Papers in Evolutionary Economic Geography

08.19

Revealed Relatedness: Mapping Industry Space

Frank Neffke & Martin Svensson Henning

Utrecht University
Urban & Regional research centre Utrecht

Revealed Relatedness: Mapping Industry Space

Version 2, 30.04.2008

Frank Neffke
Utrecht University, Erasmus Uiversity Rotterdam
The Netherlands
fneffke@gmail.com

Martin Svensson Henning
Lund University
Sweden
martin@keg.lu.se

Abstract
In this paper we measure technological relatedness between industries using a dataset on product portfolios of plants. For this purpose we first develop a general methodology to extract data on co-occurrences of classes (e.g. industries) in a single entity (e.g. a plant) to construct estimates of the relatedness between the classes. The core assumption, in line with the concept of economies of scope, is that if two products are produced in the same plant, this is an indication of relatedness between the industries the two products are a part of. Unlike earlier methods, we arrive at a Revealed Relatedness (RR) index that can be interpreted on a ratio scale, allows for the use of indirect (i.e. not directly observed) information on industry relatedness, and conceptualizes relatedness as being asymmetric or directed. Direction of relatedness provides information on, for example, the most likely direction of spillovers between two classes. We also graph the RR matrices using methods borrowed from social network analysis. The result is a visualization of the "industry space" and how that changes over time with structural transformation of the economy. In order to test the validity of the framework, the industry space is used to plot structural transformation paths of regions. It is shown that the RR matrix indeed has significant explanatory power for the composition and change of a regions portfolio of manufacturing industries, in spite of the fact that regional information played no role in its derivation. This confirms the quality of our RR estimates.

Acknowledgements
The authors gratefully acknowledge financial support provided by the Bank of Sweden Tercentenary Foundation and STINT (The Swedish Foundation for International Cooperation in Research and Higher Education) (Martin Svensson Henning); and NWO (Netherlands Organization for Scientific Research) (Frank Neffke).

1: Introduction

technological paradigms (Freeman and Perez 1988, Dosi 1988), long-term structural change (Schön 2000) and general purpose technologies (Breshnahan and Trajtenberg 1995) has shown how major technological breakthroughs propagate, one industry at a time, through the economy. Furthermore, the concept of a 'cluster' of related industries (Porter, 1990, 2000) has proven a versatile research tool in investigating the value of national and regional industrial coherence. Indeed, Hidalgo et al. (2007) recently showed that countries expand their exports by moving to products that are related to their existing portfolio. This contribution places the issue of technological relatedness in the centre of contemporary trade theory as well.

Given the importance of technological relatedness in such diverse fields, it has become crucial to reliably measure how strongly related industries are to one another. We contribute to this effort by developing a novel methodology to distil relatedness relations between industries from product portfolios. The basic idea is if products from two different industries are produced in the same plant, these industries are likely to be related. The first attempts to develop measures based on such co-occurrences of industry or technological classes in one and the same entity (such as one patent filed under multiple technology classes or one firm producing different products) were made in the 1990s. The introduction of co-occurrence based relatedness measures marked an important step forward in the research of relatedness, and our new measure is constructed in the spirit of this tradition.

The method we propose can be applied to many types of co-occurrence data. An important contribution of our method is a Bayesian extension that can be used to consistently merge co-occurrence information with any other piece of information on relatedness between industries. This may prove a valuable tool when combining information derived from different sources into a single relatedness measure. Furthermore, the method allows us to show that differences in the degree of complexity of industries naturally lead us to regard relatedness as a directed, and therefore an asymmetric, relation.

In this article, we will show how our relatedness index can be used to estimate relatedness between industries. For this purpose, we use a dataset that contains product portfolios in Swedish plants between 1969 and 2002. As many plants produce multiple products, these product portfolios can serve as the basis of a co-occurrence analysis. The advantage of portfolio data is that they reflect the distributed knowledge about relatedness used in diversification decisions at the micro level of the economy. We will therefore refer to our index as a measure of *Revealed Relatedness*. By connecting industries in a network that reflects Revealed Relatedness, we can visualize the "industry space". The data also allow us to calculate and plot the transformational path of individual regions in this industry space. Our enquiries show that regional portfolios of industries are usually not random, but rather a coherent set of related industries. This coherence is preserved over time as regions are more likely to expand into industries that are closely related to their present portfolio than into industries that are very dissimilar to their main economic activities. Although this particular application is in the field of economic geography, it is easy to imagine similar – or quite different – applications in fields as varied as strategy research or international development studies.

The outline of the article is as follows. In the first section we give an overview of the literature investigating the measurement of relatedness with a focus on the co-occurrence approach. Next, in section two, we develop our methodology to derive a matrix of Revealed Relatedness. In section three, we apply this method to our product portfolio data to estimate yearly relatedness matrices for all manufacturing industries in the Swedish economy. Section four describes the outcomes of this procedure. Section 5

applies the matrix to the structural transformation paths of two old industrial regions in Sweden, and section five concludes and describes challenges for future research.

2: Relatedness as co-occurrence

The present investigation is certainly not the first to set out on a search for a reliable relatedness measure. Historically, particularly the business literature has shown a profound interest in the issue of industry relatedness, as it clearly connects to the understanding of diversification processes and strategic decisions within firms (Gort 1962, Berry 1975, Rumelt 1982, Prahalad and Bettis 1986, Grant 1988). Earlier measures of diversification have been based on a large number of different principles. The probably most widely used measure of relatedness uses the hierarchy of the Standard Industrial Classification (SIC) system to assess similarity between industries. The lower the class two industries share in the hierarchy, the more similar they are thought to be. According to this logic, industries in the same 5 digit class are more related than industries that only share the same 3 digit class. The value of this measure has been questioned, as it is rather rigid and theory void.

This has given rise to a number of alternative approaches. In the 1980s, Scherer (1982) constructed relatedness matrices for almost 50 industries based on estimated technology flows. Later, Farjoun (1994) based relatedness on similarity in the human expertise, as proxied by occupations employed in different industries. Yet another approach is the use of input-output tables to measure relatedness in terms of similarities in input-output profiles (Fan and Lang 2000). Whereas Scherer's method has to investigate individual firms in detail and is very time consuming, the latter two measures can be readily derived from information at the level of the industry aggregate.

In the 1990s, a group of scholars devised methods to exploit the information enclosed in micro-level data. Engelsman and Van Raan (1991) used the fact that some patents are filed in multiple technology classes as evidence for the technological relatedness between these classes. The use of patents as a source of information has the advantage that it stays very close to the notion of technological relatedness. However, knowing which patent classes are related does not provide us immediately with an index of relatedness between industries. This problem is amplified by the fact that only a limited number of industries rely heavily on patenting. As a consequence, any patent based measure will be restricted to a rather small part of the economy.

The prime piece of information in any co-occurrence analysis of relatedness is the number of times a combination of classes co-occurs in a single entity. For example, in Teece et al. (1994), the analysis centres on the number of times a combination of two industries (the co-occurrence) is found in one and the same firm (the entity). Hidalgo et al. (2007) use the number of times two industries display revealed comparative advantage (the co-occurrence) in a single country (the entity). However, some co-occurrences are more likely than others, because the involved classes are more likely to occur in an entity. For example, some industries are bigger than other industries, and will therefore, a priori, tend to be found in more firms. The most important difference between the various analyses of co-occurrences, therefore, is the way authors control for the overall tendency of a class to engage in linkages with *any* other class. Authors that apply the methodology pioneered in Engelsman and Van Raan (1991) often limit the analysis to all entities that host a co-occurrence, i.e. those entities that host at least two classes. In Teece et al. (1994), for instance, the authors restrict themselves to firms that own plants in multiple industries. Next, the authors the overall number of times an industry participates in any co-occurrence as given. It is then quite natural to view the linking process as draws from a hypergeometric distribution. Let us call the number of plants producing in industry i, N_i, and the number of plants producing in industry j, N_j. Furthermore, let N be the total number of plants. The process of establishing links

between industry i and industry j can now be depicted as randomly drawing N_i plants from a population of plants that can be split in two parts: N_j plants that are producing in industry j and $N - N_j$ plants that are not producing in industry j.

Using properties of the hypergeometric distribution, it is straight forward to derive the expected number of links for each industry pair (see, e.g., Bryce and Winter, 2006). We can then compare these expected values to the actual number of co-occurrences. Accordingly, relatedness gets to be defined as the t-value of observing the actual (or more extreme) number of links, given the mean and variance of the random draws. This yields a good control for the overall tendency of classes to engage in co-occurrences. As a matter of fact, the methodology does its job too well. It controls for the total number of co-occurrences for a given class. But while some industries may have an overall tendency to "participate" in many co-occurrences because they are very attractive from a general point of view (e.g. they usually generate high profits), other industries participate in many co-occurrences because they are *similar* to many other industries. The hypergeometric approach cannot distinguish between those two forces. As a result, if we take centrality to mean the number of other classes a class is strongly related to, the approach cannot tell us anything about how central a class is. The average relatedness of a class to all other classes may vary between classes, but it is hard to interpret what this average of t-values means.

The method laid out in Hidalgo et al. (2007) differs from the Engelsman and Van Raan approach, as it defines distance between product categories in terms of conditional probabilities. The estimate of a conditional probability is equal to the number of co-occurrences of classes i and j, divided by the number of times either class i or class j is observed in the sample - depending on whether the conditioning is on i or on j. For practical reasons - and to arrive at a symmetric distance measure - the authors set relatedness equal to the minimum of the two conditional probabilities.
The general logic behind this method is that some co-occurrences are more likely than others, because one of the involved classes is larger than average. However, also here, the correction is not beyond debate. The size of a class corresponds to the number of countries that develop a relative specialization in a product, and this depends on the evenness with which production in the industry is distributed across the world economy. It is hard to assess how such a correction should be interpreted.

We will describe a way to control raw co-occurrence links that has a more natural interpretation. As we implied above, for any application there may be a number of factors that influence the general propensity of classes to participate in co-occurrences. To return to the example of multi industry firms, some industries are more likely to participate in co-occurrences not because they are bigger than other industries, but rather because they are more attractive to diversify into. This attractiveness could depend on a number of factors, such as the average profitability in the industry, wage levels and the fierceness of competition. In our methodology, we are able to control for any of such factors as long as information on them is available at the level of the class aggregate. Furthermore, the resulting index can be regarded as a probability. This makes it easy to interpret, and allows comparisons on a ratio scale. Another consequence is that there is a consistent way to gather additional, indirect, information for the relatedness of pairs of classes that cannot be estimated with great precision. This is for instance useful when the involved classes are very small. The resulting index can also be asymmetric, where the direction of relatedness provides information about the complexity of the classes involved in the co-occurrence.

3: The Revealed Relatedness Method

Generation of co-occurrences

The basic proposition behind our method is that it is possible to *predict* the number of co-occurrences between two classes using only class specific variables as predictors. These predictions can then be compared to the actual number of links, which will reveal the relatedness between the classes involved.

As an illustration, we could, for the moment, think of co-occurrences as the outcome of a decision making process. However, the Revealed Relatedness method itself is not limited to the depicted situation, but rather general. Let us assume that there are two stages in the generation of co-occurrences. In the first stage, actors in each class decide whether it would be beneficial to link to other classes. To identify the attractive candidate classes, the actors make use of class-specific information. As a result, for each combination of two classes, there are a number of actors that would consider setting up a co-occurrence link. This number depends on both the general characteristics of the originating class and on those of the receiving class. In the industry example, firms investigate whether they need to diversify at all and, if so, which industries would be good diversification candidates. For this, they take into account many factors, such as profitability and expectations about the competitive setting in their own industry and in candidate industries.

