Technological Complexity, Valuation and Drives

September 3, 2025

Abstract

This paper measures industry-level technological complexity using the Network Diversity Score, which captures how varied the connections between knowledge components are. This approach outperforms other complexity metrics in key tests. The findings show that complexity rose steadily at both sector and industry levels until the Covid-19 pandemic, after which labor-intensive industries experienced a sharp decline. An excess complexity score is introduced to compare scores across industries and use it in our regression models. The results reveal that, in less competitive markets, complexity is driven by advancement in innovation and hurt by influence level, whereas, in more competitive markets, it's purely driven by revenue.

1 Introduction

Although technological complexity and its links to economic factors have been studied extensively, a universally applied method of quantifying the complexity score is still lacking. This paper applied a quantifying method using topology theory on U.S. patent data to compute annual complexity scores for each industry. We then compare our approach with existing measures to highlight its advantages and investigate the key drivers of high complexity scores at the industry level.

Patents are the primary output of the inventive process and are a well-established proxy for innovation. Scholars have explored patent data on its own (e.g., Mariani, Medo and Lafond (2019) use the patent citation network to identify important technologies; Kelly, Papanikolaou, Seru and Taddy (2021) use text-based method to create new indicators on important patents; and Bowen III, Frésard and Hoberg (2023) also use patent text to position startups in business cycles) and in combining with stock market data (Kogan, Papanikolaou, Seru and Stoffman (2017)). Innovation continually adds new features, components, and processes to existing technologies. Over serial waves of innovation, these additional compounds, collaborating with existing ones, transform once-simple tools into richly elaborate solutions. Ultimately, the inventing action drives technological complexity upward. Surprisingly, few studies have linked patent data to formal measures of that complexity except Broekel (2019). This paper builds on and expands his approach for a wider industry-level application and uncovers the main factors driving complexity.

In Arthur (2009), technology is defined as a system of interconnected components used for practical purposes. As described in Broekel (2019), knowledge components (10-digit CPC subclass) are turned into nodes, and their interactions (co-occurrence in a patent) into edges and formed networks. This approach outstands existing and potential complexity measures through two comparisons mentioned in this paper. First, when it is compared with other existing measures for complexity, this methods performs better when relating to R&D intensity (Broekel (2019)), which is commonly associated with technological complexity. Second, instead of relying on a single network feature, it combines four different structural indicators at once, providing the ability to classify networks according to the heterogeneity degree of the interconnections between nodes.

A two-step conversion process has been developed to merge our CPC-based complexity scores with NAICS-classified economic data. First, CPC codes are mapped to the initial three NAICS versions using the ALP probabilistic links from Nathan, Lybbert and Jason (2019). Since this study spans seven NAICS revisions, we then chain together adjacent NAICS versions using probability weights from the Census Bureau's concordance files. Because complexity scores must stay on the same scale during conversions and the patent/citation count should be preserved at an aggregated level, separate weighting schemes and formulas are applied for each measure – details are laid out below. This conversion framework is a key contribution, as it can be applied to translate any CPC-based variable to NAICS or across multiple NAICS versions.

The distributions of complexity scores—whether under CPC or NAICS – follow the same pattern: overall complexity rises, with more technologies scoring high and fewer scoring zero. This fits the cumulative nature of innovation, as each wave builds on earlier work. Breaking the trend down by sector shows a steady increase across all areas until 2020, followed by a drop. At the industry level, that decline is concentrated in labor-intensive fields, while high-tech industries see no significant fall. However, the industry-level graph also shows that complexity scores are not comparable across industries, making comparisons and direct regression at the industry level in future analysis challenging. To fix this, a benchmark (the average score over time for each industry) is necessary to be subtracted from the initial complexity score. The resulting "excess" complexity score is then comparable across all industries. this paper is the first to propose the industry-specific benchmark concept for industrial structural diversity score.

To uncover what drives complexity, distinct determinants are found for three distinct samples: the full dataset and two subsamples defined by the industry HHI. First, in the full sample, both accounting factors and patent activity matter: larger industries with higher profits, lower reliance on debt, and faster patent growth show greater complexity. It is worth noticing that the scientific influence hurts the industrial complexity score, which shows that areas with high patent citations would form uniform frameworks, and citations can not be regarded as an index of complexity level. In less competitive industries, only industries with higher revenue score higher in technological complexity. The allocation of more resources, a higher risk tolerance level, or both can cause this. By contrast, in highly competitive markets, patent activity is the key to earning a higher score

– inventing better technologies or filing patents strategically are the main ways firms distinguish themselves.

The rest of this paper is organized as follows. Section 2 explains how we construct the technological complexity score, compares it with other measures, and describes the conversions to and between NAICS versions. Section 3 presents the trends and key properties of the complexity scores over time. Section 4 details our data and regression models and explores the drivers of complexity across three sample splits. Finally, Section 5 offers our conclusions.

2 Construction of Complexity Score

2.1 Idea

In network theory, networks are broadly classified into three types: ordered networks, which exhibit predictable connections between nodes and present low heterogeneity; random networks, in which the connections are determined purely by chances, so the heterogeneity degree is the largest; and complex networks, which lie between those extremes, is a mixture of ordered & random networks. Building on this classification, the industrial complexity score in this paper is calculated using the Network Diversity Score (NDS) — initially proposed by Emmert-Streib and Dehmer (2012) in a biological context. Unlike other network-based metrics, NDS simultaneously uses four structural indicators, enabling it to distinguish between ordered, complex, and random networks.

Emmert-Streib and Dehmer (2012) originally proposed NDS as a purely theoretical metric to quantify the complexity of a network. Broekel (2019) later adapted this measure to the patent world. Technologies have been viewed as combinations of interconnected elements (Arthur (2009), McNerney, Farmer, Redner and Trancik (2011)). These elements — ranging from physical parts and functional modules to individual steps in an industrial process — are considered linked whenever a change in one alters the behavior or performance of another. In Broekel (2019), a technology is described as a recipe, in which the information includes the knowledge components and the combination of how these components are combined. For example, the same basic inputs (flour,

eggs, milk, oil) can produce something as simple as scrambled eggs or something as elaborate as a multi-layer cake. Metaphoring knowledge components as ingredients and technologies as recipes illustrate that complexity depends less on the ingredients themselves and more on how knowledge components are linked with each other, specifically on the heterogeneity of their connections. By applying NDS to technological networks, we can rigorously quantify how diversely knowledge components interact.

2.2 Construction

By modeling each 4-digit CPC code as a network, the NDS represents the diversity score of the networks, and each technology (4-digit CPC code) is regarded as a combination of knowledge components (Broekel (2019)). Therefore, the complexity of each depends on how these knowledge components are linked to each other. The more diverse and complex the direct or indirect connection, the more complex this technology is. The author considered the lowest level of CPC classes (10-digit CPC subclasses) as knowledge components and their co-occurrence on patent as linkages (Fleming and Sorenson (2001), Sorenson and Fleming (2004)).

