Shifts in Comparative Advantage and Industrial Structures when Production is Internationally Fragmented

Xianjia Ye (University of Groningen, Netherlands)

Paper Prepared for the IARIW 33rd General Conference

Rotterdam, the Netherlands, August 24-30, 2014

Session 4C

Time: Tuesday, August 26, Afternoon

Shifts in Comparative Advantage and Industrial Structure when Production is Internationally Fragmented

Xianjia Ye

Global Economics and Management Department,
Faculty of Economics and Business, University of Groningen
Email: x.ye[#]rug.nl

This version: 2014-JUL-20

Abstract

This paper analyses the determinants of structural change by investigating the relatedness between different skilled activities of each industry. I use a similar framework as Hidalgo et al (2007) where the relatedness between activities is proxied by the co-existence of their revealed comparative advantage. I control for offshoring and production fragmentation by using data on value-added export of each activity, derived from the World Input-Output Database. By visualising the relatedness in a network graph, I find activities in different industries at a same skill level are quite close with each other, while within-industry skill upgrading is more difficult. Using these relatedness indices, I perform probit regressions and show that countries tend to gain new comparative advantages in activities that have higher potential for economic growth and have proximity with the current economic structure.

1. Introduction

Industrial structural change refers to the re-allocation of labour, capital and other resources between different activities. It plays an important role in accelerating economic growth, especially for developing countries where there is a considerable productivity gap between traditional and modern economic activities. But what are the determinants of structural change? This question has since long attracted attention from both economists and the government. In the literature structural change is characterized as a complex and highly endogenous progress. The notable driven forces are savings and the accumulation of capital, and the changes in consumer preferences towards industrial products, which are interwoven with the growth in income per capita (see, e.g. Chenery 1960, Kuznets 1957, and more recent models like Echevarria 1997 and Laitner 2000).

However, a practical question remains unanswered, namely why different countries have very different paths and speeds of structural change, and what determines a country's direction of industrial upgrading. This paper tries to answer it by investigating the *relatedness between various activities* in each industry at different skill levels. Briefly, I will first illustrate the structure of relatedness between different type of activities, which gives information concerning the potential upgrading paths for countries with different initial conditions. Then I will show the activities that have high potential for growth and have close relatedness with current economic structure tend to gain comparative advantage.

My analysis of activity relatedness follows the "product space" framework by Hidalgo et al (2007). The basic idea in Hidalgo and this paper is that the changes in industrial structure are path-dependent; the cost and feasibility in developing a certain product (or activity in this paper) is dependent on what kind of activity the country currently have. It is easier to develop an activity that has "proximity" with the current economic structure, namely new activities that has similar labour, capital, and technological requirements as those activities that are already in place. The bilateral relatedness index between products (activities) can be calculated, and visualised in a network map. Hidalgo et al give an analogy that compares the progress of structural change with monkey's life in a forest. Products (activities) are the trees, and the density of fruits in trees represents the potential of each product (activity). Firms are the monkeys that live in the trees. Structural change is then reflected by monkeys jumping towards the trees that give more fruits. Monkeys must also consider the distance between two trees; if two trees are too far-away then it is not possible to make a jump.

My paper improves upon current literature by providing a better measurement of relatedness which is based on value-added export of activities. The aspect of structural change in the current literature is based on industries or products: Hidalgo et al (2007) investigate the probability whether a country have revealed comparative advantage (RCA) in both products. Teece et al. (1994), Bryce and Winter (2009) and Neffke et al (2011) examine the co-occurrence of products that belong to different industries within a firm's portfolio, or within a plant. These measurements rely on the assumption that the output or export of industrial/product properly reflect the actual tasks that take place in a country or a company. But in recent decades, due to the reduction in shipping and coordinating costs, different activities within a production chain can be unbundled to different countries, to make use of their comparative advantages at task-level (Baldwin, 2006, Grossman and Rossi-Hansberg 2006). The "basic unit" of trade changes from finished good, to each task in the production process; the measurement based on products becomes inconsistent in capturing the actual activities.

Particularly, gross export of products, as in Hidalgo et al (2007), becomes unsuitable for the analysis. Firstly it is now no longer suitable to judge the development of a country by looking at what type of goods it export. Poor countries may handle low-skill intensive tasks in the value chain of high-tech product, and vice versa rich countries may focus on high-skilled procedures in traditional industries. More importantly, gross export data includes double-counted value-added components therefore is misaligned with actual value-added creation under global production fragmentation. The value-added of those imported intermediate goods is not generated by the host country but is nevertheless included in gross exports. Gross export systemically exaggerate the role of those countries that handles downstream tasks in the global value chain. Take the Chinese electronics industry as an example, as the exporter for finished electronics products, the full factory price of products is registered as Chinese gross export. But in many cases China only handles final assembly which add little values. Koopman et al (2008) show that Chinese domestic content in electronics export is only around one-third in 2007. And as an extreme example, Dedrick et al (2010)'s case study shows that China captures only around 3% of the factory cost of a 30GB IPod Classic or HP laptops. Given those large discrepancies, the RCA indices based on gross export give a biased picture of country's actual comparative advantage for the recent decades when offshoring is pervasive, and Koopman et al (2014) shows the discrepancy between gross export RCA and value-added export RCA can be enormous.

My measurement for relatedness originates from the value-added export created by each activity, which is expected to give more consistent proximity between activities that is robust to the changes in offshoring. Using WIOD data, I show quite different results compared with previous literature based on the relatedness of products or industries. Hidalgo's product space suggests that the productions of different products within an industry are close with each other, while the shifts across industries, for example the shift from textile production to electronics, is difficult and long-winding. Instead I find that the activities are to large extent clustering in terms of their skill level. Some activities in electronics and textile are quite similar with each other, e.g. the assembly/weaving activities carried out by low-skilled labour. These results have different implications for industrial policies compared with Hidalgo's product space.

In the later part of my paper I show relatedness indices have predicting power on the direction of structural change. Using regression analysis, I confirmed the hypothesis that country tends to gain new comparative advantages in those activities that have high potential for economic growth, and I find that my activity relatedness indices have significant explanatory power on the probability of the evolvement of comparative advantages of new activities. This is related with, for example Neffke et al (2011) and Boschma et al (2012) who show that regions are more likely to develop new industries that has proximity with current existing ones. Kali et al (2013) show that a country would have high probability of sustained growth acceleration if it has an advantageous initial industrial structure in the product space that is close to more advanced new products. Neffke and Henning (2013) show that firms tend to diversify towards the activities that has skill relatedness with its current core activity, which confirms the "activity of monkeys" on micro-level.

The rest of the paper is organised as follows. Section 2 gives the methodology in constructing and deriving my "activity space", i.e. the relatedness indices between activities. Section 3 describes the data source, and report the structural of activity proximity by visualising activity relatedness into a network graph. Section 4 uses regression to analyse the role of activity relatedness in determine the direction of structural change, and in section 5 I will discuss my finding and concluding remarks of the paper as well as future direction of research.

2. Methodology

Construction of the Activity Space

Instead of RCA based on gross export of different products, I calculate RCA indices of value-added export contributed by different skilled activities in each industry. Similar as the conventional RCA for products, the value-added export RCA for an activity x in country a in year t is defined as:

$$RCA_{a,x,t} = \frac{VAE_{a,x,t} / \sum_x VAE_{a,x,t}}{\sum_a VAE_{a,x,t} / \sum_{a,x} VAE_{a,x,t}},$$

$VAE_{a,x,t}$ is the value-added export of activity x in country a, the derivation of VAE will be explained later. An larger than unity $RCA_{a,x,t}$ means that country a's value-added export share of x is higher than world average level, which implies that country a has a comparative advantage in x. The calculation is same for each year, so unless necessary the time subscript t is skipped in the following equations.

