Laboratory of Economics and Management

Sant'Anna School of Advanced Studies
Piazza Martiri della Libertà, 33 - 56127 PISA (Italy)
Tel. +39-050-883-343 Fax +39-050-883-344
Email: lem@sssup.it Web Page: http://www.lem.sssup.it/

LEM

Working Paper Series

Measuring Industry Relatedness and Corporate Coherence

Giulio Bottazzi°
Davide Pirino°

° CAFED and LEM, Scuola Superiore Sant'Anna, Italy

2010/10 **July 2010**

ISSN (online) 2284-0400

Measuring Industry Relatedness and Corporate Coherence *

Giulio Bottazzi[†°] and Davide Pirino[°]

[°]CAFED and LEM, Scuola Superiore Sant'Anna, Pisa, Italy

June 16 2010

Abstract

Since the seminal work of Teece et al. (1994) firm diversification has been found to be a non-random process. The hidden deterministic nature of the diversification patterns is usually detected comparing expected (under a null hypothesys) and actual values of some statistics. Nevertheless the standard approach presents two big drawbacks, leaving unanswered several issues. First, using the observed value of a statistics provides noisy and nonhomogeneous estimates and second, the expected values are computed in a specific and privileged null hypothesis that implies spurious random effects. We show that using Monte Carlo p-scores as measure of relatedness provides cleaner and homogeneous estimates. Using the NBER database on corporate patents we investigate the effect of assuming different null hypotheses, from the less unconstrained to the fully constrained, revealing that new features in firm diversification patterns can be catched if random artifacts are ruled out.

JEL codes: C1, D2, L2

Keywords: corporate coherence; relatedness; null model analysis; patent data

1 Introduction

The relevance of corporate diversification structure in determining firm's performance has, since long, received ample recognition inside the industrial economics literature (in a vast body of contributions, see: Rumelt, 1974; Berry, 1975; Teece, 1980; Rumelt, 1982; Teece, 1982). Quite soon, scholarly contributions suggested that it is not only *how much* firms diversify to be important in determining their performances, but also *how* they do it. Firms able to diversify their operations across related fields can enjoy the advantage of economies of scope, likely generated by an increased utilization of incumbent investments or by technological spillovers, which are clearly not attainable through a random diversification of activities. Obviously, the econometric assessment and the empirical foundation of this statement require

*We thank Giovanni Dosi, Alessandro Nuvolari and Federico Tamagni for inspiring discussions and useful suggestions. Any mistake, substantial or formal, is our own responsibility.

[†]*Corresponding Author*: Giulio Bottazzi, Scuola Superiore Sant'Anna, *E-mail*: bottazzi@sssup.it.

the identification of a measure of relatedness across fields. Devising it is not an easy task. First of all, the identification of an empirical notion of economic proximity goes beyond the simple identification of a broad taxonomy of productive activities (like for instance the one in Pavitt, 1984). By definition, taxonomies classify, not relate, the possible fields of operation. Then, they do not provide any direct notion of similarity across fields. Moreover, even if such a notion can be indirectly derived, the reliance on ex-ante (and often introspective) assessment of the particular characteristics of the different *taxa* is likely to ignore more structured, and subtle, sources of complementarity among fields, often hidden to the "bird-eye" approach of the researcher.

A slightly different case is constituted by hierarchical measures of relatedness based on industrial sectors or technological classes. These classifications do in general provide a rather fine distinction in a large number of fields. They are defined by national or international bodies and avoid the risk of idiosyncrasies implicit in individual assessments. They might be useful, and have been used, to obtain a measure of the (operating or technological) "scope" of the firm, just by counting in how many sectors the firm is active (Montgomery, 1982) or how many classes are spanned by the firm's patents portfolio. However, their usefulness in investigating the relationship existing between different sectors or classes is doubtful. Indeed, these classifications focus exclusively on one, or few, aspects characterizing the different fields. Industrial classifications are usually based on the nature of the input goods (oil, steel, etc.), or on the nature of the final markets (precision instruments, furniture,...). Technological classifications, like the ones used by patent offices, consider the technological fields in which the invention can potentially be applied, without any consideration or reliance on economic aspects. In both cases, these taxonomies do not identify the bundle of resources or competences specific to a given field and, as such, are not able to capture the economic advantages (or disadvantages) associated by the combined presence, inside a productive unit, of different activities. So the fact that two industrial sectors share the first three digits of the SIC classification does not tell much about the economic advantage faced by a firm active in both sectors, nor the fact that two patents share the first three digit of the IPC classification reveals much about the increased value of owning them both.

Starting from similar considerations, in their seminal work, Teece et al. (1994) introduce an endogenous notion of relatedness, based on the "survivor principle" and derived from the observed diversification patterns. The intuition is that firms diversified in more related fields, due to positive economies of scope, enjoy on average higher competitive advantages, and thus an higher probability to survive the competitive struggle. As a consequence, activities in related fields should appear with an higher frequency inside surviving firms. This suggests to directly measure the relatedness of fields using the diversification patterns of firms themselves, adopting the number of firms simultaneously active in a pair of fields (co-occurrencies) as a proxy for the relatedness of the two fields. The actual degree of relatedness is finally obtained by comparing the observed number of co-occurrencies with what would have been obtained under the absence of any relatedness among the fields of activity. As suggested in Bryce and Winter (2010), this endogenous notion of proximity can be applied to a wide range of issues in strategic management, corporate finance and industrial economics and possesses several advantages. First, while it does not identify the "basket of resources" associated with each field, it does directly address the question of the existence of some complementarity among different baskets associated with different activities. Second, the idea of proximity that emerges from this measure, being directly based on economic considerations, is not limited exclusively to the existence and strength of technological spillovers. It can, equivalently, capture other business aspects of the joint operation of different fields, like the sharing of managerial competences

or financial advantages (Pehrsson, 2006). Third, being probabilistic in nature, this measure allows for idiosyncratic elements and path-dependent constraints which can hinder the optimal exploitation of resources in one particular firm. The averaging procedure across multiple firms, implicit in the measure, should wash away these idiosyncratic hindrances and preserve the goodness of the result. To illustrate the merit of this kind of measure, Bryce and Winter (2010), using plant level data on the U.S. manufacturing sector, show that the simple counting of co-occurrences (with some due corrections, see the next section) is able to identify "hidden" relationships among SIC sectors which are several digits apart.

Once an underlying notion of relatedness is established, one can use some appropriate averaging procedure across activities to obtain an empirical notion of corporate "coherence", measured as the degree of relatedness among the constituent businesses of a firm. This approach was proposed and effectively applied in Teece et al. (1994) to analyze the relation between corporate coherence and firm's scope. The same idea is adopted by Breschi et al. (2003) to obtain a measure of coherence of patent portfolios. The computed measure is later exploited in regression analysis investigating the determinants of corporate performance. A similar approach is followed by Piscitiello (2004), who uses technological fields and output markets to measure corporate coherence. This line of research is further investigated by Nesta and Saviotti (2006) who find that the coherence of the knowledge base within firms is a significant explanatory variable of firms stock market value. In Valvano and Vannoni (2003) the relatedness measure is applied to a modified coherence index which takes into consideration the notion of principal activity.

The present paper is mainly intended as a methodological contribution in the research line described above. We show that the approach proposed by Teece et al. (1994), and adopted as a standard methodology by a large portion of the literature, is affected by two potential drawbacks.

The first drawback is associated to the measure of relatedness itself. Teece et al. (1994) adopt a measure, or *statistics*, that catches how much the observed relatedness moves away from its expected value. The expected value is computed using a well-defined mechanism of random assignment between firms and industrial sectors. The discrepancy of the observed statistics to its expected value is measured in unit of expected standard deviation. We will show that this choice is biased in nature, providing nonhomogeneous and noisy estimates. We propose to solve this problem adopting a new quantile-based estimator, which essentially is the p-score of the observed relatedness statistics. Moreover, the use of a measure of pair relatedness based on p-score allows for a simple and straightforward introduction of the notion of "anti-relatedness". This can be used to identify new features in firm diversification structure and also answer the question raised in Bryce and Winter (2010) on the need to associate some economic content not only to the presence, but also to the *absence*, of co-occurrencies.

The second drawback is linked with the random association mechanism adopted in Teece et al. (1994) to compute the expected co-occurrence and its variance. Such kind of random association mechanism will be referred as *null model* or *null hypothesys*. The mechanism they adopt assume that the number of firms active in each industrial sector is fixed, and equal to the value observed in actual data. Then they randomly assign firms to sectors, and compute the probability that two sectors appear in the same firm. Obviously, in this way no constraints are imposed on firms' scope, i.e. on the number of industrial sectors in which each firm is active. In principle, the mechanism allow any firms to be active in *all* sectors. As a direct consequence of this assignment process, the implied distribution of firm scope converges to a binomial. This contrasts with the Paretian shape of the scope distribution often observed in industrial data. In these cases, very high levels of relatedness can be obtained because

of spurious artifacts generated by the discrepancy between the implied distribution and the observed one.

In order to investigate the possible emergence of spurious relatedness, and its effect on the measure of corporate coherence, using data from the NBER Patent Data research project, we analyze different association mechanisms between firms and industrial sectors/patent classes. Specifically, four null models are identified and labeled in increasing order of the total number of constraints they account for. The fully constrained null model, which takes into consideration both sector occupancies and firm scopes, turns out to be remarkably more effective in revealing the existence of patterns in the coherence structure of firms.

This paper is organized as follows: in Section 2 we will briefly review the original measure proposed by Teece et al. (1994), highlighting the mentioned drawbacks. Section 3 discusses several possible measure of relatedness and the advantage of using p-scores. Section 4 describes our data and investigates the effect of different null hypothesis on the measure of relatedness. Section 5 extends this analysis to corporate coherence and Section 6 concludes.