At the end of this first step, each combination of classes is under consideration by a number – possibly zero – of agents. However, not all class-combinations are equally feasible. As actors have to build on their existing strengths, only classes that are sufficiently similar to the original class can yield stable co-occurrence links. Therefore, in the second stage, agents investigate how related the candidate class is to their own class.

Direction of co-occurrences
In many applications, the classes that co-occur in an entity can be ordered according to some sort of hierarchy. Patents are classified in a main technology class and some supplementary technology classes. In firms, some activities belong to the core business, whereas other activities take a more marginal position. Often, the temporal dimension can be used to derive information on the direction of relatedness. For example, moving from one decade to another countries add new industries to their existing portfolio. In all these cases, it is possible to interpret a co-occurrence as going from an originating class to a receiving class. The direction of a co-occurrence can often be chosen in such a way that it can be interpreted as an indication of varying degrees of complexity. For example, in a firm, non-core activities should normally be expected to take place in fields that do not require large investments in capabilities that are not used in the core-activity. The reason for this is that capabilities are costly to maintain, and should therefore contribute substantially to the performance of the firm.

If we now turn back to our model, let us call the class of the investigating actor – the originating class – class i, and the candidate class – the receiving class – class j. The closer j is related to i, the easier it is for the agent in class i to establish a co-occurrence from i to j. This will translate into a higher proportion of agents that are already interested in the combination of i and j to actually build a co-occurrence link from i to j. This proportion is what we call the *Revealed Relatedness from i to j*.

In our example of industries, the firms in industry i that are investigating industry j as a diversification candidate will have to decide whether it is feasible to add industry j to their portfolio. If most of the production processes, raw materials and skills used in industry i can also be used in industry j – or if they at least have close analogues in industry j – a firm can add j to the portfolio without great investments or changes. The Revealed Relatedness of i to j is equal to the share of the firms in industry i that actually expand into industry j as a percentage of the firms that were investigating the link between industry i and j.

First step estimates:
In sum, our measure of Revealed Relatedness is equal to the observed number of co-occurrence links, corrected for class-specific characteristics that influence the overall propensity of the involved classes to participate in co-occurrences. It is therefore a good measure of the ease with which industry j can be added to a firm in industry i, and thus of the relatedness between industries.

In terms of probability theory, we can say that the Revealed Relatedness from i to j is the probability of a co-occurrence of classes i and j, given the number of agents investigating the link from i to j. Let us introduce some notation to formalize the above:

C_{ij} : number of agents that consider participating in a co-occurrence from i to j

L_{ij} : number of realized co-occurrence links from i to j

RR_{ij} : Revealed Relatedness of i to j, i.e. probability that an actor in i establishes a co-occurrence link to j given that the actor is investigating j.

By definition, if a link is being investigated, it will be established with probability equal to its Revealed Relatedness. The number of observed links can therefore be regarded as the outcome of the following binomial process:

$$(1) \qquad P\left(L_{ij} = l_{ij} \middle| C_{ij} = c_{ij}\right) = RR_{ij}^{\ l_{ij}} \left(1 - RR_{ij}\right)^{c_{ij} - l_{ij}} \binom{c_{ij}}{l_{ij}}$$

The conditional mean of L_{ij} of this binomial distribution is:

$$(2) \qquad E\left(L_{ij} \middle| C_{ij} = c_{ij}\right) = RR_{ij} c_{ij}$$

C_{ij} is assumed to be determined by a number of class specific characteristics that we call v_i and w_j. We use these characteristics to predict the number of links between i and j.

First, we have to estimate the relation between class characteristics and co-occurrence links. A way to do this is to treat the problem as a count data regression problem. As most of the class pairs typically have zero links, a Zero-Inflated Negative Binomial regression analysis is most appropriate, although the validity of our framework does not depend on the exact formulation of the regression model:

$$(3) \qquad E\left(L_{ij} \middle| v_i, w_j, \varepsilon_{ij}\right) = \left[1 - \Pi_0\left(\gamma + v_i'\delta_i + w_i'\delta_j\right)\right] e^{\alpha + v_i'\beta_i + w_i'\beta_j + \varepsilon_{ij}}$$

As in any zero-inflated count data regression, the equation consists of two parts (e.g. Greene, 1994). The exponential, right most, part is the count data part. This determines the number of links that are generated, given our regressors. The left most part is usually called the regime selection equation, and is a way to cope with the overwhelming number of zeros in the observed data.

Both the regime selection equation, and the count data equation depend on industry characteristics only. With ^ indicating fitted values, we can calculate the predicted number of links as follows:

$$(4) \qquad \hat{L}_{ij} = \left[1 - \Pi_0\left(\hat{\gamma} + v_i'\hat{\delta}_i + w_j'\hat{\delta}_j\right)\right] e^{\hat{\alpha} + v_i'\hat{\beta}_i + w_j'\hat{\beta}_j + \varepsilon_{ij}}$$

As only the number of investigating agents, C_{ij}, is supposed to vary with the general industry characteristics, the variations in predicted outcomes, $\hat{L}_{ij}$, can be fully attributed to variations in C_{ij}. Assuming that the model is correctly specified, remaining variation must be contributed to differing degrees of relatedness between classes. Thus, it is tempting to interpret the residuals, $e^{\hat{\varepsilon}_{ij}}$, as the estimated Revealed Relatedness of i to j and to split the model as follows:

(5) $E\left(L_{ij}|v_i,w_j,\varepsilon_{ij}\right)=C_{ij}e^{\varepsilon_{ij}}=C_{ij}RR_{ij}$, where

(6) $E\left(C_{ij}|v_i,w_j\right)=\left[1-\Pi_0\left(\gamma+v_i'\delta_i+w_j'\delta_j\right)\right]e^{\alpha+v_i'\beta_i+w_j'\beta_j}$

However, as we include a constant in our model, this amounts to setting the average value of ε equal to zero and, as a consequence, e^{ε}, or the RR in the typical class combination[1], equal to 1. As the RR is a probability, this is obviously impossible. We must therefore account for the fact that the RR for the typical class combination is a part of the term

$$\left[1-\Pi_0\left(\gamma+v_i'\delta_i+w_j'\delta_j\right)\right]e^{\alpha+v_i'\beta_i+w_j'\beta_j}.$$

Let us assume that the RR will not affect the regime selection equation. In this case, we can split the term α in two, one part capturing the average propensity to investigate co-occurrence links ($\tilde{\alpha}$) and the other part capturing the typical Revealed Relatedness across all industry combinations (η):

(7) $E\left(L_{ij}|v_i,w_j,\varepsilon_{ij}\right)=\left[1-\Pi_0\left(\gamma+v_i'\delta_i+w_j'\delta_j\right)\right]e^{\tilde{\alpha}+v_i'\beta_i+w_j'\beta_j+\varepsilon_{ij}+\eta}$

This breaks the regression equation down in two parts:

(8) $$\begin{aligned}E\left(L_{ij}|v_i,w_j,\varepsilon_{ij}\right) &= \left[1-\Pi_0\left(\gamma+v_i'\delta_i+w_j'\delta_j\right)\right]e^{\tilde{\alpha}+v_i'\beta_i+w_j'\beta_j}e^{\varepsilon_{ij}+\eta}\\ &= C_{ij}e^{\varepsilon_{ij}+\eta}\\ &= C_{ij}RR_{ij}\end{aligned}$$

where $C_{ij}=\left[1-\Pi_0\left(\gamma+v_i'\delta_i+w_j'\delta_j\right)\right]e^{\tilde{\alpha}+v_i'\beta_i+w_j'\beta_j}$ and $RR_{ij}=e^{\varepsilon_{ij}+\eta}$

As this implies:

(9) $\hat{L}_{ij}=\hat{C}_{ij}e^{\eta}$

We get the following expression for the predicted number of links:

(10) $\hat{C}_{ij}=e^{-\eta}\hat{L}_{ij}=k\hat{L}_{ij}$

where ^ indicates fitted values and k is a constant.

Using equation (2), we get the following estimate for RR_{ij}:

[1] As the expectation of a function is not equal to the function of an expectation, it would be a mistake to call this the average RR across all class combinations.

(11) $$\hat{RR}_{ij} = \frac{L_{ij}^{obs}}{\hat{C}_{ij}} = \frac{L_{ij}^{obs}}{k\hat{L}_{ij}}$$

where L_{ij}^{obs} is the observed number of co-occurrences from i to j.

(6) and (8) are different in that equation (8) shifts part of the intercept of the regression equation into the RR term. However, we do not observe the actual investigation process, but only the variations in it across class combinations that can be attributed to the general class characteristics v and w.[2] There is no straightforward empirical way to determine how large this part of the intercept should be. However, we do know that the RR terms should be smaller than 1 (by construction, they will always be larger than 0). As the number of industry combinations is typically large, there will also be some outliers for the RR estimates. Nevertheless, the vast majority of RR estimates should be smaller than 1. This can be achieved by choosing a suitable value for k. In this paper, we use such a value that 90% of all class combinations that have at least 1 link[3] are smaller than 1:

(11) $$k \equiv perc\left(\frac{L_{ij}^{obs}}{\hat{L}_{ij}}, 0.90 \middle| \frac{L_{ij}^{obs}}{\hat{L}_{ij}} > 0 \right)$$

where $perc\left(x_{ij}, p \middle| x_{ij} > a\right)$ denotes the pth percentile of variable x_{ij} across all i and j for which x_{ij} is larger than a.

Now, for the vast majority of cases, RR estimates are between 0 and 1. The remaining values for which $\frac{L_{ij}^{obs}}{\hat{C}_{ij}} > 1$ will be dealt with later on, when we develop a Bayesian extension of the method.

Indirect links

With the method outlined above, we can now calculate an entire matrix of relatedness indices for all possible combinations of classes. If such a matrix is to be a good representation of the relatedness between classes, its elements must maintain a certain level of transitivity with respect to each other: if class j is very similar to class m and class m is very similar to class i, then class i is probably also very similar to class j. To our knowledge, the first authors to exploit such information inherent in all relatedness matrices are Bryce and Winter (2006). Like Teece et al. (1994), they use the hypergeometric approach to calculate a relatedness matrix based on industrial portfolios of firms. This gives non-zero estimates for many industry combinations. Bryce and Winter then use shortest path analysis to calculate the relatedness for all pairs of industries that had zero co-occurrences.