The data files used to construct the networks are from USPTO's PatentsView as of July 2024. For each 4-digit CPC code, a network is built annually using all patents classified into this 4-digit CPC class with a moving window of 3 years. The nodes are the 10-digit CPC subclasses belonging to this 4-digit class and have existed in the patent files of the time range. The edges between two subclasses are defined as existing if these two subclasses ever co-occurred in one of the patent files. For each 4-digit CPC class, we first build its network and extract the giant component — the largest connected cluster of nodes and edges. Following Broekel (2019), we then randomly pick 50 nodes from that component. From each chosen node, we run a 150-step random walktrap to generate a subnetwork. Computing the iNDS on each of these 50 subnetworks by using four structural indicators and averaging the results gives the annual complexity score for the CPC class. The equations for iNDS and NDS are the followings with $S = 50$:

$$iNDS(G_T) = \frac{\alpha_{module} \cdot r_{graphlet}}{v_{module} \cdot v_\lambda} \tag{1}$$

$$NDS(\{G_T^S | G_M\}) = \frac{1}{S} \sum_{G_T \in G_M}^{S} iNDS(G_T) \tag{2}$$

$$Score(G_M) = -\log NDS(\{G_T^S | G_M\}) \tag{3}$$

The four network structural indicators in equation 1 represent different properties of a network. $\alpha_{module} = \frac{M}{n}$ with M being the number of modules (densely connected subgroups) and n being the number of nodes. This measurement represents the module density of a network, namely, how many densely connected groups there are compared to the overall size. It gives a higher value for a less heterogeneous network and helps distinguish between ordered, complex, and random networks. $v_{module} = \frac{var(m)}{mean(m)}$ with m being a vector containing sizes of all modules, showing how different in size the modules are from each other. This variable mainly distinguishes the networks within the three types of networks and gives higher values for higher heterogeneity. $V_\lambda = \frac{var(\Lambda(L))}{mean(\Lambda(L))}$ with $\Lambda(L)$ representing the eigenvalues of Laplacian matrix for the network tells how well a graph is connected. So V_λ is also higher for more heterogeneous networks within the three types of networks. The last indicator $r_{graphlet} = \frac{N_{graphlet(3)}}{N_{graphlet(4)}}$ and $N_{graphlet(a)}$ is number of graphlets of size a. This captures how do small subgroup in the network growth, or compares the number to patterns with three nodes to the number of patterns with 4 nodes. It helps to classify networks into the three types since $r_{graphlet}$ is smallest for random networks, largest for ordered networks and moderate for complex networks.

Figure 1 shows the changes in complexity score at the level of 1-digit CPC industry by averaging the 4-digit level score. The complexity score shows a tendency of increase across years for all industries before 2020. In contrast, decreases occur in all industries after the pandemic, and the sharpest drop happens in Fixed Constructions. This is mirrored by an increase in work-from-home cases and a decrease in the need for transportation and office buildings. Considering the fact that patent application is a revenue-driven activity from the perspective of firms, uncertainty about future policies and market demand might cause fewer patents to be applied and the technological complexity score to decline correspondingly.

The change in the big-picture of complexity score is clearer in figure 2. It shows the distribution of the industrial complexity scores of almost 672 technologies every year from 1982 to 2024. The shading darkens as time goes on. For each year, the distribution is bimodal with a peak at zero and bell-shaped at moderate values. The peaks at zero are caused by the fact that many technologies have too few patents to construct a complicated network and to calculate a score. As years pass, growing patent numbers lead the density at zero to become lower. Furthermore, the maximum density moving to the right side is consistent with technologies being built upon existing ones. This figure shows a similar trend with EU technologies' changes as described in Broekel (2019) using EU patent data.

2.3 Comparison

The advantages of NDS become clear both when compared with traditional complexity measures that ignore network topology and in head-to-head comparisons with other network-based metrics, where it consistently outperforms.

The first part was already discussed and studied by Dr. Broekel in Broekel (2019). The author carried out a direct comparison of NDS against Fleming's modular complexity (FS modular; Fleming and Sorenson (2001)) and Balland's technological complexity index (KCI; Balland and and (2017)) to see how each relates to R&D intensity. NDS correlates positively with patent counts and with collaborative R&D (measured by inventor numbers), assigns higher scores to high-tech fields in every period, and gives older technologies lower complexity scores. In contrast, FS modular increases complexity for older technologies, which contrasts the stylized fact. KCI not only relates negatively to patenting and collaborative R&D but also ranks older technologies as more complex. This comparison highlights that of these three, the structural diversity method (NDS) reflects certain features commonly associated with technological complexity.

The comparison between structural diversity (NDS) and other network-based metrics is theoretically mentioned in Emmert-Streib and Dehmer (2012). Other 16 network complexity measures are also studied in Emmert-Streib and Dehmer (2012) and compared with NDS. These measures differ

from NDS in two main ways. First, NDS combines four distinct structural indicators, whereas the others each rely on a single principle. Second, it constructs the estimation by sampling a finite set of networks from the population and applying the central limit theorem, which helps to eliminate ambiguity from one simple network. The authors contended that NDS is the only system capable of distinguishing ordered, complex, and random networks unambiguously by capturing the heterogeneity in how nodes are connected. In this paper, three traditional network diversity measures are applied to patent data and evaluated alongside the NDS. The three measures are the Bertz index, Wiener index, and mean distance deviation.

Bertz index expresses the total structural information content of a graph and takes the formula (Bertz (1983)):

$$B(G) = 2n\log(n) - \sum_{i=1}^{k} |N_i| \log(|N_i|),\ n = \text{number of nodes} \tag{4}$$

Wiener index is one of the most widely used topological indices in graph theory. It captures the overall compactness by summing up the pairwise length of the shortest path between nodes. The formula for this measurement is (Wiener (1947)):

$$W(G) = \sum_{\{u,v\}\subset V} d(u,v),\ d(u,v) = \text{distance between nodes } u \text{ and } v \tag{5}$$

Mean distance deviation quantifies how spread out the pairwise shortest-path distances are around their average. It complements the Wiener index (which sums distances) by capturing dispersion rather than the total sum using the distances between nodes. In consequence, it offers insight into heterogeneity—information that neither the Bertz index nor the Wiener index can provide. It is defined as (Skorobogatov and Dobrynin (1988), Todeschini and Consonni (2008)):

$$MDD(G) = \frac{1}{n}\sum_{i=1}^{n} |\mu(v_i) - \frac{2W(G)}{n}| \tag{6}$$

where

$$\mu(v_i) = \sum_{j=1}^{n} d(v_i, v_j) \tag{7}$$

When complexity over time for two areas – environmental area (Figure 3) and AI tech (Figure 4) – are reported, apparent differences can be shown. In the environmental sector, the structural diversity score, Bertz index, and Wiener index all follow a similar, steadily increasing path, while the mean distance deviation (MDD) jumps around with no clear pattern. In contrast, for AI technologies, the structural diversity score rises gently up to 2020 and then spikes up sharply from 2021 onward, matching the recent AI boom. Of the other three metrics, only the Wiener index shows a noticeable uptick after 2021 and a gradual rise before that.

The contrasts become even clearer when we aggregate the data up to the 1-digit CPC level (Figure 1 for structural diversity score). The Wiener index (Figure 5) and MDD (Figure 6) stay flat over the years, and their scales differ broadly across sectors to be comparable. The Bertz index (Figure 7) does increase across every sector, but is not yet comparable, because the scores in the electrical sector significantly exceed the rest.

2.4 Conversion to NAICS

The classification system used by the USPTO's dataset is the Cooperative Patent Classification (CPC). When calculating the industrial level complexity score using the patent file from USPTO, the classification is documented under CPC, so one score is generated for each four-digit CPC code in each year. Conversion to other forms of classification is needed since the CPC is used to manage patent documents. The North American Industry Classification System (NAICS) is the standard code used by federal agencies to classify business establishments for the purpose of analyzing statistical data related to the U.S. business economy. Therefore, in order to link the industrial complexity score with Compustat and CRSP data, the score needs to be converted from the CPC to the NAICS form. The detailed conversion processes for the complexity score can be found in Appendices A.1 and B.1.

The North American Industry Classification System (NAICS) was first adopted in 1997 to replace the Standard Industrial Classification (SIC) system. Since then, NAICS has been scheduled to be reviewed every five years for potential revisions, and will be for the foreseeable future, so that it can keep pace with the emerging industries and make clarifications for select industry definitions. Since the time range in my dataset is 1985 - 2023, the relevant NAICS versions include NAICS 1997, NAICS 2002, NAICS 2007, NAICS 2012, NAICS 2017, and NAICS 2022. Historical NAICS codes (naicsh) from Compustat, which represent each firm's NAICS classification for each fiscal year, were used, applying the relevant NAICS version for that year. The crosswalks from CPC to NAICS are only feasible for the first three NAICS versions, which are the NAICS 1997, NAICS 2002, and NAICS 2007, so the conversion between different versions of NAICS is also necessary. In this paper, two types of conversion will be applied. If the analysis needs to link the industrial complexity score with Compustat, conversion to the corresponding version will be applied. If only comparison within complexity score data across years is needed, conversion to the NAICS 2007 version is exercised for easier interpretation.