2 Null Models and Previous Approaches

The notion of corporate coherence rests upon some underling measure of relatedness among the different fields of corporate activities. In order to asses how much related are the activities carried over by a firm, one needs a topology over the different fields of operation, which quantifies their relative degree of proximity.[1] The measure of relatedness can be made endogenous by observing how active units actually distribute their activities between the various fields. The precise definition of active unit can vary. It can be a plant or a firm. Analogously, different classifications of the fields of operation has been explored, like industrial sectors (Piscitiello, 2004; Valvano and Vannoni, 2003), specific groups of similar products (Teece et al., 1994) or patent classes (Breschi et al., 2003; Nesta and Saviotti, 2006). For practical purposes, however, the scenario is similar in all these cases. One has N units and I fields. The distribution of the activities is described by the adjacency matrix $C_{n,i} \in \mathbb{N}_{N\times I}$ defined as:

$$C_{n,i} = \begin{cases} 1 & \text{if unit } n \text{ is active in field } i, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

This binary matrix is also known as *presence-absence matrix*. Such kind of matrices have been used in psychometry (Snijders, 1991) and spatial econometrics (Anselin, 1988) but they have been most intensively studied in bio-geography and ecology (Connor and Simberloff, 1979; Roberts and Stone, 1990; Sanderson et al., 1998; Zaman and Simberloff, 2002; Gotelli, 2001, among others). The ecological problem deals with the detection of interactions among species (e.g. of birds) in a given sample. The non-zero (resp. the zero) elements of the adjacency matrices are interpreted as the presence (resp. the absence) of a specie in a geographical area (e.g., an island of an archipelago as in Roberts and Stone, 1990). With an evolutionary

[1] In Bryce and Winter (2010) it is suggested that a topological notion of proximity should be superseded by a stricter metric notion. Substituting proximity with distance could be in fact too much, and can bring unwarranted consequences. Think for instance to two fields A and B, which use different technologies and a third field, which we call AB, in which the two technologies partially overlap. It is plausible to think that firms active in A or B are likely to diversify their operations by developing (or buying) part of the missing technology, so that an high degree of relatedness between A and AB and between B and AB can be observed. Differently from a metric notion, the topological idea of proximity does not imply that the field A and B should be related. Indeed it could be that no firms are active in both the original "pure" technologies since no advantages are associated with their joint operation.

argument analogously to the ecological literature, Teece et al. (1994) suggest that the co-evolution of economic units and the selection process driving the market, lead to the survival of those units characterized by the more efficient mix of activities. As a result, activities that are more related tend to appear together, inside the same unit, with higher frequency. Thus they propose as a measure of association between two industrial sectors i and j the co-occurrences matrix $J_{i,j}$, defined as the number of firms that are simultaneously active in both sectors

$$J_{i,j} = \sum_{i=1}^{N} C_{n,i}\, C_{n,j} = C^T\, C\ . \tag{2}$$

Given an observed adjacency matrix $C \in \mathbb{N}_{N\times I}$, if a particular relationship among activity fields (or species) exists, the observed value of the co-occurrences matrix is expected to be non-random. For this reason, in order to assess the strength of the association, it is necessary to build benchmark values for the co-occurrencies, representing the expected outcomes of a random matching of units and fields, and compare them with the observed ones. The random association mechanism between the N units and the I fields is usually referred as the "null model" (see Gotelli, 2001; Gotelli and Graves, 1996). Teece et al. (1994) propose to consider as the random benchmark the distribution of J generated by randomly assigning the N firms to the I sectors. More precisely, they assume that the number of active units in each field i, $u_i = \sum_{n=1}^{N} C_{n,i}$, is fixed and equal to the actual number observed. Then, they imagine to assign u_i firms, randomly selected from the population of N firms, to each sector i. Under this simple assignment procedure it is straightforward to derive the hypergeometric probability distribution of the co-occurrences

$$\text{Prob}\left\{J_{i,j} = x\right\} = \frac{\binom{u_i}{x}\binom{I-u_i}{u_j-x}}{\binom{K}{u_j}}, \quad x \leq \max\{u_i, u_j\}. \tag{3}$$

The expression for the mean $\mu_{i,j}$ and standard deviation $\sigma_{i,j}$ of the above distribution can be easily derived (Feller, 1976). The first is a measure of the number of co-occurrencies expected between two unrelated fields. The second, instead, measures the deviation from this level due to the random nature of the matching. Finally, Teece et al. (1994) propose to detect couples of related fields identifying those having large levels of the t-statistics

$$\hat{t}_{i,j} = \frac{\hat{J}_{i,j} - \mu_{i,j}}{\sigma_{i,j}}\ , \tag{4}$$

which measures how much standard deviations away the observed values are from their expected value under the null hypothesis. From now on we will denote with $\hat{a}$ the value of a generic quantity a **observed** in the dataset. Since large values of $\hat{t}$ are very unlikely under the null, their observation implies that some "deterministic" mechanisms are forcing the two fields to appear together so often, whence their large relatedness.

A first drawback of the proposed approach is that the matrix $\hat{t}_{i,j}$ is not, in general, a valid device to detect possible deterministic effects. It can indeed attain abnormally high levels. Let us clarify this point with a simple example. Suppose to have $N = 140$ firms. Consider two pairs of industrial sector. The first pair is composed of sectors with equal number of firms, $u_{i_1} = 8$,$u_{j_1} = 8$. The second pair has hugely differing numbers of firms, $u_{i_2} = 100$, $u_{j_2} = 7$. Moreover, suppose that two pairs of sectors have the same level of relatedness according to the t-statistics, that is $t_{i_1,j_1} = t_{i_2,j_2} = 0.5$. Simple computations based on the hypergeometric distribution density (3) show that

$$\text{Prob}\left[t \geq 0.5 | N = 140, u_i = 8, u_j = 8\right] = .0675\ ,$$

while

$$\text{Prob}\left[t \geq 0.5 | N = 140, u_i = 100, u_j = 7\right] = .3546 \ .$$

Even if the value of the t-statistics is the same, it constitutes, under the considered null, a very unlikely outcome for the symmetric case, while it is a near-to-average value for the case with unequal occupancies. The reason is that with very heterogeneous occupancies u_i, the implied distribution of the J's becomes very skew and the t-statistics is no longer a reliable measure of likelihood.

The second drawback of the discussed approach is somehow deeper, and concerns the choice of the null hypothesis. It is clear that assuming a constant number of firms per sector, u_i, and assign this exact number of firms to it, is not the unique random association mechanism between firms and activity fields. Let $v_n = \sum_{i=1}^{I} C_{n,i}$ be the observed number of fields in which firm n is active (the firm scope). Instead of the previous approach, one can imagine to keep these numbers fixed, and assign to each firm n, exactly v_n activity fields randomly selected from the I available. This random assignment procedure of sectors to firms (instead of firms to sectors) will lead to a new probability distribution, in general different from the one obtained under the previous null. A new distribution will in turn implies different levels for the $\hat{t}$ and different assessment of the degree of relatedness.

In general, one can see the null model as a way of randomly distributing the $M = \sum_n v_n = \sum_i u_i$ occupancies, that is the number of 1's in the original matrix, among the $N \times I$ entries of the adjacency matrix C. Given the problem at hand, one can naturally identify four main null hypotheses:

- $\mathcal{H}_1$: **Full Randomness**.

  Random assignment of the M occupancies in the $N \times I$ entries of C. In this case only the total number of occupancies is a fixed quantity. Consequently, the row and column sums, u_i with $i \in \{1, ..., I\}$ and v_n with $n \in \{1, ..., N\}$, are random variables.

- $\mathcal{H}_2$: u_i **fixed,** v_n **random**.

  Random assignment of u_i firms to activity field i, with $i \in \{1, ..., I\}$. The total number of links M and column sums u_i are given quantities, while row sums v_n are random variables.

- $\mathcal{H}_3$: v_n **fixed,** u_i **random**.

  Random assignment of v_n activities to firm n, with $n \in \{1, ..., N\}$. This case is the symmetric case of $\mathcal{H}_2$. The firm scopes v_n are given quantities while the industrial occupancy numbers u_i are random variables.

- $\mathcal{H}_4$: u_i **fixed,** v_n **fixed**.

  Random assignment of the M occupancies in the $N \times I$ entries of C, preserving both firms scope and the number of firms per field. This null corresponds to the most conservative case, where both column and row sums are assinged by the dataset.

The hypotheses are labelled in increasing number of constraints. $\mathcal{H}_1$ has only one constraint. $\mathcal{H}_2$ and $\mathcal{H}_3$ have a number of constraints equal to the number of fields and the number firms, respectively. $\mathcal{H}_4$ has $N + I$ constraints. The approach proposed in Teece et al. (1994) corresponds to hypothesis $\mathcal{H}_2$. It is rather intuitive that the more constraints one considers, the more adherent the null hypothesis is to the actual data. Consequently, $\mathcal{H}_4$ should be the better choice, if there are no specific reasons to presume that the other nulls, with

less constraints, are more adequate. This null, however, entails a slightly increased complication in the computation of the relevant statistics. On the other hand, the distorting effect of assuming a too lax null can be easily made apparent. In Section 5, the results obtained under the different null hypotheses are compared, using a publicly available database on firms patents. Before performing our exercises, we have to spent some words on the actual definition of relatedness we use.

3 Measures of Relatedness

The main problem in the direct use of the number of co-occurrencies $J_{i,j}$ as a measure of relatedness is that its spectrum $[0, \min(u_i, u_j)]$ is pair-dependent. It is in general better to deal with a normalized quantity. For this purpose one can simply consider the normalized co-occurrences matrix $N_{i,j}$ defined as

$$N_{i,j} = \begin{cases} \frac{J_{i,j}}{u_i + u_j - J_{i,j}} & \text{if } J_{i,j} > 0 \\ \\ 0 & \text{if } J_{i,j} = 0. \end{cases} \tag{5}$$

Note that $N_{ij} \in [0, 1]$, moreover $N_{i,j} = 1$ if and only if $u_i = u_j = J_{i,j}$, that is if and only if every firm active in i is also active in j (and vice versa).

As a further candidate for relatedness statistics, we consider the odds-ratios, largely exploited in social science, medical research and ecology (see for example Bishop et al., 1975; Mehta et al., 1985; Rudas, 1985; Zaman and Simberloff, 2002). In fact, this statistic is related to contingency table analysis. Consider two fields i and j and let $n_{i,j}$ be the number of firms active in both i and j (i.e. $n_{i,j} = J_{i,j}$), while let $n_{0,0}$ indicate the number of firms not in i nor in j. Define $n_{0,i} = u_i - J_{i,j}$ as the number of firms in i only and, similarly, $n_{0,j} = u_j - J_{i,j}$ as those in j only. Assume that the "treatment" of the firm is represented by being active in field j and that the success is achieved if the firm is present in field i. Consequently, the fraction of success among the treated is $n_{i,j}/n_{0,j}$, while among the untreated amounts to $n_{i,0}/n_{0,0}$, thus the odds-ratio becomes $R_{i,j} = n_{0,0}\, n_{i,j}/n_{0,j}\, n_{i,0}$. Notice that this expression is symmetric under the interchange i and j. If $u_j = J_{i,j}$, then every treated firm achieves success (it is also present in i) and, simultaneously, every untreated firm does not achieve success (it is absent in i). As a consequence the treatment reaches its maximum efficiency, i.e. $R_{i,j} = 1$. Therefore the definition of the odds-ratio statistics reads:

$$R_{i,j} = \begin{cases} \frac{n_{0,0}\, n_{i,j}}{n_{0,j}\, n_{i,0}} & \text{if } (n_{0,i} > 0 \text{ and } n_{0,j} > 0) \\ \\ 1 & \text{if } (n_{0,i} = 1 \text{ or } n_{0,j} = 1). \end{cases} \tag{6}$$

The simple co-occurrences matrix $J_{i,j}$, the normalized co-occurrencies $N_{i,j}$ and the odd-ratios $R_{i,j}$ are all non-decreasing functions of the strength of association among sectors. For this reason they all provide an acceptable measure of relatedness. Nevertheless, they are essentially different in nature. To see it, suppose to have an adjacency matrix like this:

$$C = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix},$$

i.e. there are four firms and two classes with $u_1 = 3$, $u_2 = 3$. In this case $n_{0,0} = 0$ and the three proposed measures are ordered as follows:

$$R_{1,2} = 0 \quad < \quad N_{1,2} = \frac{1}{2} \quad < \quad J_{1,2} = 2 \; . \tag{7}$$

The fact that there are no "untreated" firms with no success makes the odds-ratio measure vanish, while according to $N_{1,2}$ and $J_{1,2}$ a positive association is present, due to the first two rows of the adjacency matrix.

