



Paper to be presented at the

DRUID Society Conference 2014, CBS, Copenhagen, June 16-18

## **Technological Distance Measures: Theoretical Foundation and Empirics**

**Florian Stellner**

Max Planck Institute for Innovation and Competition  
Munich Center for Innovation and Entrepreneurship Research  
f\_stellner@gmx.de

### **Abstract**

Measuring technological distance between companies is frequently a concern for researchers dealing with spillovers, alliances and M&A. I provide a discussion of the concept of technological distance and how it relates to the potential for technology flows such as spillovers. I present a comprehensive set of distance measures, including several measures new to the literature. Two of these measures take into account the relatedness of the technology fields. Microeconomic foundations for the use of bottom-up measures are established and implemented in the form of the aggregated patent-to-patent angular separation and a text based similarity measure using the information contained in the patent title and abstract. I also provide additional support for the use of the Jaffe covariance as a more suitable distance measure than the Jaffe angular separation when it comes to measuring spillovers. A set of axioms is developed and the distance measures are assessed on this basis. In an empirical analysis of the pairwise distances of 30 companies in the chemicals and pharmaceuticals & biotechnology sectors, the correlation of these measures is established at various levels of the IPC system and evaluated. I find that the within correlation at various levels of the IPC system is typically higher than the correlation between most distinct measures. Bottom-up distance measures are found to result in significantly different distance estimates compared to standard firm-based distance measures. The text based measure is found to have some correlation with the IPC based measures, particularly the Jaffe covariance and the patent-to-patent angular separation.

# Technological Distance Measures: Theoretical Foundation and Empirics

DRAFT, February 2014

**ABSTRACT:** *Measuring technological distance between companies is frequently a concern for researchers dealing with spillovers, alliances and M&A. I provide a discussion of the concept of technological distance and how it relates to the potential for technology flows such as spillovers. I present a comprehensive set of distance measures, including several measures new to the literature. Two of these measures take into account the relatedness of the technology fields. Microeconomic foundations for the use of bottom-up measures are established and implemented in the form of the aggregated patent-to-patent angular separation and a text based similarity measure using the information contained in the patent title and abstract. I also provide additional support for the use of the Jaffe covariance as a more suitable distance measure than the Jaffe angular separation when it comes to measuring spillovers. A set of axioms is developed and the distance measures are assessed on this basis. In an empirical analysis of the pairwise distances of 30 companies in the chemicals and pharmaceuticals & biotechnology sectors, the correlation of these measures is established at various levels of the IPC system and evaluated. I find that the within correlation at various levels of the IPC system is typically higher than the correlation between most distinct measures. Bottom-up distance measures are found to result in significantly different distance estimates compared to standard firm-based distance measures. The text based measure is found to have some correlation with the IPC based measures, particularly the Jaffe covariance and the patent-to-patent angular separation.*

Keywords: Patent data, patent statistics, technological distance, technological relatedness, technological similarity, angular separation, Jaffe distance, technology flows

## I. Introduction

Technological distance has been studied and used extensively in the field of economics and innovation management (Jaffe, 1986; McNamee, 2013; Stuart & Podolny, 1996). Technological distance (and its opposite technological proximity) is one core aspect of technological relatedness and is used in at least three contexts. First, it can refer to the distance of *firms* (or other entities such as countries or industries) in terms of their technological focus or profile. Unsurprisingly, patent data, and in particular classifications and citations have been used to measure a company's technological profile and its distance to other companies. A seminal contribution in this field is the angular separation of firms' technological profiles introduced by Jaffe (1986), which is still one of the most widely used measures in applied research on this topic. Second, technological distance can be established between individual *patents* (Natterer, forthcoming). This distance is of relevance, inter alia, for patent examiners and lawyers to facilitate prior art searches and for researchers looking for a measure of technological coherence of entities such as firms. Third, it can refer to the distance of technological *fields* (e.g. between "Pharmaceutics" and "Organic Chemistry"), i.e. the extent to which two fields build on the same knowledge bases (Breschi, Lissoni & Malerba, 2003; Teece, Rumelt, Dosi & Winter, 1994).

The focus of this study is the measurement of the technological distance between different firms. Hence, when I speak of technological distance, I refer to the distance between the technological *portfolios* of distinct companies, unless otherwise stated. Measures of technological distance feature in a wide range of applied research areas, including the study of spillovers (Bloom, Schankerman & van Reenen, 2013; Harhoff, 2000; Jaffe, 1986, 1989), mergers and acquisitions (Ahuja & Katila, 2001; Bena & Li, in press; Cloudt, Hagedoorn & van Kranenburg, 2006; Hussinger, 2010), alliances (Gilsing, Nooteboom, Vanhaverbeke, Duysters & van den Oord, 2008; Mowery, Oxley & Silverman, 1998; Nooteboom, Vanhaverbeke, Duysters, Gilsing & van den Oord, 2007; Rosenkopf & Almeida, 2003) and co-movements in stock returns (Fung, 2003). I will focus on a wide range of existing and newly developed measures based on the IPC (International Patent Classification) system and one measure based on textual patent data, using the information contained in the title and abstract.

As will be argued below, the distance between technology *fields* contains information that is relevant for the assessment of the distance between companies. Hence, some of the measures discussed below will incorporate this information. The distance between individual *patents* will also be used and aggregated to measure the distance between firms. It will be shown that aggregating the pairwise distances between the individual patents of distinct companies provides a measure that has stronger microeconomic foundations for capturing spillovers than most existing distance measures.

This paper builds on and contributes to the research that investigates the use and interpretation of patent data and statistics in applied research in the fields of economics and management (Griliches, 1990). More specifically, it contributes to the research that develops and contrasts different technological distance measures, an area that has gained significant momentum recently (Bar & Leiponen, 2012; Benner & Waldfogel, 2008; Bloom et al., 2013; McNamee, 2013; Nemet & Johnson, 2012). Technological distance is not consistently used in applied research, owing in particular to the imprecise definition of the term, meaning different things to different researchers, and the lack of distinction between cause, definition ("position in knowledge space") and effect ("potential for technology flows or transactions"). Technological distance is inextricably bound to concepts such as R&D spillovers, learning

and absorptive capacity and it is thus necessary to discuss the relationship between distance and these concepts. Technological distance becomes an empty concept if it is defined and assessed without consideration to the specific research question in which it is used. Hence, the concept of technological distance has to be sketched wide enough so as to be useful in management and economic research.

The study is organized as follows: In section II. I define and discuss the concept of technological distance and where it is applied in economic and management research. Section III. outlines the various ways to describe a company's technology profile, with a particular focus on patent data. Section IV. presents existing and new measures of technological distance. In section V. I propose a set of criteria that should be used to assess distance measures. Section VI. deals with the data used in the empirical analysis and presents descriptive statistics. Section VII. reports and discusses the empirical properties of distance measures, using patent data for 30 companies in the pharmaceuticals & biotechnology and chemicals industries. Section VIII. summarizes the key findings of this research.

## **II. Technology distance between companies: Theoretical foundations**

Conceptually, technological proximity and its opposite proximity are concerned with the overlap of knowledge base (methods of search, sources of knowledge, area of application, etc.) between actors, i.e. the more overlap there is between firms the less distance there is between them. More formally, companies can be described as a series of vectors in a multidimensional technology or knowledge space (Benner & Waldfoegel, 2008; Jaffe, 1986; Olsson & Frey, 2002). These vectors constitute the position of companies like the longitude and latitude constitute the position of cities. In addition, the constituent parts of a company's technological knowledge (e.g. its scientists, patents or technological clusters) can be positioned in knowledge space and be set in relation to another company's constituent parts. By aggregating the distances of the constituent parts one can also establish the distance between firms. Hence, I define technological distance between companies as the *length of technological space between two companies or its constituent parts*.<sup>1</sup>

To measure this length, the *dimensions* of the space have to be defined and companies have to be *positioned* in technology space. Technology or knowledge space is much more complex than geographic space, first because it has more dimensions and second because the dimensions are typically related. Latitude, longitude and altitude are orthogonal and span the three-dimensional space. In contrast, technology dimension such as Physics (IPC section G) and Electricity (IPC section H) are related. What typical distance measures such as the Euclidean distance and the angular separation assume is that the dimensions in technology space are unrelated. Hence, only "overlap" in the same dimension of knowledge space results in an increase of overall proximity. I will develop a measure which accounts for the relatedness of technology fields.

---

<sup>1</sup> In analogy to the definition of distance in the Oxford dictionary as "the length of space between two points" (<http://www.oxforddictionaries.com/definition/english/distance>, last accessed on 26.02.2014)

### *Use of distance measures in applied research*

Distance measures such as the angular separation or the correlation of revealed technological advantage are used, inter alia, to measure (potential) spillovers (Griliches, 1992; Jaffe, 1986; Mohnen, 1996) and in the context of M&A transactions (Bena & Li, in press; Hussinger, 2010) and co-operations and alliances (Gilsing, Nooteboom, Haverbeke, Duysters & van den Oord, 2008; Mowery, Oxley & Silverman, 1998; Stuart, 1998). While technological distance can be considered a distinct concept (i.e. an abstract idea) in management and economics research, it is typically used as an “auxiliary” concept for other important economic and management concepts, e.g. as a proxy for the potential spillovers between firms. Given that distance measures are inextricably related to these applied research topics, I give an overview of how distance is used and the potential weaknesses of using distances in a specific context. As will be argued below, technological distance measures should ultimately be assessed based on their suitability for the applied topic analyzed.