3 Properties

From this section on, the complexity score will be shown at 4-digit NAICS level by averaging the affiliated 6-digit NAICS codes' scores. This conversion is necessary because further analysis will be based on industry-level by aggregating firm-level data from Compustat or CRSP, and the firm number categorized in each 6-digit NAICS industry is limited in each fiscal year. In order to increase the research power at the industry level, 4-digit NAICS are the better choice because they include more firms each year, facilitating a more comprehensive industry-level view.

Figure 8 shows the distribution of the complexity scores at the level of 4-digit NAICS, in which the outlier represents the year 2024 with incomplete granted patent data. It shows a similar trend with figure 2; majority density moves to the right side, and the left tail becomes thinner, indicating that the overall technologies are becoming more complex over the years. The disappearance of peaks at zero is caused by the conversion from CPC to NAICS, and there are few cases where there are

not enough patents to calculate for all 4-digit CPC classes one NAICS code mapped to.

Compared to figure 8, which presents the change in the technological complexity score on a larger scale, figure 9 selects eight industries and shows the annual complexity scores and employment for each industry. They are, from left to right and top to bottom: Agriculture Group; Air and Rail Transportation; Computer and Electronic Product Manufacturing; Data Processing; Motor Vehicle Manufacturing; Oil and Gas Extraction; Software; and Textile Mills. The chosen industries, ranging from labor-intensive to high-tech areas, all exhibit climbing scores until 2020. After the Covid-19 pandemic struck, labor-intensive industries suffered reductions in complexity scores, while high-tech ones remained stable. Given that labor is one of the main input capital resources, in order to find out whether the drops in labor-intensive industries are paralleled with a decrease in employment, employment data from the Bureau of Labor Statistics's OEWS dataset is used and represented by red lines in the graphs. It turns out that the employment statistics remained steady for these industries and can not account for the drop in complexity scores.

Another observation from Figure 9 is that the complexity score varies across industries, making direct comparisons challenging. For example, the Agriculture Group industry has scores ranging from 12.79 (in 1984) to 33.82 (in 2018), while the Data Processing industry spans from 25.60 (in 1984) to 46.97 (in 2020). High-tech industries' technological complexity levels from forty years ago are estimated to be similar to those of labor-intensive industries today (Agriculture Group scored 25.90 in 2023). This difference arises because each industry is inherently distinct, and the initial scoring system captures this because of the variances in the patent data. Consequently, it is necessary to establish an industry-specific benchmark to enable comparisons of technological complexity scores across industries. Each benchmark is determined by averaging the industry's scores over the entire time period, and the excess score is calculated by subtracting this benchmark from the initial complexity score. Figure 10 presents the same eight industries' excess scores from 1982 to 2023, and after adjusting by subtracting the benchmark, the scores become more comparable across industries.

Table 1 summarizes excess score statistics at the 2-digit NAICS level (the system's broadest category). Column 1 is the 2-digit NAICS sector. Column 2 reports, for each 2-digit NAICS sector,

the total number of panel observations in the data set from 1982 to 2023. Column 3 shows the total count of 4-digit NAICS codes according to the Census Bureau, and Column 4 reports how many of those have valid excess scores over the same period. Finally, Columns 5 and 6 present the median and standard deviation of the excess scores for each sector. There are substantial gaps between Columns 3 and 4 in several sectors, most notably 42 (Wholesale Trade), 48 (Transportation), 53 (Real Estate and Rental and Leasing), and 81 (Other Services). In these industries, fewer 4-digit codes have measurable complexity scores because they generate relatively few patents – reflecting inherently limited opportunities for innovation, either because technological improvement isn't central to their processes or because they've largely reached an innovation plateau. Figure 11 presents a scatter plot of each 2-digit NAICS sector over time, serving as a graphical counterpart to Table 1. The consistent upward trends in every sector further confirm that technological complexity has increased over the years.

4 Drives of Complexity

4.1 Data

In this section, regressions are used to discuss the determinants of complexity at the industry level (4-digit NAICS). Patent activity are incorporated by counting, for each CPC code in year t, the number of patents filed ("*patent_count*") and the citations they received within one, two, and three years of filing (named "*citation_1year*", "*citation_2years*", and "*citation_3years*", respectively). These data come from the USPTO's PatentsView database. The annual growth rates of those measures are calculated, along with the three-year growth rate of patent count. In the regressions, lagged versions of these variables are employed – for example, "*L.citation_1year*" in year t measures the one-year citations accrued by patents filed in year $t-1$ (up to the end of t for patents filed at the end of $t-1$), and similarly for "*L2.citation_2years*" and "*L3.citation_3years*" – so that all patent measures reflect prior activity when explaining current complexity.

For this section, all analyses will be shown and discussed in conversion to the corresponding

NAICS version since the patent-related data (complexity score, patent count, and citation count) need to be linked with Compustat. For the data within 1985-2011, the ALP probabilistic linkages from Nathan et al. (2019) are used to map data to the relevant NAICS version. For 2012–2023 – when later NAICS versions lack direct CPC links – data are first converted NAICS 2007 first using the CPC-NAICS07 linkages, then, conversions between adjacent NAICS versions are applied. The full details on converting patent and citation counts from CPC to NAICS and between adjacent NAICS versions are in Appendices A.2 and B.2, respectively. Once patent data are mapped to 6-digit NAICS codes, we roll them up to the 4-digit level by summing the values for all 6-digit codes that fall under each 4-digit industry.

Compustat provides annual firm-level accounting data along with each firm's historical 6-digit NAICS code. The accounting variables considered include employer numbers, sales, total assets, net income, leverage ratio, and the Herfindahl-Hirschman Index (HHI). Industry-level yearly accounting figures are then obtained by aggregating the data of all affiliated firms. A panel of accounting data for each 4-digit NAICS ranging from 1985 (the first year with historical 6-digit NAICS records) to 2023 was constructed. Then, a one-to-one merge was performed with excess complexity scores and patent activity indicators to prepare for the regressions.

Table 2 provides summary statistics for all potential explanatory variables, showing each in both its original units and in its natural-log form for use in the regressions. Tables 3 and 4 show robustness checks – using various lags – for our logged accounting and patent indicators on the 4-digit NAICS complexity score. All variables remain robust, except for the Herfindahl-Hirschman Index (HHI), which we will use to categorize industries as "competitive" or "less competitive." Because sales, total assets, and employee counts are all measurements of size at the industry level and correlated with each other, including them together risks multicollinearity and makes it difficult to determine which variables are statistically significant. To address this, a correlation matrix is also presented in Table 5 for the key variables used in our subsequent regressions. The table shows that the three size measures (employees, sales, total assets) are highly correlated (correlations above 0.75), which matches the concerns. Still, they have a weak relationship with the other two accounting variables. Likewise, accounting measures and patent indicators generally show weak or no linear relationship, except that patent counts are moderately correlated with size, which makes sense because patent

counts essentially reflect the scale of an industry's patenting activity. Within the patent indicators, the count measures are always highly correlated with each other, while the growth measures exhibit correlations ranging from moderate to strong.