All the measure of relatedness discussed above suffer from the problem of non-homogeneity and noise presented in the previous section: when the distribution of activity fields across firms, or of firms across fields, is skewed, as often occurs in real data, all these statistics have highly skewed distributions and the consequent statistics-based inference becomes unreliable.[2] This problem is solved when the p-score associated with the chosen statistics[3] is used. Consider a measure A, defined in terms of the adjacency matrix, and a null hypothesis $\mathcal{H}$. For each couples of fields i and j one can consider the probability $p_{i,j}$ that, under the chosen null, the value of the statistics $A_{i,j}$ would be below the observed one $\hat{A}_{i,j}$:

$$p_{i,j}(A, \mathcal{H}) = \text{Prob}\left[A_{i,j} \leq \hat{A}_{i,j} | \mathcal{H}\right] \; . \tag{8}$$

The p-score $p_{i,j}$ depends on both the adopted statistics A and the considered null $\mathcal{H}$. Its value can be obtained from a theoretical distribution, when available, as in the case of $\mathcal{H}_2$ and J statistics, or, more generally, by Monte Carlo distribution. In the latter case, one considers a given number of random occupancy matrices, called "replications", all fulfilling the rows and column constraints associated with the chosen null, and for each matrix computes the relevant statistics. For each couple (i, j), one keeps record of the fraction of times the statistic computed on the random matrix is below the statistic originally computed with the empirical matrix. When a sufficiently large number of replications is considered, for the Law of Large Number, the fraction converges toward the p-score defined in (8). In the case of $\mathcal{H}_1$, $\mathcal{H}_2$ and $\mathcal{H}_3$, the generation of random matrices is easy, and can be obtained with a simple fire-and-place algorithm, described in Appendix A.1. Conversely, in the case of $\mathcal{H}_4$, because of the number of constraints, the generation of random matrices is more problematic and require a different approach. The issue and the employed algorithm are discussed in Appendix A.2. It is possible to give a straightforward interpretation of the p-score $p_{i,j}$ as a measure of relatedness: a value of $p_{i,j}$ near to one means that the associated value of $A_{i,j}$ is much larger than the one expected under the null. As a consequence the two fields are strongly related.

As we will see in the next sections, the use of a p-score makes the actual choice of the relatedness statistics basically irrelevant. Moreover the use of p-scores as a proxy for relatedness lead naturally to the idea of positive and negative association. Assume that under a given null $\mathcal{H}$, for a couple of fields i and j, we obtain $p_{i,j} = .5$. This means that, according to $\mathcal{H}$, half of the possible value of $A_{i,j}$ are below $\hat{A}_{i,j}$ and half are above it. In other words, the probability to find a value less than $\hat{A}_{i,j}$ equals the probability to find a grater value. Since the degree of association of i and j is fully explained by the random model, one concludes that the relatedness between i and j is zero. Conversely, pairs that show an association with a p-score

[2]This is true in general for all the statistics based on the adjacency matrix, like for instance the cosine index. For finiteness we limit our comparison to the three example presented in this section.

[3]This was partially done in (Breschi et al., 2003). Indeed they considered a p-score based measure in their investigation of the effect of firm size threshold on the average degree of relatedness across sectors. They however revert to cosine index in their regression analysis.

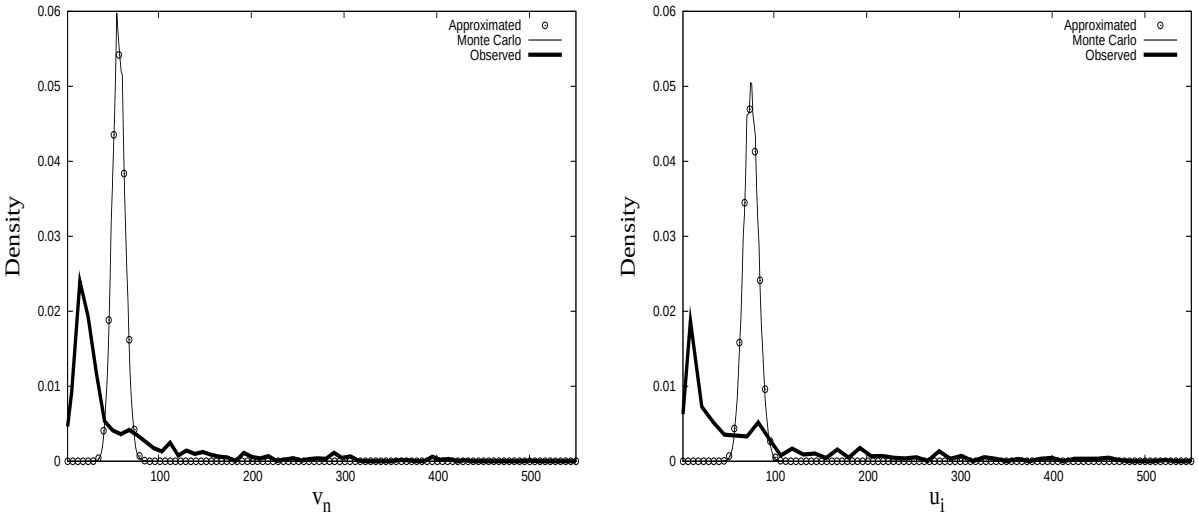


Figure 1: The left panel reports the approximated (empty circles) and the Monte Carlo (thin black line) implied distribution of firm scopes under $\mathcal{H}_2$. The thick black line represents the empirical distribution. The right panel reports the approximated (empty circles) and the Monte Carlo (thin black line) implied distribution of columns occupancy numbers under $\mathcal{H}_3$, together with the empirical distribution (thick black line).

grater (resp. less) than .5 must be interpreted as positively (resp. negatively) correlated. In this spirit, the measures of positive and negative relatedness are defined respectively as

$$p^{+}_{i,j}\left(A,\mathcal{H}\right)=\max\left(2\,p_{i,j}-1,0\right) \tag{9}$$

and

$$p^{-}_{i,j}\left(A,\mathcal{H}\right)=-\min\left(2\,p_{i,j}-1,0\right)\ . \tag{10}$$

Both $p^{+}_{i,j}$ and $p^{-}_{i,j}$ take values in $[0,1]$. The quantity in (9) is a proxy for (positive) relatedness: it is equal to 1 for fully positively associated pairs ($p_{i,j}=1$) and it is zero when no association is found ($p_{i,j}=0.5$). On the contrary the quantity in (10) is a proxy for negative relatedness (or anti-relatedness): it equals 1 for fully negatively associated pairs ($p_{i,j}=0$) and equals zero when there is no association at all.

4 Data Description and the Effect of Null on Relatedness Measures

The empirical exercises of the present paper are based on patent data collected and published on-line by the NBER Patent Data project. The database matches information on patent assignees from U.S. patent office with firms appearing in the COMPUSTAT database. The description of the matching procedure can be found in Bessen (2009).[4] The dataset covers the period from 1976 to 2006. It is very large and contains milions of lines. In order to have a manageable dataset we consider only firms with more than 50 patents. Thereafter we discard

[4]The data and the documentation are publicly available at:

`https://sites.google.com/site/patentdataproject/Home/downloads`

all industrial sectors that have not been chosen by the remaining firms. The final dataset is composed of $N = 1289$ firms and $I = 975$ industrial sectors (classified according to four digit IPC). The observed adjacency matrix has $M = 73598$ elements different from zero ($\approx 5.8\%$ of the total entries). For the sake of clarity hereafter we will refer to firms and patent classes.

The first element to consider in valuating the null hypotheses introduced in the previous section is the implications they have in terms of the distribution of non-constrained variables. Suppose to assume hypothesis $\mathcal{H}_2$. In this case the distribution of the number of classes per firm v_n is not fixed, and follows the same distribution for any n:

$$\text{Prob}\left[v_n = k | \, \mathcal{H}_2\right] = \sum_{\{l_i=0,1\}} \sum_{i=1}^{I} \binom{I}{k} \left(\frac{u_i}{N}\right)^{l_i} \left(1 - \frac{u_i}{N}\right)^{1-l_i} \delta_{\sum_{i=1}^{I} l_i, k} \, , \tag{11}$$

where δ is the Kronecker delta function, and the sum is performed over all the possible vectors $(l_1, ..., l_I)$, with $l_i = 0, 1$. The sum does in fact contains 2^I terms and since $I = 975$, it is unfeasible. This problem can be circumvented by an approximation: setting all column occupancy constant and equal to their average value:

$$u_i \approx \langle u \rangle \stackrel{\text{def}}{=} \frac{\sum_{i=1}^{I} u_i}{I},$$

one obtains the expression:

$$\text{Prob}\left[v_n = k | \, \mathcal{H}_2\right] \approx \binom{I}{k} \left(\frac{\langle u \rangle}{N}\right)^{k} \left(1 - \frac{\langle u \rangle}{N}\right)^{1-k} \, . \tag{12}$$

Figure 1 reports the implied density for the v_n under $\mathcal{H}_2$. Both the approximated expression (12) and the original one (11), computed with Monte Carlo techniques, are reported.[5] As can be seen, the approximation is very good across the entire support of the distribution. Conversely, the comparison with the empirical density differs in a noticeable way. Indeed the implied distribution is much more peaked. This suggests that hypothesis $\mathcal{H}_2$ implies an almost uniform distribution of sectors across the different firms, while in the data firm scopes are highly heterogeneous.

The same reasoning could be applied to $\mathcal{H}_3$. In this case the implied distribution of the u's, that is the number of firms per class, would be peaked around its mean value. The right panel of Figure 1 reports the implied distribution of the columns occupancy numbers under the null of fixed firms scopes, both approximated through (12) and computed via Monte Carlo, together with the empirical one. The latter appears to be Pareto-like and completely disagrees with the implied one.

The choice of null $\mathcal{H}_2$ or $\mathcal{H}_3$ implicitly assumes an uniform distribution of firms across fields, or of fields across firms. None of the two assumptions is verified in our data. This disagreement has a direct effect on economic inference. Consider for instance the analysis performed in Breschi et al. (2003) of the degree of non-randomness in firms' patent diversification structure. They classify firms according to the number of patents they own and look at the degree of relatedness among the different fields when firms of different classes are considered. They adopt $\mathcal{H}_2$ as a null and find that the fraction of positive related fields increase dramatically when

[5] Like any other statistics, the occupancy density can be computed by replications of the co-occurrencies matrix, under the chosen null, as described in the previous section.