The relatedness, or more precise the "revealed easiness for the switch from activity x to y", is represented by the conditional probability of having a comparative advantage in y, given that the country already has comparative advantage in x:

$$\varphi_{x\text{->}y}=\text{Prob}(RCA_{a,y}>1|RCA_{a,x}>1),$$

This conditional probability can be calculated from empirical data by:

$$\varphi_{x\text{->}y} =(\text{number of countries that } RCA_y>1 \text{ and } RCA_x>1)/(\text{number of countries that } RCA_x>1).$$

Same as Hidalgo et al (2007) and Neffke et al (2011), I calculate a "pooled" activity space by taking the simple average of φ values from 1995 to 2007. A high $\varphi_{x\text{->}y}$ means most country with comparative advantage in x also have comparative advantage in y, which implies the environment and abundance that nurse activity x also fit the requirements of y. So when x is already in place, y would evolve with higher probability.

In Hidalgo et al (2007), the relatedness between two industries x and y is the minimum of $\varphi_{x\text{->}y}$ and $\varphi_{y\text{->}x}$:

$$\varphi_{x,y}=\min(\text{Prob}(RCA_{a,y}>1|RCA_{a,x}>1), \text{Prob}(RCA_{a,x}>1|RCA_{a,y}>1))$$

Their proximity between two activities are assumed to be symmetric (i.e. $\varphi_{x,y=}\varphi_{y,x}$) which facilitates the visualization of the relatedness matrix[1]. In this paper I will also provide a visualised network graph of the relatedness in activities using the same method as Hidalgo et al. But in principle the shifts "from x to y" and "y to x" are two different processes; skilled workers can also handles unskilled works easily, but skilled works is troublesome for unskilled workers. Therefore in the regression analysis about

[1] If the proximity between two nodes is asymmetric, in topology it is a directed graph, and there are two edges between each pair of nodes. If the number of nodes is large it is hard to visualize a directed graph in a meaningful way.

relatedness' role in structural change, I will use asymmetric $\varphi_{x\text{->}y}$ and $\varphi_{y\text{->}x}$; the symmetric indices will be used for robustness check.

Derivation of Value-Added Export of Each Activity

I rely on multi-regional input-output analysis to calculate the value added export of each activity ($VAE_{a,x}$). Here I assume that the activities are represented by different skilled works in each industry[2]. If there are m countries in the world and each country has n industries, the so-called technology matrix $\boldsymbol{A}$ in the multi-regional input-output analysis is an m x m block matrix with each $\boldsymbol{A}_{ij}$ of dimension n x n:

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \cdots & \boldsymbol{A}_{1m} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} & & \boldsymbol{A}_{2m} \\ \vdots & & \ddots & \vdots \\ \boldsymbol{A}_{m1} & \boldsymbol{A}_{m2} & \cdots & \boldsymbol{A}_{mm} \end{bmatrix},$$

Each element in matrix $\boldsymbol{A}$, for example $A_{(i,x),(j,y)}$, is the value of intermediate goods from industry x of country i that is used in the production process of \$1 output of industry y in country j. So the sub-matrix $\boldsymbol{A}_{ij}$ captures the information about intermediate outputs produced by country i that are used in the production process of country j, and each diagonal matrix $\boldsymbol{A}_{ii}$ captures the input-output structure within country i. The input-output structure of the whole world can be written as:

$$\boldsymbol{y} = \boldsymbol{A}\,\boldsymbol{y} + \boldsymbol{d}$$

Where $\boldsymbol{y}$ and $\boldsymbol{d}$ are two vectors with nm elements, representing the total production and total final demand of each industry in each country; $\boldsymbol{Ay}$ is the total amount of intermediate goods that are used in deliver $\boldsymbol{d}$ units of final products. Re-arrange the equation, and the relationship between total production

[2] To be more precise, in this type of analysis activities are better represented by the job titles. For example, how many working hours in programming, designing, material processing, manufacturing, assembly, cleaning, … contribute to the value-added export of each country. But such analysis requires data on job decomposition of each industry in each country, which is yet unavailable.

and final demand can be calculated via the Leontief inverse (Leontief 1949):

$$\boldsymbol{y} = (\boldsymbol{I}\text{-}\boldsymbol{A})^{-1}\boldsymbol{d},$$

where $\boldsymbol{I}$ is an identity matrix with dimension (nm x nm).

I use the decomposition methodology proposed by Johnson and Noguera (2012) and Los *et al*. (2012) to calculate the value added export of activities, which is defined as the value-added created by a country's certain skilled worker in a particular industry but finally ends up in the final consumption in foreign countries. For instance in calculate VAE of country i, we first decompose the final demand in the world, $\boldsymbol{d}$, in to two parts: $\boldsymbol{d}^i$ is the final use in country i, and $\boldsymbol{d}^{-i}$ equals $\Sigma_{j\neq i}\boldsymbol{d}^j = \boldsymbol{d} - \boldsymbol{d}^i$ which is the final demand for all other countries in the world. The total production in the world can therefore be decomposed in to two parts: those production that takes place to satisfy final demand of i, and all foreign countries:

$$\boldsymbol{y} = (\boldsymbol{I}\text{-}\boldsymbol{A})^{-1}(\boldsymbol{d}^i+\boldsymbol{d}^{-i}) = (\boldsymbol{I}\text{-}\boldsymbol{A})^{-1}\boldsymbol{d}^i + (\boldsymbol{I}\text{-}\boldsymbol{A})^{-1}\boldsymbol{d}^{-i} = \boldsymbol{y}^i + \boldsymbol{y}^{-i}.$$

The exported value-added of country i are related with those production for the final demand of other countries (i.e. $\boldsymbol{y}^{-i}$). The amount of VAE from industry x of country i is given by:

$$VAE_{i,x} = v_{i,x}\, y^{-i}_{(i,x)},$$

Where $v_{i,x}$ is the value-added to gross output ratio in the production process of industry x in country i (i.e. $v_{i,x}$=1-$\Sigma_{j,y}A_{(j,y),(i,x)}$, it equals \$1 minus total intermediate goods usage in producing \$1 output in country i industry x). I further decompose the exported value-added in each industry to the contributions by capital and different skilled labour, using their income share of value added in each industry:

$$VAE_{i,x,k} = VAE_{i,x}\, ks_{i,x}$$
$$VAE_{i,x,l} = VAE_{i,x}\, ws\_L_{i,x}$$
$$VAE_{i,x,m} = VAE_{i,x}\, ws\_M_{i,x}$$
$$VAE_{i,x,h} = VAE_{i,x}\, ws\_H_{i,x}$$

Here *ks* refers to the capital share, *ws_L*, *ws_M* and *ws_H* are the wage share earned by low-, medium-, and high-skilled labour as a percentage of total payment to factors in industry *x* of country *i*.

I use the those VAEs from three skilled-level activities in each industry to calculate the RCA indices, and subsequently derive the proximity between activities. The VAE contributed by capital does not enter in calculating the RCA indices, since capital goods do not perform activities on themselves.

3. Data and the Structure of Activity Space

This paper uses the recently developed World Input Output Database (WIOD, Timmer et al., 2012) as the data source. WIOD provides multi-regional input output tables about the world economy, as well as other socio-economic and environmental indicators at country-industry level. This database contains 40 countries, covering most developed countries and emerging markets that cover more than half of the world output. The database uses multiple data sources like national supply and use tables, UN trade data at disaggregate levels, census data, and household surveys to give information about the structures of inputs, production, outputs and uses for 35 industries at 2-digit level for a 15-year period from 1995 to 2009. The industry classification across countries are harmonised so the industries in different countries are comparable.

