IAM Discussion Paper Series #030

When Category-based Indices Encounter Non-independent Categories: Solving the Taxonomy Issue in Resource-based Empirical Studies

2014 年 2 月

Post-doc researcher,
Department of Technology Management for Innovation(TMI),
School of Engineering, University of Tokyo
Dong Huo

Professor,
Department of Technology Management for Innovation(TMI),
School of Engineering, University of Tokyo
Kazuyuki Motohashi

IAM
Intellectual Asset-Based Management

東京大学 知的資産経営総括寄付講座
Intellectual Asset-Based Management Endorsed Chair
The University of Tokyo

When Category-based Indices Encounter Non-independent Categories: Solving the Taxonomy Issue in Resource-based Empirical Studies

ABSTRACT

Category-based measures such as the cosine index, Herfindahl index, entropy index, Euclidean distance, and Pearson's correlation coefficient are widely applied in resource-based studies. However, when adopting these indices in empirical studies with respect to the categories of industry, technology, or knowledge, the inherent relatedness between the categories gives rise to a taxonomy issue, as the categories are not naturally inter-independent (as they are supposed to be in the application of the original indices) owing to inaccuracies in classification or categorization. Therefore, category-based indices may fail to produce a valid measurement, and this unmeasured relatedness can result in endogeneity in empirical analyses. To solve this issue, this study proposes new indices that can harness the relatedness information of categories. An example of mathematical constructive proof for the cosine index is given, and the development process provides a rigorous solution framework. In addition, this study thoroughly examines validity such as content validity, convergent validity, discriminant validity, internal consistency, and criterion-related validity. The results reveal that this approach performs as expected in empirical analyses. Finally, the mathematical interpretation and generality of this approach for other category-based indices such as the Herfindahl index, entropy index, Euclidean distance, and Pearson's correlation coefficient are discussed. Interestingly, other proposed measures from existing literature, such as the concentric index, can also be incorporated into this solution framework.

Keywords: measurement, taxonomy, category, diversity, relatedness, validity, metric

INTRODUCTION

The craftsman who wishes to do his work well must first sharpen his tools.
— Confucius, 551–479 BC

Achilles cannot overtake the tortoise so long as their progress is considered piecemeal, endlessly having the distance between them. However, as it is not Achilles but the method of measurement which fails to catch up with the tortoise, so it is not man but his method of thought which fails to find fulfillment in experience.
—Alan Wilson Watts, 1958

Numerous scholars have focused on innovation strategies and competitive advantages of firms according to the resource-based view (RBV). In addition, a considerable number of studies have examined the extent to which the composition and structure of a firm's resources can influence its performance. When testing hypotheses with empirical data, scholars preferred adopting category-based indices for measurement,[1] which include, but are not limited to, the cosine index, Herfindahl index, entropy index, Euclidean distance, and Pearson's correlation coefficient. In applications of these indices, a category is usually defined to represent one specific type of taxon, such as industry (Neffke & Henning, 2012), technology (Sampson, 2007), knowledge (Van Der Vegt & Bunderson, 2005), and business function (Bunderson & Sutcliffe, 2002), which reflects the corresponding resource.

As these indices are constructed upon categories, the taxonomy regarding how these categories and classification schemes are established should be an extremely important issue. Numerous classification schemes are constructed in different ways to meet various requirements. The old fashioned method of constructing a classification scheme usually relies on qualitative expert assessment and conventions. For example, in the taxonomy of the patent system,

[1] The "category-based index" in the scope of this study refers to an index that involves dot product operations on vectors. They are sometimes referred to as "objective categorical measure" or "continuous measure" of diversification, rather than as the "subjective categorical measure" or "categorical measure" of diversification originally conducted by Wrigley (1970) and refined by Rumelt (1974).

International Patent Classification (IPC) and U.S. Patent Classification (USPC) are popular and established conventional schemes to classify technologies. Another example is Standard Industrial Classification (SIC), which classifies industries. Of course, there are some limited attempts that adopt quantitative approaches to adjust the scheme generated from the qualitative classification (Schmoch et al., 2003).

Regardless of which schemes are employed in empirical research, they must be in agreement with the underlying assumption that all categories must be inter-independent. For instance, in the aforementioned study by Bunderson and Sutcliffe (2002), nine categories (i.e., sales and marketing, manufacturing, distribution, service, personnel, R&D, accounting, administration, and general management) are implicitly and logically independent (or nearly independent) functional units in firms. However, this assumption has often been violated in most empirical studies, which involve non-independent categories such as industry, technology, expertise, and knowledge. For example, in the study by Sampson (2007), technological fields are believed to be related to one another, indicating possible dependency among the categories.

This research question is extremely important because such inter-dependency may incur measurement error, and therefore potential endogeneity. The unmeasured relatedness among these categories correlates with the constructs employed as independent variables in empirical studies. In addition, such relatedness might influence dependent variables of interest. Therefore, when running a regression model, it will be absorbed by the error term and result in bias.

Dahlin, Weingart, and Hinds (2005) were aware of this problem when employing the Herfindahl index in their study. However, they still assumed that the categories were proximately inter-independent. Rumelt (1982) also noticed the problem, but instead of using a Herfindahl-based index, he proposed a new approach. In general, research that involves the utilization of

category-based indices on non-independent categories may face such an issue (Breschi, Lissoni, & Malerba, 2003; Dahlin et al., 2005; Tallman & Li, 1996; Van Der Vegt & Bunderson, 2005)

In the remainder of this study, we select one typical category-based index—cosine index—to demonstrate the proposed solution framework and correspondent constructive proof.[2] Further, empirical data is employed to validate the new construct. Finally, the mathematical interpretation and generality of this approach for a variety of category-based indices such as the Herfindahl-based index, entropy index, Euclidean distance, and Pearson's correlation coefficient are discussed along with limitations and alternative solutions.

CATEGORY-BASED INDICES

Original and proposed category-based indices

Despite the different definitions of category-based indices such as the cosine-based index, Herfindahl-based index, entropy index, Euclidean distance, and Pearson's correlation coefficient, they are generally self-consistent when the categories are inter-independent. Such a dependency-free condition is the underlying assumption, which is based on the idea that the indices provide a valid measurement of diversity. Studies in which categories with respect to focal attributes are perfectly independent or nearly independent may not suffer this issue. Such attributes include, but are not limited to, gender, ethnicity, age, tenure, and functional background (Bunderson & Sutcliffe, 2002; Jackson et al., 1991; Putnam, 2007).

However, this dependency-free assumption is not always supported in other management-related research, which involves utilizing these indices with respect to the categories of industry,

[2] Please note that the remainder of this paper adopts the concept of "dissimilarity" for the cosine-based index. Alternatively, it can be replaced by the opposite concept of "similarity In addition, for simplicity of symbolic representation in mathematical equations, we use the term "cosine-based diversity index" and thus d_{Cosine} is taken to symbolize the index.

technology, and knowledge (Breschi et al., 2003; Dahlin et al., 2005; Tallman & Li, 1996; Van Der Vegt & Bunderson, 2005). The following example demonstrates what occurs if the dependency-free condition is violated.

First, the definition of the cosine-based diversity index d_{Cosine} is provided as follows. Alternatively, it can be written in a vector operation form in which k_{1i} and k_{2i} are elements in column vectors $v_1(k_{11}, \ldots\ k_{1i}, \ldots k_{1n})$ and $v_2(k_{21}, \ldots\ k_{2i}, \ldots k_{2n})$ with regard to an n-dimensional real number vector space $\boldsymbol{R^n}$. The "cosine theta" part in the definition of the d_{Cosine} is an extension of the angular separation from $\boldsymbol{R^3}$ to $\boldsymbol{R^n}$, whereas "theta" is the angle between the two vectors (Gower, 1967). In application, the technology portfolio can be taken as the vector in computation:

$$d_{Cosine} = 1 - \frac{\sum k_{1i}\, k_{2i}}{\sqrt{\sum k_{1i}^2 \sum k_{2i}^2}} = 1 - \frac{v_1^T v_2}{\sqrt{v_1^T v_1 * v_2^T v_2}}$$

Following the method utilized in the study by Sampson (2007), we assume two firms in a strategic alliance that conduct collaborative R&D. Their technology portfolios are represented as $v_1(1, 10, 2)$ and $v_2(10, 1, 2)$, respectively, and each dimension of vector v_1 and v_2 refers to each technology category that the firms involve. If all three technology categories are inter-independent, then the cosine-based diversity index d_{Cosine} is equal to $\frac{81}{105}$. However, if the first technology category is considerably similar to the second, then the inherent *actual* technology diversity for the two firms should be very close to zero because their technology portfolios might be regarded as identical,.

The non-independent issue of the category-based index indicates that the dependences between the categories need to be distinguished and then adjusted accordingly. We introduce the relatedness to measure such dependencies. This concept refers to relational characteristics

between certain objects, and it is usually quantified to measure the extent to which objects are related. Relatedness has been widely discussed, including technological relatedness (Robins & Wiersema, 1995), industry relatedness (Teece et al., 1994), knowledge relatedness (Breschi et al., 2003), skill relatedness (Neffke & Henning, 2012), and product relatedness (Rumelt, 1974).