One area where technological distance is used frequently is the measurement of spillovers. While there are many different types and definitions of spillovers, I follow Griliches (1992) in defining technological or knowledge spillovers as non-pecuniary externalities that accrue from one entity (such as a firm) to another and which are due to the non-rival nature and only partial excludability or appropriability of a technology. These spillovers can occur, inter alia, via informal communication, flow of inventors or the publication of findings in journals.

Measuring these spillovers can broadly be split between flows based measures and distance based measures (Cincera, 2005; Griliches, 1979, 1992; Jaffe, 1986; Mohnen, 1996; Scherer, 1982). With regards to the latter, it is a standard procedure to calculate spillovers from a pool or cluster  $X$  of firms to a focal company  $A$  as the product of R&D conducted by company  $i \in X$  and the angular separation of the technology profiles of company  $A$  and  $i$ , summed over all firms  $i$  in the pool (Bloom et al., 2013; Griliches, 1979; Harhoff, 2000; Jaffe, 1986).<sup>2</sup> Absent the possibility to measure actual spillovers, distance measures (in combination with R&D expenditures) serve as a proxy for flows, building on the idea that technological distance affects the *potential* for such spillovers. Using such a measure for spillovers is based on the strong assumption that spillovers are higher between companies that are close in technology space. Clearly, arguments can be made in favor of this assumption, namely that closeness in technology space implies a higher relevance of the technology for a focal company as well as a higher absorptive capacity due to the similarity in the search procedures and sources of knowledge (Cohen & Levinthal, 1989). While distance and spillovers are likely to be related, the non-linear positive relationship assumed in most spillover measures is questionable. It is certainly worthwhile to further empirically investigate the relationship between distance and spillovers as well as to provide sound microeconomic foundations for using distance measures when measuring spillovers. Writing about distance in the context of alliances, Rosenkopf and Almeida (2003) write that “[e]mpirical studies of patent data suggest a relationship between technological similarity and knowledge flows *without testing directly for this relationship*” (p. 752, emphasis added). With regards to spillovers in particular, Harhoff (2000) notes that while the theoretical literature has made significant advances in understanding spillovers, the empirical literature has not kept pace. In their review of the spillover literature, Cincera and van Pottensberghe de la Potterie (2011) also lament the lack of micro level analysis of spillovers.

---

<sup>2</sup> The original suggestion to use distance measures to capture spillovers was made by Griliches (1979)

While the use of distance measures in capturing spillovers is not undisputed, the alternatives may not perform better. Patent citations, for example, also suffer from a series of problems if used to measure spillovers (Alcacer et al., 2009). Many citations are inserted by patent examiners or lawyers without the inventor being aware of the cited technologies (Jaffe et al., 2000). Patent examiners may cite patents they are familiar with rather than those where knowledge originates. The fact that another firm's patents are cited does not mean that a positive externality arises as the citing firm may have to pay licensing fees. In addition, patent citations have many of the same patent specific disadvantages as distance based measures.

Rather than as a measure for a certain concept, distance measures have also been used as a metric in research on M&A and alliances and then interpreted with the help of concepts such as learning opportunities and absorptive capacity. On the one hand, proximity has a positive impact on innovation output due to absorptive capacity (Cohen & Levinthal, 1990; Kogut & Zander, 1992; Lane & Lubatkin, 1998): the more similar the two firms' knowledge bases, the easier it is to "recognize the value of new, external knowledge, assimilate it and apply it to commercial ends" (Cohen & Levinthal, 1990, p.128). On the other hand, too much proximity constrains the acquirer's opportunities for learning (Sapienza, Parhankangas & Autio, 2004) or novelty gain (Nooteboom et al., 2007). These two opposing forces have been used to explain an inverse U-shaped relationship between distance and innovation output post M&A transactions that some authors have found (Ahuja & Katila, 2001).

### **III. Technological position of companies**

There are several ways to measure empirically the dimensions and positions of firms in technology space. I broadly distinguish patent based measures, which I will look at more closely below, and non-patent based measures. With regards to the latter, the position in knowledge space can be determined by inventor or scientist characteristics (Adams, 1990; Farjoun, 1994). Another possibility is to analyze the R&D profile. For example, Goto and Suzuki (1989) measures technological distance between 50 sectors based on the spending of R&D into 30 product areas. Sapienza et al. (2004) use questionnaires asking Finnish CEOs of spin-offs about their assessment of the technological distance to their previous parent company (asking questions such as: "The technology developed within the spin-off firm is based upon the technological strengths of the parent firm." (p. 818)).

#### *Patents and the hierarchical IPC system*

The International Patent Classification (IPC) is a hierarchical system consisting of over 70,000 different codes. The primary purpose of the IPC system is "the establishment of an effective search tool for the retrieval of patent documents by intellectual property offices and other users, in order to establish the novelty and evaluate the inventive step or non-obviousness (including the assessment of technical advance and useful results or utility) of technical disclosures in patent applications" (p.1, Guide to the IPC (2013)).<sup>3</sup>

The first letter of the code refers to the section, of which there are eight in total (e.g. E - Electricity). The class level is determined by the first three digits of the code (e.g. H01 - Basic Electric Elements) and the first four digits define the subclass level (e.g. H01S - Devices using stimulated emission). This is followed by an even more fine-grained group and

---

<sup>3</sup> [http://www.wipo.int/export/sites/www/classifications/ipc/en/guide/guide\\_ipc.pdf](http://www.wipo.int/export/sites/www/classifications/ipc/en/guide/guide_ipc.pdf) (18.11.2013)

subgroup level. Patents are assigned to at least one classification by the patent examiner but usually a patent is assigned to more than one patent class (Fleming & Sorenson, 2001). The classes and subclasses are further grouped into 35 technology areas according to the ISI-OST-INP classification established by the Fraunhofer Gesellschaft-ISI, Observatoire de Sciences et des Techniques and the Institut National de la Propriété Nationale (Schmoch, 2008). Patents are typically assigned multiple classifications and these classifications may be in distinct subclasses, classes, areas or sections.

In analogy to the Linnaean taxonomy, the IPC system is hierarchical (also called “IS-A” relationship) in that lower hierarchical levels “inherit” the properties/contents of their respective higher hierarchical level (McNamee, 2013). For example H01L and H01G are subclasses that both contain the technological characteristics of class H01. In choosing the appropriate dimension to use for patent measures the following tradeoff has to be made: by using a level further down the hierarchy (e.g. the subclass level), there is more information higher up the hierarchy that is discarded; by using a level further up the hierarchy (e.g. section level), there is more information further down the hierarchy that is discarded. This results from the fact that standard distance measures look at the “overlap” at one particular level and do not take into account the information at other levels. For example, determining distance at the subclass level, H01G will be equally distant from H01L as A47C, although H01L and H01G share the same class H01. Likewise, determining distance at the section level A61F will be equally distant from A61G as A47C although A61F and A61G share the same class A61 while A61F and A47C do not.

So as to use the IPC hierarchy more holistically, I propose a distance measure that accounts for the relatedness of the IPC fields within one level. The fact that H01L and A47C are more distant than H01L and H01G can be determined, inter alia, by the co-occurrence of the classes within patent documents: e.g. H01L and H01G are more likely to appear as patent classifications on the same patent than H01L and A47C. Clearly, one needs to account for the size of the patent classes, which can be achieved for example through the use of the angular separation of co-classification of fields. In contrast to the structural relatedness established by the IPC system, this relatedness is empirical.

#### *Advantages and disadvantages of patent data*

Distance measures based on patent data in general and the IPC system (or the USPTO classifications) in particular are among the most widely used distance measures, primarily due to their ease of calculation and the ready availability of the data covering a very long time span. In his 1990 survey on patent statistics, Griliches highlighted the usefulness of patent data to position firms in technology space. The IPC classifications of a company's patent portfolio provide a very detailed look into the technologies developed and used by a company. Classifications are technology based rather than product market based (Jaffe, 1986).<sup>4</sup> The patent requirements of novelty and inventive step imply that a company has advanced competencies in the technological fields in which a patent is classified. Similar to publications being artifacts of scientific progress, patents represent the development of technologies (Engelsman & van Raan, 1993). The costs associated with patents imply a

---

<sup>4</sup> The degree to which the hierarchical systems are truly about technologies rather than product markets differs between the systems (e.g. the USPTO system is considered more product market based than the IPC system)

certain expectation regarding marketability (Griliches, 1990). It has also been shown that patents are closely related with new product introduction and innovation counts (Comanor & Scherer, 1969; Hagedoorn & Cloudt, 2003). It is important to stress that we use patent statistics here as a measure for technological competencies rather than as a proxy for innovation performance (e.g. through simple patent counts), the latter having frequently been criticized (Pavitt, 1988).