The world input-output tables (WIOTs) in WIOD can be used to derive the value-added export by each industry in each country. The socio-economic accounts in WIOD provide information on industry-level employment and factor payment shares classified by skill levels, therefore I can further decompose the industrial value-added export into the VAE contributed by activities at different skill level. The skill level in WIOD is employees' educational attainment. Low-skill workers refer to the employees who have only lower-secondary or lower level of education. Medium-skilled labour's highest education level is (upper-) secondary education, or other post-secondary but non-tertiary education, for example in vocational schools. High-skilled workers are university graduates with tertiary or higher level of education.

I calculate the RCA indices for each activities in each country and then their bilateral relatedness indices in the way as describe above. I focus on 28 industries in the WIOD (out of 35[3]), each industry are decomposed to activities at low-, medium- and high-skill segments; so in this paper we have 28*3=84 nodes. Subsequently I obtain my "activity space": an 84 x 84 matrix containing the relatedness index between each pair of activities.

I explore the structure of activity proximity by visualise the correlation matrix. For the purpose of visualisation I follow Hidalgo et al (2007) and impose the symmetry constrain $\varphi_{x,y}=\varphi_{y,x}=\min(\varphi_{x\text{->}y}, \varphi_{x\text{->}y})$. Figure 1 shows the heat map of activity space. The horizontal and vertical axis refers to a particular activity; activities are organized in three groups by their skill level; 1 to 28, 29 to 54, and 55 to 84 correspond with each low-, medium- or high-skilled ones. Detailed correspondence table for the numbers can be found in Appendix. The colour of each small square in the heat map represent the level

[3] I exclude public sector and some service industries, where their outputs are largely for final consumption goods/services and does not contribute to the value-added in international trade. Technically it is feasible also to include these sectors, but firstly no or only little value-added by these sectors finally ends up in the exported goods/services, and secondly statistical standard concerning the intermediate usage from those sectors seems to be different across countries, for example in some countries employees' recreating activities paid by employer can be registered as operational cost which ends up as intermediate inputs in the IO table, but for other countries it is registered as consumption. The meaning for those RCAs can be therefore ambiguous. Details can be found in the appendix.

of proximity between two activities; hotter colour (red) represent a higher relatedness value. Colours are particularly hot within the diagonal blocks, especially for low-skilled/low-skilled block, which suggest that compared with the different skilled activities within an industry, an activity is in general more closely related same skill-level activities in other industries. This result can be also illustrated by converting the proximity matrix into a hierarchical clustering dendrogram, as shown in Figure 2. The dendrogram is derived from the Euclidian distance based on the values in the proximity matrix[4]. It is frequently used in biology literature in analysing kinships between different species/genetics. Here it has a similar interpretation; highly related activities tend to be in a same branch. We see that almost all low-skilled activities (i.e. 1~28) are in a cluster located in the left part of figure 2. Similar but less obvious clustering patterns are also observed for medium- and high-skilled activitiess. Only few industries, like mining, has a different pattern that the activities at different skilled within the industry cluster together.

[insert figure 1 about here]

[insert figure 2 about here]

To have a more comprehensive view of proximity, in Figure 3 I further convert activity relatedness matrix into a network graph using the method introduced in Hidalgo et al (2007). Each activity is

[4] The Euclidean distance between two elements (assume 1 and 2) in an N x N correlation matrix is given as follows: $\mathrm{Dist}_{1,2} = \sqrt{\Sigma_{i=1}^{n}(\mathrm{Cor}(1,i) - \mathrm{Cor}(2,i))^2}$, where Cor(i,j) is the correlation value between element i and j that is given in the correlation matrix. The intuition is to assume an N^{th} order pseudo Euclidean space (though strictly speaking this is not a Euclidean space because the dimensions are not orthogonal with each other), and each element represent a single dimension with unity length (because the correlation with an element itself, i.e. Corr(i,i)=1). After deriving the distance measurement, elements are clustered by the so called "complete" method which can provide compact clusters. This is the most standard method in hierarchical clustered dendrongram in many software like R and MatLab. More details about these can be found in Chapter 3 and 4 in Everitt et al. (2011).

represented by a node, and there is a line between two nodes if the relatedness between them is higher than critical value. I use the same cut-off value as in Hidalgo et al; two activities are consider as "highly related" if their relatedness is larger than 0.55[5].

[insert figure 3 about here]

Except some "isolated" activities in mining and real estate, most activities have several close neighbours, and form relatively large clusters. Similar as the heat map, network graph shows that most low-skilled activities are highly related with each other. Another large cluster consists of medium- and high-skilled activities in traditional industries like agriculture and textile, and infrastructure sector like construction and water/electricity supply. High-skilled wholesale activity, medium and high-skilled finance activities together with all activities in sea and air transport form a cluster which is possibly related with the handling of international trade.

Medium- and high-skilled activities in "modern" manufacturing, like chemical, electronics and car manufacturing, form two separated clusters. While the medium-skilled modern manufacturing activities are relatively closely related with other medium- and high-skilled activities, high-skilled modern manufacturing is quite isolated from the rest of the economy and is also distant with medium-skilled modern manufacturing activities. Potential explanation might be that a large proportion of medium skilled manufacturing activities are administrative works and machine operating that requires certain

[5] I first draw a "maximum spanning tree" of the matrix, and then add all other edges with relatedness larger than 0.55. A small number of activities do not have higher than 0.55 relatedness with any other activity, they are still connected to the rest of the graph via the maximum spanning tree in the graph but I make dashed lines for these edges. Detailed information concerning the drawing procedure can be found in the appendix.

level of manufacturing knowledge, while high-skilled activities in manufacturing are probably more about innovation, designing and R&D which require high engineering capability that is not much needed in other activities.

My “activity space” reveals a quite different relatedness pattern compared with works like Hidalgo et al, and it has different implications for the literature concerning paths of economic upgrading. Flying geese paradigm and product cycles, for example, suggest that the industrial upgrading of country is the upward movement from already matured products to similar but more advanced ones. Hildalgo et al show a high relatedness between products within an industry, so their outcome indeed supports the product cycle argument and strategies like infancy industry production: one may first start the production of matured products of modern industries by some external force, then follow the proximity firm would upgrade easily towards more advanced products in that industry. But I find somewhat opposite. Many activities at same skill level in different industries are quite similar with each other, regardless of whether they are in a traditional or advanced industry. While skill upgrading, in contrast, seems to be far more difficult than the jump to another activity in a “more advanced” industry (given that both activities are at same skill level). I will discuss about the reasons for the differences between my results and previous literature and give potential policy implications in the concluding part.

4. Testing the Role of Activity Relatedness in Structural Change

In this section I test the role of activity relatedness in the direction of structural change. Note that the proximity indices are derived from the observed co-existence of comparative advantage of each activities, which is static. I will analyse the probability of gaining comparative advantage in new activities, to show relatedness does have prediction power for the actual direction of structural change.

The concept of activity relatedness and the “jumping monkey” analogy imply that country tends to gain comparative advantage in activities that is has high potential for growth and have proximity with the current economic structure. Naturally these lead to two hypothesis:

1. The probability that country gain comparative advantage in an activity is positively associated with the activity's potential for growth.

2. The probability that country gain comparative advantage in an activity is positively associated with the proximity of this activity and current economic structure of the country.

The hypotheses will be tested by means of probit regression. The dependent variable for the baseline regression is a binary variable about whether an activity that has a comparative disadvantage in 1995 gains comparative advantage in 2009. And for robustness check I will also use other measurement about changes in RCA as the dependent variable, and also perform a test using OLS. Comparing with other structural change indicators like output growth or changes in employment shares, RCA has unique advantage. It is “a comparison of two relative terms”; the effects of both overall growth speed of different countries, and the changes in global demand structure for each activity are cancelled-out and will distort RCA. Therefore there is no necessity to include country/industry fixed effects in the analysis.