4.2 Models

Based on the correlation matrix, the base regression model is set to be as Equation **??** to avoid the multicollinearity risk, in which, the dependent variable $C_{i,t}$ is the nature logged excess complexity score at year t for the industry (4-digit NAICS) i. Three logged accounting variables are included: total assets, net income, and leverage ratio, all with lagged terms. $g(patent)^1_{i,t-1}$ is the logged annual growth rate of the patent count at year $t-1$ (the growth rate of patent from year $t-2$ to year $t-1$, and $Citation^1_{i,t-1}$ is the logged number of citations the patent filed in year $t-1$ received within one year after filing.

$$C_{i,t} = \beta^0_{i,t} + \beta^A_{i,t} \cdot Asset_{i,t-1} + \beta^{NI}_{i,t} \cdot NI_{i,t-1} + \beta^L_{i,t} \cdot Leverage_{i,t-1} + \beta^g_{i,t} \cdot g(patent)^1_{i,t-1} + \beta^C_{i,t} \cdot Citation^3_{i,t-1} + \gamma_i + \alpha_t + u_{i,t} \quad (8)$$

For one version of the robustness checks, the choice of explanatory variables is the same, except for the patent activity indicators, and the regression model is as shown in Equation 9. $g(patent)^3_{i,t-3}$ is the logged three-year growth rate of the patent count at year $t-3$ (the growth rate of patent from year $t-6$ to year $t-3$, and $Citation^3_{i,t-3}$ is the logged number of citations the patent filed in year $t-3$ received within three years after filing, so that the citation counting stops at the end of year t for patents filed at the end of year $t-3$. Using this lagged citation measurement ensures we capture past patent activity when examining its link to the industry's current complexity.

$$C_{i,t} = \beta^0_{i,t} + \beta^A_{i,t} \cdot Asset_{i,t-1} + \beta^{NI}_{i,t} \cdot NI_{i,t-1} + \beta^L_{i,t} \cdot Leverage_{i,t-1} + \beta^g_{i,t} \cdot g(patent)^3_{i,t-3} + \beta^C_{i,t} \cdot Citation^3_{i,t-3} + \gamma_i + \alpha_t + u_{i,t} \quad (9)$$

4.3 Results

Table 6 summarizes six regressions – three sample splits and two model specifications. In the full-sample results, all coefficients remain significant in both the baseline and robustness checks. From the accounting side, a 1% rise in total assets, net income, and leverage ratio is associated with -0.029%, $+0.006\%$, and -0.263% changes in the complexity score, respectively. Indicating that smaller the size of the industry, higher the revenue generated in this field, and lower reliance on debt as financial support all boost complexity – the reduction in debt reliance has the largest effect. There are significant relationships between patent activity indicators and industrial complexity level, as expected, considering that the complexity score is calculated using the patent data. The patent growth rate is a key indicator of innovation and technological advancement in the industry. The results show that a 1% increase in this growth rate is linked to a 0.242% rise in the complexity score. This positive relationship supports the idea that advances in technology drive higher complexity. The number of citations received represents the scientific value of a patent (Kogan et al. (2017), and Arora, Belenzon, Ferracuti and Nagar (2024)). Aggregating it to the industry level shows the influence of this area's inventions on future ones. The results indicate that a 1% rise in citation counts is associated with a 0.114% drop in the complexity score, suggesting that industries with more influential patents tend to have simpler technology networks. The possible explanation could be that widely cited patents tend to be built on well-known, standardized components or frameworks, which makes it intuitive to deduce that the areas with high citations would form uniform networks (and thus less heterogeneous). In other words, citation counts capture influence rather than the novel combinations that drive network complexity.

After splitting the sample at the median Herfindahl–Hirschman Index, Table 6 shows that the drivers of complexity differ across the two groups. In industries with HHI below the 50th percentile (the less competitive group), the industrial complexity is primarily driven by patent-related activities, with effect sizes close to the full sample, and all four are significant at 1% level. There, a 1% increase in patent growth raises the complexity score by 0.264%, and a 1% rise in citations lowers it by 0.199% – compared to the 0.242% and -0.114% in the overall sample. The effects of size, revenue, and reliance on debt are no longer significant. In fiercely competitive industries,

technological complexity tends to be related to patent activity rather than balance-sheet strength for several reasons: 1) For firms that are similar in market shares (measured by share of sales), the method to stand out is through better or more inventive technologies, which is captured by patent activity indicators and the industrial complexity score; 2) Patent will be filed strategically by firms from the motive of defensive or offensive blocking (Lerner (1995), and Blind, Cremers and Mueller (2009)), and those patent webs boost the measurement of complexity score.

By contrast, in industries with HHI at or above the 50th percentile (the more competitive group), net income is the only consistent driver of higher complexity, and its coefficient is stable across both the base and robustness models. a 1% increase in net income is linked to a 0.020% rise in the complexity score – over two times larger than the 0.006% effect seen in the full sample. This suggests that in markets with lower competition, complexity increases primarily in areas that generate high profits. The channels might be that: 1) Industries with strong profits can allocate more resources to research and development, which helps in generating innovative technologies; 2) Industries that yield health earnings have higher risk tolerances which give the firms affiliated financial foundation to support projects with uncertain payoffs – usually those that push the boundary of existing technologies – without interfering the core business.

Robustness is checked again by splitting industries at the first and second tertiles of the Herfindahl–Hirschman Index using the two regression models 8 and 9. Table 7 presents these results, which remain consistent with our prior findings.

5 Conclusion

Using patent data's CPC classification from 1985 to 2023, networks representing the knowledge component's connection can be built annually for each technology, and the structural diversity score measuring the heterogeneity degree of the network is proved to be better than other existing network metrics or complexity measurements in quantifying industrial complexity level.

This paper lays out a step-by-step method for converting complexity scores, patent counts, and

citation counts from CPC to NAICS and between NAICS versions. The same procedure can be used to translate any other variables from CPC to NAICS or across different NAICS versions.

By digging into the industrial complexity score, this paper reveals that complexity rose steadily in all sectors until 2020. When the Covid-19 pandemic hit, complexity fell in labor-intensive industries – compared with the stable scores in tech-related areas, though employment didn't drop by the same amount. To make scores comparable across industries, we set each industry's own benchmark and calculated excess complexity. It has also been found that for some sectors, more affiliated industries do not have measurable complexity scores because they lack patents in the beginning. Our analysis also reveals clear contrasts in what drives complexity across industries. In competitive sectors, complexity depends mainly on patent activity – specifically, higher rates of new patents and fewer citations to older patents – however, in less competitive markets, only strong profits boost complexity.

References

Arora, Ashish, Sharon Belenzon, Elia Ferracuti, and Jay Prakash Nagar, "Revisiting the Private Value of Scientific Inventions," Working Paper 33056, National Bureau of Economic Research November 2024.

Arthur, W. Brian, "The Nature of Technology : What It Is and How It Evolves," *Free Press*, 2009.

Balland, Pierre-Alexandre and David Rigby and, "The Geography of Complex Knowledge," *Economic Geography*, 2017, *93* (1), 1–23.

Bertz, S.H., "On the complexity of graphs and molecules," *Bltn Mathcal Biology*, 1983, *45*, 849–855.

Blind, Knut, Katrin Cremers, and Elisabeth Mueller, "The influence of strategic patenting on companies' patent portfolios," *Research Policy*, 2009, *38* (2), 428–436.

Broekel, Tom, "Using structural diversity to measure the complexity of technologies," *PLOS ONE*, 05 2019, *14* (5), 1–23.

Emmert-Streib, Frank and Matthias Dehmer, "Exploring Statistical and Population Aspects of Network Complexity," *PLOS ONE*, 05 2012, *7* (5), 1–17.

Fleming, Lee and Olav Sorenson, "Technology as a complex adaptive system: evidence from patent data," *Research Policy*, 2001, *30* (7), 1019–1039.

III, Donald E Bowen, Laurent Frésard, and Gerard Hoberg, "Rapidly evolving technologies and startup exits," *Management Science*, 2023, *69* (2), 940–967.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy, "Measuring technological innovation over the long run," *American Economic Review: Insights*, 2021, *3* (3), 303–320.

Kogan, Leonid, Dimitris Papanikolaou, Amit Seru, and Noah Stoffman, "Technological Innovation, Resource Allocation, and Growth*," *The Quarterly Journal of Economics*, 03 2017, *132* (2), 665–712.

Lerner, Josh, "Patenting in the Shadow of Competitors," *The Journal of Law Economics*, 1995, *38* (2), 463–495.

Mariani, Manuel Sebastian, Matúš Medo, and François Lafond, "Early identification of important patents: Design and validation of citation network metrics," *Technological forecasting and social change*, 2019, *146*, 644–654.