Before turning to the determination of a company's technological profile, it is necessary to point to the problems that arise from the use of patent data (Griliches, 1990). A few problems relevant to this research topic are:

- a) No technological profile can be established for a company that does not patent. Hence, the service sector is not amenable to this type of analysis.
- b) The propensity to patent varies significantly between technologies. Hence, technologies with a lower propensity to patent are underrepresented (Jaffe, 1989).
- c) Patent values are highly skewed, with most patents being of little value and few patents being of very high value (Harhoff, Scherer & Vopel, 2003).
- d) The technology profile of companies and the resulting distance measure depend on the level in the hierarchy (i.e. class, subclass, etc.) that is chosen for the comparison and it is frequently not clear which level is most appropriate in a given context.
- e) The classification of patents is subject to errors on the part of patent examiners and there are potential biases, in that patent examiners may classify patents in fields they are familiar with.

Despite these problems, patent data is arguably the best source of information for the measurement of technological distance, given its fine-grained split of technological categories and its ready availability.

#### *Considerations in establishing the IPC profile*

Going forward, dimensions of the technology space are defined in terms of IPC technology fields (e.g. using IPC classes), and the coordinates are determined in the form of classifications of a company's patents into these fields.

What is required by most measures as an input is a matrix  $F$  of dimension  $N \times T$ , where  $N$  denotes the number of companies and  $T$  the number of IPC fields. Element  $f_{nt}$  of this matrix indicates the share of technology  $t$  in company  $n$ . Each row  $f_n$  of this matrix is established by condensing the information of  $M_n$  patents owned by company  $n$ . Let  $K_n$  of dimension  $M_n \times T$  constitute the matrix of all patents owned by firm  $n$ , whose rows  $k_{nm}$  refer to the technology profile of one patent. For each patent, there is at least one IPC category, but mostly several categories, some of which appear several times within the same patent document.

In order to establish the matrix  $F$ , the following decisions have to be made:

- a) The level of the hierarchy to consider (e.g. IPC class or IPC subclass)
- b) The patents which are to be considered (e.g. last 5 years or last 20 years)
- c) For each patent, the weighting of each category when a patent is classified into one category more than once

Arguably, decision a) is the most important one. When choosing between different levels of the IPC system to base the distance measure on, one is facing the following trade-off, that is particularly pronounced when using basic distance measure such as the angular separation which assumes that the proximity between different fields within a certain level is 0 (McNamee, 2013). By using a very fine resolution (say subclass level), one is "ignoring" a larger number of higher hierarchy levels. In a geographic analogy, consider an attempt to establish the distance between citizens within Europe. If one were to look at the city level, then citizens would only be close if they live in the same city. If they do not live in the same city, then it does not matter whether or not they live in the same province or country as all cities are presumed to be equidistant from each other. By using a very coarse resolution (say section level), one is "ignoring" the levels below. Hence, one would look at whether citizens live in the same country, but no insight would be gained as to how close they live within the country (Benner and Waldfogel, 2008).

Decision b) is also important in that two opposing forces have to be reconciled. On the one hand, companies may make a technological shift over the years so that patents filed ten years ago may not represent the current technological focus. It has been argued that a patenting activity in the past five years is suitable to determine a company's technological profile (Henderson & Cockburn, 1996; Stuart & Podolny, 1996). On the other hand, the sample size increases as more years are taken into account and this increases the statistical accuracy of the distance measures as found by Benner and Waldfogel (2008). Following the advice in Benner and Waldfogel (2008), I include patents over a 10 year time period in the empirical analysis.

Regarding point c), there are various ways to weigh the different IPC categories if one or more categories appear more than once on the patent document. Consider the case where technology  $u$  appears once and  $v$  appears 7 times. The simplest method is to use an indicator equal to 1 if an IPC category appears on the patent document and 0 otherwise (e.g. Benner & Waldfogel, 2008). Another method would be to count the number of times each IPC category appears, so in the above case,  $v$  would count seven times as much as  $u$ . I have opted for the second approach in the empirical analysis.

Some studies have used only the primary classification of the USPTO classification system, the so-called Original Classification ("OR"), to characterize a patent (e.g. Rosenkopf & Almeida, 2003) while others have used all classifications (Benner & Waldfogel, 2008). McNamee (2013) notes that using all classifications is relatively uncommon among researchers. Both from a theoretical as well as empirical perspective, it is advisable to use all classifications rather than only the "OR".

#### **IV. Measures for technological distance**

Technology distance measures have been used to analyze the distance of portfolios of patents and the distance of technology fields. The focus of this study is to assess the distance of patent portfolios of distinct companies. As the relatedness of technology fields is included in some of the distance measures used, I will first discuss the two main measures used to establish the relatedness of technology fields.

### *Relatedness of technology fields*

Assessing the relatedness of technology fields can broadly be split between the *ex ante* and *ex post* approach (Cantwell & Noonan, 2001). In the *ex post* approach, by looking at the presence of firms in several technology classes, Teece et al. (1994) and Engelsman and van Raan (1991) argue that firms' technological diversification is not random. The underlying assumption is that related activities will be more frequently combined within firms, and that those which combine related activities are more likely to survive (the "survivor principle").

In the *ex ante* approach, the relatedness is an intrinsic feature of the technology and has been measured by the co-occurrence of classifications in patent documents (Breschi et al., 2003; Nesta & Saviotti, 2005; Schmidt-Ehmcke & Zloczynski, 2008). If certain IPC categories appear together frequently in patent documents, they are presumed to be related. One advantage of the *ex ante* approach relative to the *ex post* approach is that using each patent as a data point rather than each company results in a number of observations that is larger by a multiple equal to the average number of patents per firm. Especially when comparing a small number of companies and when using a finer resolution of the IPC system (e.g. the subclass level), this advantage is very compelling.

The *ex ante* approach can be implemented either by measuring the deviation of the actual co-occurrences from its expectation under the random hypothesis (Nesta & Saviotti, 2005, Teece et al., 1994), or via the angular separation of the co-occurrence vectors (Breschi et al., 2003; Engelsman & van Raan, 1994). I present both *ex ante* approaches below.

Following Nesta and Saviotti (2005), the steps to calculate the *ex ante* relatedness of technology fields are as follows. Let there be  $T$  technology fields and a sample of  $K$  patents. Let the total number of patents assigned to field  $u \in T$  be denoted  $C_u$ . Let  $J_{uv}$  denote the count of the number of patents that are assigned to both technology fields  $u$  and  $v$  at the same time. If there are a lot of patents that are classified in either technology  $u$  or  $v$ , then co-occurrences of  $u$  and  $v$  are more likely (by chance). Hence, Teece et al. (1994) suggested comparing  $J_{uv}$  with its expectations. Under the assumption of joint random occurrences and a hypergeometric distribution we obtain the following first two moments:

$$\mu_{uv} = \frac{C_u C_v}{K}$$

$$\sigma_{uv}^2 = \mu_{uv} \left( \frac{K - C_u}{K} \right) \left( \frac{K - C_v}{K - 1} \right)$$

The relatedness between technology fields  $u$  and  $v$  is then calculated as:

$$r_{uv} = \frac{J_{uv} - \mu_{uv}}{\sigma_{uv}}$$

Nesta and Saviotti (2005) argue that this measure should be interpreted as complementarity of technologies  $u$  and  $v$ , as it measures the degree to which both technologies in combination are required.

In contrast, Breschi et al. (2003) calculate the ex ante relatedness of technology fields  $u$  and  $v$ ,  $\Omega_{uv}^{EA}$ , as the angular separation of the co-occurrence vectors:

$$\Omega_{uv}^{EA} = \frac{\sum_{t=1}^T J_{ut} J_{vt}}{\sqrt{\sum_{t=1}^T J_{ut}^2 \sum_{t=1}^T J_{vt}^2}}$$

### *Distance of patent portfolios*

In the following, I will focus – with one exception – on measures that build on the IPC system. The analysis will comprise the angular separation measure introduced by Jaffe (1986), the the Euclidean distance as used by Rosenkopf and Almeida (2003) and the Jaffe covariance (Bloom et al., 2013). Other common measures comprise the Pearson correlation coefficient (Benner & Walfogel, 2008) and the correlation of revealed technological advantage (CRTA) (Nooteboom et al. 2007). Two more recent distance measures are the Min-complement measure developed by Bar and Leiponen (2012) and the Mahalanobis distance developed by Bloom et al. (2013). Departing from these six existing measures, I introduce the weighted angular separation and the aggregated patent-to-patent angular separation. In addition, I present a bottom-up text based similarity measure based on a vector space model using the information contained in the patent title and abstract following Natterer (forthcoming).