All explanatory variables are calculated for the initial year (i.e. 1995). In order to test the first hypothesis, I use PRODY index in Hausmann, Rodrik and Wang (2005) as the measurement for each activity's potential for economic growth. Their PRODY for activity *x* is defined as a weighted average of GDP per capita across country, using x's share in each country *i*'s total value-added export basket ($vaes_{i,x}$) as weight:

$$PRODY_{x,95} = \frac{\sum_{i=1}^{n} vaes_{i,x,95} y_{i,95}}{\sum_{i=1}^{n} vaes_{i,x,95}}$$

Here y_i is the GDP per capita in country *i*; I use PPP adjusted real GDP per capital, provided by Penn World Table 8.0. The value of PRODY (say, $10,000) for an activity *x* can be interpreted as "a country with GDP per capita $10,000 is the representative exporter of *x*". Activities with higher PRODY indices are the those that have high export shares in rich countries, and Hausmann, Rodrik and Hwang show that country with an overall level PRODY (i.e. their EXPY index) higher than its actual GDP would have higher rate of growth, therefore PRODY can be viewed as a revealed measurement for activity's potential for growth, and shifting to activity with higher PRODY is desirable for a country.

There are also other measurements for desirability of activities, among which the hourly wage or profitability of an activity are the straightforward choices. But PRODY is a preferred measurement in capturing the potential of activities. Firstly, under market imperfections wage or profitability does not correctly reflect the marginal return of factors. For example in Lewis two-sector model (Lewis 1954), in the initial phase of development labour force will shift from agriculture to industry sector when the capitalists set industrial wage a bit higher than what a worker can get from agriculture; the relative level of wage does not represent the true differences in activity's potential. Secondly, in those sectors with high degree of increasing return to scale/scope, a set of activities can have high potential for economic growth but initially when the size of the sector is small those activities may not have high wages/profitability.

To test the second hypothesis, I need the measurement of the proximity between the new activity and country's current economic structure. In this paper I analyse this in two aspects. Firstly, I measures how close is an activity related with the current comparative advantage of a country. For the proximity with current RCA, I use a similar strategy as in Kali et al (2013) and compute a "max link" variable each activity in each country. The max link for x in country i is the maximum relatedness of x with those activities that country i holds a comparative advantage in 1995:

$$MaxLink_{i,x} = \max_y \varphi_{y \to x}\ ; \quad y \in \{\text{the set of activities that } RCA_{i,y,1995} > 1\}$$

This measurement directly follows the concept of measuring activity proximity by co-existence of comparative advantage: the environment of the country should fit those activities that currently have a comparative advantage, and new activities that have close relationship with those advantageous activities should have similar requirements therefore tend to evolve more easily. The difference from Kali et al is that I use asymmetric relatedness index instead of symmetric one. I estimate the probability of evolvement of x, so the "max link" corresponding to the relatedness values about concerning shift from other activities into x.

My second measurement focus on the new activity's proximity with the current employment. The variable $RE_{i,x}$ captures the share of employment in country i that are highly related with activity x:

$$RE_{i,x,95} = \sum_{y} emps_{i,y,95}\,, \;\; y \in \{\text{the set of activities that } \varphi_{y \to x} > 0.55\}$$

Variable $emps_{i,y}$ is employment share of an activity y in the total employment of country i, in term of working hours. Here I use the same criteria as Hidalgo, a high relatedness refers to φ>0.55.

The baseline regression is a probit model on the probability that an activity with comparative disadvantage in 1995 gains comparative advantage in 2009. I add RCA index of 1995 as control variable, since the difficulty in getting comparative advantage is different for activities with initial RCA=0.5 or 0.95. The regression equation (for country i activity x) is given by:

$$\text{Prob}(\text{RCA}_{i,x,09}>1 \mid \text{RCA}_{i,x95}<1) = \Phi(\text{ const.} + \beta_1 \text{PRODY}_{x,95} + \beta_2 \text{MaxLink}_{i,x,95} + \beta_3 \text{RE}_{i,x,95} + \beta_4 \text{RCA}_{i,x,95} + \varepsilon_{i,x})$$

Where Φ(.) is cumulative distribution function of standard normal distribution. The baseline regression result is presented in column (1) of table 1:

[insert table 1 about here]

The baseline regression shows that the probability of gaining comparative advantage in an activity is significantly positively related with PRODY and RE, the variable maxlink has the positive sign but not significant. The regression in (1) include all activities that has a smaller-than-one RCA in 1995. The result may be affected by the fluctuations or measurement errors of the "marginal observations", i.e.

those activities that has an 1995 RCA value that is slightly below one. In column (2) I perform the same probit regression, but I restrict the sample size to those observations whose RCA is smaller than 0.8 in 1995; the criterion of "getting comparative advantage" remain the same (RCA_{09}>1). There are no significant differences between (1) and (2)[6]. In specification (3) I change the measurement of RE and maxlink, and use the values from a symmetric relatedness matrix. The notable change with column (1) and (2) is that the magnitude of coefficient for maxlink becomes larger and statistically significant, while the coefficient for RE becomes smaller and insignificant. So in general, the results confirms both two hypothesis above.

In (4) to (6), I change the dependent variable. Specification (4) is a probit regression with the dependent variable on whether the RCA index of 2009 is more than 30% higher than the RCA in 1995. This specification shows that all three variables (lnprody, maxlink and RE) are quite significant. The difference with (1) to (3) is that here the control variable initial RCA in 1995 becomes negative significant, this should be relevant with the "catch-up" across different activities. (5) and (6) are OLS regressions using the growth of RCA value between 1995 and 2009 as dependent variable[7], and (6) has the same sample size as (2) that activities with RCA_{95}>0.8 are excluded. Both (5) and (6) show that activities' potential and two measurement for the proximity with current activities are highly significant.

[6] The magnitude of marginal effect in (2) is smaller than in (1), that is because (1) and (2) have a different sample and the marginal effect is evaluated at a different point.

[7] Specifically, I use $\log(RCA_{09}/RCA_{95})$, the reasoning for taking logarithm instead of using (RCA_{09}/RCA_{95}-1) is that the distribution of RCA_{09}/RCA_{95} is highly skewed.

Implicitly, specification (4) to (6) attach higher weight to the growth of low-initial-RCA activities because those activities that has big growth in their RCA but yet have not reach RCA>1 are also captured by the dependent variables in (4) to (6). These results also confirm both hypothesis that the growth of an activity is significantly related with its potential for economic growth and the proximity with current economic conditions.

The marginal effect corresponding to the probit regression shows variables are both statistically and economically significant. Take specification (1) as example, on average, an activity that has 10% higher PRODY would have around 2.5 percentage points higher possibility of getting a comparative advantage. And if an activity x is closely related 25% of total current employment while y is related with 15%, such 10 percentage points differences gives x around 1.3 percentage points higher chance in getting a comparative advantage. The effect of maxlink, although insignificant, is still economically large. Comparing two activities that has at maximum 0.6 or 0.8 relatedness with current comparative advantageous activities, the latter one has around 0.2*0.066=1.32 percentage points higher probability in getting a comparative advantage. Take into consideration that 271 out of 1976 observations has RCA>1 in 2009; the overall probability of getting comparative advantage is 13.7%, so those effect above are quite large.

I further explore the potential reason why activities' proximity with current employment is positively associated with the possibility of getting comparative advantage. There might be two potential channels.

First, the employment in highly related activities have similar qualification as the labour that are necessary for the new activity. Therefore, the proximity with a high share of employment accelerate the development of the new activity by providing a large base of potential labour supply. The second possible channel is economy of scale and agglomeration effect. A set of activities may have high degree of economy of scale and/or backward and forward linkages, the existence of related activities raises the return for the new activity, so the incentive in investment of the new activity is related with the size of the cluster of related activities.