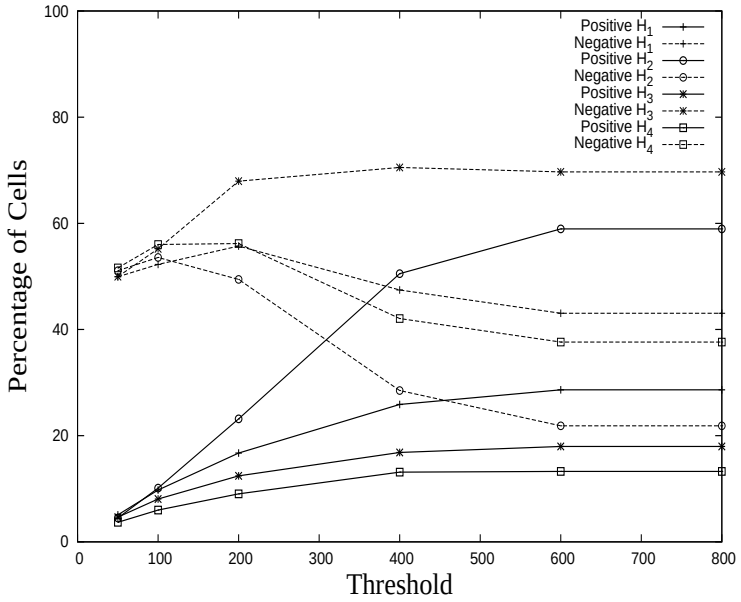


Figure 2: Percentage of pairs with a p-value grater than 0.90 (darker lines) or less than 0.10 (lighter lines) under the different null as a function of the larger allowed firm scope.

larger firms are considered, from less than 5% to more than 80%. Conversely the fraction of negatively related fields decreases from almost 50% to less than 4%.[6] We repeat their exercise in our data. We consider different groups of firms, taking successively only those firms that have patents in less than 50, 100, 200, 400, 600 and 800 different patent classes. Analogously to what done in Breschi et al. (2003), for each choice of the threshold we count how much pairs of patent classes display a p-value greater than 0.90 or less than 0.10, according to the four null hypotheses, using the t-statistics defined in (4). Results are shown in Figure 2. As can be seen, we replicate the Breschi et al. (2003) findings when using $\mathcal{H}_2$: the firms' scope has a large effect on both positive and negative relatedness. Conversely, using $\mathcal{H}_4$ as a null, one obtains much more stable levels. The stronger dependence of relatedness levels on firms size, generated by $\mathcal{H}_2$, is a spurious phenomenon. It has to do with the fact that when larger firms are considered, the scope of the firms becomes, by definition, more heterogeneous, and the disagreement of $\mathcal{H}_2$ with the data increases.

This suggests that leaving some data constrain "free" produces a spurious overestimation of the deterministic nature of the adjacency scenario with respect to that obtained if all constraints were taken into account.

The discussion above and the examples in this sections suggest to consider $\mathcal{H}_4$ as the more reliable null and discard all other hypotheses. But so far we only analyzed the structure of relatedness among fields. In the next section we will reinforce this impression by showing how and to what degree the choice of different nulls impact on the analysis of corporate coherence.

5 Corporate Coherence: Measures and Findings

We start by repeating on our database the analysis originally suggested in Teece et al. (1994). As a first measure of corporate coherence we consider the *weighted average relatedness* (WAR)

[6]Breschi et al. (2003) use the EPO-CESPRI database containing patent assignees (firms or individual) of patents grated by the European Patent Office.

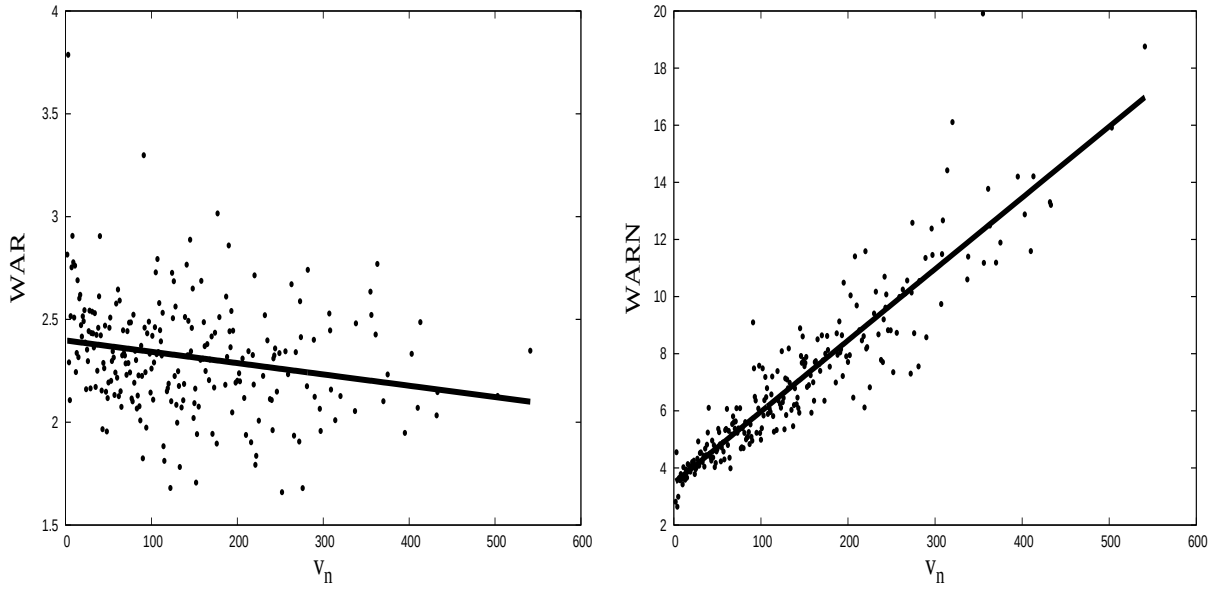


Figure 3: WAR_n (left) and WARN_n (right), computed according to (13) and (14), as a function of v_n.

defined according to

$$\mathrm{WAR}_n = \frac{1}{v_n} \sum_i C_{n,i} \left(\frac{\sum_{j \neq i} q_{n,j}\, \hat{t}_{i,j}}{\sum_{j \neq i} q_{n,j}} \right) . \tag{13}$$

where $q_{n,j}$ is the number of patents in class j owned by firm n and where $\hat{t}_{ij}$ is defined in (4). For each firm n only those patent classes i in which the firm is active, i.e. such that $C_{n,i} = 1$, are considered. Thereafter one computes the mean relatedness between class i and all other classes $j \neq i$ within the firm, weighted by the number of patents $q_{n,j}$ held by the firm itself. A final averaging is performed through the v_n patent classes in which firm n is present.[7]

A complementary definition of firm coherence suggested by Teece et al. (1994) is the *weighted average relatedness of neighbors* (WARN). Consider the n-th firm with a total number of patent classes equal to v_n. There are $v_n\,(v_n - 1)/2$ possible pairs of such classes. Nevertheless only $v_n - 1$ have to be chosen in order to produce a graph that connect all the firm's activities. Such graph becomes a weighted graph once each pair has been associated with the corresponding relatedness. For these reasons it is usally referred as a *weighted spanning tree*. Its total weight is defined as the sum of the relatednesses of all its pairs. The maximum spanning tree of firm n is the weighted spanning tree whose total weight is greater or equal than those of all other weighted spanning trees. Let $M^n_{i,j}$ be the adjacency matrix representation of the maximum spanning tree defined as:

$$M^n_{i,j} = \begin{cases} 1 & \text{if } i,j \in \text{maximum spanning tree,} \\ 0 & \text{otherwise,} \end{cases}$$

where $i, j = 1, ..., I$. Now one can compute the weighted average relatedness between a patent class i and its nearest neighbors in the maximum spanning tree, and take the mean values

[7]Note that in our case the "weight" of a patent class is proxyed by the number of patents the firms has obtained in that class, while Teece et al. (1994) consider the total workforce employed in the sector.

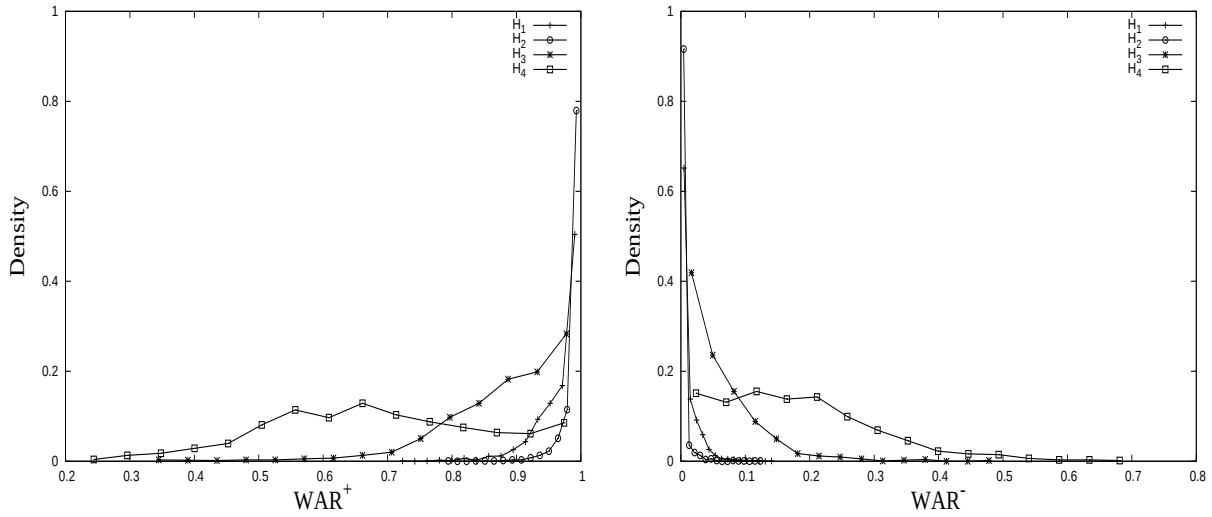


Figure 4: Empirical probability density of $\mathrm{WAR}_n^+(N)$ (left panel) and $\mathrm{WAR}_n^-(N)$ (right panel) for different null hypotheses.

including only those pairs that belong to the MST^n:

$$\mathrm{WARN}_n = \frac{1}{v_n} \sum_i C_{n,i} \left(\frac{\sum_{j \neq i} q_{n,j}\, M^n_{i,j}\, \hat{t}_{i,j}}{\sum_{j \neq i} q_{n,j}\, M^n_{i,j}} \right), \tag{14}$$

In Figure 3 we report a scatter plot of WAR_n and WARN_n vs. firm scope v_n. The progressive reduction of WAR when the scope of the firm increases is broadly in accordance with Teece et al. (1994). If one looks only at the activities constituting the core of the company, that is those activities in which the firm is more specialized, as captured by the WARN, the picture changes. We find that the relatedness of the core activities increases the broader the scope of the firm. In other terms, as a firm get more diversified, the coherence across all its activities decreases, but, at the same time, the coherence of its core increases. Conversely, the result in Teece et al. (1994) seems to suggest a constant level of core coherence. Consider however that the span of firms' scope is much lower in their data than in our case. Estimating a linear relationship[8]:

$$y_n = \alpha\, v_n + \beta + \epsilon_n, \tag{15}$$

with OLS gives $\alpha = -8.85\,10^{-4}$ for WAR and a much higher $1.63\,10^{-2}$ for WARN, both significant at 1%. The root mean squared error (RMSE) is 0.19 for the regression on WAR and 0.72 for WARN.