To simplify the comparison between the measures, I have decided to adjust the two measures that are an increasing function of distance, namely the Euclidean distance and the Min-Complement by subtracting the original measure from one. As a result, all measures as presented below are decreasing functions of distance, and thus really represent "proximity" measures rather than distance measures.<sup>5</sup>

**Angular separation:** Introduced to the literature by Jaffe (1986), this is the measure most widely used to measure technological distance. It is also called the cosine distance or uncentered correlation index. It measures to what degree the vectors point in the same direction, controlling for the length of the vector. Only patenting in the same category will increase this measure. The angular separation between firm  $i$  and firm  $j$  is calculated as follows:

$$AN_{ij} = \frac{f_i f_j'}{\sqrt{(f_i f_i') * (f_j f_j')}}}$$

Where  $f_i$  is the  $i^{th}$  row of  $F$ , representing firm  $i$ 's technology profile.

**Euclidean distance (rescaled):** Rosenkopf and Almeida (2003) proposed the use of the Euclidean distance to compare the firms' technology vectors. It compares for each

---

<sup>5</sup> Some authors also use the term technological similarity

technology category the squared difference of the share that technology category has in firm  $i$  and the share that technology class has in firm  $j$ . I subtract the Euclidean distance from one so as to obtain a measure which is decreasing in distance:

$$EU_{ij} = 1 - \sqrt{\sum_{t=1}^T (f_{it} - f_{jt})^2}$$

**Jaffe covariance:** This measure is similar to the angular separation with the exception that no adjustment is made to the length of the vectors (Bloom et al., 2013).

$$JC_{ij} = f_i f_j'$$

The normalization in the angular separation is not a simple rescaling that preserves the order of the pairwise distances, but leads to conceptually different results. For example, the angular separation identifies two fully technologically diversified and two firms focused within the same patent class, respectively, as perfectly close. In contrast, the Jaffe covariance assigns two firms focused on the same technology class a higher level of similarity than two fully diversified firms.

**Correlation:** The Pearson correlation coefficient between the technology profiles of firm  $i$  and  $j$  is the ratio of the covariance between the two vectors divided by the product of the two standard deviations:

$$CO_{ij} = \frac{Cov(f_i, f_j)}{SD(f_i) * SD(f_j)}$$

The correlation coefficient ranges between -1 and 1, where 1 denotes a high degree of proximity and -1 a large distance.

**Correlation of revealed technology advantage (CRTA):** Following Nooteboom et al. (2007) and Gilsing et al. (2008), I calculate the RTA for firm  $i$  as the number of patents firm  $i$  has in technology category  $u$  relative to all patents in category  $u$  assigned to  $N$  firms, all divided by the total number of patents owned by the firm relative to all patent filings (by all  $N$  firms in all  $T$  categories). Values larger than 1 indicate that a company has a revealed technology advantage in that category. I then repeat this for all categories, resulting in the revealed technology advantage vector for firm  $i$ . Let  $g_{iu}$  denote the number of patents owned

by firm  $i$  that are classified in technology  $u$ , then the revealed technological advantage is defined as:<sup>6</sup>

$$RTA_{iu} = \frac{g_{iu} / \sum_{n=1}^N g_{nu}}{\sum_{t=1}^T g_{it} / \sum_{n=1}^N \sum_{t=1}^T g_{nt}}$$

$$RTA_i = (RTA_{i1} \dots \dots \dots RTA_{iT})$$

The correlation of revealed technological advantage (CRTA) is the pairwise correlation of the RTA vectors for the companies:

$$CR_{ij} = Corr(RTA_i, RTA_j)$$

**Min-Complement:** Bar and Leiponen (2012) suggested the use of the Min-Complement, which only takes into account the overlap in relevant technology fields. It measures the share of overlapping patent categories of firm  $i$  and firm  $j$  as:

$$MC_{ij} = \sum_{t=1}^T \min(f_{it}; f_{jt})$$

**Mahalanobis distance:** Bloom et al. (2013) used this measure in their analysis of spillovers. This measure is more fine-grained as not only patenting in the same technology field increases the relatedness, but also patenting in fields that are related. The relatedness of technology areas is established by the presence of firms in multiple technologies (“survivorship method”).

Starting from the matrix  $F$  ( $N \times T$ ) and its rows denoted  $f_n$ , the matrix  $\tilde{F}$  ( $N \times T$ ) is calculated as follows:

$$\tilde{F} = \begin{pmatrix} f_1 / \sqrt{(f_1 f_1')} \\ \dots \\ f_N / \sqrt{(f_N f_N')} \end{pmatrix}$$

Note that  $\tilde{F} \tilde{F}'$  is equal to the angular separation described above (Jaffe, 1986). Let  $f_{(t)}$  denote the  $t^{th}$  column of  $F$  and define  $\tilde{X}$  ( $N \times T$ ) as follows:

---

<sup>6</sup> Note that in the empirical analysis I have used fractional counts, i.e. a patent with two classifications is given a count of 0.5 when determining  $g_{iu}$

$$\tilde{X} = (f_{(:,1)}/\sqrt{(f_{(:,1)}f'_{(:,1)})} \quad \dots \quad f_{(:,T)}/\sqrt{(f_{(:,T)}f'_{(:,T)})})$$

Let us define  $\Omega = \tilde{X}'\tilde{X}$ , which is the angular separation between technology fields (rather than firms). Each element of  $\Omega$  ranges between 1 (for fields that always co-occur within firms) and 0 (for fields that never co-occur within firms).

The symmetric Mahalanobis distance matrix of dimension  $N \times N$  is defined as:

$$MA = \tilde{F}'\Omega\tilde{F}$$

Element  $[i, j]$  of this matrix denotes the distance between firms  $i$  and  $j$ .

**Weighted Angular Separation:** Similar to the Mahalanobis distance, this measure incorporates the relatedness of technological fields. I propose to use the co-occurrence of patent classification in patent documents as a measure of relatedness. Specifically, I have used the angular separation between the co-occurrences (Breschi et al., 2003).

The relatedness of technology fields is considered to be a general feature of the technologies and independent of the specific pair of firms for which a distance measure is to be calculated. This means that when comparing firms  $i$  and  $j$  or  $g$  and  $m$ , the technology field relatedness is always the same and is calculated based on the entire set of patents in the sample. Let  $\Omega^{EA}$  denote the  $T \times T$  ex ante technology field relatedness matrix (Breschi et al., 2003), then the weighted angular separation is defined as:

$$AW_{ij} = \frac{f_i\Omega^{EA}f_j'}{\sqrt{(f_i\Omega^{EA}f_i')*(f_j\Omega^{EA}f_j')}}}$$

#### *Bottom-up distance measures*

The above distance measures compare the aggregate technological profiles of companies. An alternative proposed in this paper are bottom-up measures, which look at the distance between individual patents in the companies' portfolios and aggregate this information for all combinations of pairs. I propose two measures below which establish the distance between the parts (here, the individual patents) that make up the technology profile of a company. One measure is based on the IPC classification and the other measure is based on the textual similarity of patent documents established through an algorithm by Natterer (forthcoming).

It is important to note that the top down approach is not identical to the bottom up approach. Consider two companies' technology profiles along 30 technology fields. Let us assume for now that the technology fields are unrelated, i.e. orthogonal. As a first example, assume that both companies  $i$  and  $j$  have 3.33% of their patents in each field (no multiple classifications are allowed for simplification). As a second example, assume that both companies have all

their patents in one technology field. Measures such as the angular separation and the Euclidean distance would consider both companies as perfectly close. One can argue, however, that in the second example, the patent-to-patent ties are likely to be stronger: In the first example, each patent of firm  $i$  is "close" to only 3.33% of firm  $j$ 's patents whereas in the second example each patent of firm  $i$  is close to all of firm  $j$ 's patents.

I argue that bottom-up distance measures are better suited than existing measures in the context of spillovers and learning. Microeconomic foundations for the proposed bottom-up patent-to-patent measure can be established by changing one component of the communication model of Bloom et al. (2013) that was used to establish microeconomic foundations for the Jaffe covariance. Let firm  $i \in (1, J)$  have  $n_i$  scientists which are active in  $\tau \in (1, Y)$  technologies. Denote  $k_i^t$  the technology profile of the  $t^{th}$  scientist at firm  $i$ . Each scientist can potentially be active in several technology fields. One can conceive of one scientist as standing behind one patent at the firm he works for and the technologies he/she is active in is designated by the technology classifications of the patent. Rather than assuming that one unit of knowledge is transferred between overlapping scientists with probability  $\omega$  (as done by Bloom et al.), I assume that the (potential for a) spillover is *proportional* to the distance between scientists and weighted with a constant probability  $\omega$ . If scientists are closer in knowledge space, the expected spillover is larger than between distant scientists. Underlying this model is the assumption that the spillovers between scientists are non-overlapping. As a result, the spillover from firm  $j$  to firm  $i$  is established as follows:

$$SPILLOVER_{ij} = \omega * \sum_{t=1}^{n_i} \sum_{s=1}^{n_j} dist(k_i^t, k_j^s) = \omega * n_i * n_j \frac{\sum_{t=1}^{n_i} \sum_{s=1}^{n_j} dist(k_i^t, k_j^s)}{n_i * n_j}$$

where  $\frac{\sum_{t=1}^{n_i} \sum_{s=1}^{n_j} dist(k_i^t, k_j^s)}{n_i * n_j}$  is the average pairwise distance between the scientists (or the patents they stand for) of the two firms. Note that as in the case of the model of Bloom et al. (2013), the distance function is a *decreasing* function of distance, i.e. truly a proximity measure.