McNerney, James, J. Doyne Farmer, Sidney Redner, and Jessika E. Trancik, "Role of design complexity in technology improvement," *Proceedings of the National Academy of Sciences*, 2011, *108* (22), 9008–9013.

Nathan, Travis J. Lybbert, and Zolas Nikolas Jason, "An 'Algorithmic Links with Probabilities' Crosswalk for USPC and CPC Patent Classifications with an Application Towards Industrial Technology Composition," *Economics of Innovation and New Technology*, 2019.

Skorobogatov, VA and AA Dobrynin, "Metrical analysis of graphs," *Commun. Math. Comp. Chem*, 1988, *23*, 105–155.

Sorenson, Olav and Lee Fleming, "Science and the diffusion of knowledge," *Research Policy*, 2004, *33* (10), 1615–1634.

Todeschini, Roberto and Viviana Consonni, *Handbook of molecular descriptors*, John Wiley & Sons, 2008.

Wiener, Harry, "Structural Determination of Paraffin Boiling Points," *Journal of the American Chemical Society*, 1947, *69* (1), 17–20.

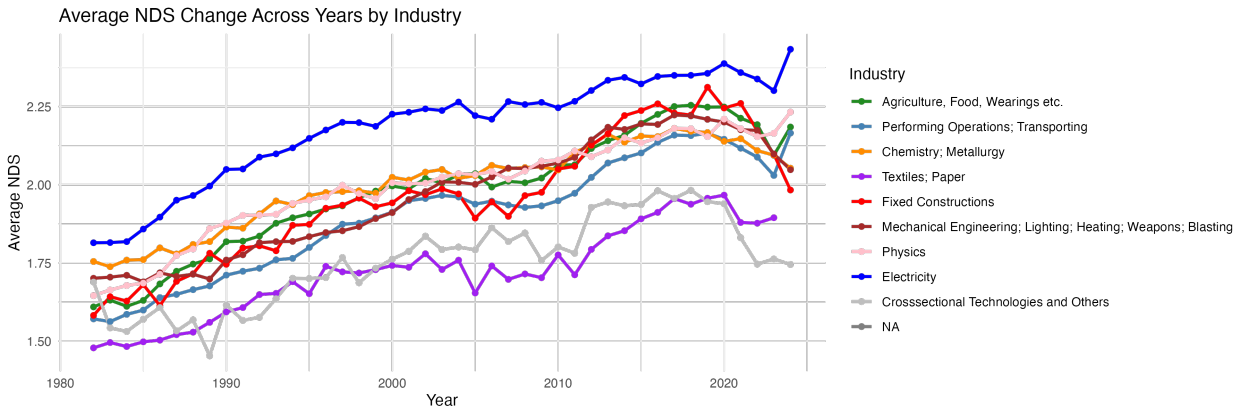


Figure 1: Complexity Score Across Years by 1-digit CPC Industry

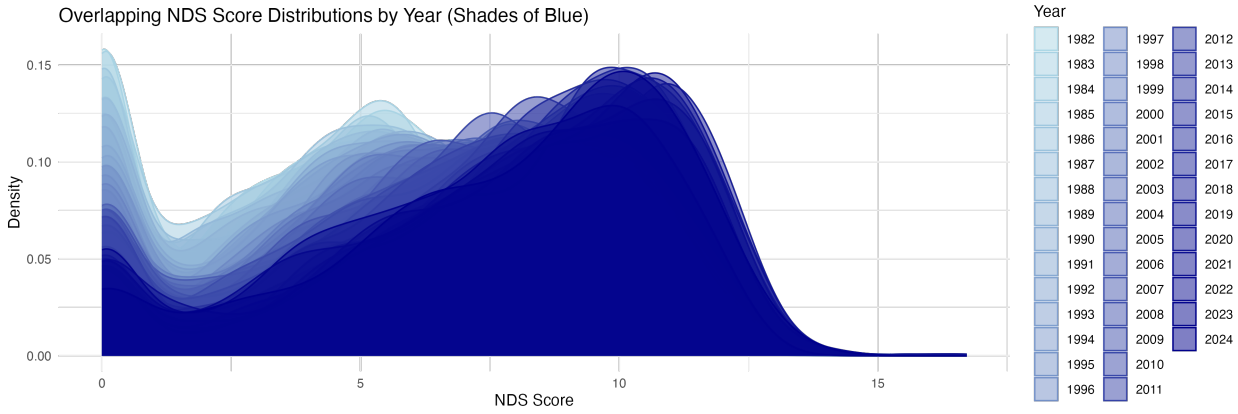


Figure 2: Overlapping NDS Score Distributions by Year

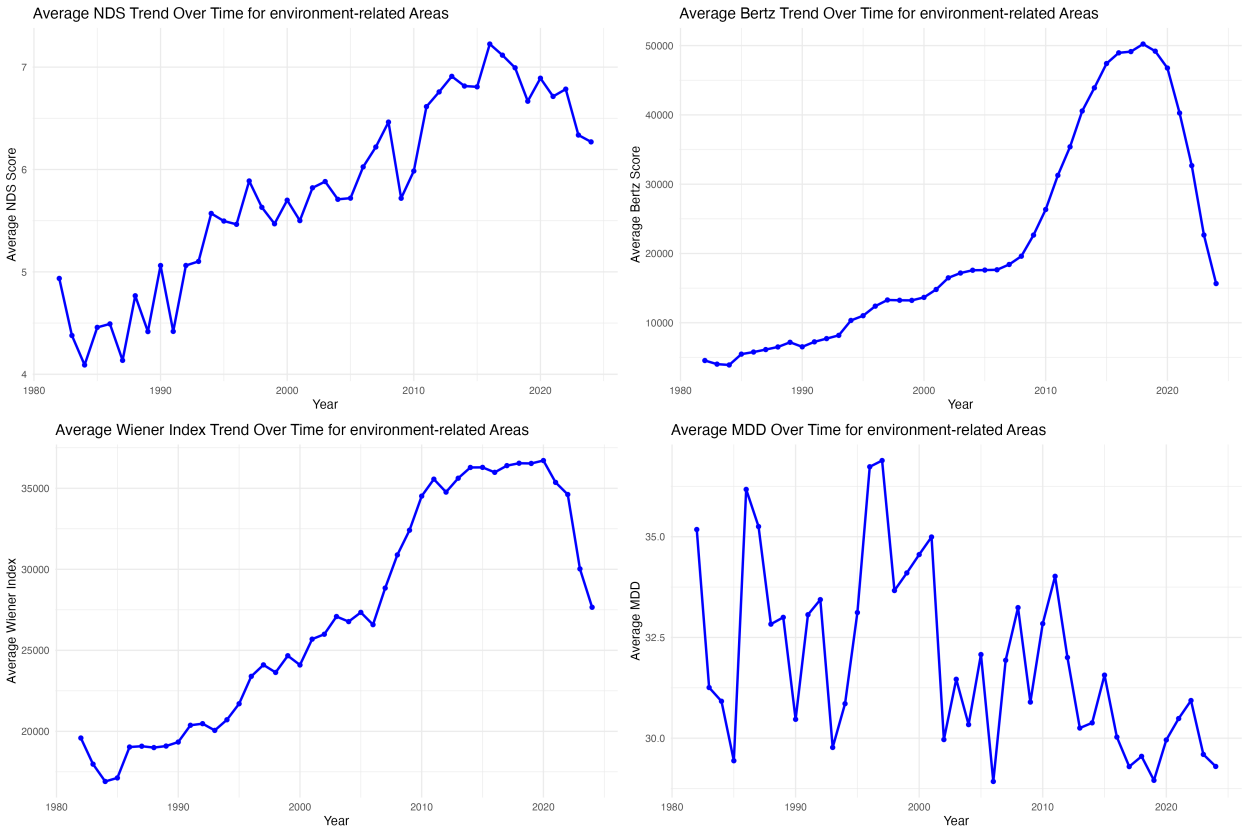


Figure 3: Annual Complexity Score of Environmental-related Areas

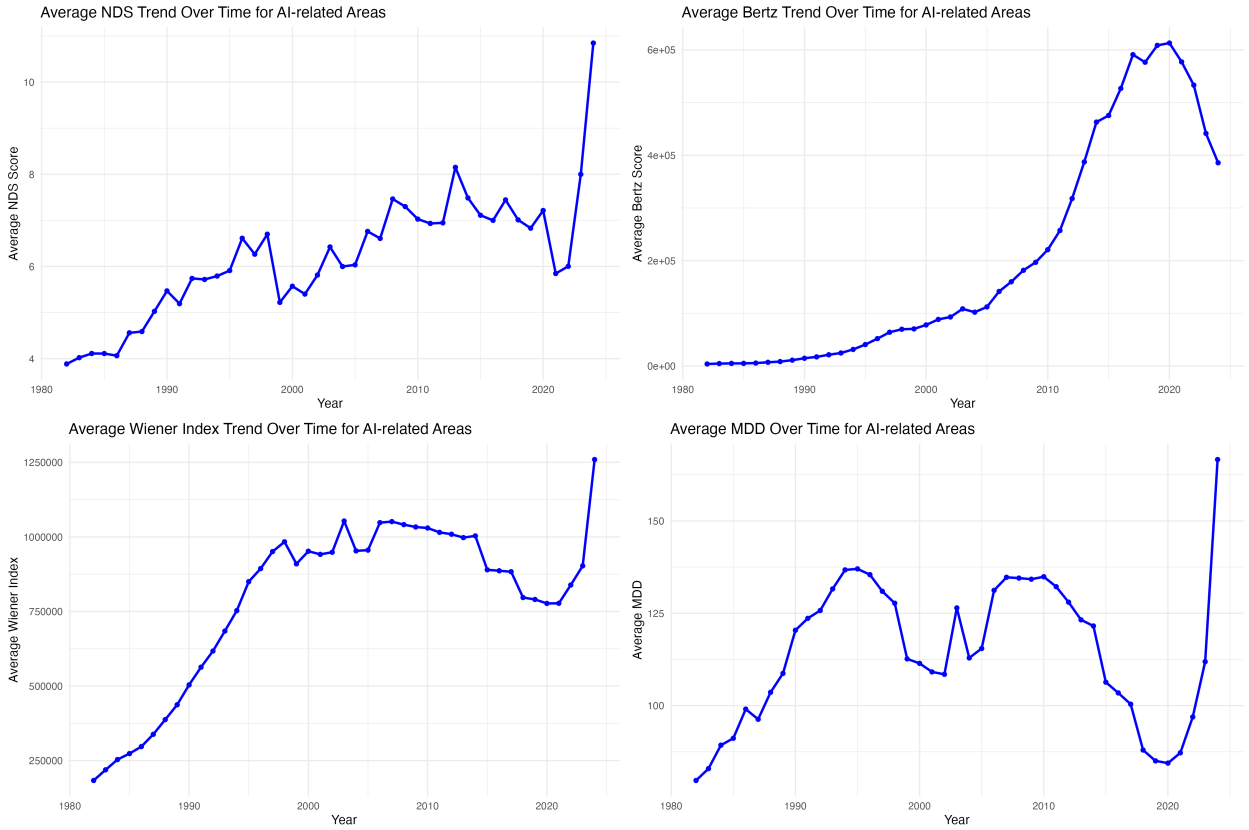


Figure 4: Annual Complexity Score of AI-related Areas

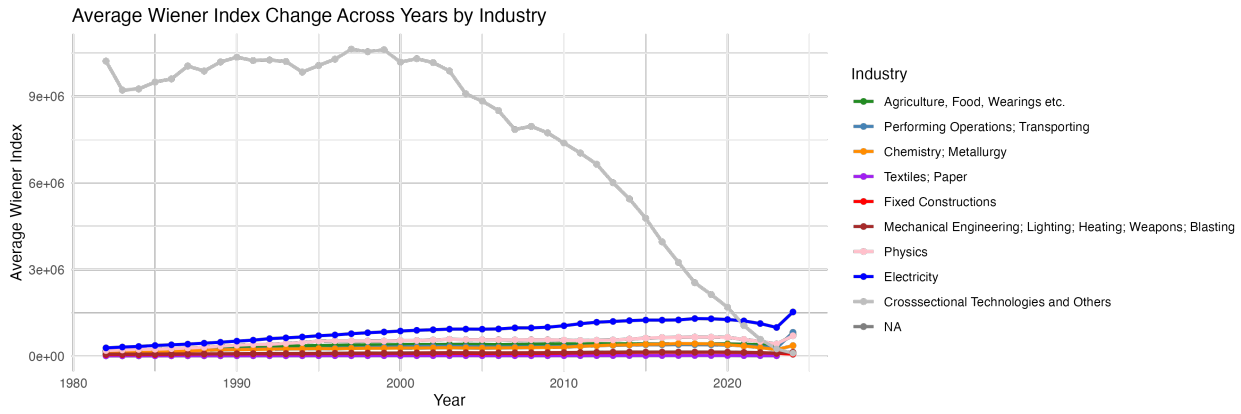


Figure 5: Complexity Wiener Index Across Years by 1-digit CPC Industry

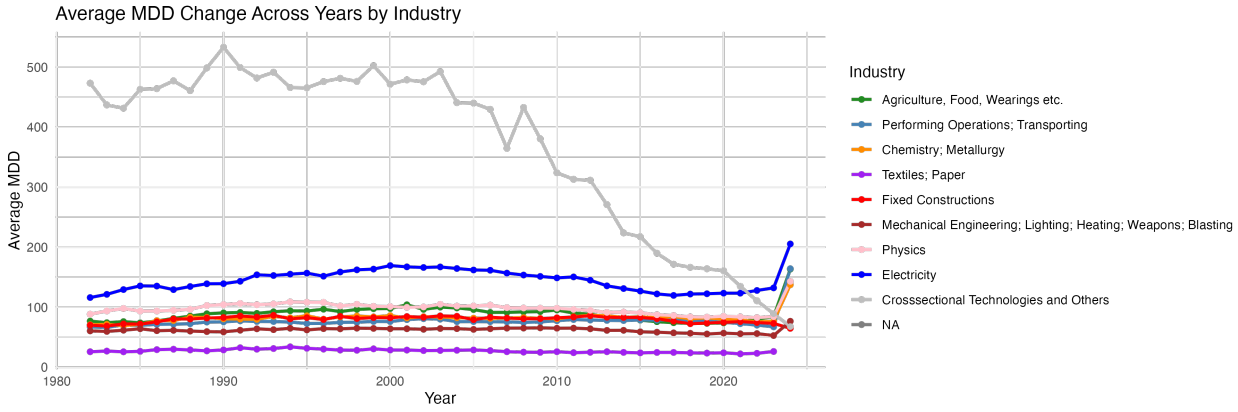


Figure 6: Complexity MDD Across Years by 1-digit CPC Industry

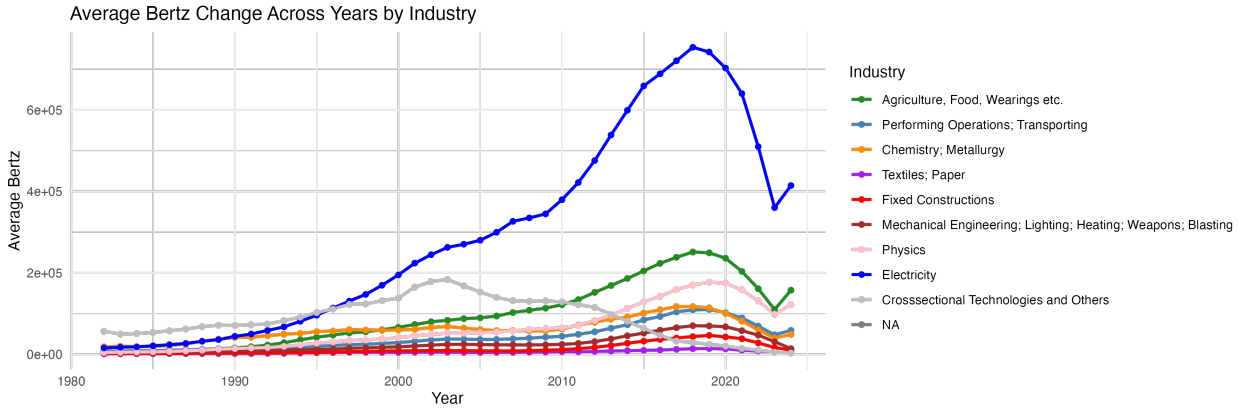


Figure 7: Complexity Bertz Index Across Years by 1-digit CPC Industry

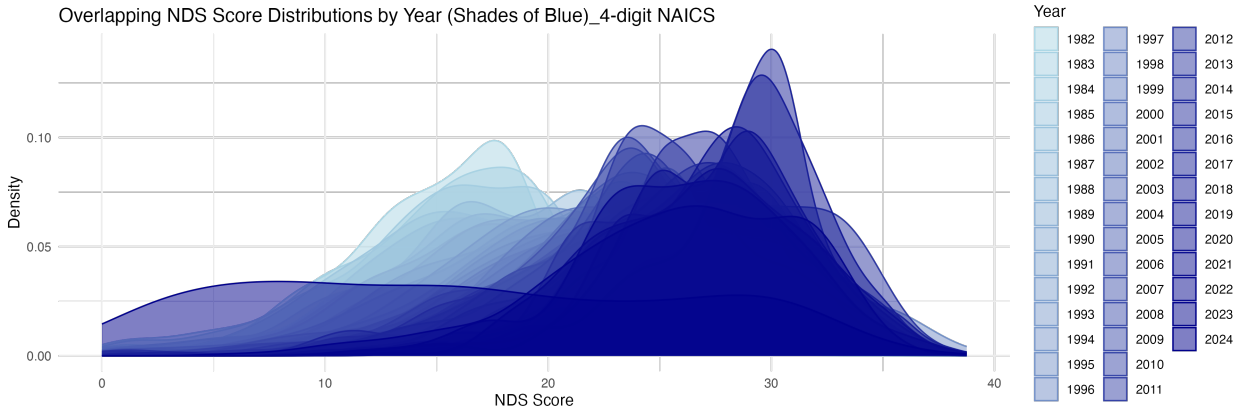


Figure 8: Overlapping NDS Score Distributions by Year at the level of 4-digit NAICS