To address the above-mentioned issue, we propose a new category-based index that incorporates information on relatedness for each pair of non-independent categories as well as enables adjustment for these categories. Such additional information of relatedness can enhance measurement validity of diversity, and the endogeneity owing to unmeasured relatedness in original indices can be eliminated or reduced. Consider the following adjusted cosine-based index as a demonstration:

$$d_{Cosine} = 1 - \frac{v_1^T M v_2}{\sqrt{v_1^T M v_1 * v_2^T M v_2}}$$

The relatedness matrix M consists of all relatedness scores for each pair of categories. Each non-diagonal element r_{ij} (that represents the relatedness between categories i and j) in M includes its scores in [0, 1], and the diagonal consists of n ones (n refers to the number of categories):

$$M = \begin{bmatrix} 1 & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & 1 \end{bmatrix}$$

Finally, owing to the participation of the relatedness matrix M, the values of the above proposed indices will shrink even though the minimum value of zero is maintained. Furthermore, to meet the preferences of some empirical samples, they can be further min–max normalized to [0, 1].

Constructive proof

This section illuminates the process of constructing this proposed index in greater detail, thus also mathematically indicating its rigor and validity. At this stage, we assume that the relatedness matrix M is already in possession. The next section discusses the approach for deriving this matrix.

The first step in constructing this proposed index is to establish an *n*-dimensional oblique (i.e., non-orthogonal) coordinate system according to the relatedness matrix M. The n dependent categories could be regarded as *n* non-perpendicular axes in an *n*-dimensional oblique coordinate system. In linear algebra, the n categories form a family of n linearly independent but non-orthogonal vectors in an *n*-dimensional Euclidean space. The angle separation between any two axes reflects the extent of their geometrical proximity. It also implies the relatedness between the two correspondent technology categories when extended to this proposed application. Instead of the degree of an angle, the cosine is applied to make it consistent with the normalized relatedness scores generated above. Through these means, a system of non-linear equations can be established, and the n linearly independent but non-orthogonal vectors in the *n*-dimensional Euclidean space can be derived after the equations are solved. The system of equations is given in a matrix expression, as follows:

$$A^T A = M$$

in which A represents *n* column vectors A_i of the desired solution:

$$A = A(A_1, \cdots, A_n) = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}$$

To verify if A is the solution, any two columns of A_i and A_j can be selected from A. The dot product $A_i^T A_j = r_{ij}$ and magnitude $\|A_i\| = 1$, meeting the requirement for an oblique coordinate system. Note that in the scope of this study, the relatedness matrix is symmetric so that $r_{ij} = r_{ji}$.

The second step includes a conversion between the *n*-dimensional oblique coordinate system and the *n*-dimensional Cartesian coordinate system through orthogonalization and normalization. A series of distinguished QR decomposition approaches can be applied to orthonormalize the oblique coordinate system, such as the Gram–Schmidt process (Trefethen & Bau III, 1997), Householder reflections (Householder, 1958), and Givens rotations (Givens, 1958). After being processed through any of these approaches, a transition matrix R is produced. The matrix R mathematically represents the transition matrix for a transition from basis $Q(Q_1, \cdots, Q_n)$ to basis $A(A_1, \cdots, A_n)$. The resulting orthonormal basis can be regarded as a set of linearly independent virtual categories, and each axis in the basis represents one virtual category:

$$A = QR$$

Different QR decomposition approaches may lead to a different orthogonal matrix Q and a corresponding transition matrix R. However, any approach produces the same proposed index. Thus, it can be explained that a cosine-based diversity index captures the angular separation of the two technology portfolio vectors, which would not change in different Cartesian coordinate systems with respect to different orthonormal basis. In other words, a certain orthogonal matrix Q might be seen as an outcome from a series of rotations and shift transformations of the standard basis I (I is unit matrix) in a Cartesian coordinate system. Moreover, in accordance with this, a correspondent transition matrix R can be derived.

The third step is to transform original coordinates in accordance with a new orthonormal basis. This is achieved by left-multiplying the original coordinates with the transition matrix R:

$$v' = Rv$$

This cosine-based diversity index can directly take technology portfolios as vectors in either an oblique or a Cartesian coordinate system. That is, each element $k_i^{'}$ in a new technology

portfolio $v^{'}(k_1^{'}, \ldots k_i^{'}, \ldots k_n^{'})$ represents a virtual number of resources in the specific virtual category i, analogous to the original technology portfolio. As a result, a new cosine-based diversity index is shown in the following form:

$$d_{Cosine} = 1 - \frac{(Rv_1)^T(Rv_2)}{\sqrt{(Rv_1)^T(Rv_1) * (Rv_2)^T(Rv_2)}} = 1 - \frac{(v_1^T R^T)(Rv_2)}{\sqrt{(v_1^T R^T)(Rv_1) * (v_2^T R^T)(Rv_2)}}$$

Furthermore, this expression can be simplified. With $A^T A = M$ and $A = QR$, it is possible to derive $R^T Q^T QR = M$. Due to the associative property of matrix multiplication and $Q^T Q = I$ (where Q is the orthogonal matrix and I is the unit matrix), $R^T R = M$ is obtained. As a result, the expression of the new index becomes the form proposed at the beginning.

It is important to note that the value of the proposed index becomes smaller than that of the original ones, because the angle between any two non-perpendicular axes in an original oblique system is an acute angle. This means that the corresponding categories *i* and *j* are related to the extent of r_{ij}. Utilizing the Pythagorean equation, $a^2 + b^2 = c^2$ (where c is the length of the hypotenuse, and a and b represent the lengths of the other two sides), for the distance will not hold in an oblique coordinate system and for non-independent categories.[3] Accordingly, $\frac{v_1^T M v_2}{\sqrt{v_1^T M v_1 * v_2^T M v_2}}$ will produce a greater value than will $\frac{v_1^T v_2}{\sqrt{v_1^T v_1 * v_2^T v_2}}$. In other words, if the coordinates are used in the original manner, in which dependency-free conditions cannot be secured, then the resulting cosine diversity index will be overestimated. This argument is consistent with what was addressed earlier: the diversity of interest is not supposed to be as diverse as the original diversity index measures, in the cases in which categories are related.

[3] More mathematical details can be referred to on the basis of a generalization of the Pythagorean theorem (Sayili, 1960).

RELATEDNESS MATRIX

Approaches to measure relatedness

In existing literature, there are three major approaches for measuring relatedness scores. The first is to manually assign fixed values that are derived directly from an established conventional hierarchical classification system (e.g., SIC and IPC). For practical use in empirical studies, scholars have employed an *n*-digit prefix of the classification code (the value of *n* depends on the level or granularity of the categories in need) by which the relatedness scores are determined on the basis of a principle that categories under the same hierarchical super-categories are more closely related than those under different super-categories. Examples of this type (either used alone or as a distance weight for other indices) are given in the literature on firm diversification (Chatterjee, 1990a; Finkelstein & Haleblian, 2002; Lien & Klein, 2006; Robins & Wiersema, 1995) based on *n*-digit SIC codes.[4] Furthermore, this type of method encounters a problem in that it indiscriminately assigns weights that merely depend on the hierarchical structure of the established classification system. However, such hierarchy itself is usually not reliable enough. For example, the frequently employed SIC system is believed to be vulnerable due to "varying degrees of breadth in the SIC classes" (Rumelt, 1982).

The second approach is the economic regression-based approach, which is based on the co-occurrence method pioneered by Engelsman and van Raan (1991). Neffke and Henning (2008) developed an index called, "Revealed Relatedness," which can be interpreted as the ratio of the observed value over the estimated co-occurrence of two classes in a single entity (e.g., the co-occurrence of two industries in a single plant for deriving industry relatedness). The predicting model is obtained after regression for factors such as average profitability in the industries, wage

[4] For different granularities (i.e., *n*-digit), distance weight is assigned correspondingly different values, and it decreases when the granularity becomes finer (i.e., *n* becomes greater) as the relatedness scores are computed.

levels, and fierceness of competition. In other words, the predicting model controls for economic factors that result in the attractiveness of diversification toward a specific industry; thus, the implying inherent relatedness can be revealed by establishing the fitted and predicted co-occurrence as a benchmark (i.e., denominator in the ratio). This method can also be utilized to compute asymmetric relatedness in which the relatedness from *i* to *j* differs from the relatedness from *j* to *i*. For example, in a study by Neffke and Henning (2012), an examination regarding the dataset of labor flow among industries is conducted in this manner to investigate the relationship between skill relatedness and firm diversification.