**Aggregated patent-to-patent angular separation:** For this measure I establish the profile of one particular patent of firm  $i$  ( $k_{ix}$ ) and one patent of firm  $j$  ( $k_{jy}$ ), compute the angular separation, repeat this for all possible combination of patents and then take the average. Let  $X$  denote the number of patents owned by firm  $i$  and  $Y$  denote the number of patents owned by firm  $j$ , the distance measure is calculated as follows:

$$PP_{ij} = \frac{1}{X} \frac{1}{Y} \sum_{x=1}^X \sum_{y=1}^Y \frac{k_{ix} k_{jy}'}{\sqrt{(k_{ix} k_{ix}') * (k_{jy} k_{jy}')}}$$

**Text based distance using vector space model:** In contrast to the measures above, this measure is based on the analysis of the words contained in the title and abstract section of the patent document. The similarity measure was developed by Natterer (forthcoming) to

assess the similarity of individual patents. It uses pre-processing methods of patent documents including error correction and language harmonization, generic and specific stop-word elimination and word-stemming methods. After applying local and global word weighting algorithms, each patent is characterized by a vector whose elements refer to the weights of the words appearing in the patent document. Finally, distance between patents is determined using the cosine index (which is identical to the angular separation). The algorithm has been trained by maximizing the hit rate of finding those patents in a pool of patents which were cited as prior art by a focal patent. Similar to the IPC based patent-to-patent measure I aggregate the measure for all patents in the companies' portfolios.

## V. Assessing distance measures

Extending the criteria proposed by Bloom et al. (2013) to assess distance and spillover measures, the following are criteria for distance measures which I argue are desirable:

- (1) **Microeconomic foundations:** The distance measure should have sound microeconomic foundations for the research topic analyzed. The only distance measures which have been given microeconomic foundations to date are the Jaffe covariance (Bloom et al., 2013) and the two bottom-up measures proposed in this research. It is important to note that microeconomic foundations are typically with respect to the ultimate area of application, such as spillovers.
- (2) **Own distance:** The distance of one company to itself is the same for all companies and the value constitutes the highest level of proximity possible. Hence, no pair of distinct companies can be closer to each other than a company is to itself. While this is a very natural criterion, some measures do not satisfy this property. For example, the Mahalanobis distance (Bloom et al., 2013) does not satisfy this property. Also, the bottom-up measures of distance do not satisfy it as when relating a company's individual patents to each other, they are not necessarily closer to each other than the patents of different companies. Note that this depends on aspects of diversification and coherence of the patent portfolio of a company.
- (3) **Symmetry:** The distance between two companies should be symmetric. This is a natural assumption to make and all measures proposed here satisfy this criterion.
- (4) **Independence of scaling:** Consider two companies which are identical in terms of the fields in which they patent, but company  $i$  has twice as many patents in every field as company  $j$ . Then a compelling argument can be made that the measure should rank them as identical (Bloom et al. 2013). As the measures presented above take the *shares* in each field as input, this criterion is generally satisfied.
- (5) **Robustness to the propensity to patent:** Benner and Waldfoegel (2010) have shown, using bootstrap methods, that subsamples of the entire patent portfolios result, on average, in biased estimates of the true distance. As a result a company which has a lower propensity to patent is, in expectation, likely to be more or less distant to a particular firm than a firm with a higher propensity to patent (but otherwise identical). This bias is a result of the measure itself, and better measures should have less of a bias. In related research, I have shown that only the Jaffe covariance and the bottom-up measures do not have a bias in small samples.

- (6) **Same field patenting:** The measure should be an increasing function of patenting in the same IPC field. The more firms patent in the same IPC fields, the more similar are their technology positions and thus the closer they are in technology space. Formally, holding constant the technology profile of firm  $j$  which is active in technology field  $\tau$ , firm  $i$  is closer to firm  $j$  the larger is the share of field  $\tau$  in firm  $i$ , *holding constant the shares in firm  $i$  of all other technologies that are also used by firm  $j$* , i.e. the higher share of  $\tau$  in firm  $i$  comes at the expense of fields which are not relevant for firm  $j$ . This criterion is generally satisfied by the proposed measures. While it is not directly applicable to the textual measure, the latter satisfies the criterion if we replace IPC fields with “words”.
- (7) **Account for technology field relatedness:** McNamee (2013) states that the most fundamental problem of existing distance measures is that the proximity between different subclasses or classes is assumed to be nil: If the distance measure is calculated at the subclass level, then IPC subclass A21B is assumed to be as unrelated to A21C as C01D, even though the first pair shares the same class. Hence, I propose to take into consideration in how far the relatedness between distinct technology fields at one particular level of the IPC system is taken into account. This should lead to a much more valid and fine grained relatedness measure. Only the weighted angular separation and the Mahalanobis distance take this into account. The textual based method does not account for the relatedness of the words in the patent document.
- (8) **Insensitivity to the level of aggregation:** Bloom et al. (2013) argue that the index should not be very sensitive to the level of aggregation. This is a valid criterion when it is unclear which level of aggregation is the most suitable one to use. What is of particular concern is the impact that the level of aggregation has in the between firm dimension, e.g. whether moving from area to subclass level changes the cardinal ranking of between firm distances. To assess this, I compute the correlation coefficient of the pairwise distance measures established at the area and class and subclass level. As will be shown, the correlation coefficient of specific measures established as the area, class and subclass level is typically quite high and mostly above 0.95. Only the CRTA performs poorly on this criterion.
- (9) **Independence of irrelevant patent fields:** If the technology profile of firm  $i$  and  $j$  is identical in all categories in which firm  $k$  has a positive share (and is thus relevant for firm  $k$ ), then firm  $i$  and  $j$  should be equidistant from firm  $k$  (Bar & Leiponen, 2012).

As argued by Bar and Leiponen (2012) this criterion is especially relevant when a large diversified company and a small focused company are compared to a focal company. If both have equal shares in the relevant categories, but the small company has one large share in an irrelevant category while the diversified company has a positive share in many other irrelevant categories, then both companies should be regarded as equidistant from the focal firm.

Bar and Leiponen (2012) show that the angular separation, the Euclidean distance and the Pearson correlation coefficient fail to satisfy this criterion, while the Min-Complement measure satisfies it. Another measure that satisfies the criterion is the Jaffe covariance, as the multiplication with the 0 of the irrelevant categories of the focal companies makes the measure independent of how the firms split their technologies in these categories.

(10) **Ease of calculation:** In applied research, the calculation of the distance measure frequently forms a small element of the entire research project. This is why many authors resort to easy to calculate measures such as the angular separation. I argue that more complicated distance measures are preferable to simple measures only to the extent that they are better in fulfilling the above criteria. Measures based on textual analysis score low here as the development and testing of algorithms (if not publicly available) as well as the requirement to obtain patent text is typically associated with significant preparation and computer time. Also, the aggregated weighted patent-to-patent angular separation is associated with significant computer power.<sup>7</sup>

Table 1 summarizes the axiomatic assessment of the distance measures. It is worth noting that the importance of the axioms differs, and they have to be assessed in light of the specific context in which the distance measures are used.

	(1) Micro-economic foundations	(2) Own distance	(3) Symmetry	(4) Independence to scaling	(5) Robustness to the propensity to patent	(6) Same field patenting	(7) Account for field relatedness	(8) Insensitivity to level of aggregation	(9) Independence of irrelevant fields	(10) Ease of calculation
Angular separation		✓	✓	✓		✓		✓		✓
Euclidean distance		✓	✓	✓		✓		✓		✓
Jaffe covariance	✓		✓	✓	✓	✓		✓	✓	✓
Pearson correlation		✓	✓	✓		✓		✓		✓
CRTA		✓	✓	✓		✓				✓
Min-complement		✓	✓	✓		✓		✓	✓	✓
Mahalanobis distance			✓	✓		✓	✓	✓		(✓)
Weighted angular separation		✓	✓	✓		✓	✓	✓		(✓)
Patent-to-patent angular separation	✓		✓	✓	✓	✓		✓		
Patent-to-patent textual distance	✓		✓	✓	✓	(✓)		n/a	n/a	

**Table 1:** Axiomatic assessment of distance measures. A tick signifies that the criterion is satisfied, a tick in parenthesis signifies that the criterion is partly satisfied while an empty cell signifies that the criterion is not satisfied.