Are the previous result robust with respect to the use of different measure of relatedness? What if one consider a different null hypothesis? As described in the Section 3, we will consider the p-score associated to different statistics (co-occurrences, odds-ratio or normalized co-occurrencies) as a measure of association and we will build both positive and negative relatedness. Starting from a measure A and a null model $\mathcal{H}$ one can define both a positive measure of coherence and a negative one (anti-coherence) by considering:

$$\mathrm{WAR}_n^{\pm}(A, \mathcal{H}) = \frac{1}{v_n} \sum_i C_{n,i} \left(\frac{\sum_{j \neq i} q_{n,j}\, p^{\pm}_{i,j}(A, \mathcal{H})}{\sum_{j \neq i} q_{n,j}} \right). \tag{16}$$

[8]In expression (15) the dependent variable y_n can be either WAR_n or WARN_n.

The distribution of $\mathrm{WAR}_n^{\pm}(N, \mathcal{H})$ for the population of firms are reported in Figure 4 taking normalized occurrencies $N_{i,j}$ in (5) as the relatedness statistics. The impact of the choice of the null is huge: while in the cases of $\mathcal{H}_1$ and $\mathcal{H}_2$ firm's coherence distribution is peaked around its maximum value, a more diversified structure appears when testing against $\mathcal{H}_3$ and this result is drastically amplified in the full constrained hypothesis $\mathcal{H}_4$. In the latter case firm's distributions for both positive and negative coherence are spread through nearly the entire range $[0, 1]$. This confirms the idea suggested in Section 2 that neglecting some data constraints overestimates pair relatednesses, pushing the firm's coherence distribution toward its maximum value.

Keeping the same statistics N, one can investigate the relationship between coherence level and firm's scope. The result are reported in Figure 5 for positive coherence and in Figure 6 for the negative version. Inspection of Figure 5 reveals that, in the case of the partial constrained hypotheses $\mathcal{H}_1$-$\mathcal{H}_2$-$\mathcal{H}_3$, the linear regression (15) still produces a good agreement with the observed WAR^+, with a highly significant and negative slope (see Table 1 for details). A clear advantage in using measure of coherence based on p-scores is that the relationship appears more clearly. Indeed the regression is less noisy having a RMSE which is one order of magnitude lower than those obtained using t-statistics. Even in this case, however, we observe a constant de-coherence rate. According to $\mathcal{H}_1$, $\mathcal{H}_2$ and $\mathcal{H}_3$, the effect of diversification on firm's coherence is essentially scale invariant. A diffrent picture emerge if one uses the fully constrained null model $\mathcal{H}_4$. In this case a linear fit would poorly describe the behaviour of WAR^+ as a function v_n. Conversely the logarithmic regression

$$y_n = \alpha \log v_n + \beta \tag{17}$$

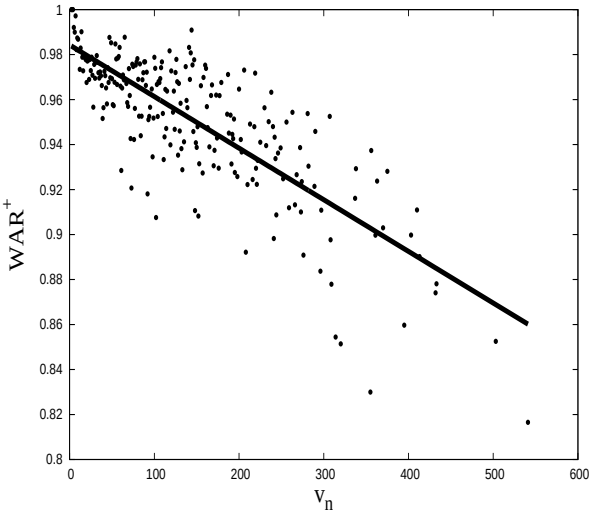
fits surprisingly well with the data. This empricial result has a natural economic interpretation: if heterogeneity in firm's scope and patent classes' size is properly accounted for, then the addition of a new activity to small firms reduces coherence much more than in large firm.

A specular behaviour is found for the mean weighted average negative relatedness (or anti-relatedness) reported in Figure 6. A pair of patent classes is strongly anti-related whenever the presence of a firm in one class of the pair strongly reduces the probability that the same firm is active in the other class. As a consequence, an high level of WAR_n^- for firm n corresponds to a diversification strategy that requires a large number of capabilities. Not surprisingly, Figure 6 shows that WAR_n^- is an increasing function of the firm scopes, with a linear behaviour when relatedness is measured against $\mathcal{H}_1$, $\mathcal{H}_2$ or $\mathcal{H}_3$ and a logarithmic trend for $\mathcal{H}_4$. As expected small firms maintain a low level of mean anti-relatedness, while larger firms tend to invade classes that are based on very different knowledge. This diversification pattern is again saturated in hypothesis $\mathcal{H}_4$, revealing that such a mechanism is increasingly reduced when increasing the firm size.

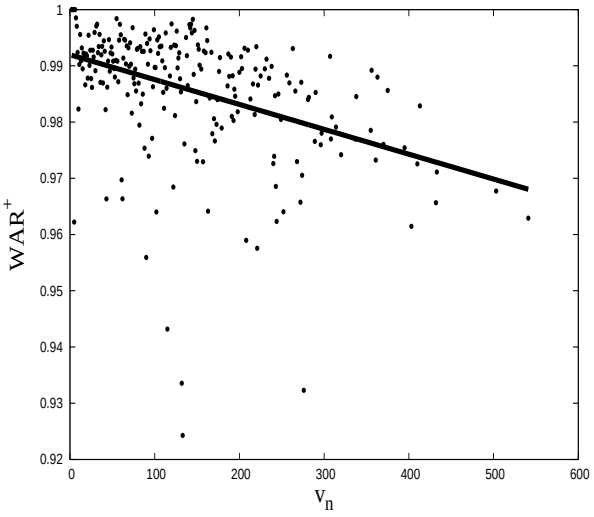
Concerning the use of different statistics, as long as one bases the definition of coherence on the associated p-score, the choice does not seem important. Figure 7 reports the relationship between firm's scope v_n and positive coherence WAR_n^+ computed, under $\mathcal{H}_4$, using normalized co-occurrencies, co-occurrences or odds-ratio together with their log-linear fit (17). As can be seen the curves are basically identical. The same applies irrespectively of the chosen null model. [9]

A positive and negative version of WARN can be defined, similarly to what done for WAR,

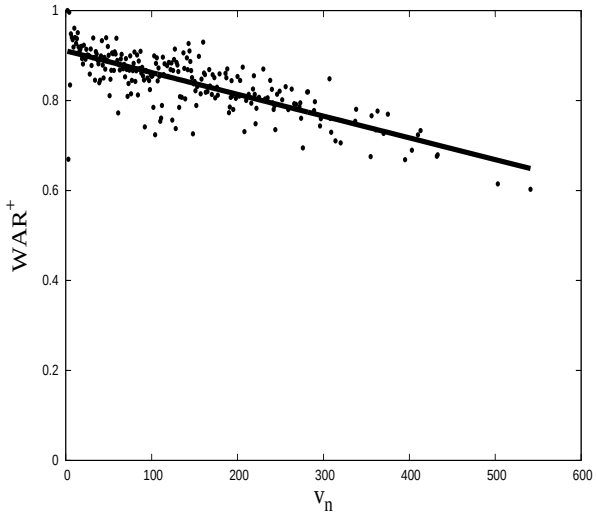
[9] The plots relative to other nulls and the associates estimates are available upon request.



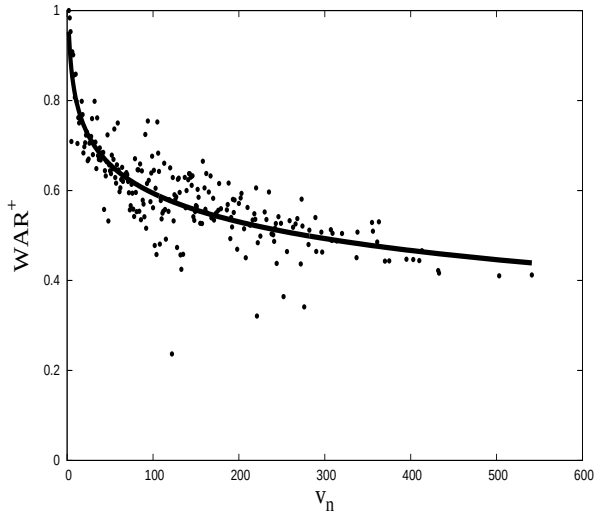
(a) $\mathcal{H}_1$: Random column and row sums.



(b) $\mathcal{H}_2$: Fixed column sums.