The third approach, a co-occurrence-based approach, follows Teece et al. (1994) and the relevant improved version by (Bryce & Winter, 2009). Teece et al. assumed that *K* firms are active in two or more industries, and the number of active firms in industry *i* represents the industry's size. The co-occurrence of industries *i* and *j* partially reflects the propensity to participate and the extent to which the two industries are related. However, an adjustment is required since the raw co-occurrence counts rely not only on inherent relatedness but also on the industry sizes of *i* and *j* (i.e., appearances of industry in co-occurrence data). This adjustment is accomplished by assuming a random model to form a combination of firms in which industries *i* and *j* overlap. The random selection, as described above, leads to a hypergeometric distribution, and its mean and standard deviation can be derived. Eventually, the *t*-statistic style value is taken as the relatedness score for industries *i* and *j*. In other words, the essence behind this method is that the random model, which is under the assumption of no relatedness, is presumed, and the resulting mean μ and standard deviation σ are taken as the benchmarks in order to correct for industry size. Improving on this approach by Teece et al., Bryce and Winter (2009) added two consecutive steps: weighting the scores by economic importance, and generating the relatedness

scores for industry combinations which were not observed in the focal data. Therefore, owing to its predictive validity, the present study's proposed diversity index is based on the approach by Teece et al.

However, applying Teece et al.'s relatedness scores to this proposed diversity index without modification is problematic. The "survivor principle" (Stigler, 1983) implies that information regarding co-occurrence reflects the relatedness of the relevant industries. This argument is logically reliable in the original application of Teece et al.'s approach on corporation establishment data (Teece et al., 1994). However, when it is extended to patent data, as in the study by Bryce and Winter (2009), the co-occurrence method abandons patent documents that are assigned to only one technology field. Therefore, it may incur information loss when applying the survivor principle as these patents also play the role of survivor. For example, consider IPC class *C12M* (described as "apparatus for enzymology or microbiology") and *H01R* (described as "electrically-conductive connections; structural associations of a plurality of mutually-insulated electrical connecting elements; coupling devices; current collectors"). Although both have almost the same number of co-occurrences (1,313 vs. 1,387), the numbers of their single-technology patents (i.e., patents with only one technology field) differ significantly: single occurrences account for approximately 25.5 percent versus 88.1 percent of total occurrences (including both single occurrences and co-occurrences). That is, *C12M* is apt to associate with other technologies, whereas *C12M* is not.

As a result, the "co-occurrence matrix" is replaced with the resulting "occurrence matrix" which also includes single-technology patents. And then we utilize it to develop the relatedness scores. The benefit is that completeness of the data's information can be preserved while there is only a minor change in the construction of the relatedness.

Construction of the relatedness matrix

The first step in constructing the relatedness matrix is to establish the occurrence matrix. Unlike the method elaborated in existing literature about the co-occurrence matrix (Breschi et al., 2003; Bryce & Winter, 2009; Teece et al., 1994), we introduce concepts and tools from social network analysis and matrix algebra to explain the establishment of the matrix. One concept is the two-mode network, in which the edges are always between two different sets of nodes. For example, "member–affiliation" data includes "member" as the first mode and "affiliation" as the second. It is also known as a "bipartite graph" in graph theory. In this case, the two-mode network for the patents data is established and stored in a two-mode matrix in which the rows represent technology categories and columns represent patents. Then the two-mode network is transformed into a one-mode network (technology–technology network), and by conducting matrix algebra multiplication, $O_{\text{tech}-\text{tech}} = O_{\text{tech}-\text{patent}} O^{\text{T}}_{\text{tech}-\text{patent}}$ is obtained, in which O^{T} represents the transpose of matrix O. As a result, each cell in the one-mode matrix $O_{\text{tech}-\text{tech}}$ is:

$$o_{ij} = \begin{cases} co-occurrence\ of\ technology\ i\ and\ j, & if\ i \neq j \\ single-occurrence\ of\ technology\ i, & if\ i = j \end{cases}$$

where o_{ij} is the observed co-occurrence or single occurrence of specific technologies. In addition, the row sum $n_i = \sum_k o_{ik}$ refers to the size of each technology i.

The second step is to adjust the matrix by following the approach developed by Teece et al. (1994). We assume there is an equal likelihood that a technology tag is assigned to a patent. As a result, the extent x to which one patent is assigned both the technology i and j tags follows hypergeometric distribution. In this regard, the probability mass function is given as:

$$P[X_{ij} = x] = \frac{\binom{n_i}{x}\binom{K-n_i}{n_j-x}}{\binom{K}{n_j}}$$

where random variable X_{ij} refers to the number of patents that have both the technology *i* and *j* tags. In addition, n_i, n_j, and K represent the fixed sizes of technologies *i* and *j* and the total number of patents (given as parameters), respectively. More specifically, n_j from K patents are selected and assigned the technology *j* tag; hence, there are $\binom{K}{n_j}$ ways of achieving this. On the basis of this selection, *x* out of n_i technology *i* tags are selected and assigned to *x* patents which already have the technology *j* tag. Furthermore, $n_j - x$ out of $K - n_i$ other tags (i.e., except technology *i*) are selected and assigned to $n_j - x$ patents, which already have the technology *j* tag. Hence, there are $\binom{n_i}{x}\binom{K-n_i}{n_j-x}$ possible ways of achieving this. Then, the probability, as shown in the above probability mass function is obtained. Note that in this proposed approach, single-technology patents are taken into account so that n_i, n_j, and K consist of both co-occurrences and single occurrences in matrix $O_{\text{tech}-\text{tech}}$. That is, the single-technology patents have a likelihood equal to that of the multiple-technology patents in the selection described above. In this new assumption, the technology tags are randomly assigned to patents, including cases in which only one technology is assigned. Therefore, the expression $P[X_{ij} = x]$ does not change. As a result, the mean of the distribution $\mu_{ij} = E(X_{ij}) = \frac{n_i n_j}{K}$ and the standard deviation $\sigma_{ij} = \sqrt{\mu_{ij}\frac{(K-n_i)(K-n_j)}{K(K-1)}}$ are derived, and the *t*-statistic $t_{ij} = \frac{O_{ij}-\mu_{ij}}{\sigma_{ij}}$ is taken as the output relatedness score in this second step.

The third step is to normalize the relatedness scores. To establish dependency-free categories, the relatedness scores should be normalized to the range [0, 1]. The value of zero between two technologies means that they are independent, whereas the value between one means that they are inherently identical. Zero is assigned to all non-adjacent pairs of technologies because there

is no co-occurrence in the data for non-adjacent pairs, which implies that they may not be related. Besides, the least related adjacent pairs after correction for n_i, n_j, and K should be greater than zero (0.1 is assigned to these pairs) because they have shown some co-occurrence. Therefore, each non-diagonal element r_{ij} (which represents the relatedness of technologies *i* and *j*) in the relatedness matrix *M* includes scores in the range [0, 1], while the diagonal consists of *n* ones (*n* refers to the number of technologies):

$$M = \begin{bmatrix} 1 & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & 1 \end{bmatrix}$$

Three examples of relatedness scores can be given to demonstrate if the relatedness scores actually reflect the relatedness among technologies. First, besides the zero-value pairs of technologies (i.e., independent pairs), the least related pair (valued at 0.1) is *G06F* (described as "electric digital data processing") and *H01L* (described as "semiconductor devices; electric solid state devices not otherwise provided for"). This pair has the greatest number of patent occurrences (74,149 and 54,162, respectively) and single occurrences (61,038 and 47,634, respectively). The shares of single occurrences for these two technologies are impressively large (82.3 percent and 87.9 percent, respectively), which implies that they do not generally connect to other technologies in patents. Thus, the relatedness score for this pair of technologies is expected to be low. Second, the most related pair (valued at 1) is *A61K* (described as "preparations for medical, dental, or toilet purposes") and *C07D* (described as "heterocyclic compounds"). These two technologies include relatively large shares of overall co-occurrences (59.2 percent and 72.0 percent, respectively), and they are frequently assigned to relevant patents at the same time (30.5 percent of overall co-occurrences for *A61K* and 61.3 percent for *C07D*). Even in the documentary notes of *A61K*, *C07D* is mentioned as "attention is drawn to the notes in class C07, for example the notes following the title of the subclass C07D, setting forth the rules for

classifying organic compounds in that class, which rules are also applicable, if not otherwise indicated, to the classification of organic compounds in A61K," which suggests their close relationship. Third, even if the co-occurrences of two technologies are equivalent, they may differ in their relatedness scores owing to the different preferences of single-technology patents. As exemplified in the previous section, *C12M* and *H01R* have a similar number of co-occurrences (1,313 and 1,387, respectively). However, their single occurrences differ substantially (450 and 10,298, respectively), which implies that *C12M* is comparably apt to associate with other technologies. As a result, their relatedness scores with the third-party technology *B65D* (described as "containers for storage or transport of articles or materials, e.g., bags, barrels, bottles, boxes, cans, cartons, crates, drums, jars, tanks, hoppers, forwarding containers; accessories, closures or fittings therefor; packaging elements; packages") are different (0.30 and 0.28). In sum, these examples partially assure the validity of the relatedness scores by which the present study's proposed cosine-based index can be further developed.[5]

CONSTRUCT VALIDATION

Validation methods

A thorough examination of construct validation is usually exerted on five types of validity: content validity, convergent validity, discriminant validity, internal consistency, and criterion-related validity (Campbell & Fiske, 1959; Hoskisson, Hitt, Johnson, & Moesel, 1993; Venkatraman & Grant, 1986). Each type reflects a distinct aspect of the measurement.