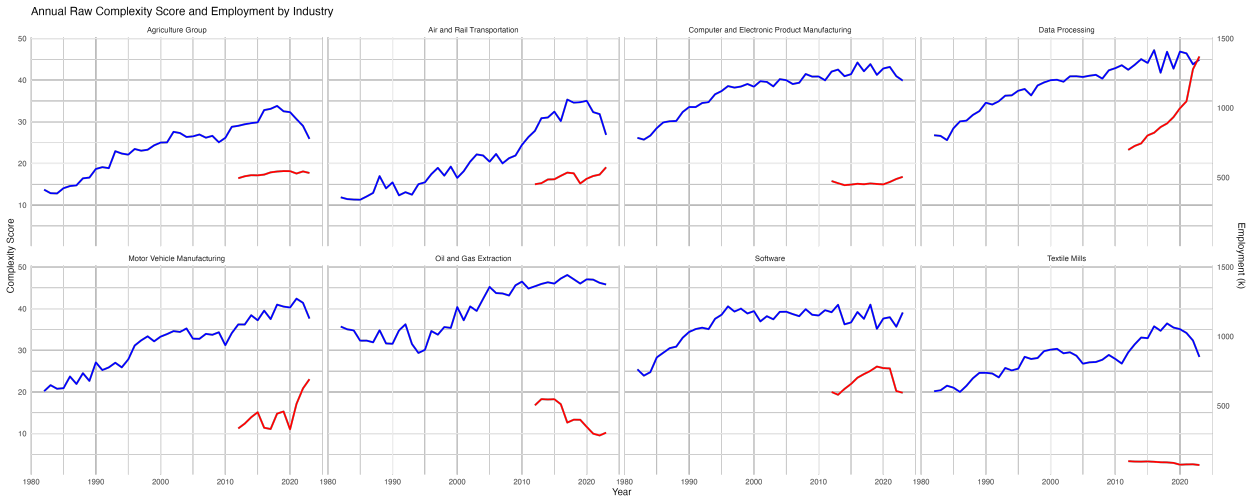


Figure 9: Annual Complexity Score and Employment by Industry

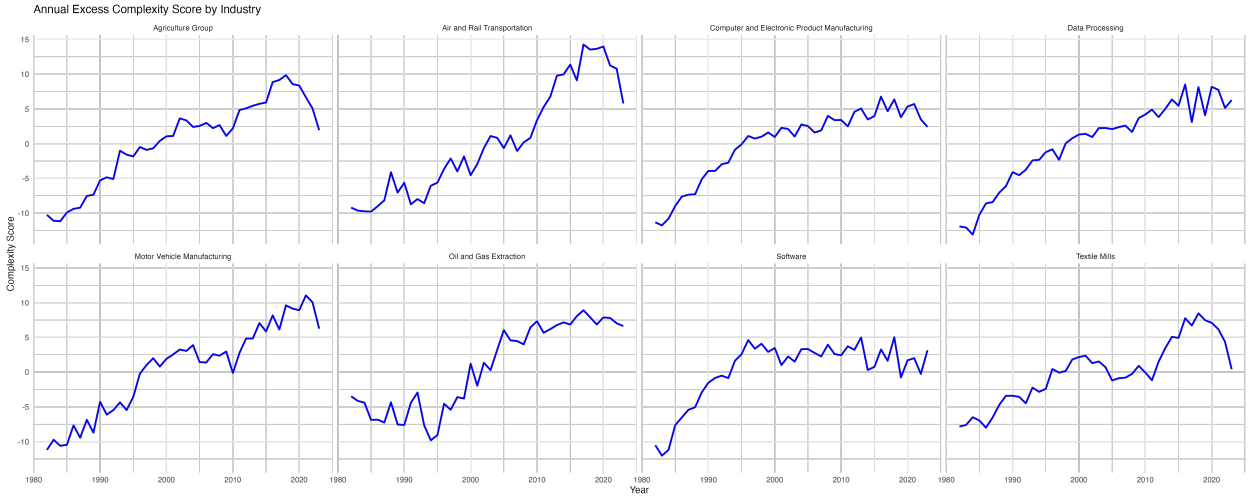


Figure 10: Annual Excess Complexity Score by Industry

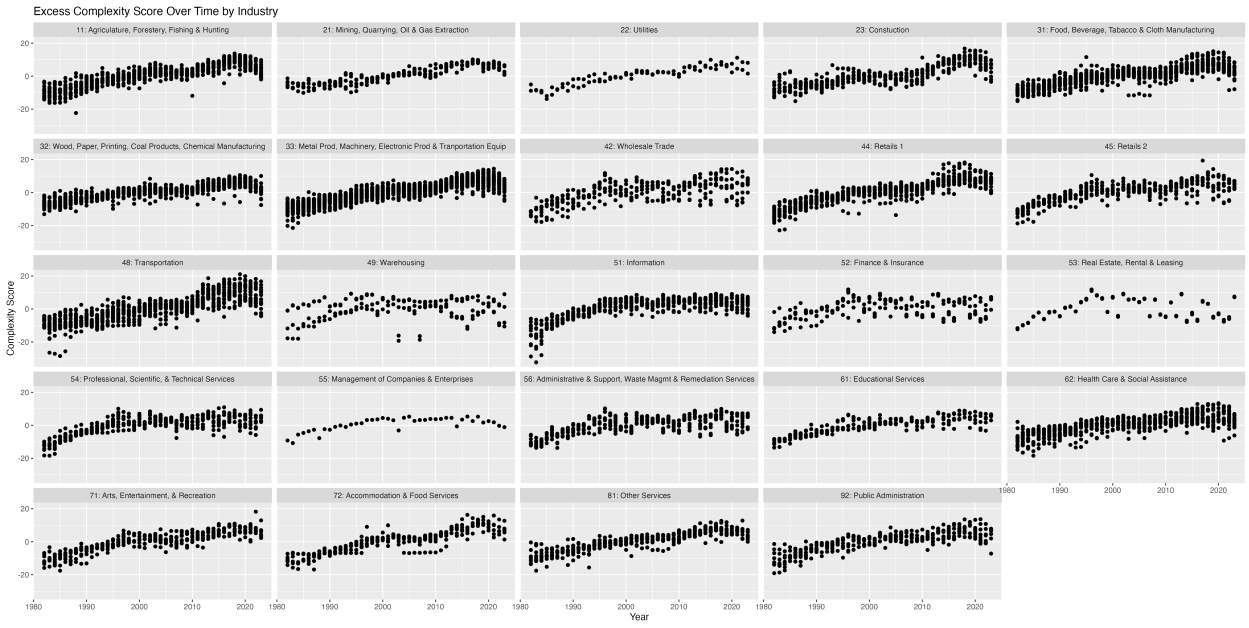


Figure 11: Annual Excess Complexity Score Over Time by Industry

Table 1: Statistics for Excess Score at 4-digit NAICS level

2-digit NAICS	Obs	Count of 4-digit NAICS	Count w/ score data	Median	Std. dev.
11 (Agriculture, Forestry, Fishing & Hunting)	773	19	18	0.83	6.57
21 (Mining, Quarrying, Oil & Gas Extraction)	215	5	5	0.51	5.79
22 (Utilities)	86	3	2	1.16	5.69
23 (Construction)	430	10	10	-0.47	6.57
31 (Food, Beverage, Tabacco & Cloth Manufacturing)	938	22	22	0.59	5.95
32 (Wood, Paper, Printing, Coal Products, Chemical Manufacturing)	901	21	21	0.54	4.80
33 (Metal Prod, Machinery, Electronic Prod & Transportation Euip)	1,803	43	42	0.58	6.11
42 (Wholesale Trade)	429	19	10	0.04	6.50
44 (Retail 1)	686	16	16	0.69	7.07
45 (Retail 2)	387	11	9	0.97	6.30
48 (Transportation)	902	25	21	0.07	7.66
49 (Postal Service & Warehousing)	171	4	4	1.57	6.61
51 (Information)	473	12	11	1.75	6.76
52 (Finance & Insurance)	473	11	11	-0.26	6.02
53 (Real Estate & Rental & Leasing)	129	8	3	-1.56	5.99
54 (Professional, Scientific & Technical Services)	387	9	9	0.87	5.49
55 (Management of Companies & Enterprises)	43	1	1	1.85	6.23
56 (Administrative & Support, Waste Mgmt & Remediation Service)	387	11	9	0.40	5.24
61 (Educational Services)	215	7	5	0.76	4.95
62 (Health Care & Social Assistance)	687	18	16	0.83	5.78
71 (Arts, Entertainment, & Recreation)	344	9	8	-0.94	6.51
72 (Accommodation & Food Services)	258	7	6	1.14	7.27
81 (Other Services)	386	14	9	0.58	5.89
92 (Public Administration)	344	8	8	0.90	6.05
24 Sectors	11,847	313	276	0.64	6.23

Notes. This table summarizes excess score statistics at the 2-digit NAICS level (the system's broadest category). Column 1 is the 2-digit NAICS sector. Column 2 reports, for each 2-digit NAICS sector, the total number of panel observations in the data set from 1985 to 2023. Column 3 shows the total count of 4-digit NAICS codes according to the Census Bureau, and Column 4 reports how many of those have valid excess scores over the same period. Finally, Columns 5 and 6 present the median and standard deviation of the excess scores for each sector.

Table 2: Statistics for Accounting Variables and Patent Activity Indicators

Name	Obs	Mean	P25	Median	P75	Std. dev.
Panel A: Reported in their original units at industry level						
Accounting Variables:						
Employee Count (Thousands)	7,696	179.8	8.4	50.2	165.8	383
Sales (Millions)	7,696	47,339	1,421	8,453	35,731	121,793
Total Asset (Millions)	7,696	226,440	1,496	8,389	38,142	1,681,689