In order to test which channel is dominant, I divide highly related employment share (RE) into two parts: RE_lower$_{i,x}$ is the part of highly related employment in the activities that has lower potential, while RE_higher$_{i,x}$ is in those activities with higher potential than x:

$$RE\_lower_{i,x,95} = \sum\nolimits_{y} emps_{i,y,95}\,,\ \ y \in \{\text{set of activities that } \varphi_{y\to x} > 0.55 \textit{ and } \text{Prody}_y < \text{Prody}_x\}$$

$$RE\_higher_{i,x,95} = \sum\nolimits_{y} emps_{i,y,95}\,,\ \ y \in \{\text{set of activities that } \varphi_{y\to x} > 0.55 \textit{ and } \text{Prody}_y > \text{Prody}_x\}$$

If the effect is only caused by labour supply effect, then only employment in those activities that have lower potential would be attracted towards the new activity, and the higher-potential activities are more attractive than the new activity soemployment there will not have incentive to re-allocate. In comparison, under the agglomeration effect, employment in both related higher- and lower-potential activities are relevant in raising the attractiveness of the new activity, so both RE_lower and RE_higher would be positive significant.

I run three groups of regressions: a baseline probit regression using whether an activity gains comparative advantage as dependent variable, an OLS on the changes in RCA values, and another probit regression on whether the RCA index increase by more than 30%. The results are shown in table 2.

[insert table 2 about here]

Results in column (4) to (6) show that proximity to the employment of both higher- and lower-potential activities correlate positively with the growth in RCA values, which implies both potential labour supply effect and agglomeration effect play an role in the growth of the new activities. But interestingly specification (1) to (3) suggest that only the proximity with lower-potential activities increase the possibility of obtaining an comparative advantage, while proximity with higher-potential activities turns out to be insignificant, and similar pattern is observed for (7) to (9). These results may suggest that the re-allocation of labour from highly-related activity is a more important factor in radical development of new activities.

5. Concluding Remarks and Discussion

This paper applied a similar framework as Hidalgo et al (2007) and analysed the relatedness between different activities. I find the pattern of activity relatedness is quite different from the relatedness of products in the previous literature. I show that under global production fragmentation, many activities at same skill level are closely related with each other, while the within industry upgrading of skillness is a more difficult process.

I show that the transition path of a country's comparative advantage is in line with the "jumping monkey" analogy, countries tend to gain new comparative advantages in activities that have high potential for economic growth and have proximity with current comparative advantage and employment.

The differences in the proximity structure between my activity space and product space is believed to originate from the new pattern of trade-in-tasks and the increasing degree of offshoring. Understanding the different structure between activity space and product space has important implications in thinking the structural changes strategies for the past and the present. Hidalgo et al's product space does not consider offshoring, and is derived from Feenstra et al (2005) trade data for 1962 to 2000 when the degree of offshoring is much lower than the time period in this paper (1995~2009). Their results support the targeted industry protectionism strategies that was used by many developing countries. It is feasible that "first start production in modern industries by external force, then follow the proximity firms would upgrade on its own towards more advanced products in the industry". This strategy was successful in the history, for example Korea, where initially many firms enter a new industry by producing low-end products that is already mature in Japan or U.S., either for domestic market under the help of trade protection policies or OEM production for foreign firms. During the process firm gain knowledge, and firms gradually have their own capabilities and eventually become major players in the world market (see Amsden 1989).

Under offshoring, these strategy might not work longer. According to my results, if a developing country follows the strategy above, it can easily start the low-skilled activities in some "advanced industries", since almost all kinds of low-skilled activities are quite related with each other. But based on a growth perspective, it is questionable that how much a developing country can learn from handling these low-skilled tasks in advanced industries when offshoring is pervasive. When offshoring is impossible, if underdeveloped countries decide to start business in low-end part of the advanced industries like car making and electronics, in many cases they have high level of domestic content. There are some requirements on technological transfer, and firms are able to get insights and accumulate knowledge from some high-skilled activities in the production process of those low-end products. But under offshoring, tasks being offshored to developing countries are not only low-skilled, but also "specially designed", i.e. standardized and routinized which is easy for monitoring and coordinating and can be deployed to most developing countries quickly. Those higher-skilled, non-routine and cognitive tasks, which are very different from offshored tasks and often have higher potentials of innovation, largely stay within developed countries.

Participating in assembling of mobile telephones for example, are not that different from traditional activities like textile and weaving. If it does give higher return, in the short run the government may find it beneficial. However, it does not teach the country much about the engineering, design, and the knowledge in making chips; future upgrading is not likely to happen spontaneously after such industrial policies. Horizontal policies like education and training which increase the overall skilled labour abundance of a country and generate labour supply suitable for more advanced activities, are believed to be more important determinants for long run sustainable structural upgrading.

There are several limitations with this paper. A notable problem is that the database I use measures skill level of an activity by skill attainment. In principle, the proper measurement for skill for this research is skill *requirement*, i.e. the proportions of employment in an industry that require low-, medium-, high-skilled workers. Using skill attainment therefore potentially overestimate the share of high-skilled activitiess in an industry, since a highly educated person may also do low- or medium-skilled works, but not vice versa. Another problem would be the treatment of capital. In this paper I calculate the RCAs based on the value-added exports contributed by labour and capital is not included in the analysis. But activities need particular kinds of capitals, so capital is indivisible from workers and in principle we should also contribution of capital goods that work in combination with labour. This would require the decomposition of industrial level capital stock into different categories that work in combination with different occupation; the data is yet not available. I assume that capital is mobilized within an country, therefore the capital return in all activities are equalized and capital re-allocation passively accompanies labour re-allocation. But for developing countries where capital goods for particular industrial sectors are scarce and labour's bargaining power is low, it does not hold, which may affect the usefulness of the activity-level RCA indices of developing countries.

Due to data limitation, this paper uses different skilled labour in each industries to proxy different activities. This is different from "concrete" activities like assembly, programming, cooking, … etc. Future researches may be carried out in combining multiregional IO table with occupational data in each country (when it becomes available) and to analyse the relatedness and structural upgrading between each "real" activities.

Reference

[1] Amsden, A. H., (1992), "Asia's Next Giant: South Korea and Late Industrialization", Oxford University Press.

[2] Baldwin, R. (2006), "Globalisation: the Great Unbundling (s)", Economic Council of Finland.

[3] Boschma, R., Minondo, A., and Navarro, M., (2012), "Related variety and regional growth in Spain", *Papers in Regional Science*, Vol. 91(2), pp.241-256.

[4] Bryce, D. J., and Winter, S. G., (2009), "A general interindustry relatedness index", *Management Science,* Vol. 55(9), pp.1570-1585.

[5] Chenery, H. B., (1960), "Patterns of Industrial Growth", *American Economic Review*, Vol. 50(4), pp. 624-654.

[6] Dedrick, J., K. L. Kraemer, and G. Linden, (2010), "Who Profits from Innovation in Global Value Chains? A Study of the iPod and Notebook PCs", *Industrial and Corporate Change*, Vol. 19(1), pp. 81-116.

[7] Echevarria, C., (1997), "Changes in Sectoral Composition Associated with Economic Growth", *International Economic Review*, Vol. 38(2), pp. 431-452.

[8] Everitt, B. S., S. Landau, M. Leese, and D. Stahl, (2011), "Cluster Analysis", 5th Edition, John Wiley & Sons Ltd, West Sussex.

[9] Feenstra, R. C., Lipsey, R. E., Deng, H., Ma, A. C., and Mo, H., (2005), "World trade flows: 1962-2000", NBER Working Paper #11040.

[10] Grossman, G. M., and E. Rossi-Hansberg, (2008), "Trading Tasks: A Simple Theory of Offshoring", *American Economic Review*, Vol. 98(5), pp. 1978-1997.

[11] Hausmann, R., Hwang, J., and Rodrik, D, (2007), "What you export matters", *Journal of economic growth*, Vol. 12(1), pp. 1-25.

[12] Hidalgo, C. A., Klinger, B., Barabási, A. L., and Hausmann, R., (2007). "The Product Space Conditions the Development of Nations", *Science*, Vol. 317, No. 5837, pp. 482-487.