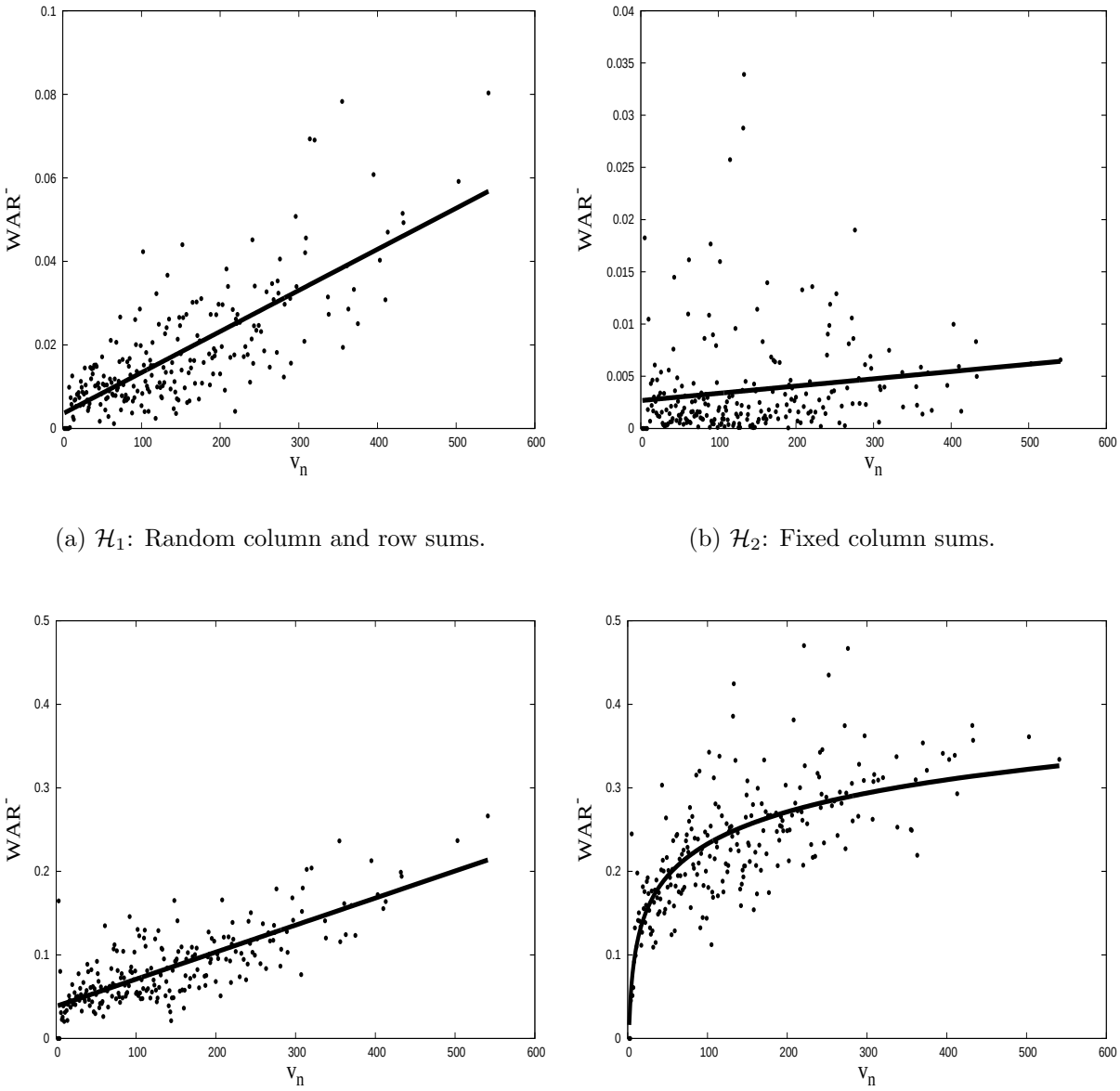


(c) $\mathcal{H}_3$: Fixed row sums.



(d) $\mathcal{H}_4$: Fixed column and row sums.

Figure 5: Scatter plot of $\mathrm{WAR}_n^+(N)$ vs. v_n for different null hypothesis together with the estimated regression. The number of normalized co-occurrencies N is used as a relatedness statistics. Regressions estimates are reported in Table 1.



(a) $\mathcal{H}_1$: Random column and row sums.

(b) $\mathcal{H}_2$: Fixed column sums.

(c) $\mathcal{H}_3$: Fixed row sums.

(d) $\mathcal{H}_4$: Fixed column and row sums.

Figure 6: Scatter plot of $\mathrm{WAR}_n^-(N)$ vs. v_n for different null hypothesis together with the estimated regression. The number of normalized co-occurrencies N is used as a relatedness statistics. Regressions estimates are reported in Table 1.

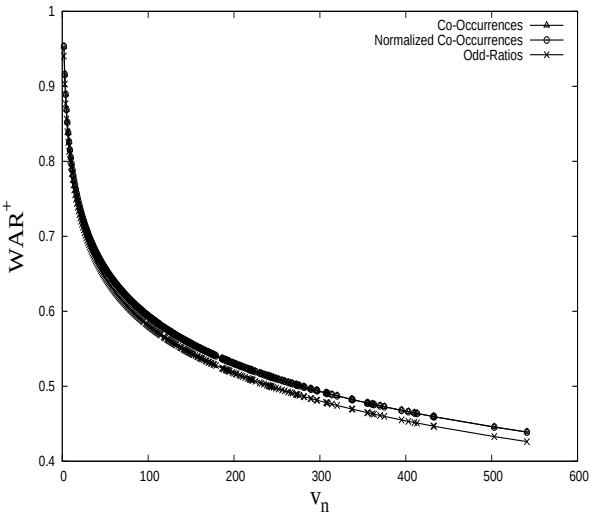


Figure 7: WAR^{+} as a function of v_n in the case of $\mathcal{H}_4$ computes using the three proposed relatedness statistics.

in the following way:

$$\mathrm{WARN}_n^{\pm}\left(A,\mathcal{H}\right)=\frac{1}{v_n}\sum_i C_{n,i}\left(\frac{\sum_{j\neq i}q_{n,j}\,M_{i,j}^{(\pm,n)}\,p_{i,j}^{\pm}\left(A,\mathcal{H}\right)}{\sum_{j\neq i}q_{n,j}\,M_{i,j}^{(+,n)}}\right), \tag{18}$$

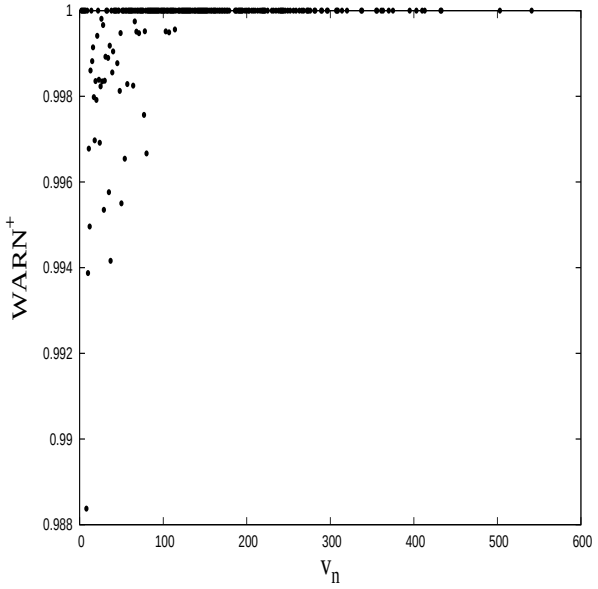
for given relatedness statistics A and null hypothesys $\mathcal{H}$. Notice that in general the maximum spanning tree $M^{(\pm,n)}$ associated with the diversification structure of a firm is different for positive and negative relatedness.

The observed behaviour for $\mathrm{WARN}^{\pm}$ is reported in Figures 8 and 9. As shown in the former figure, firms display a bunch of core activities where they are completely positively coherent (i.e. with a maximum WARN^{+} level) independently of their scope. Essentially we find a constant level of positive coherence which is well in tune with the intuition proposed in Teece et al. (1994) and with their findings. The difference with respect to the results obtained with the t-statistics, and reported in the left panel of Figure 3, is also related to the compact nature of the p-score. It is worth to notice that a measure of relatedness must be, in fact, bounded. When every possible value of the relatedness statistics generated by the null happens to be below the observed one, we have to conclude that the pair has reached its maximum achievable level of association: an higher value would be meaningless from a null-analysis perspective.

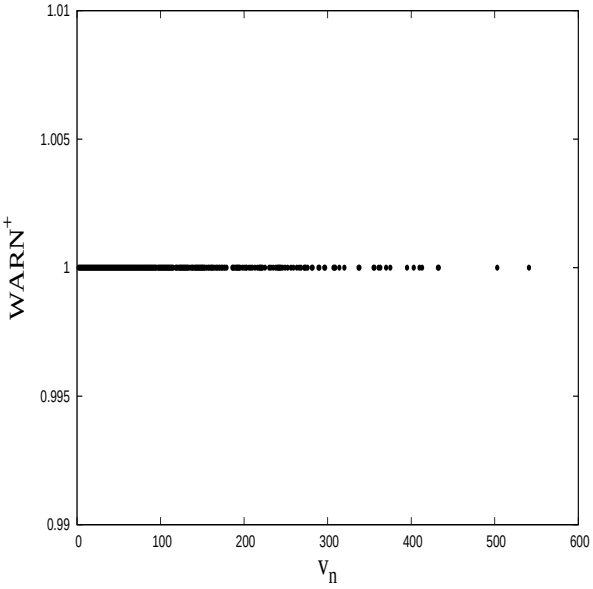
Finally, we find that weighted average anti-relatedness of neighbors (WARN^{-}) is well described, for all the null models, by an exponential law of the type

$$\mathrm{WARN}_n^{-}=\alpha\,\left(1-\exp\left(-\beta\,v_n\right)\right). \tag{19}$$

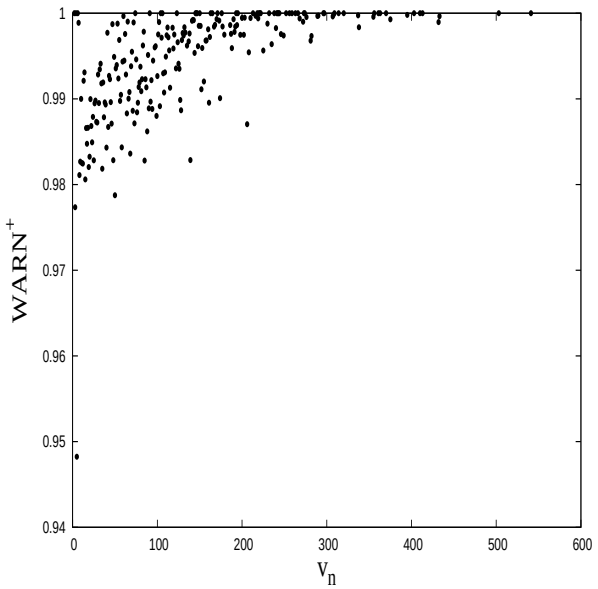
This result suggests the existence of an activation threshold. For small firms both WARN^{-} and WAR^{-} are very low. When the scope of the firm increases, WAR^{-} slowly increases with it, while WARN^{-} is characterized by a rapid saturation. In the $\mathcal{H}_4$ case, its maximum value is already reached for $v_n\sim 100$.



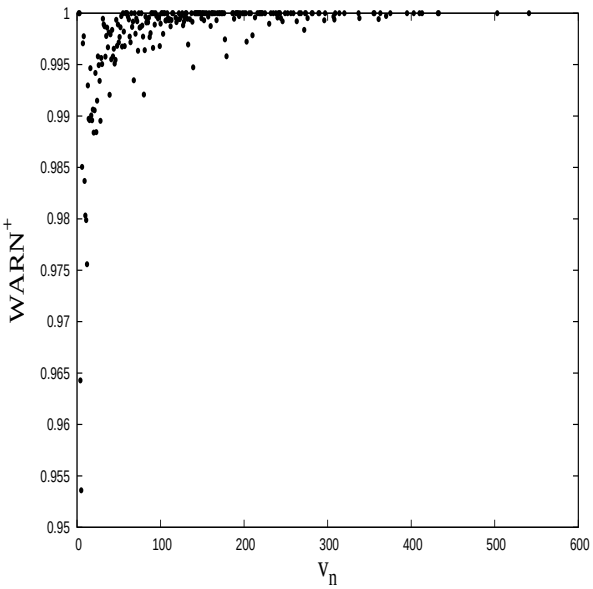
(a) $\mathcal{H}_1$: Random column and row sums.