We share the opinion that a relatedness matrix contains a lot of valuable "indirect" information. However, we take a different approach to put this information to use. Let us assume that we have a situation where many of the estimated RR indices are equal to zero. If we concentrate on one of the class pairs that yield zero relatedness, it is not necessarily the case that this zero is a consequence of a lack of information. An estimate

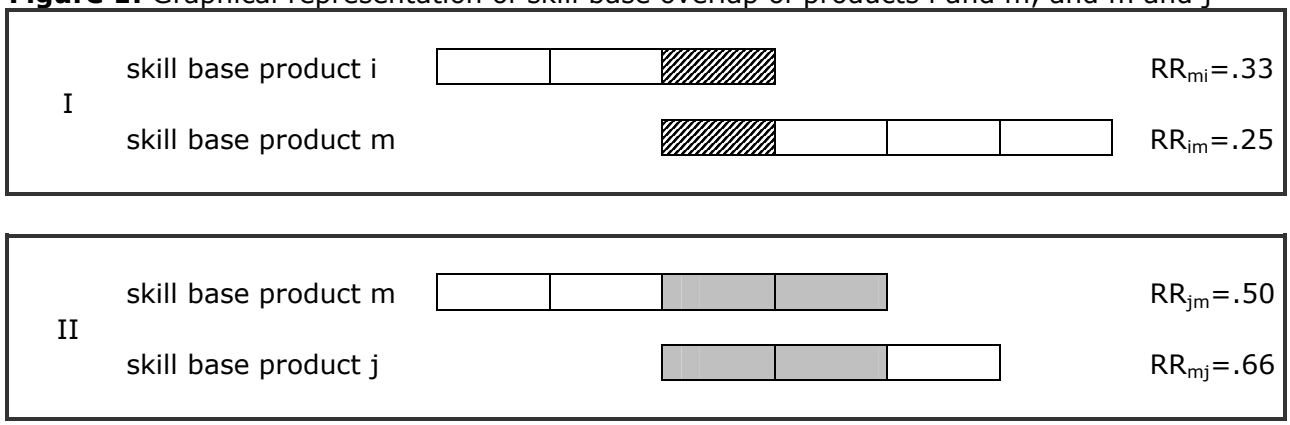
[2] As a matter of fact, we define RR in such a way that we can predict these variations without any error. All remaining variation is attributed to variation in the RR. In practice however, this means that we will measure RR with some error. In this sense our method is not different from any of the other methods encountered in the literature.

[3] As the vast majority of class combinations has zero links, using a rule that takes the 90th percentile of all class combinations, including the ones with zero links, will most likely end up with k equal to zero.

of zero may just indicate that the classes are unrelated. Therefore it is also not obvious that we should change all zero relatedness estimates in favour of an indirect estimate of relatedness. Nevertheless, it may still be the case that we do not observe any co-occurrences because we only expect a very small number of actors to investigate this specific combination of classes.The count data procedure may even predict the number of investigating agents to be smaller than 1, suggesting that the zero relatedness is caused by a lack of investigating agents. In such a case, an outcome of zero relatedness becomes questionable. As a matter of fact, this reasoning applies to all estimates for which the predicted number of investigating agents is small.

For combinations of classes with a low predicted number of investigating agents, it therefore does make sense to use the indirect relatedness information that is contained in the matrix as a whole. Let us depict a class as consisting of a number of elements. For example, an industry can consist of a number of production processes, or use a number of skills. These skills are not necessarily unique to the industry, but could be used in other industries as well. The greater the overlap in the elements of two classes, the more the classes are related to each other. Figure 1 shows the set of skills used in three different industries, i, m and j.

Figure 1: Graphical representation of skill base overlap of products i and m, and m and j



Let us now assume that we are unable to directly estimate the relatedness from i to j, because the predicted number of agents investigating this combination is very small. However, let us also assume that it is possible to assess the relatedness between m and j and between m and i respectively. The outcomes are depicted in the figure (to the right). Product m is quite similar to product j, as most of j's elements are also part of m: RR_{mj} = 66%. We can therefore use m as a proxy for j.[4] According to our estimate of RR_{im}, 25% of m's elements are also present in i. Our best guess is that the same must be true for j, because m is very similar to j. Therefore, we would guess that RR_{ij} = 25%. The closer m is to j as measured by RR_{mj}, the more confident we will be. Hence, a sound strategy to derive indirect information on a pair of classes, (i,j), is to look for a class m, for which relatedness with both i and j is reliably measured and that is very related *towards j*. In this case, the relatedness *from i* to m can be used as a proxy for the relatedness from i to j: $R_{ij} = R_{im}$.

Adding second step estimates: Bayesian updating

By now, we have developed two different ways of estimating the RR index of (i,j). Nonetheless, in empirical situations there might also be other relatedness matrices available that were constructed in quite different ways. For example, we may use expert opinions. In other applications, there may be an input-output based relatedness matrix.

[4] Note that we use RR_{mj} because this expresses the number of elements that are found in both m and j as a percentage of the number of elements of j. The reason for this is that we assume that the decision to effectuate a link from i to j depends on how many of the elements of j are shared by i. Therefore, the denominator of the percentage should be the number of elements of j.

The main difference with the indirect estimates is that it is not clear how the numeric values can be compared to the frequentist RR indices. All the same, it is possible to use regression analysis to relate the first step RR indices to such an external relatedness matrix.[5] The main question would however be how to merge this information with the information that is present in the observed co-occurrences. For this, we will reformulate the estimation model in a Bayesian framework. As notation becomes somewhat cumbersome, we drop the subscripts in this section.

Let us return to equation (1) in which we express the probability of observing a co-occurrence conditioned on the number of investigating agents. Making it explicit that we do not know RR and dropping subscripts, our likelihood function – the probability of observing l co-occurrences, given RR and C – is:

$$(12) \qquad P\left(L=l \middle| RR=r, C=c\right) = (r)^{l}(1-r)^{c-l}\binom{c}{l}$$

where we define $\binom{c}{l} \equiv 0$ if $c < l$

We are, however, interested in the probability distribution of RR conditional on C and L. Using Bayes' law we get:

$$(13) \qquad P\left(RR=r \middle| L=l, C=c\right) = P\left(L=l \middle| RR=r, C=c\right)\frac{P\left(RR=r \middle| C=c\right)}{P\left(L=l \middle| C=c\right)}$$

According to our model, the number of investigating agents, C, is not dependent on RR. Therefore, we can drop the conditioning on C in the numerator in (13):

$$(14) \qquad P\left(RR=r \middle| L=l, C=c\right) = P\left(L=l \middle| RR=r, C=c\right)\frac{P\left(RR=r\right)}{P\left(L=l \middle| C=c\right)}$$

As always in Bayesian inference, we have to specify our prior beliefs about the variable we are interested in. Given the binomial likelihood function, we choose the unconditional – or prior – probability of RR to be distributed according to a $BETA(a,b)$ distribution.[6] a and b are parameters that reflect our prior knowledge of RR.[7] The prior expected value of a $BETA(a,b)$ distributed variable is:

$$(15) \qquad E\left(RR\right) = \frac{a}{a+b}$$

[5] The main problem is that we need relatedness estimates that are measured in the same units as the RR index. One way to do that is to use a regression analysis of the first step estimates on the external relatedness matrix. The predicted values of such an estimation would be expressed in RR units. However, as the RR matrix contains proportions data which are always between 0 and 1, standard methods are not appropriate. However, methods have been developed for such data (e.g. Papke and Wooldridge 1996).

[6] Due to the size of the relatedness matrix, which for 200 different products would contain 39 800 elements, we choose to use the conjugate prior. That way, we can arrive at an analytical expression and do not have to engage in time-consuming posterior simulation.

[7] If a and b are both equal to 1, the BETA distribution simplifies to a uniform distribution on the interval [0,1]. Taking a=b>1 will give rise to a symmetric distribution with maximum probability density on mean ½. a=b<1 leads to a symmetric distribution with asymptotically increasing probability densities towards the edges, 0 and 1. For values of a and b such that a≠b, asymmetric distributions arise, with sometimes asymptotic increases towards one edge and sometimes maximum probability densities at (a-1)/(a+b-2).

At this stage, we can use the intermediate RR estimates. As we did not use any information on the number of links from i to j, but only information about the links between i and m and m and j, these estimates can be used as prior information. Let us choose a and b in such a way that $E(RR)$ equals our indirect estimate.[8] The prior probability of RR is then equal to:

(16) $$P(RR = r) = r^{a-1}(1-r)^{b-1}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}, \text{ with } a,b > 0$$

Filling in (12) and (16) in (14), we arrive at the following expression (details can be found in Appendix A):

(17) $$P(RR = r | L = l, C = c) = (r)^{l+a-1}(1-r)^{c-l+b-1}\frac{\Gamma(a+c+b)}{\Gamma(l+a)\Gamma(c-l+b)}$$

We have now expressed the probability distribution of RR in terms of the number of observed co-occurrences, the information present in the intermediate estimates (as captured by a and b) and the number of investigating agents, C, is unobserved. Therefore, in a final step, we have to remove the conditioning on C. To achieve this, we treat C as the outcome of another binomial process. The number of agents in the originating class, class i, is a known variable. We also know how many agents we would *expect* to investigate the co-occurrence from i to j. If we for the moment reintroduce the subscripts, and use the expression for the mean of a binomial process, we get:

(18) $$E(C_{ij}) = N_i q_{ij}$$

where N_i is the number of agents in class i, and q_{ij} is the probability that they will investigate the co-occurrence to j.

Setting this expected value equal to our estimates for C_{ij}, we can calculate q_{ij}. The probability distribution of C_{ij} is binomial. Dropping subscripts again, we get:

(19) $$P(C = c) = (q)^c(1-q)^{N-c}\binom{N}{c}$$

We can use this expression together with the law of total probability, to get rid of the conditioning on C in equation (17). Appendix B shows a detailed derivation. The end result of this is that we can calculate a Bayesian variant of Revealed Relatedness: the expected RR given the number of observed links.[9]

[8] The shape of the BETA distribution can be quite peculiar. In particular, if any of the values a or b are chosen smaller than 1, the probability density near one of the edges explodes. This tends to put a lot of probability mass near the edges. We would, however, rather like the mode of the distribution to be near the value of the indirect estimate. After some experimentation, we decided that this is best accomplished by choosing the minimum of a and b to be equal to 2 and at the same time to have a/(a+b) equal to the value of the indirect estimate.

[9] We can now also calculate RR indices for which $L_{ij}^{obs} > \hat{C}_{ij}$. In our Bayesian estimates, L_{ij} can take any value between 0 and N_i, without causing any problems.

$$(4.21)\quad E\left(RR\middle|L=l\right)=\sum_{c=l}^{N}\left\{\frac{l+a}{a+c+b}\cdot\frac{\beta(a,b,c,l)q^{c}(1-q)^{N-c}\binom{c}{l}\binom{N}{c}}{\sum_{\tilde{c}=l}^{N}\beta(a,b,\tilde{c},l)q^{\tilde{c}}(1-q)^{N-\tilde{c}}\binom{\tilde{c}}{l}\binom{N}{\tilde{c}}}\right\}$$

where: $\beta(a,b,c,l)=\frac{\Gamma(l+a)\Gamma(c-l+b)}{\Gamma(a+b+c)}$, with $\Gamma(.)$ the Gamma function.

4: Empirical application: calculation of industry RR using product portfolios of plants

Diversification as a branching process

A better understanding of industry relatedness is instrumental in many economic applications. In strategy research, an extensive literature argues that multi-industry firms are active in industries that share certain commonalities (e.g. Teece 1982, Teece et al. 1994, Breschi et al. 2003). This suggests that firms tend to expand their business by moving into industries that are related to their current activities. As a result, the growth of a firm's portfolio resembles a branching process, in which future expansion builds on competences that were gathered in the past as in Penrose (1959). As coherent portfolios are thought to exploit economies of scope, they should be more efficient than incoherent portfolios. Accordingly, studying the portfolios of surviving firms sheds light on which industries share many commonalities. Therefore, Teece et al. (1994), invoke the survivor principle according to which economic competition will weed out inefficient organizations, and surviving organizations display, as a consequence, efficient practices. Our application is similar to the one in Teece et al. (1994). However, where Teece and his colleagues look at *firms* that own plants in different industries, we look at different products that are produced in one and the same *plant*.