[5] The data for constructing the relatedness matrix is the patent dataset from the National Bureau of Economic Research (NBER), which includes utility patents granted between 2000 and 2006 by the United States Patent and Trademark Office (USPTO) and 610 four-digit IPC codes chosen as categories of interest. This is similar to the one used in the following construct validation section.

Content validity refers to the extent to which empirical measurement falls within a specific content domain. In other words, it reflects whether the measurement is suitable for content. For example, a measurement of mass, such as kilogram, cannot be applied to measure length even though it may have a strong correlation (e.g., in case of a fixed density and cross section area); therefore, content validity is not supported. In the context of non-independent categories, application of the proposed cosine-based index eliminates measurement error incurred by the non-independent condition, as elaborated in the aforementioned constructive proof. Therefore, the proposed diversity index resides in the valid content domain.

Convergent validity reflects the degree to which different methods of measuring the same concept are in agreement. Discriminant validity, however, is the opposite of convergent validity; it reflects the extent to which a concept differs from others. An analytical tool for evaluating convergent and discriminant validity is the "multitrait-multimethod matrix" created by Campbell and Fiske (1959), by which we examine these two types of validity. Internal consistency represents the overall consistency of a focal measure. In general, a reliable measure should generate consistent results. Here Cronbach's alpha (Cronbach, 1951) is utilized to investigate internal consistency.

In examinations of convergent validity, discriminant validity, and internal consistency, two categorization schemes are adopted to represent parallel forms (or methods) of measurement: IPC and USPC. Both schemes are available in the U.S. patent dataset and are adaptive in a wide range of levels. More specifically, the analyses in the present study are based on four-digit IPC and three-digit USPC, which have been widely acknowledged as representatives of technology fields (Corrocher, Malerba, & Montobbio, 2007; Lanjouw & Schankerman, 2004; Lerner, 1994).

Criterion-related validity, also known as nomological or predictive validity, is based on the degree to which the construct of interest functions in predicted relationships, as suggested in theory. In this regard, an example of empirical analysis is provided, which corresponds to the cosine-based diversity index. More specifically, regression models are employed to examine the relationship between "inter-citations" and "inter-firm technological diversity." In this case, it is hypothesized that for any firm dyad (*A*, *B*), the likelihood of citing or being cited by patents from firm *B* for patents from firm *A* is negatively influenced by the extent to which the two firms differ with respect to their technological portfolios. Similarly, the number of citing or being cited patents is also hypothesized to be negatively associated to such diversity. Accordingly, the logistic and negative binomial models are applied to the examinations, respectively. The citations made and received between firms *A* and *B* are counted to first generate the dependent variable in negative binomial models, and then the resulting continuous variable is dichotomized to binary values (i.e., non-zero values as one) as dependent variables in the logistic models.

Five constructs (number of overlapped categories, ratio of overlapped categories, original cosine diversity, hierarchy-adjusted cosine diversity, and proposed cosine diversity) are used as independent variables to proxy inter-firm technological diversity. The first four are introduced for comparison with the present study's proposed measure. Number of overlapped categories refers to the number of overlapped four-digit IPCs (or three-digit USPCs) for a firm dyad. The ratio of overlapped categories, ranging between zero and one, is defined as the number of overlapped categories divided by the number of categories in the two technology portfolios from two collaborating firms. They are negatively correlated to inter-firm technological diversity. Hierarchy-adjusted cosine diversity is also weighted by the relatedness matrix. However, this matrix is derived from the hierarchical structure of the IPC scheme, which is similar to that

utilized by the concentric index (Caves, Porter, & Spence, 1980). That is, each entry in the matrix is assigned a fixed relatedness score regarding hierarchical distance.

Data

The data for constructing the relatedness matrix are the well-known patent dataset from the NBER, which include utility patents granted between 2000 and 2006 by the USPTO. The granularity or level of aggregation of categorization in this study is with regard to technology fields, i.e., four-digit IPC codes or three-digit USPC codes, as mentioned above. Two distinct relatedness matrices are generated in accordance with the IPC and USPC categorization schemes. The former includes 610 IPCs and the latter comprises 699 USPCs. In addition, the relatedness matrix for computing hierarchy-adjusted cosine diversity is generated under the IPC scheme, which follows the approach used for the SIC scheme in its original application.[6] For example, the relatedness scores between *A01B* and *A01B*, *A01B* and *A01C*, *A01B* and *A21C*, and *A01B* and *B01B*, are assigned 1, $^{2}/_{3}$, $^{1}/_{3}$, and 0, respectively.

To examine convergent validity, discriminant validity, internal consistency, and criterion-related validity, a sample for the cosine-based index is created as follows. First, we identify 12,841 distinct firms which were granted utility patents by the USPTO in 2003. Next, 1,143 of these are selected on the basis of the rule that each firm has applied for 30–99 patents in 2000–2002. There are two reasons why these firms are confined: 1) the technology portfolio of a firm may not be well represented if the firm has too few prior patents and 2) the size of the resulting sample equals the number of combination dyads of firms so that the number of firms cannot be too large; otherwise, the resulting sample would be too costly in computation. Finally, the

[6] The four-digit IPC code includes three levels. For example, for IPC *A01B*, "*A*" refers to the top level, "*01*" refers to the second level, and "*B*" represents the third level. However, the three-digit USPC is usually regarded as "flat," i.e., all three-digit codes lie on the same level. Thus, the relatedness matrix for hierarchy-adjusted cosine diversity is only generated under the IPC scheme.

resulting sample consists of 652,653 firm dyads, which is created by combining the selected 1,143 firms. In addition, the five proxies of independent variables (inter-firm technological diversity, number of overlapped categories, ratio of overlapped categories, original cosine diversity, hierarchy-adjusted cosine diversity, and proposed cosine diversity) are computed. The time window for evaluation of the technological portfolio is 2000–2002, when each observed firm applied for patents before the sampling year, 2003.

In addition, as for an examination of criterion-related validity, dependent variables and control variables are included. The dependent variable "number of inter-citations" is the sum of citations made and received between firms *A* and *B* by the end of 2006. It is worth noting that the citations assigned by examiners in the USPTO[7] are applied rather than all the citations by both examiners and applicants because the examiner citations are believed to be more valid for reflecting actual technological linkages to prior art (Cotropia, Lemley, & Sampat, 2013). Finally, the characteristics of each firm in a dyad are controlled for separately; thus, "number of prior patents 1," "number of prior patents 2," "experience years 1," and "experience years 2" are included.[8]

Results

Table 1 summarizes all the measures. It is worth noting that the proposed index is less skewed compared to other proxies. The majority of observations under contrastive proxies are concentrated toward one end; however, the proposed cosine diversity substantially reduces skewness (by approximately 90 percent).

[Insert Table 1 here]

[7] Data regarding examiner citations became available from 2001 as the USPTO adopted new reporting procedures. This study utilizes examiner citation data provided by Sampat (2012).

[8] Counting experience years is based on the NBER dataset. According to the dataset, the starting point is the year when a firm applies for its first patent. The earliest patent in the dataset was applied for in 1901.

Tables 2 and 3 present the Pearson and Spearman correlation coefficients, respectively, for the cosine-based index.[9] The bold and underlined values are validity diagonals (i.e., the monotrait-heteromethod), and they indicate that these measures under different methods (i.e., categorization schemes IPC and USPC) are convergent due to significant correlations.

Discriminant validity is examined when validity diagonals for the cosine-based index are compared with heterotrait-heteromethod triangles (i.e., non-formatted values) or heterotrait-monomethod triangles (i.e., italicized values). The higher correlations of validity diagonals suggest that the measures (or traits) differ from others. For example, proposed cosine diversity differs from the number of overlapped categories.

Most importantly, the results in these tables reveal that the proposed measure has a better correlation coefficient across the two categorization schemes than did the original one. For example, in Table 3, the Spearman correlation coefficient for proposed cosine diversity was 0.8519, whereas the original cosine diversity was 0.0933 lower.

[Insert Table 2 here]

[Insert Table 3 here]

In addition, as presented in Table 4, the internal consistency for these proxies is examined using Cronbach's alpha for the resulting index under the IPC and USPC categorization schemes. All the alphas are large enough (i.e., above 0.8) to facilitate the interpretation that the value does not differ significantly across the IPC and USPC categorization schemes, regardless of which proxy is taken into account. This implies that the all measures' internal consistencies are good or even excellent. In particular, when the alpha of the proposed measure is compared with that of the others, the results suggest improvements in internal consistency. For example, Cronbach's

[9] The number of overlapped categories and ratio of overlapped categories are proxies for similarity. Therefore, their correlation coefficients with original cosine diversity and improved cosine diversity are negative. This study is only concerned with the absolute values of the coefficients.

alpha increases from 0.9082 to 0.9311 when the proposed approach is applied to the cosine-base index.

[Insert Table 4 here]

Despite increases in convergent validity, discriminant validity, and internal consistency for the proposed measures, their criterion-related validity must be examined to check for improvements. The criterion used for this examination is the influence of "inter-firm technological diversity" on "number of inter-citations." Four proxies (number of overlapped categories, ratio of overlapped categories, original cosine diversity, and proposed cosine diversity) are investigated and compared. Table 5 presents the descriptive statistics.

