3.2.1). Combining Equations (A.1.2), (A.1.3), and (A.1.5) we get:

$$d_i(t+1) - d_i(t) = \gamma m_K + d_i(t) \cdot \frac{\gamma m_N}{m} \tag{A.1.6}$$

As we can see from Equation (A.1.6), the total number of customers that a firm has is independent from location matching probability function $g(.,.)$.

A.2 Solution to Customer Acquisition Difference Equation

Equation (3.2.5) is a linear first-order autonomous difference equation, with initial condition $d_i(t_0) = 0$:

$$d_i(t+1) - d_i(t) = \gamma m_K + d_i(t) \cdot \frac{\gamma m_N}{m} \tag{A.2.1}$$

First, we find the steady-state value ($\bar{d_i}$) by setting $d_i(t+1) = d_i(t)$:

$$0 = \gamma m_K + \bar{d_i}\frac{\gamma m_N}{m} \tag{A.2.2}$$

$$\bar{d_i} = -m\frac{m_K}{m_N} \tag{A.2.3}$$

$$\tag{A.2.4}$$

To simplify notation we define $r = \frac{m_K}{m_N}$. This can also be shown to imply: $\frac{m}{m_N} = \frac{m_N + m_K}{m_N} = 1 + r$. This leads to the steady-state solution:

$$\bar{d_i} = -rm \tag{A.2.5}$$

Next, we solve for the general form ($\tilde{d_i}$) by setting all arguments not relating to d_i to zero:

$$d_i(t+1) - d_i(t) = d_i(t)\frac{\gamma m_N}{m} \tag{A.2.6}$$

$$d_i(t+1) = d_i(t)\left[1 + \frac{\gamma m_N}{m}\right] \tag{A.2.7}$$

$$\tilde{d_i} = C \cdot \left[1 + \frac{\gamma m_N}{m}\right]^t \tag{A.2.8}$$

$$\tilde{d_i} = C \cdot \left[1 + \frac{\gamma}{1+r}\right]^t \tag{A.2.9}$$

Combining steady-state and general solutions ($d_i = \bar{d_i} + \tilde{d_i}$) results in:

$$d_i(t) = C \cdot \left[1 + \frac{\gamma}{1+r}\right]^t - rm \tag{A.2.10}$$

Using the initial condition, which states that $d_i(t_0) = 0$, we can find the constant C:

$$0 = C \cdot \left[1 + \frac{\gamma}{1+r}\right]^{t_0} - rm \tag{A.2.11}$$

$$C = \frac{rm}{\left[1 + \frac{\gamma}{1+r}\right]^{t_0}} \tag{A.2.12}$$

Using the solution of C, the final solution of the difference equation takes the form:

$$d_i(t) = \frac{rm}{\left[1+\frac{\gamma}{1+r}\right]^{t_0}}\left[1+\frac{\gamma}{1+r}\right]^{t} - rm \tag{A.2.13}$$

$$d_i(t) = rm\left[\left(1+\frac{\gamma}{1+r}\right)^{t-t_0} - 1\right] \tag{A.2.14}$$

A.3 Analytical Features of Firm Distribution

Re-arranging Equation (3.3.11) for the counter-cumulative distribution and applying the logarithmic function to both sides results in:

$$\log[1-F_t(d)] = \log\left[\left(\frac{rm}{d+rm}\right)^{\log(1+\gamma)\cdot\left(\log(1+\frac{\gamma}{1+r})\right)^{-1}}\right] \tag{A.3.1}$$

$$= \frac{\log(1+\gamma)}{\log(1+\frac{\gamma}{1+r})}\cdot\log\left[\frac{rm}{d+rm}\right] \tag{A.3.2}$$

$$= \frac{\log(1+\gamma)}{\log(1+\frac{\gamma}{1+r})}\cdot\Big[\log(rm)-\log(d+rm)\Big] \tag{A.3.3}$$

Define $Z_t(d) = \log[1-F_t(d)]$. Next, focus on how $Z_t(d)$ changes with respect to $\log(d)$. This can be obtained by applying the chain rule:

$$\frac{\partial Z_t(d)}{\partial\log(d)} = \frac{\partial Z_t(d)}{\partial\log(d)}\frac{\partial d}{\partial d} = \frac{\partial Z_t(d)}{\partial d}\left(\frac{\partial\log(d)}{\partial d}\right)^{-1} = \frac{\partial Z_t(d)}{\partial d}\cdot d \tag{A.3.4}$$

Hence the derivative is equal to:

$$\frac{\partial Z_t(d)}{\partial\log(d)} = -\frac{\log(1+\gamma)}{\log(1+\frac{\gamma}{1+r})}\cdot\frac{1}{d+rm}\cdot d \tag{A.3.5}$$

The first things that can be observed is what happens when d becomes large ($d\to\infty$). Since $\lim_{d\to\infty}\frac{d}{d+rm} = 1$ the Equation (A.3.5) becomes a constant, and the distribution in the log-log case is approximately linear when d is large (right-tail).