We argue that there are a number of reasons why firms produce several different products that are not linked to the technology used in production processes. For example, portfolio construction at the firm level may reflect marketing economies or risk diversification strategies. In holding companies and large conglomerates, access to cheap capital, superior management capabilities and cross-financing may play a key role in the determination of the portfolio. However, at the plant level, such considerations are less important. It is likely that products are built in the same plant, because the production processes involved are similar. Similarity in skills and routines embodied in human capital, but also similarities in the physical capital, the machinery and raw materials that are used, will generate economies of scope at the plant level and therefore make joint production of products attractive. However, it is unlikely that two products whose production processes have no commonalities whatsoever, will be produced in one and the same production facility. Rather, it may make sense to use separate production locations to avoid that the different production process interfere with each other. Therefore, a RR matrix based on plant portfolio data is likely to reflect relatedness at the level of the production processes and the technology involved.

Data and implementation

The database we use contains information on products produced in thousands of plants in all manufacturing industries in Sweden. As we are interested in the relatedness between industries, we translate these product codes into industry codes. Over the entire sampling period, about 57% of all observed plants produce products in only one industry.

The vast majority of plants that are active in multiple industries, are only active in two industries (22 % af all plants).
In the regional analysis of section 5, we will use data on all Swedish labour market regions, or A-regions. These data we take from a database that covers all manufacturing plants in Sweden with more than 5 employees (1968-1990). From 1990 to 2002 the sample is limited to plants with more than 5 employees belonging to firms with over 10 employees.

For our zero-inflated negative binomial regressions, we use industry aggregates at the Swedish level. The database on the product portfolios provides us with total sales value for each product in each plant, which we then aggregate to the sales level of industries at the national level. The number of active plants, total profit, total valued added and total number of employees are taken from the database we also use in the regional analysis. Again, we add up all plant level data to arrive at national industry aggregates. With these data, we predict the number of co-occurrences in every single year. These predictions are then used in the way described in section 3.

Often, a plant produces some main products and a couple of minor by-products. We can rank the industries in which a plant is active according to the sales value of the products involved. The industry in which the highest sales value is generated is taken to be the core business of the plant. This hierarchy can be used to establish the direction of co-occurrences: co-occurrences are defined as links that run *from* the core industry *towards* the other industries. The implicit assumption is that non-core products are kept in a portfolio because their production processes are relatively simple extensions of the production processes used in the production of the core products. However, it may also be the case that all elements in a portfolio share commonalities. If this is true, each combination of two products in a portfolio represents a co-occurrence, but it is impossible to establish the direction of links. In total, we get between 10,000 (in the year 1970) and 2,500 (in 2000) directed (asymmetric) co-occurrences and between 63,000 (in 1970) and 17,700 (in 2000) undirected (symmetric) co-occurrences. We also create versions of our RR matrix that are based on such undirected co-occurrences.[10] In these versions, all possible combinations of two industries in a portfolio are administered as a co-occurrence.

To make sure our intermediate (indirect) estimates are reliable, we calculated them using only elements (i,j) of our first step RR matrices for which we predicted more than 10 co-occurrences.[11] Otherwise, the intermediate estimates are base don too weak evidence to improve the first step estimates. The correlation between our intermediate and first step estimates is strong: across the entire sample, the rank correlation between the elements of the intermediate RR matrix and the first step RR matrix are in every year above .40 and typically around .50.[12] However, we use the intermediate estimates only if the frequentist estimates are based on predictions of 10 or less co-occurrences. In other instances, we do not expect the intermediate estimates to improve the frequentist estimates. To smooth our relatedness matrices, we take 3 year moving averages of the RR matrices.

General properties of the relatedness industry networks: symmetric links

[10] In general, using undirected co-occurrences does not necessarily lead to symmetric RR matrices. This depends on the exact specification of the zero-inflated negative binomial model. Moreover, the indirect estimates introduce asymmetry as well. To arrive at symmetric matrices, we therefore have to symmetrize the RR matrix by imposing $RR_{sym}(i,j)=max(RR(i,j);RR(j,i))$.
[11] More over, we demand that the estimated relatedness between m, the proxy industry, and j, RR(m,j) are over twice the average of all relatedness estimates that were based on at least one co-occurrence. This ensures that m is a strong proxy for j.
[12] In the asymmetric case, correlations are considerably weaker, but there the number of elements that meet the conditions for recalculation are also considerably smaller.

Our estimated RR matrices allow us to regard industries as nodes in a network of relatedness links. We call this complete network industry space, in analogy to the product space in Hidalgo et al. (2007). Below, we visualize different instances of industry space using a spring embedded algorithm.[13] This algorithm treats the industry nodes as equally charged particles that exert a repulsive force on each other. However, the industries are attached to each other with springs the rigidity of which reflects the strength of RR links between them. This results in a field of forces that scatters industries across the entire plane, but in such a way that closely related industries are also located closely together in that plane. If two industries are visually close, this can be interpreted as that they are located close to each other in industry space.

Figure A displays industry space for Sweden in 1970 using the symmetric RR matrix. Each node (circle) represents an industry and the ties represent the 1500 strongest RR links in 1970. We have applied a colour scheme that shows the 2-digit family in the Swedish SNI69 industrial classification system of each node:

Food beverages and tobacco
Textiles wearing apparel and leather industries
Wood and wood products
Pulp paper and paper products
Chemicals and chemical petroleum coal rubber and plastic products
Non-metallic mineral products
Basic metal industries
Fabricated metal products machinery and equipment
Fabricated metal products machinery and equipment, R&D intensive
Other manufacturing

Table 1: Colour scheme used for 2-digit industry categories in the network graphs[14].

In the 1970 industry space (figure 2), the pattern of industries clearly follows the 2-digit classification. Nodes that share a colour are, in general, clustered closely together. On a first glance, we would therefore say that the RR measure classifies industries in a similar way as the SNI69 hierarchy. However, there are also quite some exceptions. For example, members of the non-metallic mineral products industry are scattered across a large part of the industry space. The quality of the match between the SNI69 hierarchy and the RR matrix can be quantified by calculating the correlation between the relatedness matrix derived from the SNI69 hierarchy[15] and the RR matrix.[16] For 1970, we get an estimated average rank correlation of .43 with a variance of .04 across the columns of the matrices. In other decades we obtain similar estimates (see table C.7 in Appendix C) which confirms our impression that the RR matrix and the SNI69 hierarchy yield similar pictures. This is quite remarkable as the RR calculations make no use whatsoever of the hierarchical relations within the SNI69 system. At the same time,

[13] The graphs are made with the Software packages UciNet and NetDraw (Borgatti et al 2002, Borgatti 2002). For a description of the spring embedded algorithm, we refer to the help file of these packages.
[14] To highlight technologically advanced industries, we adapt the 2-digit system by dividing the fabricated metal products machinery and equipment into R&D intensive industries and regular industries using the categories of Ohlsson and Vinell (1987).
[15] We construct a SNI69 relatedness matrix by counting for each combination of two industries the number of digits both industries share. For example the relatedness between an industry 311200 and 321000 is equal to 1, whereas the relatedness between industries 345209 and 345202 is equal to 5
[16] In the remainder, correlations between two relatedness matrices are calculated as the average rank correlations of their columns. Reported variances are do not refer to the variance in the estimate of this rank correlation coefficient, but rather to the variances of the rank correlations across the columns. The variance expresses therefore how much the correlation between two relatedness vectors differs across industries.

however, the correlation is low enough to conclude that there are also some marked differences.

Apart from such differences, the industry space also gives some information that cannot be derived from the SNI69 classification. For example, paper and fabricated metal products are mostly located in the first and second quadrants. Wood and textiles clusters are positioned in the second and third quadrant, whereas chemicals and food clusters are located in the fourth quadrants. This shows where the clusters around the various 2-digit industries are located relatively to each other. As an example, basic metal industries are positioned closely to the fabricated metal products, machinery and equipment industries, whereas many members of the non-metallic mineral products industry are more closely related to industries in the chemicals cluster. Zooming in on the borders between two clusters, we see that these borders are quite fuzzy. Some of the paper and paper product industries are located along the lower rim of the food and chemical clusters. This indicates that these industries are rather related to chemicals whereas this is less so for other industries, such as printing and newspapers. Finally, it is possible to move to the level of the individual industry and assess its position relative to all other industries. Here we can observe, for example, that explosives take a central position in industry space. Another example is the communication industry, that, in 1970, is located at quite some distance from the other technologically advanced industries.

For comparison, the industry space of 2000 is depicted in figure 3.[17] By and large, the clusters that existed in 1970 are also present in 2000. However, there have been quite some changes. For instance, the fabricated metals cluster has becomes more diffuse, but the R&D intensive industries have moved closer together and more to the centre of the cluster. This suggests that there has been an increased specialization within the industries of fabricated metals, but it is beyond the scope of this paper to explore this issue further.

Although there is some stability it is clear that industry space is not static but rather dynamic. As we do not expect relatedness structures to change very fast, on a yearly basis, shifts in relatedness would point out a weakness in the RR method. Therefore, it is reassuring to see that the RR matrices are highly stable in the short run. If we consider the symmetric RR matrices between 1970 and 1971, the average rank correlation is 0.83. Between 1970 and 1972 it is 0.78. The short term changes in the later part of the period are quite a bit larger (correlation 0.63 from 2000 to 2001 and 0.52 from 2000 to 2002). This may be caused by the fact that the number of plants in our sample goes down over time, reducing the quality of our estimates.[18] In the long run, in contrast, we expect that structural changes in the economy are likely to leave an imprint on industry space. Looking at the symmetric RR matrices between 1970 and 1980, we find a correlation of 0.55. Comparing 1970 to 1990 estimates this correlation is reduced to 0.32 and it falls to 0.26 when we compare 1970 and 2000 (for a comparison between other decades and the first and second step matrices, see tables C.3 to C.6 in appendix C). These numbers suggest that industry space indeed underwent some significant changes over the past decades. In this way, our RR calculations could also be used to describe the velocity of structural change in economies.

General properties of industry space: directed links

One of the strengths of our method is that it also allows for a directed interpretation of the concept of relatedness. Rank correlations comparing the columns and rows of asymmetric RR matrices are typically around 0.4 (see table C.8 in appendix C). The "outgoing relatedness" is therefore quite similar to the "incoming relatedness", yet, there are also profound differences.

[17] We initialized the spring embedded algorithm with the position of nodes in figure 2 to improve comparability.
[18] To better assess the quality of the RR estimates, tables C.1 and C.2 in appendix C contain correlation coefficients for the remaining decades and for both the first and second step estimates.

As discussed earlier, the interpretation of the asymmetric industry space is that the direction of the arrows indicate a move down the complexity hierarchy from industries producing a more complex set of products to industries with less complex output. This complexity should be interpreted primarily as complexity in production technology. As an example, figure 4 shows the RR links of the medicines industry in 1980. The medicines industry has outgoing links to a large set of industries such as soap, pesticide, resins and animal feeds. This shows that manufacturers of medicines can relatively easily venture into niches in those industries. The explanation we porpose for this is that these manufacturers already possess the complex skills needed in the production of medicines. By adding less complex skills they can manufacture new products that compete with the products of other industries. However, it does not mean that a manufacturer of medicines can easily appropriate the skills to manufacture *all* products in the new industries. Rather, the producer can move into specific niches for which the skills involved in the production of medicines are complementary. A good example is the production of soap products[19] with a therapeutic application. A manufacturer specialized in soap products, in contrast, generally cannot compete in niches of the medicines industry without large efforts and investments.

The link from other textiles to medicines is harder to explain in this light. The interpretation of the link is that producers of other textiles, a category containing a large diversity of products, can move without too much effort into niches of medicines production. The reverse move would be much harder. This seems counter intuitive. However, if we take a closer look, we find that this strong link appears because the production of bandages is classified in the other textiles sector. Yet, obviously, bandages have almost exclusively medical applications. A plant that specializes in bandages will therefore be more likely to produce medical desinfectants as a secondary product than pairs of jeans. A similar problem arises with the industry of other food, where the "nuisance products" are spices, herbs and sauces.