## VI. Data and Descriptives

The data for the empirical analysis was obtained from the 2012 EU Industrial R&D Scoreboard, Derwent World Patent Index and PATSTAT. I selected European based companies active in the pharmaceutical & biotechnology (15 companies) and chemicals industries (15 companies) based on the 2012 EU Industrial R&D Scoreboard, which ranks the top 1,500 most R&D intensive firms worldwide. Industry membership is determined based on the 3 digit ICB code. I included all European firms in these two industries which

<sup>7</sup> For example, comparing two companies with 4,000 patents each in their portfolios requires the calculation of c.4 million distances, which are subsequently averaged.

were among the top 450 firms, and added more European firms randomly from these industries until I reached 15 companies per industry. The reason for choosing these two sectors is that they are technologically related to some degree, thereby not providing a very high level of distance. When calculating the distances between the 30 firms, I expect to obtain both small distances (e.g. between two chemicals firms) as well as higher distances between firms belonging to different industries. Using these 30 firms, I can calculate 435 ( $= n * [n - 1] / 2$ ) pairwise distances.

In a following step, I obtained the patent portfolios of these companies from Derwent World Patent Index using the patent assignee code.<sup>8</sup> This allows us to obtain the patents which are assigned to the parent company or any of its subsidiaries. For example, for BASF, I compiled a list of all patents under the code BADI-C. I restricted the attention to European patents (i.e. the patent number starting with EP) and ending with B1 (i.e. those which were eventually granted). I restricted the timeframe to 10 years, covering patents with priority date ranging from 1.1.1998 to 31.12.2007. Overall, 16,041 patents were included in the analysis. Based on these patent numbers, I obtained the IPC code as well as the English title and abstract from PATSTAT.

The following table provides a list of the companies included in the analysis.

Company	Derwent code	Sector	Patents
Air Liquide	AIRL	Chemicals	587
Akzo Nobel	ALKU	Chemicals	794
AstraZeneca	ASTR	Pharma / biotechnology	807
Actelion	n/a	Pharma / biotechnology	12
BASF	BADI	Chemicals	4,541
Boehringer Ingelheim	BOEH	Pharma / biotechnology	396
Borealis	BORA	Chemicals	246
Altana	BYKG	Chemicals	132
Chiesi	CHIE	Pharma / biotechnology	42
Elan	ELAN	Pharma / biotechnology	77
Bayer	FARB	Chemicals	773
Givaudan	GIVA	Chemicals	87
GlaxoSmithKline	GLAX	Pharma / biotechnology	647
Roche	HOFF	Pharma / biotechnology	1,407
Ipsen	n/a	Pharma / biotechnology	81
Kemira	KEMH	Chemicals	120
Lonza	LONZ	Chemicals	165
Lubrizol	LUBR	Chemicals	209
Merck	MERE	Pharma / biotechnology	802
Novo Nordisk	NOVO	Pharma / biotechnology	574
Novartis	NOVS	Pharma / biotechnology	959
NeuroSearch	NURO	Pharma / biotechnology	7
SGL Carbon	SIGE	Chemicals	65
Sanofi-Aventis	SNFI	Pharma / biotechnology	733
DSM	STAM	Chemicals	714
Syngenta	SYGN	Chemicals	304
Symrise	SYMR	Chemicals	78
Shire	n/a	Pharma / biotechnology	43
UCB	UNIO	Pharma / biotechnology	106
Wacker	WACK	Chemicals	533
			16,041

**Table 2:** Companies included in the analysis

<sup>8</sup> For the companies Shire, Ipsen and Actelion no Derwent assignee code is available and therefore the names of the companies were used to obtain the patents from Derwent World Patent Index.

In the sample, patents had on average 8.5 classifications. Patents in the sample were classified in 8 sections, 35 areas, 113 classes and 434 subclasses. Each patent is assigned to an average of 1.6 distinct sections (range from 1 to 5), 1.9 areas (range from 1 to 7), 1.9 classes (range from 1 to 9) and 2.7 subclasses (range from 1 to 13).

The text based similarity measure based on a vector space model using the information contained in the patent title and abstract was calculated using the algorithm described in section IV. I then aggregated these similarity data to the company level. Due to data availability, the textual based distances were established based on a subsample of 5,245 patents (out of the 16,041 patents in the sample). The company Actelion was dropped from the dataset as only one patent remained in the portfolio. All other IPC distance measures were calculated also on the subsample to check that the results mirror the findings of the full sample, which they broadly do. The smaller dataset was used only in the section which discusses the textual based measure.

## **VII. Empirical results**

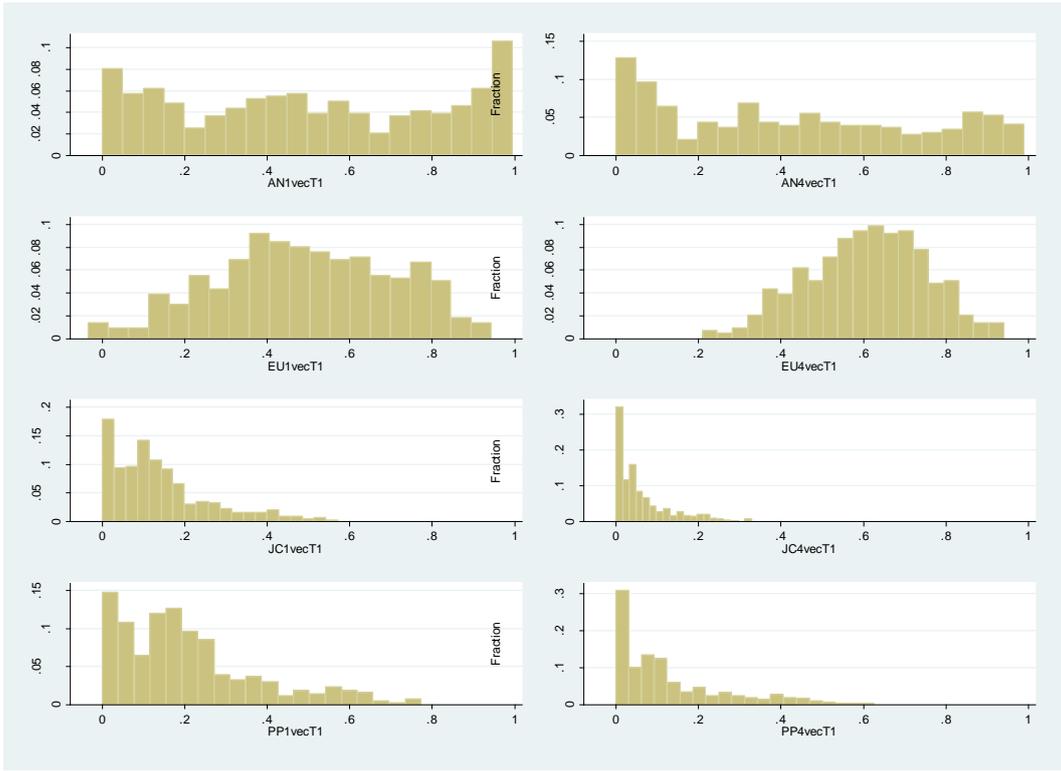
The empirical results are presented in four parts. I first present the results of the distribution of the distance measures. I then discuss the correlation between the measures established at various levels of the IPC hierarchy. This is followed by a presentation of the correlations between distinct measures. Finally, I present the findings pertaining to the textual based measure.

### *Distribution of distance measures*

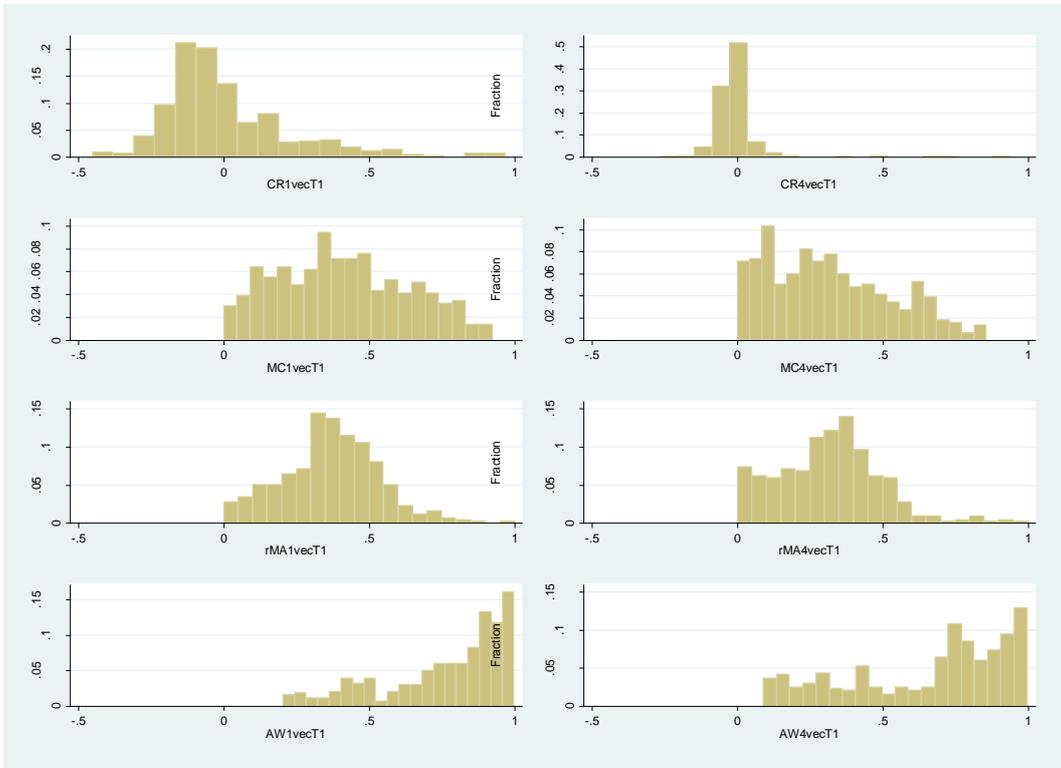
The histograms in Graphs 1 and 2 are established at the area level (graphs on the left) and the subclass level (graphs on the right). The distributions differ significantly between the measures: While the Euclidean distance, the CRTA, the Min-Complement and the Mahalanobis distance exhibit a bell shape distribution, the patent-to-patent angular separation and the Jaffe covariance exhibit distributions with a lot of mass at a very low level of proximity. The angular separation exhibits a relatively uniform distribution with spikes at the extremes. The Pearson correlation coefficient exhibits a profile which is almost identical to the angular separation and is thus not shown.