Next we can focus on what happens when d is small (left-tail). However, since $\lim_{d\to 0}\frac{d}{d+rm} = 0$ to obtain a descriptive property we focus on the concavity of the left-tail. In particular, since the functional form in Equation (A.3.3) of the distribution is differentiable for $d > rm$, we can test the concavity of the function by second-order partial derivative. First denote the result of Equation (A.3.5) as $Z_t'(d)$. Then we can use the previous chain rule to get the result:

$$\frac{\partial Z_t'(d)}{\partial \log(d)} = \frac{\partial Z_t'(d)}{\partial d} \cdot d \tag{A.3.6}$$

$$\frac{\partial Z_t'(d)}{\partial \log(d)} = -\frac{\log(1+\gamma)}{\log(1+\frac{\gamma}{1+r})} \cdot \frac{(d+rm-d)}{(d+rm)^2} \cdot d \tag{A.3.7}$$

$$= -\frac{\log(1+\gamma)}{\log(1+\frac{\gamma}{1+r})} \cdot \frac{rm \cdot d}{(d+rm)^2} < 0 \tag{A.3.8}$$

Hence the distribution is concave in the left-tail when d is small.

A.4 Solution for Difference Equation of $\hat{f}_t(\omega)$

This part covers the solution of the difference equation of Fourier transformed variable $\hat{f}_t(\omega)$:

$$\hat{f}_t(\omega) = rm \cdot \left[\left(1 + \frac{\gamma \hat{g}(\omega)}{1+r}\right)^{t-t_0} - 1 \right] \tag{A.4.1}$$

First, solving for steady-state by setting $\hat{f}_t(\omega) = \hat{f}_{t+1}(\omega)$ we get:

$$0 = \gamma m_K \cdot \hat{g}(\omega) + \frac{\gamma m_N}{m} \hat{f}_t(\omega) \cdot \hat{g}(\omega) \tag{A.4.2}$$

$$\hat{f}_t(\omega) = -rm \tag{A.4.3}$$

Next, solving for general form by setting all components not related to $\hat{f}_t(\omega)$ to zero:

$$\hat{f}_{t+1}(\omega) - \hat{f}_t(\omega) = \frac{\gamma m_N}{m} \hat{f}_t(\omega) \cdot \hat{g}(\omega) \tag{A.4.4}$$

$$\hat{f}_{t+1}(\omega) = \hat{f}_t(\omega) \left[1 + \frac{\gamma m_N}{m} \cdot \hat{g}(\omega) \right] \tag{A.4.5}$$

$$\hat{f}_t(\omega) = C \left[1 + \frac{\gamma m_N}{m} \cdot \hat{g}(\omega) \right] \tag{A.4.6}$$

$$\hat{f}_t(\omega) = C \cdot \left[1 + \frac{\gamma \hat{g}(\omega)}{1+r} \right]^t \tag{A.4.7}$$

Combining Equations (A.4.3) and (A.4.7) the solution leads to:

$$\hat{f}_t(\omega) = -rm + C \cdot \left[1 + \frac{\gamma \hat{g}(\omega)}{1+r} \right]^t \tag{A.4.8}$$

Since $\hat{f}_{t=t_0} = 0$, i.e. number of customers in a specific location when a firm is born is zero, we get:

$$0 = -rm + C \cdot \left[1 + \frac{\gamma \hat{g}(\omega)}{1+r}\right]^{t_0} \tag{A.4.9}$$

$$C = \frac{rm}{\left[1 + \frac{\gamma \hat{g}(\omega)}{1+r}\right]^{t_0}} \tag{A.4.10}$$

Using the combined expression the final solution is:

$$\hat{f}_t(\omega) = rm \cdot \left[1 + \frac{\gamma \hat{g}(\omega)}{1+r}\right]^{t-t_0} - rm \tag{A.4.11}$$

$$\hat{f}_t(\omega) = rm \cdot \left[\left(1 + \frac{\gamma \hat{g}(\omega)}{1+r}\right)^{t-t_0} - 1\right] \tag{A.4.12}$$