(b) $\mathcal{H}_2$: Fixed column sums.

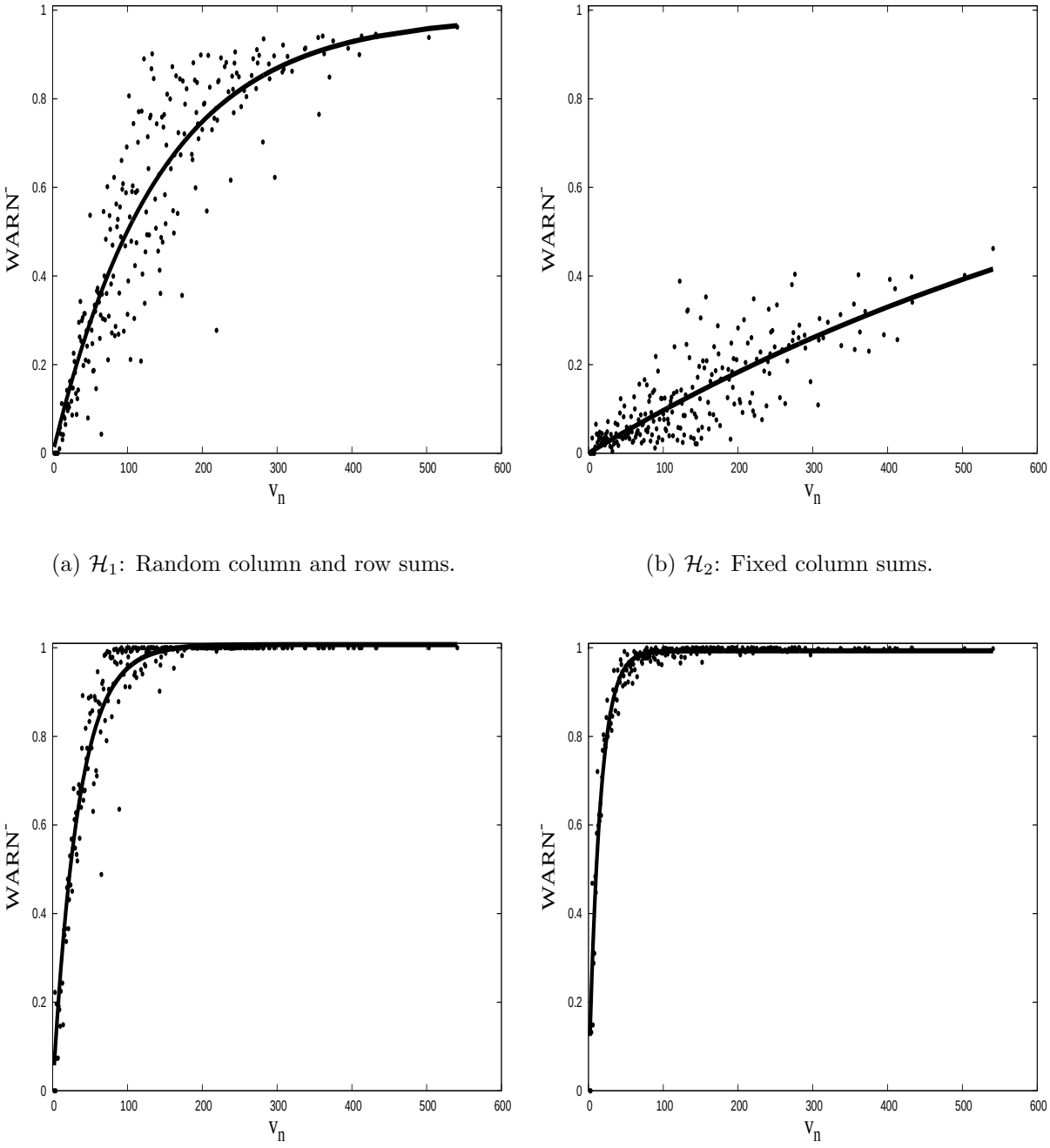


(c) $\mathcal{H}_3$: Fixed row sums.



(d) $\mathcal{H}_4$: Fixed column and row sums.

Figure 8: Scatter plot of $\mathrm{WARN}_n^+(N)$ vs. v_n for different null hypothesis together with the estimated regression. The number of normalized co-occurrencies N is used as a relatedness statistics. Regressions estimates are reported in Table 1.



(a) $\mathcal{H}_1$: Random column and row sums.

(b) $\mathcal{H}_2$: Fixed column sums.

(c) $\mathcal{H}_3$: Fixed row sums.

(d) $\mathcal{H}_4$: Fixed column and row sums.

Figure 9: Scatter plot of $\mathrm{WARN}_n^-(N)$ vs. v_n for different null hypothesis together with the estimated regression. The number of normalized co-occurrencies N is used as a relatedness statistics. Regressions estimates are reported in Table 1.

		α	β	$RMSE$	Model
H_1	WAR^+	$-2.294320e-04$ $[2.006207e+01]$	$9.842853e-01$ $[4.806972e+02]$	$1.860130e-02$	linear
	WAR^-	$9.842485e-05$ $[1.784885e+01]$	$3.538272e-03$ $[3.583651e+00]$	$8.969300e-03$	linear
	$WARN^-$	$9.862520e-01$ $[2.927760e+01]$	$7.106452e-03$ $[1.420152e+01]$	$1.191080e-01$	exp
H_2	WAR^+	$-4.422043e-05$ $[6.841834e+00]$	$9.919780e-01$ $[8.571967e+02]$	$1.051270e-02$	linear
	WAR^-	$6.948663e-06$ $[2.422512e+00]$	$2.673725e-03$ $[5.206076e+00]$	$4.665510e-03$	linear
	$WARN^-$	$9.270066e-01$ $[2.451838e+00]$	$1.099867e-03$ $[2.114068e+00]$	$6.412200e-02$	exp
H_3	WAR^+	$-4.837426e-04$ $[1.846181e+01]$	$9.105250e-01$ $[1.940800e+02]$	$4.261910e-02$	linear
	WAR^-	$3.236429e-04$ $[1.884814e+01]$	$3.862268e-02$ $[1.256244e+01]$	$2.792940e-02$	linear
	$WARN^-$	$1.007200e+00$ $[1.882406e+02]$	$2.938008e-02$ $[4.010970e+01]$	$5.768390e-02$	exp
H_4	WAR^+	$-9.181099e-02$ $[2.404566e+01]$	$1.016809e+00$ $[5.658533e+01]$	$5.882380e-02$	log
	WAR^-	$5.538015e-02$ $[1.566224e+01]$	$-2.211615e-02$ $[1.329020e+00]$	$5.447480e-02$	log
	$WARN^-$	$9.932689e-01$ $[4.080911e+02]$	$6.650294e-02$ $[5.419234e+01]$	$3.244760e-02$	exp

Table 1: Reports OLS estimates of Figures 5, 6 and 9. The last column indicates the model used for regression, linear (15), logarithmic (17) or exponential (19). For each parameter estimate the corresponding t-statistics is reported in brackets.

6 Conclusions

The analysis of the technological scope and structure of corporate activities has become increasingly common in recent times. It has been applied at very different scales, from the study of managerial behaviour pertaining to the theory of the firm to the empirical investigation of sectoral dynamics. As discussed inside a broad theoretical tradition, and shown by several empirical studies, it is not only the scope of the technological diversification of a firm that matters, but rather the degree of complementarity, or the strength of externalities, existing among the activities in which it diversifies. This idea led to the notion of corporate coherence: a company is more coherent if its activities take place (mainly) in fields which are more strictly related. Despite its relevance, the design of appropriate statistical tools apt to measure the degree of corporate coherence did not received much attention. Lacking any reliable external (and exogenous) definition of a notion of "proximity" among technical activities, the literature mainly explored the possibility of building a notion of topology starting from the observed diversification structure of the firms themselves. The approach is similar to the one used by ecologists, who measure the relatedness among different species by observing the pattern of their geographical distribution. Following the seminal work of Teece et al. (1994), this paper proposes several methodological improvements with respect to the tools presently adopted in the field. First, we show that irrespectively of the statistics chosen to asses the degree of relatedness among activities, the appropriate measure to use is the p-score of the statistics itself, as it neutralizes spurious effects generated by the nature of the distribution of the underlying variables. Indeed, irrespectively of the measure adopted (patents, products, lines of business,...) and the relative definition of technological fields, the distribution of business units across these fields, and the distribution of fields across business units, is likely to be extremely uneven. The result is that any adopted statistics will display an highly skewed distribution, making measures based on central tendency, like mean and variance, unreliable. Moreover the use of the p-score naturally leads to a notion of positive and negative coherence, allowing for the contemporaneous (and complementary) analysis of two likely asymmetric phenomena, taking place in the core of the firm: the development of competencies along related fields, facilitated by the existence of positive technological spillover, and the push toward diversification and exploration of new fields.

Second, we discuss the relevance of the correct choice of the null-hypothesis, that is the benchmark against which the observed degree of coherence is measured. We show that, lacking any specific reason not to do so, the correct choice is the fully constrained null hypothesis. The distribution of the statistics cannot be in general computed under this null, but we present efficient and easy-to-implement numerical methods which can be effectively used to obtain Monte Carlo estimates of the desired quantities.

We illustrate our methods applying them to the analysis of data from the NBER patent data project. We show that, when the appropriate null is used, the actual degree of relatedness among sectors is not strongly influenced by a possible cut-off on firm's size. Concerning the relationship between firm's scope and coherence, our empirical findings broadly confirm the original intuition in Teece et al. (1994).

The degree of corporate coherence, when measured using all the activities in which a firms is involved, tends to decrease with the scope of the activities themselves. This is testified by the contemporaneous decrease of positive coherence and increase of negative coherence, when more diversified firms are considered. The effect is however non linear: the marginal reduction of coherence due to the addition of new fields decreases with the number of fields in which the firm is active.

On the other hand, if one only considers the degree of coherence existing among the core activities of the firm, this turns out to be a non decreasing function of firm's scope. In this case we observe a clear threshold effects: while for firms active in very few sectors the degree of core coherence increases with the number of active fields, as soon as a sufficiently diversified structure is reached the effect of scope on coherence disappears.

References

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Kluwer Academic Publisher.

Berry, C. H. (1975). *Corporate growth and diversification*. Princeton University Press.

Bessen, J. (2009, May). Matching patent data to compustat firms. Nber working papers, National Bureau of Economic Research.

Bishop, Y. Y. M., S. E. Fienberg, and P. W. Holland (1975). *Discrete Multivariate Analysis*. MIT Press, Cambridge, Massachusetts.

Breschi, S., F. Lissoni, and F. Malerba (2003). Knowledge-relatedness in firm technological diversification. *Research Policy 32*(1), 69 – 87.

Bryce, D. J. and S. G. Winter (2010). A general interindustry relatedness index. *Management Science 55*(9), 1570–1585.