### *Correlation of measures*

Table 3 provides an overview of the correlation coefficient between the various measures established at the section, area, class and subclass level. The shading of the cells illustrate the level of correlation (correlations below 0.6 are not shaded). We begin by discussing the correlation of each measure as the hierarchical level changes.



**Graph 1:** Distribution of the angular separation (top), Euclidean distance (second from top), Jaffe covariance (second from bottom) and the patent-to-patent angular separation (bottom). The graphs on the left show the distributions at the area level and the graphs on the right show the distributions at the subclass level.



**Graph 2:** Distribution of the CRTA (top), Min-Complement (second from top), Mahalanobis distance (second from bottom) and the weighted angular separation (bottom). The graphs on the left show the distributions at the area level and the graphs on the right show the distributions at the subclass level.



### *Correlation at various levels of the IPC system*

The correlation of one measure at various level of the IPC system is illustrated by the 4x4 triangles below the diagonal. Several points stand out:

- a) The CRTA appears to be an outlier as it exhibits a much lower correlation when it is measured at different levels of the hierarchy. Hence, the below comments refer to the other eight measures.
- b) The distances established at the section level show a much lower level of correlation to the distances established at a finer level of aggregation. The coefficients between the section level and the area level range from 0.70 to 0.83. The correlation with distances established at the class and subclass levels is again lower (e.g. between 0.58 and 0.72)
- c) The correlation coefficient between the area, class and subclass levels are generally above 0.9 and typically around 0.95. The correlation between the class and subclass level is above 0.95 for all measures apart from the Euclidean distance (0.90) and the Mahalanobis distance (0.93).
- d) The correlation coefficient between the Jaffe covariance measures established at the area, class and subclass level is one of the highest observed (from 0.95 to 0.97). This illustrates that the issue raised by Bloom et al. (2013) that the Jaffe covariance is sensitive to the level of aggregation is not an issue when comparing many companies.

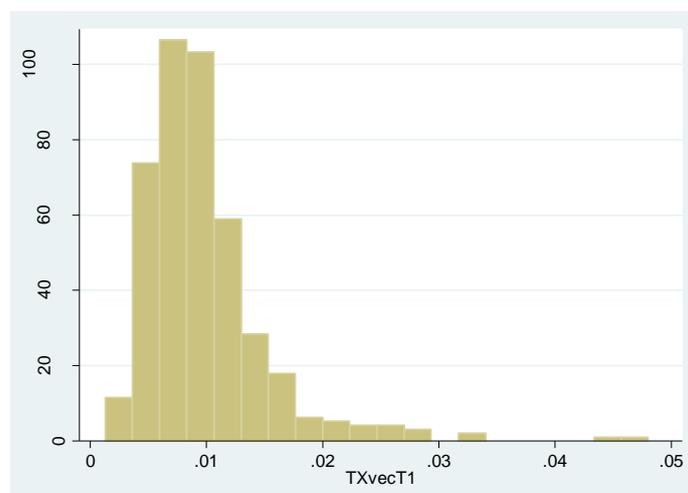
### *Correlation between distance measures*

As for the correlation between the measures calculated at one level of the classification, there are several observations to make:

- e) The correlation between measures established at one level of the IPC hierarchy is typically lower than the correlation of one measures established as distinct levels of the hierarchy.
- f) The CRTA exhibits the lowest correlation with the other measures, with correlation coefficient generally below 0.6. The Mahalanobis distance also has a lower correlation coefficient with the other measures, albeit somewhat higher than for the CRTA.
- g) The correlation between the angular separation and the correlation coefficient is very high, indicating that these two measures can be used interchangeably. The min-complement is also closely correlated (coefficient of 0.96) with these two measures.
- h) The weighted angular separation has a correlation coefficient of around 0.9 with the angular separation, which is lower than the correlation coefficient of one measures established at various levels of the hierarchy.
- i) The Jaffe covariance and the patent-to-patent angular separation are also very highly correlated with coefficient 0.99 at the class and subclass level.
- j) The correlation between the patent-to-patent angular separation / Jaffe covariance and the angular separation ranges from 0.83 to 0.88.

### *Text based distance measure*

The distribution of the text based measures is bell shaped and right skewed as illustrated in graph 3:



**Graph 3:** Distribution of the text based measure

With regards to the correlation of the text based measure with the IPC based measures, several findings stand out:

- a) The correlation between the text based measure and the IPC based measures established at the area, class or subclass level ranges from around 0.3 to 0.7.
- b) The Mahalanobis distance and the Euclidean distance have the lowest correlation with the text based measure.
- c) The Jaffe covariance and the patent-to-patent angular separation have the highest correlation with the text based measure with a coefficient ranging between 0.62 and 0.71. This coefficient is lower than that between the two measures and the angular separation or the min-complement.
- d) The correlation between the text based measure and the IPC based measures is typically highest when the IPC based measures are established at the area level (Schmoch, 2008).

The above findings suggest that the text based measure has a substantial correlation with the IPC based measures, but that the distance measures based on the text based measure differ significantly from IPC based measures.

### **VIII. Summary of findings**

This research provides theoretical foundations as well as an empirical assessment of distance measures. Technological distance and proximity are concerned with the degree to which the technological profiles of companies overlap, or, more formally, the length of technological space between companies. It has been argued that technological distance measures have to be assessed based on their suitability for the economic or management

concept, e.g. spillovers, learning and absorptive capacity. No single distance measure is necessarily better in all applications.

I have analyzed a wide range of distance measures, including three measures new to the literature. Each new measure addresses weaknesses of the existing distance measures. The weighted angular separation accounts for the relatedness of the categories within one level of the IPC hierarchy (e.g. the classes). Only the Mahalanobis distance has accounted for this relationship so far. The aggregated patent-to-patent angular separation addresses the processes that underlie concepts such as spillovers and absorptive capacity and thus constitutes a distance measure with sound microeconomic foundations for these applications. The text based similarity measures based on a vector space model (Natterer, forthcoming) is also a bottom-up measure and, in addition, transcends the IPC hierarchy completely.

In the axiomatic assessment of the distance measures, no single distance measure dominates all other measures. I argue that the Jaffe covariance, a non-normalized version of the angular separation, may be more suitable for many applications than many of the existing distance measures. In particular, it has been given microeconomic foundations and it satisfies the criterion of independence of irrelevant patent fields. Overall, when dealing with technological distance, it is advisable to check the robustness of the findings to different distance measures.

As has been shown in the empirical analysis, the choice of distance measure typically has a larger impact on the cardinal ranking of companies than the choice of the level in the IPC hierarchy. There are groups of distance measures that provide similar results. For example, the angular separation, the Pearson correlation coefficient and the min-complement are all highly correlated. This suggests that the theoretical appeal of the independence of irrelevant patent classes (Bar & Leiponen, 2012) does not lead to different results than the existing measures which do not satisfy this criterion (at least in this sample of firms). Also, the Jaffe covariance and the aggregated patent-to-patent angular separation are highly correlated, which is encouraging as they are the only measures that have been given microeconomic foundations to date. The correlation of revealed technological advantage and the Mahalanobis distance have a much lower correlation with the other measures proposed in this research. The text based measure yields results that have substantial correlation with the IPC based measures. However, the coefficients are typically lower than the coefficients between the IPC based measures.

## References

- Adams, J.D. (1990). Fundamental Stocks of Knowledge and Productivity Growth. *Journal of Political Economy*, 98(41), 673-702.
- Ahuja, G., & Katila, R. (2001). Technological acquisitions and the innovation performance of acquiring firms: A longitudinal study. *Strategic Management Journal*, 22(3), 197-220.
- Alcacer, J., Gittelman, M., & Sampat, B. (2009). Applicant and examiner citations in US patents: An overview and analysis. *Research Policy*, 38(2), 415-427.
- Bar, T., & Leiponen, A. (2012). A measure of technological distance. *Economics Letters*, 116(3), 457-459.