[Insert Table 5 here]

Logistic and negative binomial models are adopted for estimation.[10] Tables 6 and 7 present the corresponding results. All the independent variables of interest are significant at the one-percent level, and the estimates imply that inter-firm technological diversity impedes a firm from citing or being cited by the other firm in a dyad. Furthermore, in terms of similarity, the results suggest that the more similar the two firms are, the more likelihood and the greater the number of inter-citations, as the estimates of the number of overlapped categories and ratio of overlapped categories show.

[Insert Table 6 here]

[Insert Table 7 here]

The goodness-of-fit for these models are also examined. Tables 8 and 9 present the goodness-of-fit scores for comparison of models (8)–(10) in various measures. The results show that the goodness-of-fit for the cosine-based index is considerably increased. For example, McFadden's

[10] Since the observations with value at one account for only two percent in the sample, we in addition employ a rare-events logistic model (King & Zeng, 2001) for estimation instead of a logistic model. In this case, the results are almost identical to those in Tables 6.

R^2 computed from the logistic model increases by 27.3 percent, i.e., from 0.154 to 0.196. Moreover, when the proposed index is compared with the hierarchy-adjusted index, McFadden's R^2 still increases by roughly 10 percent, implying that the present study's approach for computing relatedness scores outperforms the hierarchical structure approach.

[Insert Table 8 here]

[Insert Table 9 here]

DISCUSSION AND CONCLUSION

Through a careful examination of validity, this study's proposed methodology has been proven to be valid and effective for measuring category-based diversity in the context of non-independent categories (such as technology or industry). The problem solved in this study has long been neglected and has arguably led to failure in measurement. Moreover, the universal application of category-based indices urges one to seek a solution to this problem. In general, the mathematic proof demonstrates that the development of this solution mainly consists of three consecutive steps along a constructive path: 1) establishing an oblique coordinates system according to the measured relatedness between non-independent categories; 2) converting the basis between an oblique system and a Cartesian system; and 3) transforming the vectors used in the cosine-based index from an oblique system to a Cartesian system. The first step is based on several scholars' efforts in designing the relatedness scores for related industries or technologies (Breschi et al., 2003; Bryce & Winter, 2009; Teece et al., 1994). However, the other two steps are novel in strategic management research and social science.

Mathematical interpretation and generality

Mathematically speaking, the solution framework in this study employs linear algebra techniques analogous to those used in the Mahalanobis distance (Mahalanobis, 1936) and principal component analysis (Pearson, 1901), although the latter two adopt a different relatedness matrix for weighting the original indices and serve for different purposes. For example, principal component analysis is designed for obtaining several "principal" inter-independent canonical variates out of numerous variables. However, in this study's application of similar techniques, the issue which variates or categories are of principal and they does not affect the resulting indices. In addition, the sample's covariance matrix in the principal component analysis can be regarded as a type of relatedness matrix, which takes covariance as a proxy of relatedness instead of that produced from this study. Other approaches can also produce relatedness matrix, such as the revealed relatedness economic regression model by Neffke and Henning (2008).

In terms of linear algebra, the relatedness matrix M in this study is equivalent to the "metric matrix" or "metric tensor" (Vaughn, 2008), in which any entry r_{ij} equals the dot product of basis vectors A_i and A_j. Therefore, if two column vectors x and y are assumed, then accordingly, $x^T My$ is the dot product of x and y with respect to metric M.[11] With this metric matrix M, this study's approach can be further extended to other category-based indices such as the Herfindahl index, entropy index, Euclidean distance, and Pearson's correlation coefficient, where their operations are based on the dot product between vectors. The adjustments for these indices are exemplified as follows:[12] [13]

[11] $x^T y$ can be regarded as the dot product of x and y with respect to the Euclidean metric I, where I represents unit matrix.

[12] The adjusted Euclidean distance becomes equivalent to the Mahalanobis distance if M^{-1} is the covariance matrix of the sample.

[13] The term $\ln\left(\frac{1}{p}\right)$ represents the column vector $\left(\ln\left(\frac{1}{p_1}\right), \ldots, \ln\left(\frac{1}{p_n}\right)\right)$ in the adjusted entropy index.

$$Adjusted\ Herfindahl\ index = 1 - p^T Mp$$

$$Adjusted\ entropy\ index = p^T M\ ln\left(\frac{1}{p}\right)$$

$$Adjusted\ Euclidean\ distance = \sqrt{(x-y)^T M(x-y)}$$

$$Adjusted\ Pearson's\ correlation\ coefficient\ = \frac{(x-\overline{x})^T M(y-\overline{y})}{\sqrt{(x-\overline{x})^T M(x-\overline{x}) * (y-\overline{y})^T M(y-\overline{y})}}$$

Interestingly, a similar solution has been found to correct the Herfindahl index, which is called the concentric index. It was originally defined by Caves et al. (1980) and applied in several studies (Montgomery & Wernerfelt, 1988; Robins & Wiersema, 1995; Wernerfelt & Montgomery, 1988). The definition is given below and its alternative expression is further developed in terms of vector and matrix operations:

$$d_{Concentric} = \sum_{j=1}^{n} p_j \sum_{i=1}^{n} p_i d_{ij} = p^T Dp$$

where $p(p_1, \ldots\ p_i, \ldots p_n)$ is the same as it is defined in the original and adjusted Herfindahl index. However, the resulting matrix D differs from this study's relatedness matrix M in that each element d_{ij} in the former refers to the distance (rather than the relatedness) between i and j. For example, according to its definition, when SIC is utilized in computation, d_{ij} is zero if i and j include the same three-digit SIC code (including d_{ij} when $i = j$), d_{ij} equals one if they have different three-digit codes but identical two-digit codes, and d_{ij} equals two if they include different two-digit codes (Caves et al., 1980; Montgomery & Wernerfelt, 1988; Wernerfelt & Montgomery, 1988). The distance measure can be regarded as the opposite of the relatedness measure. Thus, D can be scaled to [0, 1] by dividing the maximum value of two and then scaling the distance matrix $D' = \frac{1}{2}D$. In addition, all the distance scores d'_{ij} in D' are further subtracted from the maximum value of one to obtain the relatedness matrix M. Finally, D in the concentric

index can be replaced and represented in terms of M, and the relationship between the concentric index and the hierarchy-adjusted Herfindahl index can be derived:

$$M = \begin{bmatrix} 1 & \cdots & 1 - d'_{1n} \\ \vdots & \ddots & \vdots \\ 1 - d'_{n1} & \cdots & 1 \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix} - D'$$

$$d_{Concentric} = p^T Dp = 2p^T D'p = 2\left(p^T \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix} p - p^T Mp\right) = 2(1 - p^T Mp) = 2d_{Herfindahl}$$

The concentric index employs a hierarchical approach to measure the relatedness scores between non-independent categories. Instead, this study's approach for measuring relatedness scores in nature adopts a peer-to-peer (P2P) method in which all categories are treated as flat. This approach more closely reflects the inherent relatedness between any pair of related technology fields rather than indiscriminately assigning weights that merely depend on manually organized hierarchical structures. This is because these hierarchical categorization systems (such as SIC, IPC, and USPC) have been argued as having shortcomings of varying degrees of breadth in the classes (Rumelt, 1982). Results from the validity examination support the proposed P2P approach, in which the goodness-of-fit scores for the resulting index are improved provided that they are adjusted for relatedness. Moreover, the proposed cosine-based diversity index exhibits greater correlations and Cronbach's alpha across the IPC and USPC categorization schemes, which implies that this proposed approach can be applied, regardless of a particular categorization scheme. Moreover, this merit can be attributed to the proposed P2P approach.

An alternative perspective of understanding the benefits of this proposed approach might suggest that the relatedness is embedded in category-based indices as if they are controlled when regressions are conducted. However, these relatedness factors have drawn limited attention. But what if they are taken into consideration and controlled? In this case, is this proposed approach still attractive for empirical analysis? The answer is "yes." The introduction of these relatedness

factors may further incur multicollinearity problems as they are not inter-independent. This proposed approach provides a comprehensive solution to control for these factors, ensuring that there is no need to create dozens of control variables to capture the characteristics of all the categories.

In addition, another virtue of this study is that it provides a better way to compute the relatedness for technologies or industries. By harnessing co-occurrences as well as single occurrences together, it utilizes the relative complete information collected from empirical data. Although this study follows the approach by Teece et al., in which random technology assignment likelihood is assumed and a hypergeometric distribution is thus used to correct for occurrences, the inclusion of single occurrences is also applicable to other co-occurrences-based approaches such as revealed relatedness (Neffke & Henning, 2008).