To summarize, our RR method yields very promising empirical results. However, none of the above can count as a real quality assessment of the RR measure. We will put the RR matrix to the test by using it to predict changes in the industrial profiles of regions. As the region has not played any role in our analyses so far, finding out whether industry space has any explanatory force in this example can be considered as a litmus test for its general usefulness in economic applications.

[19] This category also includes detergents, essential oils and perfumes.

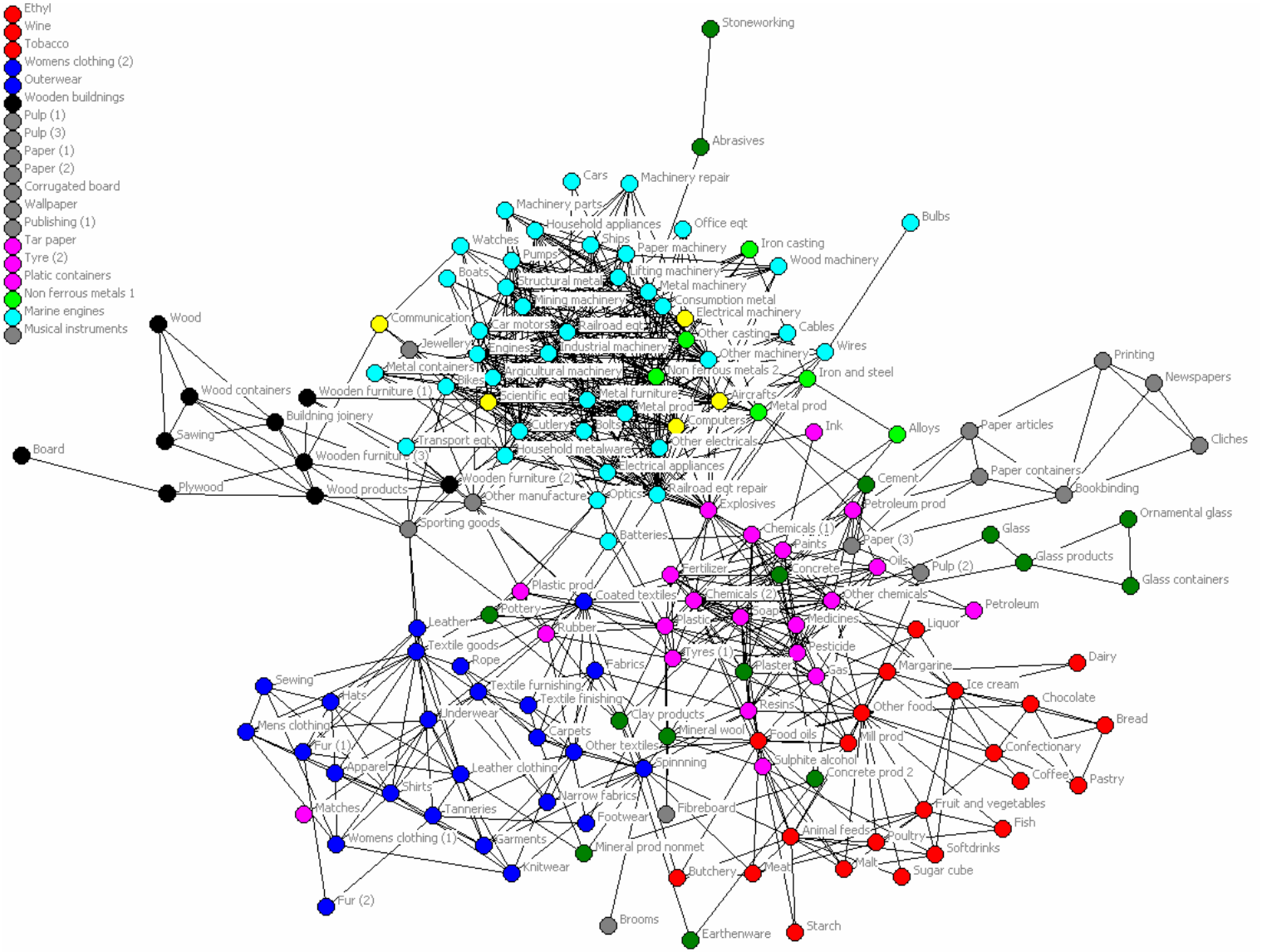


Figure 2: Industry Space in 1970. Isolated nodes (with threshold of 1500 most important links) in North-East corner.

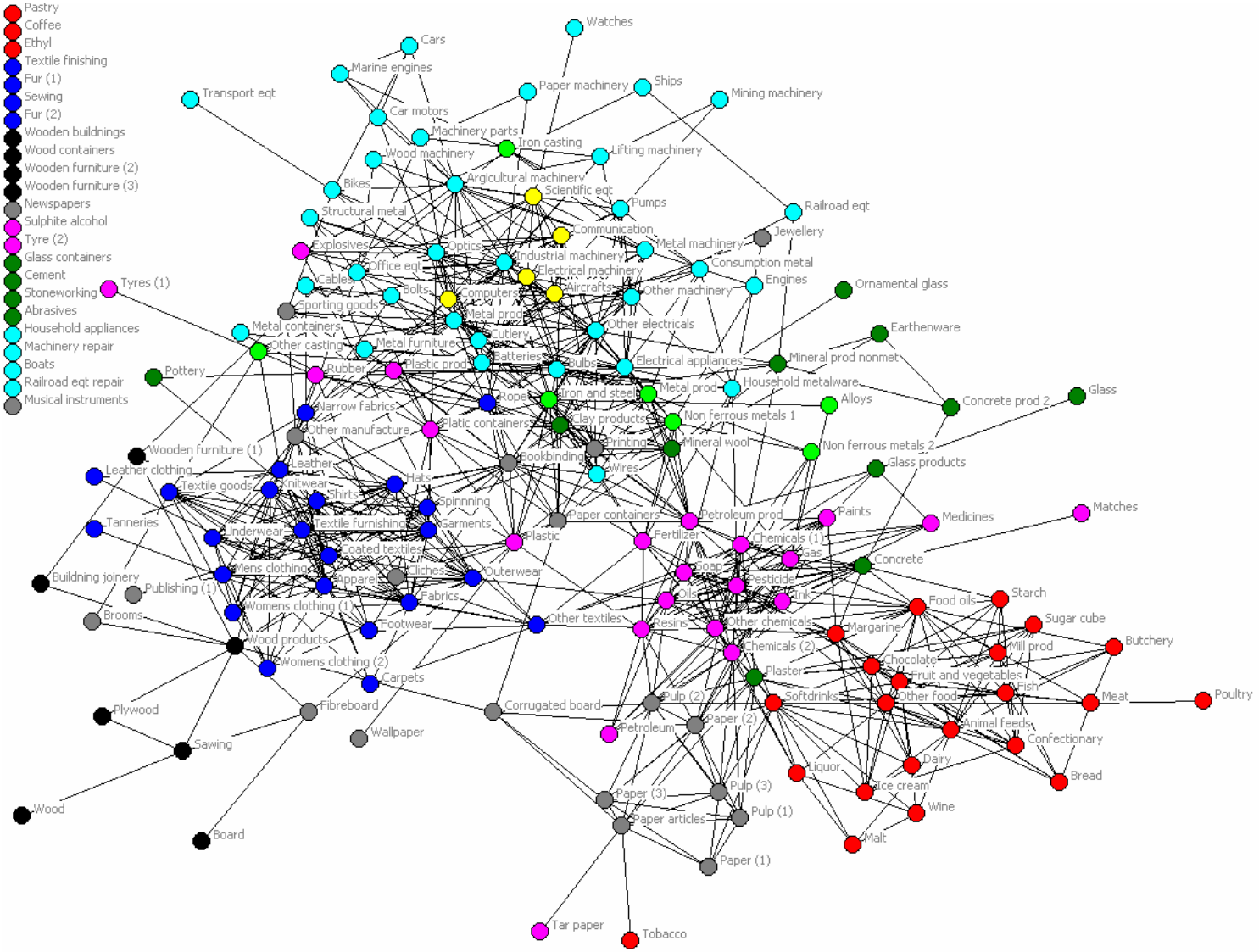


Figure 3: Industry Space in 2000. Isolated nodes (with threshold of 1500 most important links) in North-East corner.

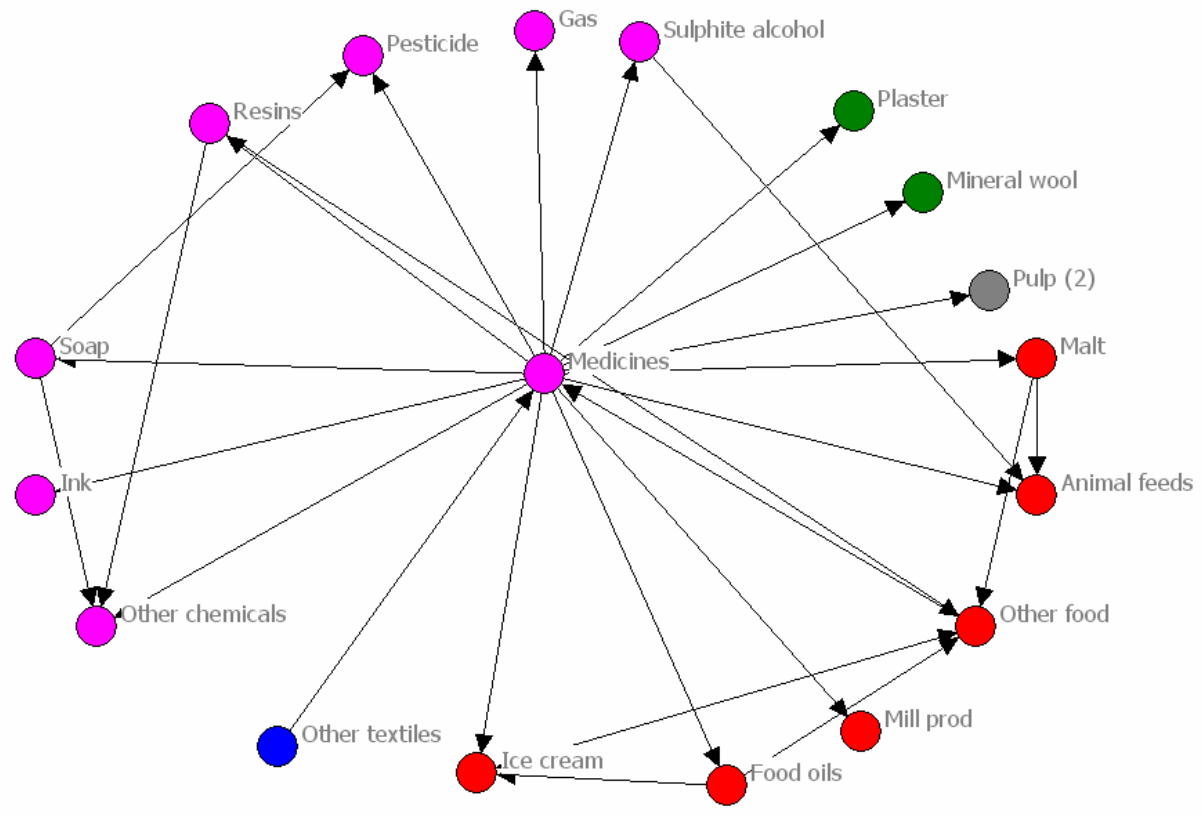


Figure 4: the ego-network for medicines in 1980.