[13] Johnson, R., and G. Noguera (2012), "Accounting for Intermediates: Production Sharing and Trade in Value Added", *Journal of International Economics*, Vol. 86(2), pp. 224-236.

[14] Kali, R., Reyes, J., McGee, J., and Shirrell, S., (2013), "Growth networks", *Journal of Development Economics*, Vol. 101, pp. 216-227.

[15] Koopman, R., Wang, Z., and Wei, S. J., (2008), “How much of Chinese exports is really made in China? Assessing domestic value-added when processing trade is pervasive”, NBER working paper #14109.

[16] Koopman, R., Wang, Z., and Wei, S. J., (2014), “Tracing Value-Added and Double Counting in Gross Exports”, *American Economic Review*, Vol. 104(2), pp. 459-494.

[17] Kuznets, S., (1957), “Quantitative Aspects of the Economic Growth of Nations: II. Industrial Distribution of National Product and Labor Force”, *Economic Development and Cultural Change*, Vol. 5(4), pp. 1-111.

[18] Laitner, J., (2000), “Structural change and economic growth”, *Review of Economic Studies*, Vol. 67(3), pp. 545-561.

[19] Leontief, W., (1949), “Structural Matrices of National Economies”, *Econometrica*, Vol. 17, Supplement: Report of the Washington Meeting, pp. 273-282.

[20] Lewis, A. W., (1954), “Economic Development with Unlimited Supplies of Labour”, *the Manchester School*, Vol. 22(2), pp. 139-191.

[21] Los, B., M.P. Timmer, and G. J. de Vries, (2012), “China and the World Economy: A Global Value Chain Perspective on Exports, Incomes and Jobs”, Research Memorandum GD-128, Groningen Growth and Development Center, Faculty of Economics and Business, University of Groningen.

[22] Neffke, F., Henning, M., and Boschma, R., (2011), “How do regions diversify over time? Industry relatedness and the development of new growth paths in regions”, *Economic Geography*, vol. 87(3), pp. 237-265.

[23] Neffke, F. and Henning, M., (2013), “Skill relatedness and firm diversification”, *Strategic Management Journal*, vol. 34(3), pp. 297-316.

[24] Teece, D. J., Rumelt, R., Dosi, G., and Winter, S., (1994). “Understanding corporate coherence: Theory and evidence”, *Journal of Economic Behavior & Organization*, Vol. 23(1), pp.1-30.

[25] Timmer, M. P., A. A. Erumban, R. Gouma, B. Los, U. Temurshoev, G. J. de Vries, I. Arto, V. A. A. Genty, F. Neuwahl, J. M. Rueda-Cantuche, A. Villanueva, J. Francois, O. Pindyuk, J. Pöschl, R. Stehrer, and G. Streicher, (2012), “The World Input-Output Database (WIOD): Contents, Sources and Methods”, www.wiod.org.

Table 1: The results of the regression

	Baseline		RCA95<0.8		Symetric Proximity		RCA Growth>30%		Growth in RCA	
	(1)	MEF	(2)	MEF	(3)	MEF	(4)	MEF	(5)	(6)
ln(prody)	1.712	0.245	1.380	0.146	1.596	0.230	2.512	0.963	1.602	1.626
	0.297***	0.041***	0.348***	0.036***	0.286***	0.040***	0.195***	0.074***	0.086***	0.096***
maxlink	0.461	0.066	0.397	0.042	0.977	0.141	0.542	0.208	0.541	0.586
	0.426	0.060	0.505	0.053	0.451*	0.064*	0.289*	0.111*	0.140***	0.157***
RE	0.925	0.131	0.774	0.082	0.601	0.086	0.910	0.349	0.443	0.394
	0.336**	0.048**	0.414*	0.044*	0.500	0.072	0.271***	0.103***	0.134***	0.157**
RCA95	2.532	0.361	2.725	0.288	2.498	0.359	-1.309	-0.502	-0.654	-0.763
	0.178***	0.026***	0.258***	0.027***	0.178***	0.026***	0.122***	0.046***	0.057***	0.079***
const	-19.7	-	-16.39	-	-18.7	-	-24.6	-	-15.6	-15.8
	2.97***	-	3.48***	-	2.86***	-	1.93***	-	0.85***	0.95***
R2	0.2361		0.1876		0.2314		0.1015		0.1779	1702
Obs	1976		1680		1976		1976		1976	1680
# Positive	271	(13.7%)	155	(9.2%)	271		800	(40.5%)	-	-

Notes: significance level: *: 0.1, **:0.01, ***:0.001. (1) to (4) are probit regression, MEF is the marginal effect evaluated at mean of independent variables using delta method. the dependent variable of (1) to (3) are whether the country gains a comparative advantage in an activity in 2009, and (4) is whether the RCA value of an industry rises more than 30% from 1995 to 2009. (5) and (6) are OLS using $\ln(RCA_{09}/RCA_{95})$ of each activity in each country as dependent variable. Sample size of (2) and (6) are restricted to the activities that has RCA<0.8 in 1995. (1) to (4) use pseudo-R^2, and (5) and (6) use adjusted R^2.

Table 2. Potential Labour Supply Effect v.s. Agglomeration Effect

	Gain an RCA in 2009 (probit)			Changes in RCA Value (OLS)			RCA growth> 30% (probit)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
ln(prody)	1.515	1.705	1.57	1.561	1.65	1.619	2.382	2.532	2.440
	0.294***	0.310***	0.313***	0.088***	0.089***	0.091***	0.195***	0.202***	0.204***
maxlink	0.569	1.073	0.479	0.687	0.67	0.536	0.683	0.994	0.561
	0.389	0.376**	0.425	0.127***	0.126***	0.141***	0.264**	0.259***	0.290*
RE_lower	1.287		1.282	0.397		0.377	1.205		1.180
	0.420**		0.421**	0.176*		0.176*	0.354***		0.355***
RE_higher		0.362	0.304		0.568	0.546		0.574	0.469
		0.561	0.566		0.220**	0.221*		0.455	0.461
RCA95	2.559	2.494	2.556	-0.647	-0.670	-0.658	1.283	-1.322	-1.293
	0.178***	0.177***	0.179***	0.058***	0.058***	0.058***	0.122***	0.122***	0.122***
const	-17.8	-19.8	-18.3	-15.2	-16.1	-15.7	-23.4	-25.0	-23.9
	3.0***	3.1***	3.1***	0.9***	0.9***	0.9***	1.9***	2.0***	2.0***
R^2	0.2371	0.2316	0.2373	0.1755	0.1762	0.1936	0.1016	0.0978	0.1020
Obs	1976	1976	1976	1976	1976	1976	1976	1976	1976

Notes: significance level: *: 0.1, **:0.01, ***:0.001. (1) to (3) and (7) to (9) are probit regression, The dependent variable of (1) to (3) is whether the country gains a comparative advantage in an activity in 2009, and (4) to (6) is $\ln(RCA_{09}/RCA_{95})$ for each activity, and (7) to (9) is whether the RCA of an activity rises more than 30% from 1995 to 2009. (1) to (3) and (7) to (9) use pseudo-R^2, and (4) to (6) use adjusted R^2.