A.5 Solution For Average Distance of Customers

First we find the first and second derivatives of the Fourier transformed variable $\hat{g}_t$:

$$\hat{g_t}'(\omega) = \frac{\gamma \hat{g}'(\omega)}{1+r}(t-t_0) \cdot (1 + \frac{\gamma \hat{g}(\omega)}{1+r})^{t-t_0-1} \cdot \left[(1 + \frac{\gamma}{1+r})^{t-t_0} - 1\right]^{-1} \tag{A.5.1}$$

$$\begin{aligned}\hat{g_t}''(\omega) =& \frac{(t-t_0)\gamma \hat{g}''(\omega)}{1+r}\left(1 + \frac{\gamma \hat{g}(\omega)}{1+r}\right)^{t-t_0-1} + \left(\frac{\gamma \hat{g}'(\omega)}{1+r}\right)^2 \cdot (t-t_0) \cdot (t-t_0-1) \\ & \left[1 + \frac{\gamma \hat{g}(\omega)}{1+r}\right]^{t-t_0-2} \cdot \left[(1 + \frac{\gamma}{1+r})^{t-t_0} - 1\right]^{-1} \end{aligned} \tag{A.5.2}$$

$$\begin{aligned}=& \left[\hat{g}''(\omega)\left(1 + \frac{\gamma \hat{g}(\omega)}{1+r}\right)^{t-t_0-1} + \frac{(t-t_0)\gamma}{1+r} \cdot (\hat{g}'(\omega))^2 \cdot (t-t_0-2) \cdot \left(1 + \frac{\gamma \hat{g}(\omega)}{1+r}\right)^{t-t_0-2}\right] \\ & \left[(1 + \frac{\gamma}{1+r})^{t-t_0} - 1\right]^{-1} \cdot \frac{(t-t_0)\gamma}{1+r} \end{aligned} \tag{A.5.3}$$

Next, note that since distance from where the firm is located is symmetric around zero, the first moment is equal to zero: $\hat{g_t}'(0) = 0$. Also note that $\hat{g}(0) = 1$, since the Fourier transformation is applied to a probability density function. Using previous definition of Δ_t as the average squared distance of a firm at time t that is of age $(t - t_0)$. This would be the second moment of C_t, which corresponds to:

$$\Delta_t \equiv \sum_{c \in C} c^2 g_t(|c|) = E[C_t^2] = \hat{g_t}''(0) \tag{A.5.4}$$

$$= \hat{g}''(0) \cdot \frac{(t-t_0)\gamma}{1+r} \cdot \left(1 + \frac{\gamma \hat{g}(0)}{1+r}\right)^{t-t_0-1} \cdot \left[(1 + \frac{\gamma}{1+r})^{t-t_0} - 1\right]^{-1} \tag{A.5.5}$$

$$= \left[\sum_{c \in C} c^2 g(|c|)\right] \cdot \frac{(t-t_0)\gamma}{1+r} \cdot \left(1 + \frac{\gamma}{1+r}\right)^{t-t_0-1} \cdot \left[(1 + \frac{\gamma}{1+r})^{t-t_0} - 1\right]^{-1} \tag{A.5.6}$$

An important result that can be obtained is how the average squared distance changes based on the number of customers d. To obtain this, we need to replace all terms containing $(t - t_0)$ with terms that depend on d. To do this, we can use previously obtained results:

$$(t - t_0) = \ln\big(\frac{d + rm}{rm}\big) \cdot \left[\ln\big(1 + \frac{\gamma}{1+r}\big)\right]^{-1} \tag{A.5.7}$$

$$\big(1 + \frac{\gamma}{1+r}\big)^{t-t_0} = \frac{d+rm}{rm} \tag{A.5.8}$$

$$\big(1 + \frac{\gamma}{1+r}\big)^{t-t_0} = \frac{d+rm}{rm} \Rightarrow \tag{A.5.9}$$

$$\big(\frac{1+r+\gamma}{1+r}\big)^{t-t_0-1} = \frac{d+rm}{rm} \cdot \frac{1+r}{1+r+\gamma} \tag{A.5.10}$$