Limitations

Despite the contributions and application of the present study's approach, there remain several limitations. First, the relatedness scores may have to be further refined if scholars need to employ category-based indices to a different domain. For example, in this study's approach, the angle between any two axes in an *n*-dimensional oblique coordinates system is non-obtuse. That is, it is assumed that categories would not be negatively associated. However, in other application domains, such a negative relationship might be needed. In this case, other methods for constructing the relatedness scores can be considered as replacements. Second, when the validity examination was conducted, only four-digit IPCs or three-digit USPCs were employed. Although applications of the proposed indices could be extended to other granularities, a cross-level application (e.g., regarding *A01B* and *B01* as two categories in application for a possible reason of miscategorization) may have difficulty determining the vectors for computation. In addition,

values calculated on different levels might differ greatly; thus, a cross-level comparison is not feasible. This flaw, which appears in both the original and proposed indices, should be resolved in future research even though there are limited needs for cross-level analyses in empirical research.

Hopefully, the new category-based indices proposed in this study can both encourage further empirical research as well as prompt scholars to reconsider taxonomy and measurement issues before applying these imperfect or even incorrect measures.

REFERENCES

Blau, P. M. 1977. Inequality and heterogeneity: A primitive theory of social structure: Free Press (New York).

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

Bryce, D. J., & Winter, S. G. 2009. A general interindustry relatedness index. ***Management Science***, 55(9): 1570–1585.

Bunderson, J. S., & Sutcliffe, K. M. 2002. Comparing alternative conceptualizations of functional diversity in management teams: Process and performance effects. ***Academy of Management Journal***, 45(5): 875–893.

Campbell, D. T., & Fiske, D. W. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. ***Psychological Bulletin***, 56(2): 81.

Caves, R. E., Porter, M. E., & Spence, A. M. 1980. Competition in the open economy: A model applied to Canada: Harvard University Press (Cambridge, Mass.).

Chatterjee, S. 1990a. Excess resources, utilization costs, and mode or entry. ***Academy of Management Journal***, 33(4): 780–800.

Chatterjee, S. 1990b. ***The gains to acquiring firms: The related principle revisited***. Paper presented at the Academy of Management Best Papers Proceedings.

Chatterjee, S., & Wernerfelt, B. 1991. The link between resources and type of diversification: Theory and evidence. ***Strategic Management Journal***, 12(1): 33–48.

Corrocher, N., Malerba, F., & Montobbio, F. 2007. Schumpeterian patterns of innovative activity in the ICT field. ***Research Policy***, 36(3): 418–432.

Cotropia, C. A., Lemley, M. A., & Sampat, B. 2013. Do applicant patent citations matter? ***Research Policy***, 42(4): 844–854.

Cronbach, L. J. 1951. Coefficient alpha and the internal structure of tests. ***Psychometrika***, 16(3): 297–334.

Dahlin, K. B., Weingart, L. R., & Hinds, P. J. 2005. Team diversity and information use. ***Academy of Management Journal***, 48(6): 1107–1123.

Engelsman, E. C., & van Raan, A. F. J. 1991. Mapping of technology: A first exploration of knowledge diffusion amongst fields of technology, ***Report to the Netherlands Ministry of Economic Affairs (1991) published in the Policy Studies on Technology and Economy (BTE) Series, Nr. 15 (ISSN 0923–3164), The Hague***.

Fasham, M. J. R. 1977. A comparison of nonmetric multidimensional scaling, principal components and reciprocal averaging for the ordination of simulated coenoclines, and coenoplanes. ***Ecology***, 58(3): 551–561.

Finkelstein, S., & Haleblian, J. 2002. Understanding acquisition performance: the role of transfer effects. ***Organization Science***, 13(1): 36–47.

Garcia-Vega, M. 2006. Does technological diversification promote innovation?: An empirical analysis for European firms. ***Research Policy***, 35(2): 230–246.

Gibbs, J. P., & Martin, W. T. 1962. Urbanization, technology, and the division of labor: International patterns. ***American Sociological Review***, 27(5): 667–677.

Givens, W. 1958. Computation of plain unitary rotations transforming a general matrix to triangular form. ***Journal of the Society for Industrial & Applied Mathematics***, 6(1): 26–50.

Gower, J. C. 1967. Multivariate analysis and multidimensional geometry. ***Journal of the Royal Statistical Society. Series D (The Statistician)***, 17(1): 13–28.

Harrison, D. A., & Klein, K. J. 2007. What's the difference? Diversity constructs as separation, variety, or disparity in organizations. ***Academy of Management Review***, 32(4): 1199–1228.

Herfindahl, O. C. 1950. ***Concentration in the steel industry***. Columbia University.

Hirschman, A. O. 1964. The paternity of an index. ***The American Economic Review***, 54(5): 761–762.

Hoskisson, R. E., Hitt, M. A., Johnson, R. A., & Moesel, D. D. 1993. Construct validity of an objective (entropy) categorical measure of diversification strategy. ***Strategic Management Journal***, 14(3): 215–235.

Householder, A. S. 1958. Unitary triangularization of a nonsymmetric matrix. ***Journal of the ACM (JACM)***, 5(4): 339–342.

Jackson, S. E., Brett, J. F., Sessa, V. I., Cooper, D. M., Julin, J. A., & Peyronnin, K. 1991. Some differences make a difference: Individual dissimilarity and group heterogeneity as correlates of recruitment, promotions, and turnover. ***Journal of Applied Psychology***, 76(5): 675.

Jaffe, A. B. 1986. Technological opportunity and spillovers of R&D: evidence from firms' patents, profits, and market value. ***The American Economic Review***, 76(5): 984–1001.

King, G., & Zeng, L. 2001. Logistic regression in rare events data. ***Political Analysis***, 9(2): 137–163.

Lanjouw, J. O., & Mark Schankerman. 2004. Protecting intellectual property rights: are small firms handicapped? ***Journal of Law and Economics***, 47(1): 45–74.

Lerner, J. 1994. The importance of patent scope: an empirical analysis. ***The Rand Journal of Economics***: 319–333.

Lien, L. B., & Klein, P. G. 2006. Relatedness and acquirer performance. ***Advances in Mergers & Acquisitions***, 5: 9–23.

Lubatkin, M., & Rogers, R. C. 1989. Research notes. Diversification, systematic risk, and shareholder return: a capital market extension of Rumelt's 1974 study. ***Academy of Management Journal***, 32(2): 454–465.

Mahalanobis, P. C. 1936. On the generalized distance in statistics. ***Proceedings of the National Institute of Sciences (Calcutta)***, 2: 49–55.

Miner, A. S., Haunschild, P. R., & Schwab, A. 2003. Experience and convergence: Curiosities and speculation. ***Industrial and Corporate Change***, 12(4): 789–813.

Montgomery, C. A. 1985. Product-market diversification and market power. ***Academy of management journal***, 28(4): 789–798.

Montgomery, C. A., & Wernerfelt, B. 1988. Diversification, Ricardian rents, and Tobin's q. ***The Rand Journal of Economics***: 623–632.

Neffke, F., & Henning, M. 2008. Revealed relatedness: Mapping industry space. ***Papers in Evolutionary Economic Geography***, 8.

Neffke, F., & Henning, M. 2012. Skill relatedness and firm diversification. ***Strategic Management Journal***: 34(3): 297–316.

Norton, E., & Tenenbaum, B. H. 1993. Specialization versus diversification as a venture capital investment strategy. ***Journal of Business Venturing***, 8(5): 431–442.

Palich, L. E., Cardinal, L. B., & Miller, C. C. 2000. Curvilinearity in the diversification–performance linkage: an examination of over three decades of research. ***Strategic Management Journal***, 21(2): 155–174.

Pearson, K. 1901. LIII. On lines and planes of closest fit to systems of points in space. ***The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science***, 2(11): 559–572.

Petruzzelli, A. M. 2011. The impact of technological relatedness, prior ties, and geographical distance on university–industry collaborations: A joint-patent analysis. ***Technovation***, 31(7): 309–319.

Putnam, R. D. 2007. E pluribus unum: Diversity and community in the twenty-first century the 2006 Johan Skytte Prize Lecture. ***Scandinavian Political Studies***, 30(2): 137–174.

Quintana-García, C., & Benavides-Velasco, C. A. 2008. Innovative competence, exploration and exploitation: The influence of technological diversification. ***Research Policy***, 37(3): 492–507.

Reyment, R. A., & Jöreskog, K. 1996. Applied factor analysis in the natural sciences: Cambridge University Press.

Robins, J., & Wiersema, M. F. 1995. A resource-based approach to the multibusiness firm: Empirical analysis of portfolio interrelationships and corporate financial performance. ***Strategic Management Journal***, 16(4): 277–299.

Rumelt, R. P. 1974. Strategy, structure, and economic performance: Harvard Business School Press.

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

Sampat, B. 2012. Examiner Citation Data. http://hdl.handle.net/1902.1/18735 UNF:5:mKUxLoeuskOler9Lq8l7cw== Bhaven Sampat [Distributor] V2 [Version].

Sampson, R. C. 2007. R&D alliances and firm performance: The impact of technological diversity and alliance organization on innovation. ***Academy of Management Journal ARCHIVE***, 50(2): 364–386.

Sayili, A. 1960. Thabit ibn Qurra's generalization of the Pythagorean theorem. ***Isis***, 51(1): 35–37.