Analyses of Regional Structural Transformation Processes
In the course of their history, regions build up tangible and intangible assets to the benefit of their local firms. Good examples of such assets are cost-effective producer-supplier networks, specialized infrastructure, a trained workforce and a strong regional brand. Moreover, it is often argued that, by cooperating, the actors in a region engage in a process of collective learning (Florida 1995, Asheim 1996). In economic geography, the benefits of these regional features have been intensively debated within the industrial districts, cluster and regional innovation system approaches (e.g. Marshall 1920, Asheim 2000, Porter 2000). Furthermore, empirical studies have shown that regions that specialize in the production of specific products are often particularly good at fostering a fertile environment for manufacturers in these main industries (Amin 2003, Asheim et al 2003).

All this leads us to expect that regions generally will show a higher propensity to diversify into technologically related industries than into technologically unrelated ones. With such a strategy, regional actors can fall back on known routines and avoid large switching costs. At the same time, industries that are not related to any other local industries may become isolated and run a higher risk of exiting the region. The industry space can be used as an analytical tool to investigate such issues of structural transformation in regions.

The regional portfolio and structural transformation
A region's industrial portfolio consists of all industries in which the region has any employment. The RR matrix can tell us how close an industry is to each other industry. However, to implement the notion of coherence of a regional portfolio, we still have to specify how we would measure the closeness of an industry to an entire portfolio of

industries. An intuitive approach is to count to how many portfolio members the industry is closely related. Let us call two industries, i and j, closely related if RR(i,j) is above a threshold value c. The closeness of an industry i to the portfolio of region r, C_{ir}, is therefore a function of c:[20]

$$C_{ir}(c) = \sum_{j \neq i} I\left(j \in PF_r \wedge RR(i,j) >= c\right)$$

where PF_r is the set of industries belonging to the portfolio of region r and $I(.)$ is an indicator function that takes the value 1 if its argument is true and 0 if not.

The choice of c is in a way arbitrary. It should not be too high, because this will treat almost all industry pairs as unrelated, nor should it be too low, as this will take many quite weakly related industries to be close to each other. We strike the balance at c equal to .3. However results are very similar for a large range of values for c. Further, as the symmetric RR matrices are estimated with a higher precision than the asymmetric RR matrices, we build on the symmetric estimates. However, also here, the asymmetric matrices yield very similar conclusions.

To test the value of our RR matrix, we can investigate whether or not the closeness index derived from it helps explain the composition and the change in regional portfolios. More specifically, we want to find out if 1) regional portfolios consist of related industries, 2) regional portfolios are more likely to expand into related industries than into unrelated industries, and 3) regional portfolios are more likely to shed industries that are related to most other industries in the portfolio than technologically more peripheral industries. The number of exits and entries of local industries in regional portfolios are counted by looking at 5-year shifts in portfolios. So, for example, we compare the regional portfolio in 1990 with the portfolio in 1995 to assess whether an industry entered or exited a region.

To answer the first question, we can calculate the rank correlation between a membership dummy variable that takes the value 1 if an industry is a member of a region's portfolio on the hand, and the closeness, C, of that industry to the rest of the region's portfolio members on the other. We indeed find that the correlation between membership and closeness is significant and positive. Averaged across all years in our sample, this correlation equals 0.24 which seems to be rather low. However, if, for each value of C between 0 and 20, we estimate the probabilities that an industry is part of a portfolio and plot that probability against C, we see a pronounced effect of relatedness. In figure 5A we show the the estimated membership probability averaged across all years fora n industry that is close to C other industries in the region. If we move from the lowest values for C in the graph to the highest, the membership probability more than quadruples. This shows that the degree of coherence in regional portfolios is actually quite high.

Figure 5B and 5C show similar graphs for estimated expansion and contraction probabilities. Again, we find a pronounced effect of our closeness variable. As expected, industries that are closely related to many portfolio members have a high probability of entering the portfolio within five years. As a matter of fact, the probability increases more than sevenfold when we move from industries that are not close to any portfolio members to industries that are close to thirteen portfolio members. The picture for

[20] We also calculated closeness as the sum of relatedness to all industries in the regional portfolio. $C'_{ir} = \sum_{j \neq i} I\left(j \in PF_r\right) RR(i,j)$. The results are very similar to the ones we discuss in the main text.

contraction is reversed. Industries in the region that are unrelated to the other members of the regional portfolio run about three times the risk of falling out of the portfolio compared to industries that are related to many other local industries. Calculations rank correlations between closeness and expansion, and between closeness and contraction confirm this picture. Averaged across all years, the correlation between an expansion dummy and closeness is significant and positive correlation at 0.13. For contraction, similar calculations yield a negative and significant correlation equal to -0.11.

The overall conclusion is that the RR matrix has substantial predictive power when it comes to the composition and changes in regional portfolios. We have shown that, in terms of our Revealed Relatedness concept, regions not only show a strong coherence in their manufacturing activities. This coherence is also sustained over time by the exit of unrelated and entry of related industries. As the RR matrix is based on plant level data without taking into consideration any regional aspects, the predictive quality is quite remarkable.

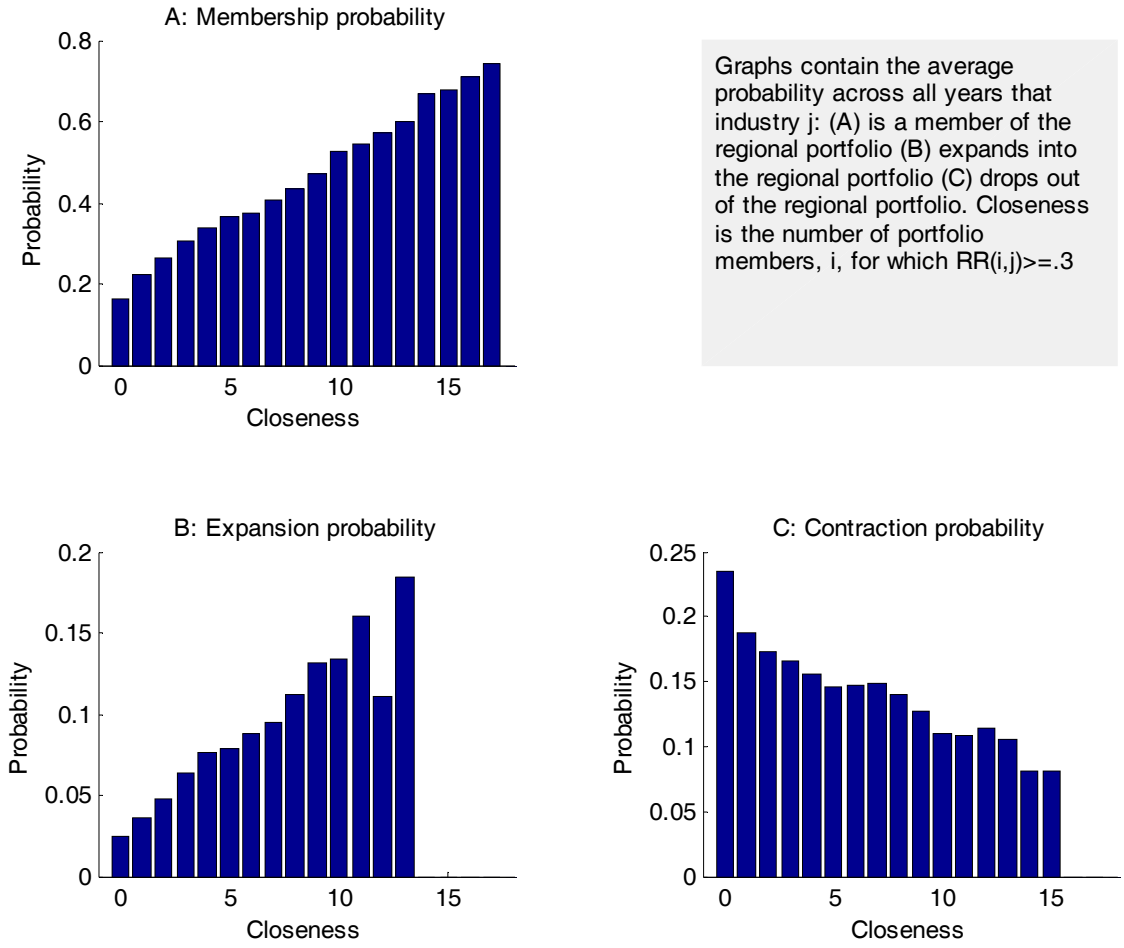


Figure 5: The influence of closeness to a regions portfolio members on membership, expansion and contraction.

5. Conclusion

Essentially, the Revealed Relatedness index presented in this paper compares raw co-occurrence counts to the number of co-occurrences that can be predicted based on knowledge about class specific attributes only. This comparison yields a meaningful relatedness index whenever co-occurrences are more likely to arise between related classes. The index has the advantage that it can be interpreted as a probability. Moreover, we have shown that relatedness as quantified this way is not necessarily a symmetric notion: its direction can express a complexity gradient of the involved classes. This feature is useful, for example, if spillovers between classes are more likely to arise from more complex to less complex classes than vice versa.

In an application to portfolio data, we have derived industry space with many reassuring properties. The industry space is stable enough in the short run to lend it credibility, but, at the same time, it is sufficiently dynamic to unveil shifts in relatedness in the long run. It is compatible with the relatedness implicit in the industrial classification system, but offers many additional, and some new insights.

The use of industry space in our economic geography application shows the versatility of our relatedness measure. Although it is primarily based on the assumption that plant portfolios contain products for which production processes share many similarities, we have shown that the industry space has substantial predictive power in a study of structural transformations of regions. The theoretical assumptions we test are admittedly simple: regions have coherent portfolios of industries. Nevertheless, given that no regional information whatsoever is used in the construction of the industry space, the explanatory force of the index is remarkable.

The graphical representations of regions in industry space is a good example of a simple, yet powerful, use of the Revealed Relatedness matrix. As such, it may support regional policy making or simply help to transparently communicate the threats and opportunities that lie ahead. However, the matrices can also be used in more complex studies. In economic geography, spillovers and agglomeration externalities can be studied in more detail if knowledge is available about the closeness between industries. In strategy research, comparable indices are already in use to quantify corporate coherence. In the field of industrial dynamics, a time series of RR matrices can become the subject of investigations, shedding light on how relatedness shifts over time. Apart from applications of the industry space, the methodology can be used in conjunction with other data bases. The first thing that comes to mind is a repetition for other countries. This would allow for comparisons of industry spaces across countries. It is by no means a given that the Swedish industry space is representative for the entire world. If no comparable data is available, the issue could be explored by investigating the predictive power of the Swedish industry space for, e.g., regional transformations in other countries. Furthermore, co-occurrences can also be based on firm, regional or national portfolios. There is also ample opportunity to use other inter-industry flow data as the basis of co-occurrences. On a more methodological level, experiments of combining such RR matrices using the Bayesian framework is an interesting challenge.

The study of relatedness is, in sum, potentially a very fruitful field of research, where many questions are still unresolved. Our hope is that the method in this paper can be of service to any study that would benefit from a reliable quantification of class-similarity.

References

Amin A (2003). Industrial Districts. In E Sheppard and T J Barnes (ed.). A Companion to Economic Geography. Oxford: Blackwell. Pp 149-168.

Asheim B T (1996). Industrial districts as 'learning regions': A condition for prosperity. European Planning Studies, Volume 4, Issue 4 August 1996 , pages 379 – 400.