Applying these expressions to Equation (A.5.6), we get moments with respect to number of customers Δ_d:

$$\Delta_d = \big[\sum_{c \in C} c^2 g(|c|)\big] \cdot \frac{\gamma}{1+r} \cdot \ln\big(\frac{d+rm}{rm}\big) \cdot \left(\ln\big(1 + \frac{\gamma}{1+r}\big)\right)^{-1} \cdot \tag{A.5.11}$$

$$\frac{d+rm}{rm} \cdot \frac{1+r}{1+r+\gamma} \cdot \left[\frac{d+rm}{rm} - 1\right]^{-1} \tag{A.5.12}$$

$$= \big[\sum_{c \in C} c^2 g(|c|)\big] \cdot \frac{\gamma}{1+r+\gamma} \cdot \left[\ln\big(1 + \frac{\gamma}{1+r}\big)\right]^{-1} \cdot \ln\big(\frac{d+rm}{rm}\big) \cdot \frac{d+rm}{rm} \cdot \frac{rm}{d} \tag{A.5.13}$$

$$= \big[\sum_{c \in C} c^2 g(|c|)\big] \cdot \frac{\gamma}{1+r+\gamma} \cdot \left[\ln\big(1 + \frac{\gamma}{1+r}\big)\right]^{-1} \cdot \ln\big(1 + \frac{d}{rm}\big) \cdot \left(1 + \frac{rm}{d}\right) \tag{A.5.14}$$

Since the interest of this study is how Δ_d, the average squared distance, changes as we increase number of customers d we can combine all components not containing d to a constant:

$$\Delta_d = A \cdot \ln\big(1 + \frac{d}{rm}\big) \cdot \left(1 + \frac{rm}{d}\right) \tag{A.5.15}$$

$$A = \big[\sum_{c \in C} c^2 g(|c|)\big] \cdot \frac{\gamma}{1+r+\gamma} \cdot \left[\ln\big(1 + \frac{\gamma}{1+r}\big)\right]^{-1} \tag{A.5.16}$$

A.6 Sectoral Likelihood of Adoption

To derive results regarding likelihood of sectoral linkage formation, consider the probability that a firm located in c_0, classified into sector $s_j$would adopt firm i as an input. Note that since essential inputs are used for classification into sectors, this can only happen via network search. Hence, this event can occur only if a new firm draws essential inputs, that define sector s_j and from that neighbourhood it chooses firm i. The likelihood of this occurring can be expressed as the number of firms that define sector s_j, that use i ($i_{s_j}(c)$), divided by the total neighbourhood of varieties that define sector s_j: $x \cdot m$. Since in expectation the new firm will use k_{s_j} inputs of the sector in which it is defined, it will have k_{s_j} draws to use

i as an input. Furthermore, this also needs to be accounted by the probability of matching in location c, starting the search from c_0 ($g(c_0, c)$). The remaining possibility of adoption would come from previously defined network search process $(m_K - k_{s_j})\frac{f_{i,t}(c)}{n_t}$. Aggregating over all locations leads to:

$$\sum_{c \in C} g(c_0, c) \cdot \left(k_{s_j} \frac{i_{s_j}(c)}{x \cdot m} + (m_K - k_{s_j}) \frac{f_{i,t}(c)}{n_t} \right) \tag{A.6.1}$$

In expectation we have that values for k_{s_j}, x, m_K, n_t and $f_{i,t}(c)$ are the same for all sectors s_j. Hence, the expression can be rearranged to:

$$\sum_{c \in C} \frac{k_{s_j}}{x \cdot m} \cdot g(c_0, c) \cdot i_{s_j}(c) + \sum_{c \in C} g(c_0, c)(m_K - k_{s_j}) \frac{f_{i,t}(c)}{n_t} = \tag{A.6.2}$$

$$\frac{k_{s_j}}{x \cdot m} \sum_{c \in C} g(c_0, c) \cdot i_{s_j}(c) + \frac{m_K - k_{s_j}}{n_t} \sum_{c \in C} g(c_0, c) \cdot f_{i,t}(c) \tag{A.6.3}$$

Since the second term does not depend on s_j, we can proceed by focusing on the first term. The above expression relates to a single firm, hence we can extend this to sectors by taking sector s_i in country c_i and adding up across all firm that define it.

$$\sum_{i \in s_i} \frac{k_{s_j}}{x \cdot m} \sum_{c \in C} g(c_0, c) \cdot i_{s_j}(c) = \frac{k_{s_j}}{x \cdot m} \sum_{i \in s_i} \sum_{c \in C} g(c_0, c) \cdot i_{s_j}(c) \tag{A.6.4}$$