Connor, E. F. and D. Simberloff (1979). The assembly of species communities: Chance or competition? *Ecology 60*(6), 1132–1140.

Feller, W. (1976). *An Introduction to Probability Theory and Its Applications*. Wiley.

Gotelli, N. J. (2001). Research frontiers in null model analysis. *Global Ecology and Biogeography 10*(4), 337–343.

Gotelli, N. J. and G. R. Graves (1996). *Null Model in Ecology*. Smithsonian Institution Press, Washinton, DC.

Mehta, C. R., N. R. Patel, and R. Gray (1985). Computing an exact confidence interval for the common odds ratio in several 2 2 contingency tables. *Journal of the American Statistical Association 80*(392), 969–973.

Montgomery, C. A. (1982). The measurement of firm diversification: Some new empirical evidence. *The Academy of Management Journal 25*(2), 299–307.

Nesta, L. and P. P. Saviotti (2006). Firm knowledge and market value in biotechnology. *Industrial and Corporate Change 15*(4), 625652.

Pavitt, K. (1984). Sectoral pattern of technical change: Towards a taxonomy and a theory. *Research Policy 13*, 343–373.

Pehrsson, A. (2006). Business relatedness and performance: A study of managerial perceptions. *Strategic Management Journal 27*(3), 265–282.

Piscitiello, L. (2004). Corporate diversification, coherence and economic performance. *Industrial and Corporate Change 13*(5), 757–787.

Roberts, A. and L. Stone (1990). Island-sharing by archipelago species. *Oecologia 83*(4), 560–567.

Rudas, T. (1985). *Odds ratios in the analysis of contigency tables*. Sage Publications.

Rumelt, R. P. (1974). *Strategy, structure, and economic performance*. Harvard University Press.

Rumelt, R. P. (1982). Diversification strategy and profitability. *Strategic Managment Journal 3*(4), 359–369.

Ryser, H. J. (1960). Matrices of zeros and ones. *Bullettin of the American Mathematical Society 66*(1), 442–464.

Sanderson, J. G., M. P. Moulton, and R. G. Selfridge (1998). Null matrices and the analysis of species co-occurrences. *Oecologia 116*(1-2), 257–283.

Snijders, T. A. B. (1991). Enumeration and simulation methods for 0-1 matrices with given marginals. *Psychometrika 56*(3), 397–417.

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

Teece, D. J. (1980). Economics of scope and the scope of an enterprise. *Journal of Economic Behavior and Organization 1*(3), 223–247.

Teece, D. J. (1982). Towards an economic theory of the multiproduct firm. *Journal of Economic Behavior and Organization 3*(1), 39–63.

Valvano, S. and D. Vannoni (2003). Diversification strategies and corporate coherence evidence from italian leading firms. *Review of Industrial Organization 23*(1), 25–41.

Zaman, A. and D. Simberloff (2002). Random binary matrices in biogeographical ecology-Instituting a good neighbor policy. *Environmental and Ecological Statistics 9*, 405–421.

A APPENDIX A

A.1 Fire-and-Place Algorithm

Suppose that both $u_i = \sum_n C_{n,i}$ (i.e. the number of firms active in sector i) and $v_n = \sum_i C_{n,i}$ (i.e. the number of industrial sectors chosen by firm n) are random quantities. In this scenario only the total number of links

$$M \stackrel{\text{def}}{=} \sum_i u_i^0 = \sum_n v_n^0 = \sum_{i,n} C_{i,n}^0, \tag{20}$$

is a fixed quantity. The random assignment coincides with the random placement of balls in boxes. Each $C_{n,i}$ represents the success ($C_{n,i} = 1$) or the failure ($C_{n,i} = 0$) of placing the i-th ball in the n-th box (both firms and patent classes can be interpreted as balls or boxes).

The generation of the random sample shows no particular difficulties in this case. For each replication we start from an empty matrix, i.e. a matrix whose entries are all set to zero. Thereafter the matrix is filled by means of a fire-and-place algorithm.

At each step a pair of indexes (n^*, i^*), with $n^* \in \{1, ..., N\}$ and $i^* \in \{1, ..., I\}$, is extracted from a flat distribution.

If the corresponding element C_{n^*,i^*} is empty (i.e., equal to zero) then a 1 is placed. Otherwise a new "bullet" is fired. The procedure is repeated until M bullets are placed. This allows to obtain a random replication generated according to $\mathcal{H}_1$.

A similar approach is taken for the generation of random paths according to $\mathcal{H}_2$ and $\mathcal{H}_3$: in the case of $\mathcal{H}_2$ (resp. $\mathcal{H}_3$) a 1-element is placed in the matrix at the uniformly extracted entry (n^*, i^*) (provided that the entry is empty). The coulum sum (resp. row sum) constraint is imposed subtracting 1 from the number $\hat{u}_{i^*}$ (resp. $\hat{v}_{n^*}$), i.e.

$$\hat{u}_{i^*} \Rightarrow \hat{u}_{i^*} - 1 \quad (\text{resp.} \quad \hat{v}_{n^*} \Rightarrow \hat{v}_{n^*} - 1\ ). \tag{21}$$

If the extracted pair (n^*, i^*) is such that $\hat{u}_{i^*} = 0$ (resp $\hat{v}_{n^*} = 0$) we ignore the extraction and the algorithm chooses another pair.

We continue the pairs extraction until the number of placed 1 is equal to M or, equivalently, until:

$$\forall i,\ \hat{u}_i = 0 \quad (\text{resp.} \quad \forall n,\ \hat{v}_n = 0\ ). \tag{22}$$

Following these procedures we generate 10^2 adjacency matrices, in order to span a large number of configurations.

A.2 Swap Algorithm

We indicate by $\mathbb{S}(\hat{\mathbf{u}}, \hat{\mathbf{v}})$ the space of $N \times I$ binary matrices whose column and row sums are given by $(\hat{u}_i)_{\{i=1,...,I\}}$ and $(\hat{v}_n)_{\{n=1,...,N\}}$ respectively.

The generation of a random sample in $\mathbb{S}(\hat{\mathbf{u}}, \hat{\mathbf{v}})$ is not a trivial problem. For large and sparse matrices the fire-and-place algorithm usually reaches a locked-in state.

Sanderson et al. (1998) propose a modification of the well-known knight's tour algorithm in order to produce a sequence of matrices in $\mathbb{S}(\hat{\mathbf{u}}, \hat{\mathbf{v}})$ such that each matrix is produced once and only once. Null-matrices are generated just iterating the fire-and-place algorithm untill a locked-in state is reached. Thereafter the algorithm is moved backward to the last unlocked state and iterated again. However this procedure is not suitable for our case where a large and

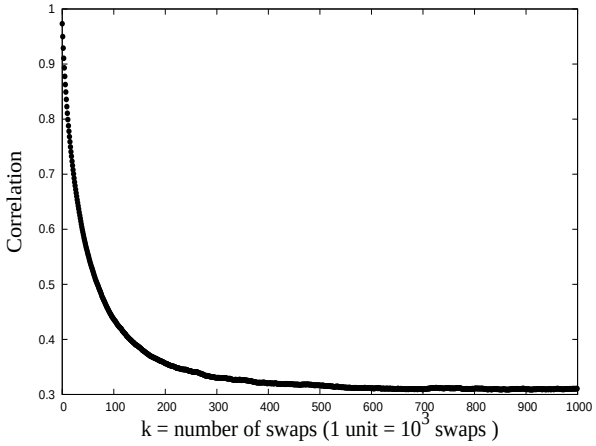


Figure 10: Reports the correlation coefficient between the observed adjacency matrix $\hat{C}$ and one obtained from it after an increasing number of random swaps.

sparse matrix must be produced. Only a relative small part of the entire matrix is completed: the algorithm moves forward and backward without reaching a final state.

In order to generate Monte Carlo replications of matrix with fixed column and row sums we adopt a *swap* algorithm. The algorithm starts with the observed matrix and looks for 2×2 diagonal or anti-diagonal sub-matrices, i.e. it looks for sub-matrices of the form

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \text{ or } \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{23}$$

and thereafter change one into the other. Note that the sub-matrix elements can belong to non-adjacent columns or rows, i.e. their distance in the original matrix can be as large as the matrix dimensions.

It is evident that the swap trasformation preserves both row and column sums.

Starting from the original observed matrix $\hat{C}$ we perform a N_{swp} number of swaps obtaining a new matrix $\hat{C}[1]$. We then re-start the algorithm with the new matrix $\hat{C}[1]$. After performing N_{rep} iterations we have at our disposal a Monte Carlo chain of matrices $\hat{C}[1], \hat{C}[2], ..., \hat{C}[N_{rep}]$, where each nodes of the chain is obtained from the previous one performing N_{swp} random swaps.

If N_{swp} is large enough the Monte Carlo chain can be considered at equilibrium and correlations among matrices can be neglected, i.e. the chain can be considered as Markovian. Similarly the larger the value of N_{rep} the larger the space spanned by the algorithm.

How much large should be chosen N_{swp} in order to consider the chain Markovian? Let $\hat{C}[k]$ be a matrix obtained form the observed adjacency matric $\hat{C}$ with k random swaps. Markovian properties of the chain can be checked computing the correlation coefficient:

$$\rho\left(\hat{C}, \hat{C}[k]\right) = \frac{\sum_{n,i}\left(\hat{C}_{n,i} - \mu\right)\left(\hat{C}_{n,i}[k] - \mu\right)}{\sqrt{\sum_{n,i}\left(\hat{C}_{n,i} - \mu\right)^2}\sqrt{\sum_{n,i}\left(\hat{C}_{n,i}[k] - \mu\right)^2}}, \tag{24}$$

where $\mu = M/(N * I)$ is the mean value of matrix, which is not modified by swaps. Figure 10 reports the correlation coefficients (24) as a function of k. We find that after approximately

3×10^5 swaps the correlation coefficient reaches its minimum value. Note that at equilibrium $\rho\left(\hat{C}, \hat{C}\left[k\right]\right) \approx 30\%$, this is due to the fact that several constraints link $\hat{C}$ with $\hat{C}\left[k\right]$.

Therefore we assume that after 3×10^5 swaps the chain is at equilibrium. Our routines are very fast and we can obtain Monte Carlo replications in reasonable time even with $N_{swp} = 10^6$, which is our final choice. Similarly for the null models $\mathcal{H}_1$, $\mathcal{H}_2$ and $\mathcal{H}_3$ we produce 10^2 replications of the adjacency matrix.

Swaps are not the unique transformations that map $\mathbb{S}\left(\hat{\mathbf{u}}, \hat{\mathbf{v}}\right)$ into itself. However they are the simplest ones. Moreover any two matrices in $\mathbb{S}\left(\hat{\mathbf{u}}, \hat{\mathbf{v}}\right)$ can be transformed one into another by swaps, as demonstraed in the paper of Ryser (1960). Therefore the entire space $\mathbb{S}\left(\hat{\mathbf{u}}, \hat{\mathbf{v}}\right)$ can be spanned by simply iteratively swapping one matrix of the set.