- Bena, J., & Li, K. (in press). Corporate innovations and mergers and acquisitions. *The Journal of Finance*. doi: 10.1111/jofi.12059.
- Benner, M., & Waldfoegel, J. (2008). Close to you? Bias and precision in patent-based measures of technological proximity. *Research Policy*, 37(9), 1556-1567.
- Bloom, N., Schankerman, M., & Van Reenen, J. (2013). Identifying technology spillovers and product market rivalry. *Econometrica*, 81(4), 1347-1393.
- Breschi, S., Lissoni, F., & Malerba, F. (2003). Knowledge-relatedness in firm technological diversification. *Research Policy*, 32(1), 69-87.
- Cantwell, J., & Noonan, C. (2001, June). Technology relatedness and corporate diversification 1890-1995. Paper presented at *Nelson Winter Conference, Denmark, 12-15 June 2001*.
- Cincera, M. (2005). Firms' productivity growth and R&D Spillovers: an analysis of alternative technological proximity measures. *Economics of Innovation and New Technology*, 14(8), 657-682.
- Cincera, M., & van Pottelsberghe de la Potterie, B. (2001). International R&D spillovers: a survey. *Cahiers Economiques de Bruxelles*, 169(1), 3-32.
- Cloudt, M., Hagedoorn, J., & Van Kranenburg, H. (2006). Mergers and acquisitions: Their effect on the innovative performance of companies in high-tech industries. *Research Policy*, 35(5), 642-654.
- Cohen, W. M., & Levinthal, D. A. (1989). Innovation and learning: the two faces of R & D. *The Economic Journal*, 99(397), 569-596.
- Cohen, W. M., & Levinthal, D. A. (1990). Absorptive capacity: a new perspective on learning and innovation. *Administrative Science Quarterly*, 35(1), 128-152.
- Comanor, W. S., & Scherer, F. M. (1969). Patent statistics as a measure of technical change. *The Journal of Political Economy*, 77(3), 392-398.
- Engelsman, E. C., & van Raan, A. F. (1994). A patent-based cartography of technology. *Research Policy*, 23(1), 1-26.
- Engelsman, E. C., & Van Raan, A. F. J. (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.
- Engelsman, E. C., & Van Raan, A. F. J. (1993). International comparison of technological activities and specializations: a patent-based monitoring system. *Technology Analysis & Strategic Management*, 5(2), 113-136.
- Farjoun, M. (1998). The independent and joint effects of the skill and physical bases of relatedness in diversification. *Strategic Management Journal*, 19(7), 611-630.
- Fleming, L., & Sorenson, O. (2001). Technology as a complex adaptive system: evidence from patent data. *Research Policy*, 30(7), 1019-1039.
- Fung, M. K. (2003). Technological proximity and co-movements of stock returns. *Economics Letters*, 79(1), 131-136.
- Gilsing, V., Nooteboom, B., Vanhaverbeke, W., Duysters, G., & van den Oord, A. (2008). Network embeddedness and the exploration of novel technologies: Technological distance, betweenness centrality and density. *Research Policy*, 37(10), 1717-1731.
- Goto, A., & Suzuki, K. (1989). R & D capital, rate of return on R & D investment and spillover of R & D in Japanese manufacturing industries. *The Review of Economics and Statistics*, 71(4), 555-564.

- Griliches, Z. (1979). Issues in assessing the contribution of research and development to productivity growth. *Bell Journal of Economics*, 10, 92-116.
- Griliches, Z. (1990). Patent Statistics as Economic Indicators: A Survey. *Journal of Economic Literature*, 28(4), 1661-1707.
- Griliches, Z. (1992). The Search for R&D Spillovers. *The Scandinavian Journal of Economics*, 94, S29-S47.
- Hagedoorn, J., & Cloudt, M. (2003). Measuring innovative performance: is there an advantage in using multiple indicators?. *Research Policy*, 32(8), 1365-1379.
- Harhoff, D. (2000). R&D spillovers, technological proximity, and productivity growth—evidence from German panel data. *Schmalenbach Business Review*, 52(3), 238-260.
- Harhoff, D., Scherer, F. M., & Vopel, K. (2003). Citations, family size, opposition and the value of patent rights. *Research Policy*, 32(8), 1343-1363.
- Henderson, R., & Cockburn, I. (1996). Scale, scope, and spillovers: the determinants of research productivity in drug discovery. *The Rand Journal of Economics*, 27(1) 32-59.
- Hussinger, K. (2010). On the importance of technological relatedness: SMEs versus large acquisition targets. *Technovation*, 30(1), 57-64.
- Jaffe, A. B. (1986). Technological opportunity and spillovers of R&D: Evidence from firms' patents, profits, and market value. *American Economic Review*, 76 (5), 984–1001.
- Jaffe, A. B. (1989). Characterizing the “technological position” of firms, with application to quantifying technological opportunity and research spillovers. *Research Policy*, 18(2), 87-97.
- Jaffe, A. B., & Trajtenberg, M. (1999). International knowledge flows: Evidence from patent citations. *Economics of Innovation and New Technology*, 8(1-2), 105-136.
- Jaffe, A. B., Trajtenberg, M., & Fogarty, M. S. (2000). Knowledge Spillovers and Patent Citations: Evidence from a Survey of Inventors. *American Economic Review*, 90(2), 215-218.
- Kogut, B., & Zander, U. (1992). Knowledge of the firm, combinative capabilities, and the replication of technology. *Organization Science*, 3(3), 383-397.
- Lane, P. J., & Lubatkin, M. (1998). Relative absorptive capacity and interorganizational learning. *Strategic Management Journal*, 19(5), 461-477.
- McNamee, R. C. (2013). Can't see the forest for the leaves: Similarity and distance measures for hierarchical taxonomies with a patent classification example. *Research Policy*, 42(2013), 855– 873.
- Mohnen P. (1996). R&D externalities and productivity growth, *STI Review*, 18, 39-66.
- Mowery, D. C., Oxley, J. E., & Silverman, B. S. (1998). Technological overlap and interfirm cooperation: implications for the resource-based view of the firm. *Research Policy*, 27(5), 507-523.
- [Natterer, forthcoming doctoral thesis]
- Nemet, G. F., & Johnson, E. (2012). Do important inventions benefit from knowledge originating in other technological domains?. *Research Policy*, 41(1), 190-200.
- Nesta, L., & Saviotti, P. (2005). Coherence of the Knowledge Base and the Firm's Innovative Performance: Evidence from the US Pharmaceutical Industry. *Journal of Industrial Economics*, 53(1), 123-142.

- Nooteboom, B., Van Haverbeke, W., Duysters, G., Gilsing, V., & Van den Oord, A. (2007). Optimal cognitive distance and absorptive capacity. *Research Policy*, 36(7), 1016-1034.
- Olsson, O., & Frey, B. S. (2002). Entrepreneurship as recombinant growth. *Small Business Economics*, 19(2), 69-80.
- Pavitt, K. (1988). Uses and abuses of patent statistics. In A.F.J. Van Raan (Ed.), *Handbook of quantitative studies of science and technology* (pp. 509-536). Amsterdam: Elsevier
- Rosenkopf, L., & Almeida, P. (2003). Overcoming local search through alliances and mobility. *Management Science*, 49(6), 751-766.
- Sapienza, H. J., Parhankangas, A., & Autio, E. (2004). Knowledge relatedness and post-spin-off growth. *Journal of Business Venturing*, 19(6), 809-829.
- Scherer, F. M. (1982). Inter-industry technology flows and productivity growth. *The Review of Economics and Statistics*, 64(4), 627-634.
- Schmidt-Ehmcke, J., & Zloczynski, P. (2008, April). *Technology Portfolio and Market Value* (No. 780). DIW Berlin, German Institute for Economic Research. Retrieved from <http://www.diw.de/documents/publikationen/73/82017/dp780.pdf>
- Schmoch, U. (2008, June). Concept of a Technology Classification for Country Comparisons. Finals Report to the World Intellectual Property Organization (WIPO). Retrieved from [http://www.wipo.int/export/sites/www/ipstats/en/statistics/patents/pdf/wipo\\_ipc\\_technology.pdf](http://www.wipo.int/export/sites/www/ipstats/en/statistics/patents/pdf/wipo_ipc_technology.pdf)
- Stuart, T. E. (1998). Network positions and propensities to collaborate: An investigation of strategic alliance formation in a high-technology industry. *Administrative Science Quarterly*, 43(3), 668-698.
- Stuart, T. E., & Podolny, J. M. (1996). Local search and the evolution of technological capabilities. *Strategic Management Journal*, 17(S1), 21-38.
- Teece, D. J., Rumelt, R., Dosi, G., & Winter, S. (1994). Understanding corporate coherence: Theory and evidence. *Journal of Economic Behavior & Organization*, 23(1), 1-30.