As before, given that summation terms are all greater than zero, but finite, we can reverse the sums to get:

$$\frac{k_{s_j}}{x \cdot m} \sum_{c \in C} \sum_{i \in s_i} g(c_0, c) \cdot i_{s_j}(c) = \frac{k_{s_j}}{x \cdot m} \sum_{c \in C} g(c_0, c) \sum_{i \in s_i} i_{s_j}(c) = \frac{k_{s_j}}{x \cdot m} \sum_{c \in C} g(c_0, c) \cdot \eta_{s_i, s_j}(c) \tag{A.6.5}$$

If we take two sectors s_j and s'_j, located respectively in locations c_j and c'_j, the only differences from Equation (A.6.5) would come from the $\sum_{c \in C} g(c_0, c) \cdot \eta_{s_i, s_j}(c)$ term. Specifically, we say that s_j, located in c_j is more likely to adopt sector s_i located in c_i as input than s'_j located in c'_j if:

$$\sum_{c \in C} g(c_j, c) \cdot \eta_{s_i, s_j}(c) > \sum_{c \in C} g(c'_j, c) \cdot \eta_{s_i, s'_j}(c) \tag{A.6.6}$$

A.7 List of Sample Countries and Industries

ID	Industry
1	Crop and animal production, hunting and related service activities
2	Forestry and logging
3	Fishing and aquaculture
4	Mining and quarrying
5	Manufacture of food products, beverages and tobacco products
6	Manufacture of textiles, wearing apparel and leather products
7	Manufacture of products of wood and cork
8	Manufacture of paper and paper products
9	Printing and reproduction of recorded media
10	Manufacture of coke and refined petroleum products
11	Manufacture of chemicals and chemical products
12	Manufacture of basic pharmaceutical products and preparations
13	Manufacture of rubber and plastic products
14	Manufacture of other non-metallic mineral products
15	Manufacture of basic metals
16	Manufacture of fabricated metal products
17	Manufacture of computer, electronic and optical products
18	Manufacture of electrical equipment
19	Manufacture of machinery and equipment n.e.c.
20	Manufacture of motor vehicles, trailers and semi-trailers
21	Manufacture of other transport equipment
22	Manufacture of furniture; other manufacturing
23	Repair and installation of machinery and equipment
24	Electricity, gas, steam and air conditioning supply
25	Water collection, treatment and supply
26	Waste collection, treatment and disposal activities; materials recovery
27	Construction
28	Wholesale and retail trade and repair of motor vehicles and motorcycles
29	Wholesale trade, except of motor vehicles and motorcycles
30	Retail trade, except of motor vehicles and motorcycles
31	Land transport and transport via pipelines
32	Water transport
33	Air transport
34	Warehousing and support activities for transportation
35	Postal and courier activities
36	Accommodation and food service activities
37	Publishing activities
38	Programming and broadcasting activities
39	Telecommunications
40	Computer programming, consultancy and related activities
41	Financial service activities, except insurance and pension funding
42	Insurance, reinsurance and pension funding
43	Activities auxiliary to financial services and insurance activities
44	Real estate activities
45	Legal, consultancy and accounting activities
46	Architectural and engineering activities; technical testing and analysis
47	Scientific research and development
48	Advertising and market research
49	Other professional, scientific and technical activities; veterinary activities
50	Administrative and support service activities
51	Public administration and defence; compulsory social security
52	Education
53	Human health and social work activities
54	Other service activities

Table A.1: **Sample Industries**. List of industries that were chosen for empirical investigation of sectoral network formation in WIOD.

ISO-3	Country	ISO-3	Country
AUS	Australia	IRL	Ireland
AUT	Austria	ITA	Italy
BEL	Belgium	JPN	Japan
BGR	Bulgaria	KOR	South Korea
BRA	Brazil	LTU	Lithuania
CAN	Canada	LUX	Luxembourg
CHE	Switzerland	LVA	Latvia
CHN	China	MEX	Mexico
CYP	Cyprus	MLT	Malta
CZE	Czech Republic	NLD	Netherlands
DEU	Germany	NOR	Norway
DNK	Denmark	POL	Poland
ESP	Spain	PRT	Portugal
EST	Estonia	ROU	Romania
FIN	Finland	RUS	Russia
FRA	France	SVK	Slovakia
GBR	United Kingdom	SVN	Slovenia
GRC	Greece	SWE	Sweden
HRV	Croatia	TUR	Turkey
HUN	Hungary	TWN	Taiwan
IDN	Indonesia	USA	United States of America