Figure 1. Heat Map corresponding to the Activity Relatedness Matrix

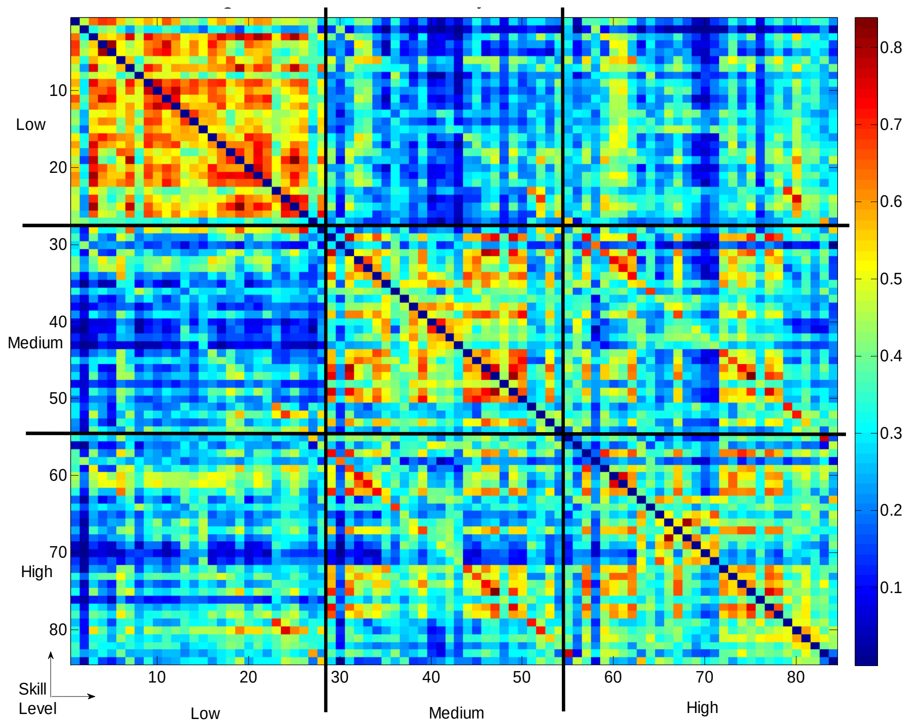


Figure 2. Hierarchical Dendrogram of Activity Relatedness

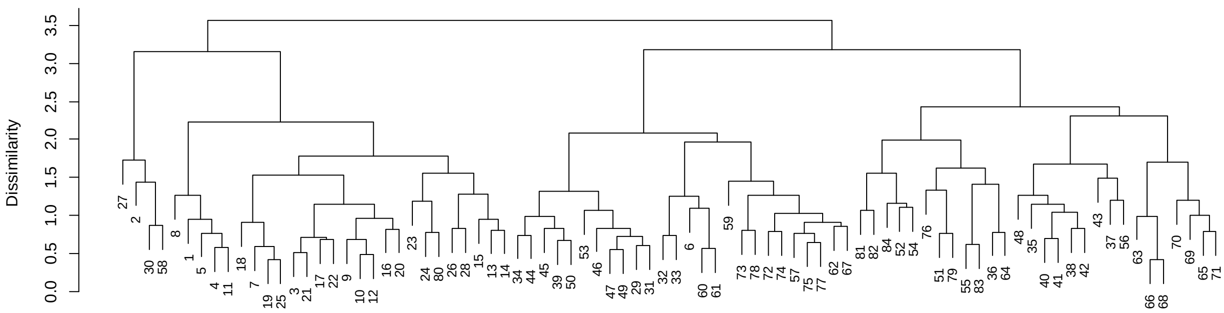
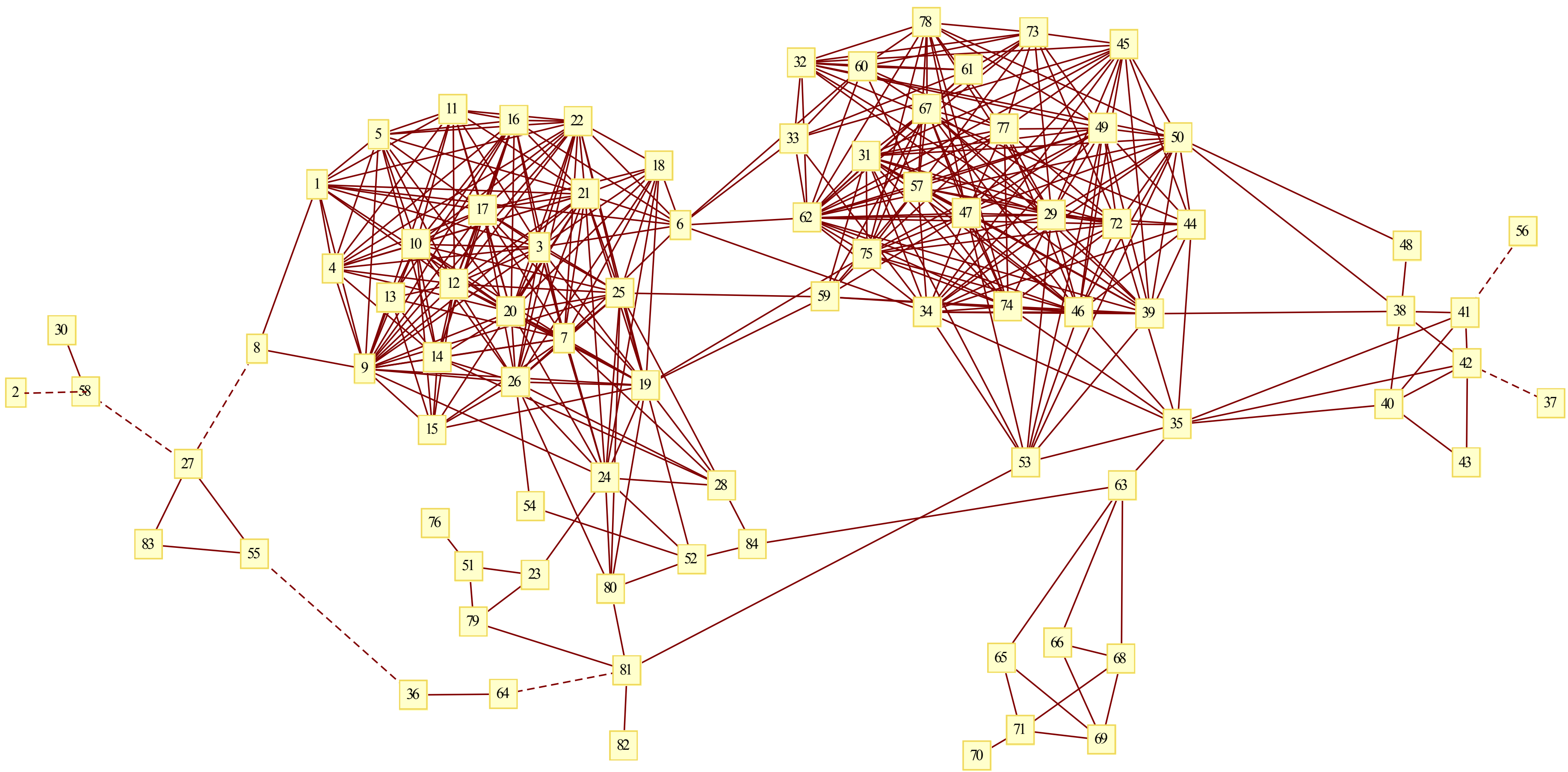


Figure 3. “Activity Space”: The Network Graph of Relatedness Between Activities



Appendix:

Industry and Activity Code List

Skill-level (educational attainment) / Industry	LOW	MEDIUM	HIGH
Agriculture, Hunting, Forestry and Fishing	1	29	57
Mining and Quarrying	2	30	58
Food, Beverages and Tobacco	3	31	59
Textiles and Textile Products	4	32	60
Leather, Leather and Footwear	5	33	61
Wood and Products of Wood and Cork	6	34	62
Pulp, Paper, Paper , Printing and Publishing	7	35	63
Coke, Refined Petroleum and Nuclear Fuel	8	36	64
Chemicals and Chemical Products	9	37	65
Rubber and Plastics	10	38	66
Other Non-Metallic Mineral	11	39	67
Basic Metals and Fabricated Metal	12	40	68
Machinery, n.e.c.	13	41	69
Electrical and Optical Equipment	14	42	70
Transport Equipment	15	43	71
Manufacturing, n.e.c.; Recycling	16	44	72
Electricity, Gas and Water Supply	17	45	73
Construction	18	46	74
Sale, Maintenance and Repair of Motor Vehicles and Motorcycles; Retail Sale of Fuel	19	47	75
Wholesale Trade and Commission Trade, Except of Motor Vehicles and Motorcycles	20	48	76
Retail Trade, Except of Motor Vehicles and Motorcycles; Repair of Household Goods	21	49	77
Inland Transport	22	50	78
Water Transport	23	51	79
Air Transport	24	52	80
Post and Telecommunications	25	53	81
Financial Intermediation	26	54	82
Real Estate Activities	27	55	83
Renting of Machine & Equipment and Other Business Activities	28	56	84