Schmoch, U., Laville, F., Patel, P., & Frietsch, R. 2003. Linking technology areas to industrial sectors. ***Final Report to the European Commission, DG Research***: 1.

Simpson, E. H. 1949. Measurement of diversity. ***Nature***, 163(4148).

Stigler, G. J. 1983. The organization of industry: University of Chicago Press.

Tallman, S., & Li, J. 1996. Effects of international diversity and product diversity on the performance of multinational firms. ***Academy of Management Journal***, 39(1): 179–196.

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.

Trefethen, L. N., & Bau III, D. 1997. ***Numerical linear algebra***: Society for Industrial Mathematics.

Van Der Vegt, G. S., & Bunderson, J. S. 2005. Learning and performance in multidisciplinary teams: The importance of collective team identification. ***Academy of Management Journal***: 532–547.

Vaughn, M. T. 2008. ***Introduction to mathematical physics***: John Wiley & Sons.

Venkatraman, N., & Grant, J. H. 1986. Construct measurement in organizational strategy research: A critique and proposal. ***Academy of Management Review***: 71–87.

Wernerfelt, B., & Montgomery, C. A. 1988. Tobin's q and the importance of focus in firm performance. ***The American Economic Review***, 78(1): 246–250.

Wrigley, L. 1970. ***Divisional autonomy and diversification***. Harvard University.

Table 1. Summary of measures

Variables	Obs.	Mean	S.D.	Skewness	Min	Max
Cosine, IPC						
Number of overlapped categories	652653	1.4824	1.8920	2.0193	0	22
Ratio of overlapped categories	652653	0.0521	0.0706	2.6945	0	1
Original cosine diversity	652653	0.9368	0.1459	-3.6048	0.0002	1
Proposed cosine diversity	652653	0.6213	0.2029	-0.4498	0	1
Cosine, USPC						
Number of overlapped categories	652653	3.0227	3.4772	1.8396	0	38
Ratio of overlapped categories	652653	0.0650	0.0741	2.0806	0	1
Original cosine diversity	652653	0.9382	0.1375	-3.6539	0.0027	1
Proposed cosine diversity	652653	0.5910	0.1910	-.0.3785	0	1

Table 2. Multitrait-multimethod matrix for a cosine-based index (Pearson)

	IPC				USPC			
	Num.	Ratio	Original	Proposed	Num.	Ratio	Original	Proposed
IPC								
Number of overlapped categories								
Ratio of overlapped categories	*0.8545*							
Original cosine diversity	*−0.6022*	*−0.7023*						
Proposed cosine diversity	*−0.6875*	*−0.6763*	*0.6872*					
USPC								
Number of overlapped categories	**0.8201**	0.6545	−0.5510	−0.6936				
Ratio of overlapped categories	0.7276	**0.7865**	−0.6614	−0.7103	*0.8750*			
Original cosine diversity	−0.6424	−0.7132	**0.8334**	0.6646	*−0.6223*	*−0.7218*		
Proposed cosine diversity	−0.6925	−0.6468	0.6099	**0.8727**	*−0.7544*	*−0.7495*	*0.7021*	

n = 652,653. All correlation coefficients are significant ($p < 0.001$).

Table 3. Multitrait-multimethod matrix for a cosine-based index (Spearman)

	IPC				USPC			
	Num.	Ratio	Original	Proposed	Num.	Ratio	Original	Proposed
IPC								
Number of overlapped categories								
Ratio of overlapped categories	*0.9665*							
Original cosine diversity	*−0.9171*	*−0.9100*						
Proposed cosine diversity	*−0.6700*	*−0.6560*	*0.7117*					
USPC								
Number of overlapped categories	**0.7623**	0.7055	−0.7189	−0.6933				
Ratio of overlapped categories	0.7322	**0.7183**	−0.7074	−0.6864	*0.9709*			
Original cosine diversity	−0.7409	−0.7108	**0.7586**	0.7301	*−0.9130*	*−0.9103*		
Proposed cosine diversity	−0.6777	−0.6377	0.6830	**0.8519**	*−0.7832*	*−0.7647*	*0.8227*	

n = 652,653. All correlation coefficients are significant ($p < 0.001$).

Table 4. Cronbach's alpha

Variables	Alpha	Alpha (std.)
Number of overlapped categories	0.8156	0.9011
Ratio of overlapped categories	0.8799	0.8805
Original cosine diversity	0.9082	0.9091
Proposed cosine diversity	0.9311	0.9320

Table 5. Descriptive statistics

Variables	1	2	3	4	5	6	7	8	9	10
1. Number of inter-citations										
2. Number of overlapped categories	0.1127									
3. Ratio of overlapped categories	0.1210	0.8545								
4. Original cosine diversity	−0.1585	−0.6022	−0.7023							
5. Hierarchy-adjusted cosine diversity	−0.1146	−0.6312	−0.6495	0.6986						
6. Proposed cosine diversity	−0.1072	−0.6875	−0.6763	0.6872	0.7536					
7. Number of prior patents 1	0.0259	0.1799	0.0694	−0.0432	−0.0479	−0.0713				
8. Number of prior patents 2	0.0226	0.1654	0.0657	−0.0214	−0.0232	−0.0438	0.0003			
9. Experience years 1	0.0231	0.0764	0.0159	−0.0186	−0.0236	−0.0311	0.0674	0.0266		
10. Experience years 2	0.0269	0.1026	0.0535	−0.0566	−0.0572	−0.0577	0.0082	0.0871	0.2921	
Mean	0.0518	1.4824	0.0521	0.9368	0.7755	0.6213	52.9935	51.6872	21.5810	11.6875
S.D.	0.8152	1.8920	0.0706	0.1459	0.2280	0.2029	19.1868	18.6492	10.4660	7.9719
Min	0	0	0	0.0002	0.0002	0	30	30	2	2
Max	169	22	1	1	1	1	99	99	101	101

n = 652,653.

Table 6. Logistic estimates for inter-citations and inter-firm technological diversity (IPC)

Variables	(1) w/o controls	(2) w/o controls	(3) w/o controls	(4) w/o controls	(5) w/o controls	(6) w/ controls	(7) w/ controls	(8) w/ controls	(9) w/ controls	(10) w/ controls
Number of overlapped categories	0.4126***					0.3852***				
	(0.0027)					(0.0029)				
Ratio of overlapped categories		9.0612***					9.1213***			
		(0.0699)					(0.0703)			
Original cosine diversity			−4.0663***					−4.0494***		
			(0.0245)					(0.0259)		
Hierarchy-adjusted cosine diversity				−4.3477***					−4.2929***	
				(0.0291)					(0.0301)	
Proposed cosine diversity					−6.0197***					−5.8567***
					(0.0387)					(0.0399)
Number of prior patents 1						0.0051***	0.0155***	0.0153***	0.0142***	0.0136***
						(0.0005)	(0.0004)	(0.0004)	(0.0004)	(0.0004)
Number of prior patents 2						0.0026***	0.0127***	0.0143***	0.0133***	0.0128***
						(0.0005)	(0.0004)	(0.0004)	(0.0004)	(0.0004)
Experience years 1						0.0201***	0.0247***	0.0248***	0.0237***	0.0204***
						(0.0007)	(0.0006)	(0.0006)	(0.0006)	(0.0006)
Experience years 2						0.0195***	0.0263***	0.0238***	0.0243***	0.0209***
						(0.0009)	(0.0008)	(0.0009)	(0.0009)	(0.0009)
Constant	−4.8023***	−4.5359***	−0.2559***	−0.9760	−0.7830***	−5.8842***	−7.0519***	−2.8084***	−3.4109***	−3.0806***
	(0.0122)	(0.0108)	(0.0211)	(0.0167)	(0.0169)	(0.0395)	(0.0399)	(0.0438)	(0.0423)	(0.0426)
Chi^2	23828	16812	27658	22259	24149	25666	22937	30753	27151	28611
df	1	1	1	1	1	5	5	5	5	5

n = 652,653. Robust standard errors in parentheses.

*** p < 0.01, ** p < 0.05, * p < 0.1.

Table 7. Negative binomial estimates for number of inter-citations and inter-firm technological diversity (IPC)

Variables	(1) w/o controls	(2) w/o controls	(3) w/o controls	(4) w/o controls	(5) w/o controls	(6) w/ controls	(7) w/ controls	(8) w/ controls	(9) w/ controls	(10) w/ controls
Number of overlapped categories	0.6566***					0.6303***				
	(0.0083)					(0.0085)				
Ratio of overlapped categories		17.6423***					16.9757***			
		(0.2120)					(0.2074)			
Original cosine diversity			−6.6685***					−6.3513***		
			(0.0804)					(0.0639)		
Hierarchy-adjusted cosine diversity				−5.3443***					−5.3712***	
				(0.0500)					(0.0504)	
Proposed cosine diversity					−7.3852***					−7.3461***
					(0.0737)					(0.0724)
Number of prior patents 1						0.0051***	0.0147***	0.0160***	0.0159***	0.0142***
						(0.0011)	(0.0008)	(0.0006)	(0.0007)	(0.0007)
Number of prior patents 2						0.0021*	0.0114***	0.0152***	0.0146***	0.0132***
						(0.0012)	(0.0008)	(0.0007)	(0.0008)	(0.0008)
Experience years 1						0.0250***	0.0420***	0.0444***	0.0388***	0.0378***
						(0.0019)	(0.0017)	(0.0016)	(0.0015)	(0.0016)
Experience years 2						0.0197***	0.0247***	0.0249***	0.0245***	0.0223***
						(0.0023)	(0.0017)	(0.0015)	(0.0015)	(0.0016)
Constant	−4.7955***	−4.7625***	2.5103***	0.1647	0.3637***	−5.9742***	−7.5126***	−0.9397***	−2.7938***	−2.3948***
	(0.0226)	(0.0203)	(0.0766)	(0.0334)	(0.0345)	(0.0942)	(0.0752)	(0.0827)	(0.0731)	(0.0721)
Chi^2	6289	6925	6876	11446	10048	6653	9084	11627	13257	12119
df	1	1	1	1	1	5	5	5	5	5

n = 652,653. Robust standard errors in parentheses.