Table A.2: **Sample Countries**. List of countries that were chosen for empirical investigation of sectoral network formation in WIOD. ISO-3 codes are abbreviations used in plotting the complete network in Figure 4.1

A.8 Summary Statistics of Explanatory Variables

	Mean	S.D.	Min.	25%	50%	75%	Max.
$net\_dist_{s_1,s_2}$	5.35	1.92	2.00	4.00	5.00	6.00	15.00
$dist_{c_1,c_2}$	5.03	4.46	0.00	1.19	2.64	8.71	18.23
$out\_degree_{s_1}$	19.78	17.82	1.00	6.00	15.00	28.00	213.00
$isolation$	4.74	2.57	3.08	3.23	3.40	4.29	13.92
$\log(EMP_{s_1})$	4.40	1.77	0.00	3.09	4.33	5.58	11.36
$\log(VA_{s_1}/EMP_{s_1})$	0.87	2.17	-4.44	-0.57	0.13	1.78	10.74
$\log(EMP_{s_2})$	3.87	1.97	0.00	2.48	3.81	5.21	11.36
$\log(VA_{s_2}/EMP_{s_2})$	0.48	2.19	-4.44	-0.87	-0.13	1.48	10.74

Table A.3: **Summary Statistics of Explanatory Variables**. Different moments of variables used for explaining the likelihood of input adoption. Note that $dist_{c_1,c_2}$ and $isolation$ are measured in thousands of kilometres. Estimated coefficients for $isolation$ and logged value of employment and labour productivity are not reported, but are part of sets of controls.

A.9 Linear Probability Model Estimation Results

	(1)		(2)		(3)		(4)	
$P(Adopt_{s_1,s_2}\|X)$	β^{OLS}	$\hat{\beta}$	β^{OLS}	$\hat{\beta}$	β^{OLS}	$\hat{\beta}$	β^{OLS}	$\hat{\beta}$
$net\_dist_{s_1,s_2}$	-0.0028***	-0.0857	-0.0009***	-0.0267	-0.0009***	-0.0291	-0.0009***	-0.0269
	(0.0000)		(0.0000)		(0.0000)		(0.0000)	
$dist_{c_1,c_2}$			0.0002***	0.0121	-0.0002***	-0.0130	-0.0002***	-0.0127
			(0.0000)		(0.0000)		(0.0001)	
$dist^2_{c_1,c_2}$			0.0000***	0.0129	0.0000***	0.0254	0.0000***	0.0243
			(0.0000)		(0.0000)		(0.0000)	
$out\_degree_{s_1} \times$			-0.0000***	-0.0586	-0.0000***	-0.0610	-0.0000***	-0.0598
$dist_{c_1,c_2}$			(0.0000)		(0.0000)		(0.0000)	
$out\_degree_{s_1}$			0.0003***	0.0726	0.0003***	0.0726	0.0002***	0.0692
			(0.0000)		(0.0000)		(0.0000)	
$avg\_dist_{s_1}$			-0.0000***	-0.0026	-0.0000***	-0.0055	-0.0000***	-0.0028
			(0.0000)		(0.0000)		(0.0000)	
$D_{c_1=c_2}$			0.0875***	0.1960	0.0870***	0.1948	0.0862***	0.1989
			(0.0011)		(0.0011)		(0.0011)	
α	0.0188***		0.0030***		0.0032***		0.0029**	
	(0.0002)		(0.0002)		(0.0006)		(0.0009)	
Fixed Effects	No		No		Yes		Yes	
Additional Controls	No		No		No		Yes	
Adoption Events	14,493		14,493		14,493		13,837	
Observations	3,720,405		3,720,405		3,720,405		3,464,643	
R^2	0.007		0.044		0.048		0.049	

Table A.4: **LPM Estimation**. Column (1) reports estimates of a model using only network proximity to predict adoption. Columns (2) report results using all previously described explanatory variables, except for controls. Column (3) reports results with all explanatory variables, fixed effect and isolation controls, whereas (4) reports results with additional employment and productivity controls. Column β^{logit} reports OLS coefficients. Terms in the brackets are robust standard errors. Column $\hat{\beta}$ report fully standardised coefficients. Symbols * indicate statistical significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.