Excluded Industries in WIOD

The following industries are in WIOD, but *not* included in the analysis of this paper, due to the reasons as mentioned above in the paper:

(1) *Hotels and Restaurants;*
(2) *Other Supporting and Auxiliary Transport Activities; Activities of Travel Agencies;*
(3) *Public Administration and Defence, Compulsory Social Security;*
(4) *Education;*
(5) *Health and Social Work;*
(6) *Other Community, Social and Personal Services;*
(7) *Private Households with Employed Persons*

Steps in Deriving the Network Graph

1. Deriving a Maximal Spanning Tree as the "Skeleton" of the Network

I follow a similar procedure as Hidalgo *et al.* (2007) in drawing the network, first deriving the skeleton of the network by means of the maximum spanning tree (MST) algorithm. A spanning tree is a graph that connects all nodes with each other with minimal number of edges; the number of edges therefore equals the number of nodes minus one. An MST is a spanning tree where the summation of the value of the edges, i.e. the sum of proximity indices, is maximized. Figure A1 gives an illustration for a maximum spanning tree. Intuitively, MST gives a compact skeleton connecting all activities, where the overall degree proximity represented in the skeleton is maximized.

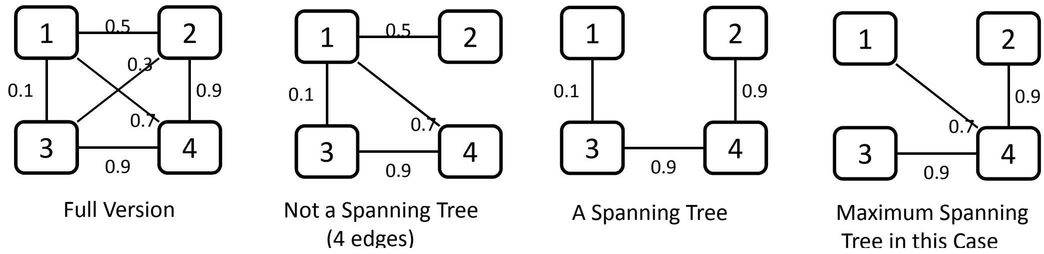


Figure A1. An illustration for maximum spanning tree

The MST of the relatedness matrix is illustrated in Figure A2.

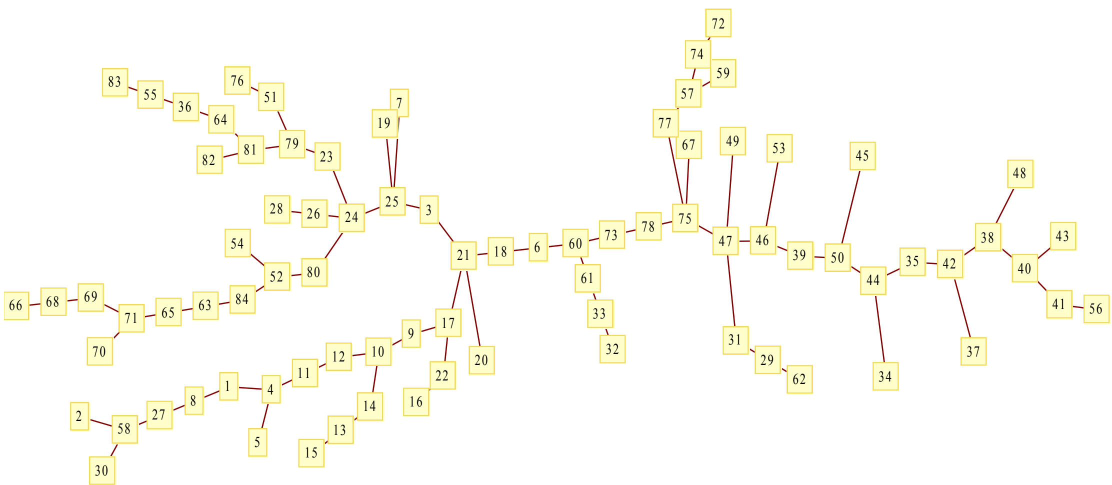


Figure A2: Maximum Spanning Tree corresponding to the Activity Relatedness Matrix

Step 3: Add other edges between the nodes that the relatedness index is higher than 0.55. And use the Spring Electrical Embedding Algorithm in Mathematica 9.0 to organize the position of the nodes. The direct outcome after running the SEEA is given in Figure A3:

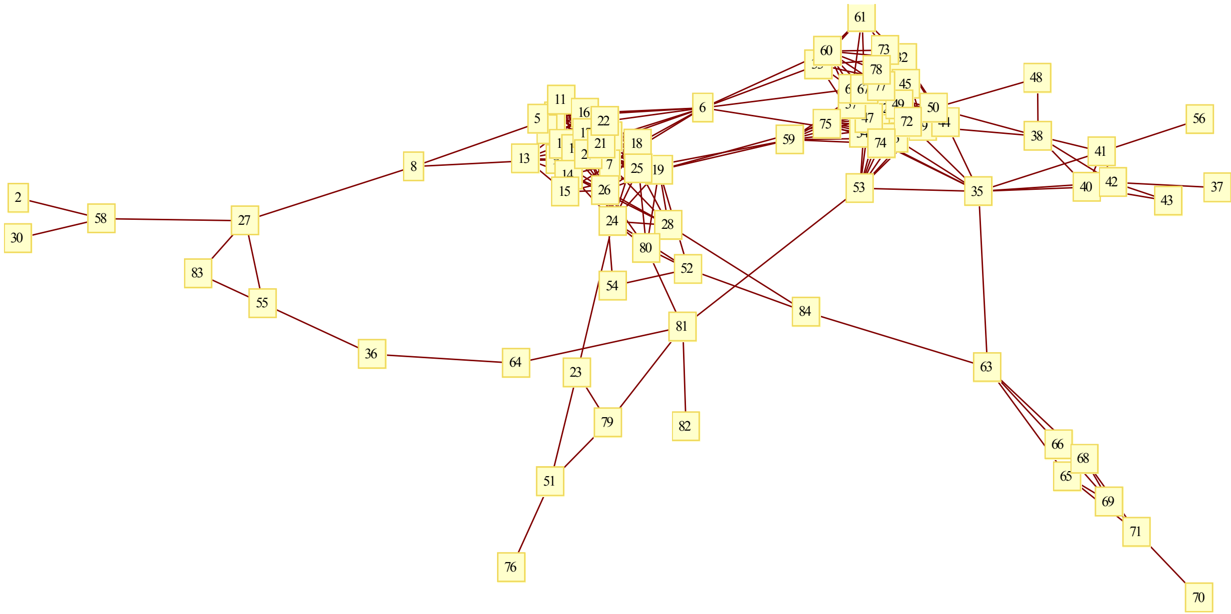


Figure A3: The Activity Space, Direct Output from Mathematica After SEEA Algorithm

4. the last step is manually drag the nodes and make all numbers in the graph readable, and replace those curves with relatedness lower than 0.55 by dashed line. This ends up to the Figure 3 above in the main part of the paper.