*** p < 0.01, ** p < 0.05, * p < 0.1

Table 8. Comparison of the goodness-of-fit for logistic models (IPC)

Model (LOGIT)	Proposed cosine index	Difference (w/ original cosine index)	Difference (w/ hierarchy-adjusted cosine index)
Log-Lik Intercept	−71053.418	0.000	0.000
Log-Lik Full	−57136.656	2972.966	1202.112
Likelihood Ratio	27833.525	5945.932	2404.224
McFadden's R^2	0.196	0.042	0.017
McFadden's Adj R^2	0.196	0.042	0.017
Maximum Likelihood R^2	1.000	0.000	0.000
Cragg & Uhler's R^2	1.000	0.000	0.000
McKelvey & Zavoina's R^2	0.348	0.172	0.063
Efron's R^2	0.068	0.018	0.004
Count R^2	0.977	0.000	0.000
Adj Count R^2	−0.001	0.016	0.001
AIC	114285.312	−5945.932	−2404.224
BIC	114353.645	−5945.932	−2404.224

Table 9. Comparison of the goodness-of-fit for negative binomial models (IPC)

Model (NBREG)	Proposed cos.	Difference (w/ original cosine index)	Difference (w/ hierarchy-adjusted cosine index)
Log-Lik Intercept Only	−91592.545	0.000	0.000
Log-Lik Full Model	−77303.332	2038.500	1159.180
Likelihood Ratio	28578.426	4077.001	2318.361
McFadden's R^2	0.156	0.022	0.013
McFadden's Adj R^2	0.156	0.022	0.013
Maximum Likelihood R^2	1.000	0.000	0.000
Cragg & Uhler's R^2	1.000	0.000	0.000
AIC	154620.664	−4077.000	−2318.361
BIC	154700.386	−4077.000	−2318.361

APPENDIX

Tables for criterion-related validation of a cosine-based diversity index under the USPC scheme

Logistic estimates for inter-citations and inter-firm technological diversity (USPC)

Variables	(1) w/o controls	(2) w/o controls	(3) w/o controls	(4) w/o controls	(5) w/ controls	(6) w/ controls	(7) w/ controls	(8) w/ controls
Number of overlapped categories	0.2291***				0.2149***			
	(0.0015)				(0.0017)			
Ratio of overlapped categories		10.1403***				9.9879***		
		(0.0762)				(0.0755)		
Original cosine diversity			−4.5119***				−4.4644***	
			(0.0257)				(0.0274)	
Proposed cosine diversity				−6.8382***				−6.6620***
				(0.0431)				(0.0449)
Number of prior patents 1					0.0031***	0.0137***	0.0159***	0.0137***
					(0.0005)	(0.0004)	(0.0004)	(0.0004)
Number of prior patents 2					0.0015***	0.0112***	0.0148***	0.0130***
					(0.0005)	(0.0004)	(0.0004)	(0.0004)
Experience years 1					0.0179***	0.0226***	0.0231***	0.0192***
					(0.0007)	(0.0006)	(0.0006)	(0.0006)
Experience years 2					0.0196***	0.0260***	0.0206***	0.0181***
					(0.0009)	(0.0008)	(0.0009)	(0.0009)
Constant	−4.8861***	−4.8115***	0.1308***	−0.5713***	−5.7431***	−7.0649***	−2.4277***	−2.8228***
	(0.0125)	(0.0123)	(0.0221)	(0.0175)	(0.0391)	(0.0396)	(0.0445)	(0.0429)
Chi^2	23068	17713	30922	25194	25247	24124	32841	28966
df	1	1	1	1	5	5	5	5

Robust standard errors in parentheses.
*** p < 0.01, ** p < 0.05, * p < 0.1.

Negative binomial estimates for number of inter-citations and inter-firm technological diversity (USPC)

Variables	(1) w/o controls	(2) w/o controls	(3) w/o controls	(4) w/o controls	(5) w/ controls	(6) w/ controls	(7) w/ controls	(8) w/ controls
Number of overlapped categories	0.3729***				0.3646***			
	(0.0047)				(0.0048)			
Ratio of overlapped categories		17.9215***				17.3663***		
		(0.1861)				(0.1830)		
Original cosine diversity			−7.3170***				−6.9101***	
			(0.0842)				(0.0672)	
Proposed cosine diversity				−8.2255***				−8.1273***
				(0.0780)				(0.0772)
Number of prior patents 1					0.0021*	0.0126***	0.0164***	0.0147***
					(0.0011)	(0.0008)	(0.0006)	(0.0007)
Number of prior patents 2					−0.0007	0.0092***	0.0157***	0.0137***
					(0.0012)	(0.0008)	(0.0007)	(0.0008)
Experience years 1					0.0246***	0.0407***	0.0400***	0.0338***
					(0.0019)	(0.0017)	(0.0015)	(0.0015)
Experience years 2					0.0214***	0.0246***	0.0228***	0.0195***
					(0.0026)	(0.0017)	(0.0015)	(0.0016)
Constant	−4.9678***	−5.1696***	3.0383***	0.5283***	−5.8633***	−7.6501***	−0.3967***	−2.1752***
	(0.0307)	(0.0226)	(0.0803)	(0.0339)	(0.0927)	(0.0709)	(0.0859)	(0.0707)
Chi^2	6400	9274	7546	11110	7453	11241	13128	13405
df	1	1	1	1	5	5	5	5

Robust standard errors in parentheses.
*** p < 0.01, ** p < 0.05, * p < 0.1.

Comparison of the goodness-of-fit for logistic models (USPC)

Model (LOGIT)	Proposed cos.	Difference (w/ Original cos.)
Log-Lik Intercept	−71053.418	0.000
Log-Lik Full	−56193.131	2848.538
Likelihood Ratio	29720.574	5697.076
McFadden's R^2	0.209	0.040
McFadden's Adj R^2	0.209	0.040
Maximum Likelihood R^2	1.000	0.000
Cragg & Uhler's R^2	1.000	0.000
McKelvey & Zavoina's R^2	0.374	0.194
Efron's R^2	0.078	0.019
Count R^2	0.977	0.000
Adj Count R^2	−0.002	0.021
AIC	112398.262	−5697.076
BIC	112466.595	−5697.076

Comparison of the goodness-of-fit for negative binomial models (USPC)

Model (NBREG)	Proposed cosine index	Difference (w/ original cosine index)
Log-Lik Intercept Only	−91592.545	0.000
Log-Lik Full Model	−76436.076	1570.170
Likelihood Ratio	30312.939	3140.340
McFadden's R^2	0.165	0.017
McFadden's Adj R^2	0.165	0.017
Maximum Likelihood R^2	1.000	0.000
Cragg & Uhler's R^2	1.000	0.000
AIC	152886.152	−3140.340
BIC	152965.874	−3140.340

Definition of the goodness-of-fit measures

Measures of goodness-of-fit	Definition
Likelihood Ratio	$2(LL_M - LL_0)$
McFadden's R^2	$1 - \frac{LL_M}{LL_0}$
McFadden's Adjusted R^2	$1 - \frac{LL_M - K}{LL_0}$
Maximum Likelihood R^2	$1 - (e^{LL_0 - LL_M})^{2/N}$
Cragg & Uhler's R^2	$\frac{1 - (e^{LL_0 - LL_M})^{2/N}}{1 - (e^{LL_0})^{2/N}}$
McKelvey & Zavoina's R^2	$1 - \frac{Var(\varepsilon)}{Var(Y^*)}$
Efron's R^2	$1 - \frac{\sum(y_i - y_i)^2}{\sum(y_i - \overline{y})^2}$
Count R^2	$\frac{n}{N}$
Adjusted Count R^2	$\frac{n - \max(observed\ successes, observed\ failures)}{N - \max(observed\ successes, observed\ failures)}$
AIC	$-2(LL_M - K)$
BIC	$-2LL_M + K * \ln N$

LL_M: log likelihood for the full model.
LL_0: log likelihood for the null model.
N: number of observations.
K: number of parameters.
n: number of correctly classified observations.
$Var(\varepsilon)$: variance of error.
$Var(Y^*)$: variance of estimated dependent variables.