Asheim B T (2000). Industrial Districts: Marshall and Beyond. In: G L Clark, M P Feldman and M S Gertler (ed.). The Oxford Handbook of Economic Geography. Oxford: Oxford University Press. Pp 413-431.

Asheim B T, Coenen L and Svensson Henning M (eds.) (2003). *Nordic SMEs and Regional Innovation Systems.* Oslo: Nordic Industrial Fund.

Berry, C H (1975). Corporate Growth and Diversification. Princeton: Princeton University Press.

Borgatti S P 2002. NetDraw: Graph Visualization Software. Harvard: Analytic Technologies.

Borgatti S P, M G Everett and L C Freeman. 2002a. Ucinet for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies.

Breschi S, F Lissoni & F Malerba (2003). Knowledge-relatedness in firm technological diversification. Research Policy 32 (2003), pp. 69-87.

Breshnahan T F and M Trajtenberg (1995). General Purpose Technologies: Engines of Growth? Journal of Econometrics 65 1, pp 83-109

Bryce D J and S G Winter (2006). A general inter-industry relatedness index. Center of Economic Studies discussion papers, CES 06-31.

Dosi G (1988). Sources, Procedures, and Microeconomic Effects of Innovation. Journal of Economic Literature 26 (3), pp. 1120-1171.

Engelsman E C and A F J van Raan (1991). Mapping of Technology, A First Exploration of Knowledge Diffusion Amongst Fields of Technology, Policy Studies on Technology and Economy (BTE) Series, No. 15. The Hague.

Fan J P H and L H P Lang (2000). The Measurement of Relatedness: An application to Corporate Diversification. Journal of Business, 73 (4) pp. 629-660.

Florida R (1995). Toward the Learning Region. Futures, 27, 5, pp 527-536.

Freeman C, and C Perez (1988). Structural crises of adjustment, business cycles and investment behaviour, in: G Dosi, C Freeman, R Nelson, G Silverberg, and L Soete. Technical change and economic theory. London: Pinter.

Gort M (1962). Diversification and integration in American industry. Princeton: Princeton University Press.

Grant R M (1988). On 'Dominant Logic', Relatedness and the Link between Diversity and Performance. Strategic Management Journal, 9 (6), pp. 639-642

Hidalgo C A, B Klinger, A-L Barabási and R Hausmann (2007). The Product Space Conditions the Development of Nations. Science, 317, pp. 482-487.

Marshall A (1920/1890). Principles of Economics. London: Macmillan.

Myrdal, G (1957). Economic theory and under-developed regions. London : Duckworth.
Ohlsson L and L Vinell. 1987. Tillväxtens drivkrafter. Stockholm: Industriförbundets förlag.

Ohlsson L and L Vinell (1987). Tillväxtens drivkrafter. Stockholm: Industriförbundets förlag.

Papke L E and J M Wooldridge (1996). Econometric Methods for Fractional Response Variables With an Application to 401 (K) Plan Participation Rates. Journal of Applied Econometrics 11 (6), pp. 619-632.

Penrose E (1959). The Theory of the Growth of the Firm. New York: John Wiley.

Porter M E (1990). The Competitive Advantage of Nations. London: Macmillan.

Porter M E (2000). Locations, Clusters, and Company Strategy. In: G L Clark, M P Feldman and M S Gertler (ed.). The Oxford Handbook of Economic Geography. Oxford: Oxford University Press. Pp 253-274.

Prahalad C K and R A Bettis (1986). The Dominant Logic: A New Linkage between Diversity and Performance. Strategic Management Journal 7 (6), pp. 485-501.

Rumelt R P (1982). Diversification Strategy and Profitability. Strategic Management Journal 3 (4), pp. 359-369.

Scherer F M (1982). Inter-industry technology flows in the United States. Research Policy 11, pp. 227-245.

Schön L (2000). Electricity, technological change and productivity in Swedish industry, 1890-1990. European Review of Economic History 4, pp 175-194.

Teece D J (1982). Toward an economic theory of the multiproduct firm. Journal of Economic Behavior and Organization 3, pp. 39-63.

Teece T D, R Rumelt, G Dosi and S Winter (1994). Understanding corporate coherence. Theory and evidence. Journal of Economic Behaviour and Organization 23, pp. 1-30.

Appendix A: Derivation of posterior distribution of RR

Substituting (12) and (16) into (14) yields:

$$P\left(RR=r\middle|L=l,C=c\right)=\frac{\left(r\right)^{l}\left(1-r\right)^{c-l}\binom{c}{l}\cdot r^{a-1}\left(1-r\right)^{b-1}\frac{\Gamma\left(a+b\right)}{\Gamma\left(a\right)\Gamma\left(b\right)}}{P\left(L=l\middle|C=c\right)} \tag{A.1}$$

Using the law of total probability, we can get an expression for the denominator of this expression:

$$P\left(RR=r\middle|L=l,C=c\right)=\frac{\left(r\right)^{l}\left(1-r\right)^{c-l}\binom{c}{l}\cdot r^{a-1}\left(1-r\right)^{b-1}\frac{\Gamma\left(a+b\right)}{\Gamma\left(a\right)\Gamma\left(b\right)}}{\int_{0}^{1}P\left(L=l\middle|C=c,RR=\tilde{r}\right)P\left(RR=\tilde{r}\middle|C=c\right)d\tilde{r}} \tag{A.2}$$

As RR and C are independent by assumption, we can fill in terms and simplify to get:

$$P\left(RR=r\middle|L=l,C=c\right)=\frac{\left(r\right)^{l+a-1}\left(1-r\right)^{c-l+b-1}}{\int_{0}^{1}\left(\tilde{r}\right)^{l+a-1}\left(1-\tilde{r}\right)^{c-l+b-1}d\tilde{r}} \tag{A.3}$$

As the $BETA\left(l+a,c-l+b\right)$ distribution must integrate to 1 over its domain, we get:

$$\int_{0}^{1}\left(r\right)^{l+a-1}\left(1-r\right)^{c-l+b-1}\frac{\Gamma\left(a+c+b\right)}{\Gamma\left(l+a\right)\Gamma\left(c-l+b\right)}dr=1 \tag{A.4}$$

Rearranging terms:

$$\int_{0}^{1}\left(r\right)^{l+a-1}\left(1-r\right)^{c-l+b-1}dr=\frac{\Gamma\left(l+a\right)\Gamma\left(c-l+b\right)}{\Gamma\left(a+c+b\right)} \tag{A.5}$$

Using this result in (A.3) gives:

$$P\left(RR=r\middle|L=l,C=c\right)=\left(r\right)^{l+a-1}\left(1-r\right)^{c-l+b-1}\frac{\Gamma\left(a+c+b\right)}{\Gamma\left(l+a\right)\Gamma\left(c-l+b\right)} \tag{A.6}$$

As $l \geq 0$ and $c \geq l$ means that $l + a > 0$ and $c - l + b > 0$ the degree of relatedness, given the number of co-occurrences (the posterior distribution of RR) is $BETA(l + a, c - l + b)$ distributed.

Appendix B: Removing conditioning on C

From (17) we can calculate the conditional expectation of RR, given that we observe l co-occurrences, and c plants investigating the co-occurrence. However, we do not observe c. We therefore remove the conditioning on C using the law of total probability:

$$\text{(B.1)} \quad P\left(RR = r \middle| L = l\right) = \sum_{c=l}^{N} P\left(RR = r \middle| L = l, C = c\right) \cdot P\left(C = c \middle| L = l\right)$$

We now have to find an expression for $P\left(C = c \middle| L = l\right)$. Bayes' law gives the following expression for the conditional probability of $C = c$, given that there are l co-occurrences.

$$\text{(B.2)} \quad P\left(C = c \middle| L = l\right) = \frac{P\left(L = l \middle| C = c\right) P\left(C = c\right)}{P\left(L = l\right)}$$

Applying the law of total probability twice in the denominator and once in the numerator, and using that RR and C are independently distributed from each other gives:

$$\text{(B.3)} \quad P\left(C = c \middle| L = l\right) = \frac{\int_0^1 P\left(L = l \middle| RR = r, C = c\right) P\left(RR = r\right) dr \, P\left(C = c\right)}{\sum_{\tilde{c}=l}^{N} \left\{ \int_0^1 P\left(L = l \middle| RR = \tilde{r}, C = \tilde{c}\right) P\left(RR = \tilde{r}\right) d\tilde{r} \right\} P\left(C = \tilde{c}\right)}$$

Let us focus on the integral:

$$\text{(B.4)} \quad \int_0^1 P\left(L = l \middle| RR = r, C = c\right) P\left(RR = r\right)$$

$$= \int_0^1 \left(r\right)^l \left(1 - r\right)^{c-l} \binom{c}{l} r^{a-1} \left(1 - r\right)^{b-1} \frac{\Gamma\left(a + b\right)}{\Gamma\left(a\right)\Gamma\left(b\right)} dr \quad \text{(see (12) and (16))}$$

$$= \binom{c}{l} \frac{\Gamma\left(a + b\right)}{\Gamma\left(a\right)\Gamma\left(b\right)} \int_0^1 \left(r\right)^{l+a-1} \left(1 - r\right)^{c-l+b-1} dr$$

$$= \binom{c}{l} \frac{\Gamma\left(a + b\right)}{\Gamma\left(a\right)\Gamma\left(b\right)} \frac{\Gamma\left(l + a\right)\Gamma\left(c - l + b\right)}{\Gamma\left(a + c + b\right)} \quad \text{(see (A.5))}$$

Let: $\beta\left(a, b, c, l\right) = \frac{\Gamma\left(l + a\right)\Gamma\left(c - l + b\right)}{\Gamma\left(a + b + c\right)}$, with $\Gamma\left(.\right)$ the Gamma function. Using the result above in (B.3) leads to:

(B.5) $$P(C=c|L=l)=\frac{\binom{c}{l}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\beta(a,b,c,l)P(C=c)}{\sum_{\tilde{c}=l}^{N}\binom{\tilde{c}}{l}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\beta(a,b,\tilde{c},l)P(C=\tilde{c})}$$

Filling in the binomial likelihood for the probability of $C=c$ and simplifying:

(B.6) $$P(C=c|L=l)=\frac{\binom{c}{l}\beta(a,b,c,l)q^{c}(1-q)^{N-c}\binom{N}{c}}{\sum_{\tilde{c}=l}^{N}\binom{\tilde{c}}{l}\beta(a,b,\tilde{c},l)q^{\tilde{c}}(1-q)^{N-\tilde{c}}\binom{N}{\tilde{c}}}$$

If we fill in (B.6) and (A.6) into (B.1) and simplify, we get:

(B.7) $$P(RR=r|L=l)=\sum_{c=l}^{N}(r)^{l+a-1}(1-r)^{c-l+b-1}\cdot\frac{\binom{c}{l}q^{c}(1-q)^{N-c}\binom{N}{c}}{\sum_{\tilde{c}=l}^{N}\binom{\tilde{c}}{l}\beta(a,b,\tilde{c},l)q^{\tilde{c}}(1-q)^{N-\tilde{c}}\binom{N}{\tilde{c}}}$$

We can now calculate the posterior expectation of RR:

(B.8) $$E(RR|L=l)=\int_{0}^{1}r\sum_{c=l}^{N}(r)^{l+a-1}(1-r)^{c-l+b-1}\frac{\binom{c}{l}q^{c}(1-q)^{N-c}\binom{N}{c}}{\sum_{\tilde{c}=l}^{N}\binom{\tilde{c}}{l}\beta(a,b,\tilde{c},l)q^{\tilde{c}}(1-q)^{N-\tilde{c}}\binom{N}{\tilde{c}}}dr$$

$$=\sum_{c=l}^{N}\left[\frac{\binom{c}{l}q^{c}(1-q)^{N-c}\binom{N}{c}}{\sum_{\tilde{c}=l}^{N}\binom{\tilde{c}}{l}\beta(a,b,\tilde{c},l)q^{\tilde{c}}(1-q)^{N-\tilde{c}}\binom{N}{\tilde{c}}}\int_{0}^{1}(r)^{l+a}(1-r)^{c-l+b-1}dr\right]$$

Using that the $BETA(l+a+1,c-l+b)$ distribution must integrate to 1 over its domain we find that:

(B.9) $$\int_{0}^{1}(r)^{l+a}(1-r)^{c-l+b-1}dr=\frac{\Gamma(l+a+1)\Gamma(c-l+b)}{\Gamma(a+1+c+b)}$$

Using (B.9) in (B.8), we get:

$$\text{(B.10)}\quad E\left(RR\middle|L=l\right)=\sum_{c=l}^{N}\left[\frac{\binom{c}{l}q^{c}\left(1-q\right)^{N-c}\binom{N}{c}}{\sum_{\tilde{c}=l}^{N}\binom{\tilde{c}}{l}\beta\left(a,b,\tilde{c},l\right)q^{\tilde{c}}\left(1-q\right)^{N-\tilde{c}}\binom{N}{\tilde{c}}}\frac{\Gamma\left(l+a+1\right)\Gamma\left(c-l+b\right)}{\Gamma\left(a+1+c+b\right)}\right]$$

From the Gamma function, we know that $\Gamma\left(x+1\right)=x\Gamma\left(x\right)$. Therefore, we can express (B.10) as:

$$\text{(B.11)}\quad E\left(RR\middle|L=l\right)=\sum_{c=l}^{N}\left[\frac{\left(l+a\right)}{\left(a+c+b\right)}\frac{\beta\left(a,b,c,l\right)q^{c}\left(1-q\right)^{N-c}\binom{c}{l}\binom{N}{c}}{\sum_{\tilde{c}=l}^{N}\beta\left(a,b,\tilde{c},l\right)q^{\tilde{c}}\left(1-q\right)^{N-\tilde{c}}\binom{\tilde{c}}{l}\binom{N}{\tilde{c}}}\right]$$

This result is reported as equation (20) in the main text.

For completeness we also derive the unconditional variance of RR. The only piece of information we still miss for this, is the unconditional expectation of RR^{2}.

$$\text{(B.12)}\quad E\left(RR^{2}\middle|L=l\right)=\int_{0}^{1}r^{2}\sum_{c=l}^{N}\left(r\right)^{l+a-1}\left(1-r\right)^{c-l+b-1}\frac{\binom{c}{l}q^{c}\left(1-q\right)^{N-c}\binom{N}{c}}{\sum_{\tilde{c}=l}^{N}\binom{\tilde{c}}{l}\beta\left(a,b,\tilde{c},l\right)q^{\tilde{c}}\left(1-q\right)^{N-\tilde{c}}\binom{N}{\tilde{c}}}dr$$

$$=\sum_{c=l}^{N}\left[\frac{\binom{c}{l}q^{c}\left(1-q\right)^{N-c}\binom{N}{c}}{\sum_{\tilde{c}=l}^{N}\binom{\tilde{c}}{l}\beta\left(a,b,\tilde{c},l\right)q^{\tilde{c}}\left(1-q\right)^{N-\tilde{c}}\binom{N}{\tilde{c}}}\int_{0}^{1}\left(r\right)^{l+a+1}\left(1-r\right)^{c-l+b-1}dr\right]$$

Again we can use the fact that a specific $BETA$ distribution, namely $BETA\left(l+a+2,c-l+b\right)$ must integrate to 1 over its domain:

$$\text{(B.13)}\quad \int_{0}^{1}\left(r\right)^{l+a+1}\left(1-r\right)^{c-l+b-1}dr=\frac{\Gamma\left(l+a+2\right)\Gamma\left(c-l+b\right)}{\Gamma\left(a+2+c+b\right)}$$

Using this in (B.12) gives:

(B.14) $$E\left(RR^2\middle|L=l\right)=\sum_{c=l}^{N}\left[\frac{\binom{c}{l}q^c(1-q)^{N-c}\binom{N}{c}}{\sum_{\tilde{c}=l}^{N}\binom{\tilde{c}}{l}\beta(a,b,\tilde{c},l)q^{\tilde{c}}(1-q)^{N-\tilde{c}}\binom{N}{\tilde{c}}}\frac{\Gamma(l+a+2)\Gamma(c-l+b)}{\Gamma(a+2+c+b)}\right]$$

If we apply $\Gamma(x+1)=x\Gamma(x)$ twice to this expression, we arrive at the following:

(B.15) $$E\left(RR^2\middle|L=l\right)=\sum_{c=l}^{N}\left[\frac{\beta(a,b,c,l)\binom{c}{l}q^c(1-q)^{N-c}\binom{N}{c}}{\sum_{\tilde{c}=l}^{N}\binom{\tilde{c}}{l}\beta(a,b,\tilde{c},l)q^{\tilde{c}}(1-q)^{N-\tilde{c}}\binom{N}{\tilde{c}}}\frac{(l+a+1)(l+a)}{(a+1+c+b)(a+c+b)}\right]$$

As $VAR\left(RR\middle|L=l\right)=E\left(RR^2\middle|L=l\right)-\left[E\left(RR\middle|L=l\right)\right]^2$, the expression for the unconditional variance of RR is:

(B.16) $VAR\left(RR\middle|L=l\right)=$

$$\sum_{c=l}^{N}\left[\frac{\beta(a,b,c,l)\binom{c}{l}q^c(1-q)^{N-c}\binom{N}{c}}{\sum_{\tilde{c}=l}^{N}\binom{\tilde{c}}{l}\beta(a,b,\tilde{c},l)q^{\tilde{c}}(1-q)^{N-\tilde{c}}\binom{N}{\tilde{c}}}\frac{(l+a+1)(l+a)}{(a+1+c+b)(a+c+b)}\right]$$

$$-\left\{\sum_{c'=l}^{N}\left[\frac{(l+a)}{(a+c'+b)}\frac{\beta(a,b,c',l)q^{c'}(1-q)^{N-c'}\binom{c'}{l}\binom{N}{c'}}{\sum_{\tilde{c}'=l}^{N}\beta(a,b,\tilde{c}',l)q^{\tilde{c}'}(1-q)^{N-\tilde{c}'}\binom{\tilde{c}'}{l}\binom{N}{\tilde{c}'}}\right]\right\}^2$$

Appendix C

Tables C.1 and C.2 describe the short-term stability of the different relatedness estimates during the first year of each decades. In most cases, the correlations (stability) for the *First step* estimates are higher than for the other estimates. The stability also shows a tendency to decrease with time. The row *Intermediate only* shows the correlation for only those second step estimates for which the first step estimates were two weak and the intermediate estimates were used. It is therefore quite remarkable that even for the elements where information of raw counts was very weak, stability is still so high. As expected, the stability of the *Directed links* is also lower than of the *Symmetric links*.

Table C.1: Average Spearman rank correlation between relatedness vectors during given years for the *Symmetric links*. Standard deviation across industries in brackets.

Symmetric links

	1970-1971	**1970-1972**	**1980-1981**	**1980-1982**	**1990-1991**	**1990-1992**	**2000-2001**	**2001-2002**
First step	.90 (.01)	.86 (.01)	.93 (.01)	.89 (.01)	.89 (.02)	.79 (.03)	.85 (.03)	.76 (.03)
Intermediate only	.82 (.03)	.78 (.03)	.86 (.04)	.79 (.06)	.82 (.06)	.61 (.2)	.65 (.11)	.63 (.08)
Second step	.83 (.01)	.78 (.02)	.85 (.01)	.81 (.01)	.72 (.03)	.59 (.04)	.63 (.04)	.52 (.04)

Table C.2: Average Spearman rank correlation between relatedness vectors during given years for the *Directed links*. Standard deviation across industries in brackets.

Directed links

	1970-1971	**1970-1972**	**1980-1981**	**1980-1982**	**1990-1991**	**1990-1992**	**2000-2001**	**2001-2002**
First step	.86 (.02)	.79 (.02)	.87 (.01)	.79 (.02)	.85 (.02)	.76 (.03)	.81 (.04)	.72 (.04)
Intermediate only	.76 (.04)	.73 (.03)	.83 (.03)	.64 (.18)	.67 (.27)	.88 (.01)	.70 (.00)	1.00 (.00)
Second step	.80 (.03)	.73 (.03)	.85 (.02)	.77 (.03)	.82 (.02)	.75 (.03)	.78 (.06)	.68 (.05)

Tables C.3-C.6 display the long-term stability of the different relatedness estimates between decades. The stability of the relatedness structures generally decreases the longer the time period, as could be expected. As in table C.1 and C.2, the *First step* estimates are generally higher than the *Second step* estimates.

Table C.3: Average Spearman rank correlation between the Relatedness vectors of decades, *Second step*, *Symmetric links*. Based on 3-year moving averages. Standard deviation across industries in brackets.

Second step, Symmetric links

	1980	**1990**	**2000**
1970	.55 (.04)	.32 (.03)	.26 (.02)
1980		.38 (.02)	.26 (.03)
1990			.4 (.03)

Table C.4: Average Spearman rank correlation between the Relatedness vectors of decades, *Second step, Directed links*. Based on 3-year moving averages. Standard deviation across industries in brackets.

Second step, Directed links

	1980	**1990**	**2000**
1970	.53 (.04)	.37 (.04)	.33 (.04)
1980		.42 (.04)	.35 (.04)
1990			.45 (.05)

Table C.5: Average Spearman rank correlation between the Relatedness vectors of decades, *First step estimates, Symmetric links*. Based on 3-year moving averages. Standard deviation across industries in brackets.

First step, Symmetric links

	1980	**1990**	**2000**
1970	.68 (.03)	.45 (.03)	.44 (.03)
1980		.46 (.03)	.43 (.03)
1990			.53 (.04)

Table C.6: Average Spearman rank correlation between the Relatedness vectors of decades, *First step estimates, Directed links*. Based on 3-year moving averages. Standard deviation across industries in brackets.

First step, Directed links

	1980	**1990**	**2000**
1970	.6 (.04)	.43 (.04)	.38 (.04)
1980		.46 (.04)	.39 (.04)
1990			.5 (.05)

A comparison between our relatedness estimates and the relatedness according to the hierarchical SNI69-system is displayed in table 7. The correlation is surprisingly low, especially compared to the network pictures displayed in the paper.

Table C.7: Average Spearman rank correlation between the Relatedness vectors and relatedness according to the SNI69-system over all industries. Based on 3-year moving averages. Standard deviation across industries in brackets.

Symmetric links

	1970	**1980**	**1990**	**2000**
First step	.43 (.04)	.39 (.05)	.35 (.04)	.35 (.04)
Second step	.31 (.05)	.27 (.05)	.26 (.03)	.25 (.04)

Table 8 displays the correlation between in- and outgoing links.

Table C.8: average Spearman rank correlations between in-and outgoing links. Based on 3-year moving averages. Standard deviation across industries in brackets.

	1970	**1980**	**1990**	**2000**
First step	.48 (.08)	.44 (.05)	.37 (.07)	.39 (.08)
Second step	.37 (.04)	.36 (.04)	.32 (.06)	.35 (.07)