Net Income (Millions)	7,696	2,842	1.88	188.3	1,285	11,783
Leverage Ratio	7,678	0.345	0.208	0.301	0.417	1.186
Herfindahl-Hirschman Index	7,696	3,945	1,677	2,916	5,387	2,982
Patent Activity Indicators:						
Patent Count	10,470	1,387	6	51	629	4,567
Citation Count in One Year	10,470	1,231	3	37	420	4,448
Citation Count in Two Year	10,470	7,733	19	215	2,368	29,373
Citation Count in Three Year	10,470	13,368	33	375	4,294	51,719
1-year Patent Growth Rate	10,144	0.021	-0.059	0.037	0.120	0.383
3-year Patent Growth Rate	9,531	0.146	-0.054	0.137	0.313	0.500
Annual Growth Rate of 1-year Citation	9,881	0.728	-0.400	-0.080	0.373	20.777
Annual Growth Rate of 2-year Citation	10,079	-0.004	-0.288	-0.064	0.137	0.951
Annual Growth Rate of 3-year Citation	10,106	-0.048	-0.266	-0.062	0.111	0.614
Panel A: Reported in their logged terms at industry level						
Accounting Variables:						
Employee Count	7,696	3.699	2.237	3.936	5.117	1.953
Sales	7,696	8.687	7.260	9.042	10.484	2.584
Total Asset	7,696	8.829	7.311	9.035	10.549	2.734
Net Income	7,696	3.676	1.058	5.243	7.159	4.937
Leverage Ratio	7,678	0.274	0.189	0.263	0.348	0.152
Herfindahl-Hirschman Index	7,696	7.939	7.425	7.978	8.592	0.990
Patent Activity Indicators:						
Patent Count	10,470	4.198	1.898	3.957	6.44	2.743
Citation Count in One Year	10,470	2.873	1.405	3.624	6.043	2.780
Citation Count in Two Year	10,470	5.413	2.989	5.373	7.770	3.085
Citation Count in Three Year	10,470	5.916	3.524	5.930	8.365	3.163
1-year Patent Growth Rate	10,144	-0.022	-0.061	0.036	0.113	0.325
3-year Patent Growth Rate	9,531	0.021	-0.055	0.128	0.272	0.569
Annual Growth Rate of 1-year Citation	9,881	-0.110	-0.490	-0.079	0.323	0.900
Annual Growth Rate of 2-year Citation	10,079	-0.171	-0.334	-0.065	0.129	0.641
Annual Growth Rate of 3-year Citation	10,106	-0.183	-0.307	-0.064	0.106	0.606

Notes. This table summarizes all the accounting variables and patent activity indicators at the industry level (4-digit NAICS). Panel A gives the statistics at their original units and the nature logged terms' results are in Panel B. The nature logged term is calculated as $ln(x) = sgn(x)\ln(1+|x|)$.

Table 3: Robustness Check for Accounting Variables

	Excess Complexity Score		
Employee Number (lag 1)	-0.081***	-0.019	-0.007
	(0.018)	(0.032)	(0.033)
Employee Number (lag 2)		-0.072***	0.013
		(0.033)	(0.042)
Employee Number (lag 3)			-0.102***
			(0.034)
Sales (lag 1)	-0.039***	0.001	0.013
	(0.012)	(0.019)	(0.019)
Sales (lag 2)		-0.045**	-0.003
		(0.019)	(0.027)
Sales (lag 3)			-0.056**
			(0.023)
Total Asset (lag 1)	-0.039***	-0.008	-0.008
	(0.012)	(0.019)	(0.019)
Total Asset (lag 2)		-0.034*	0.025
		(0.019)	(0.028)
Total Asset (lag 3)			-0.063***
			(0.022)
Leverage Ratio (lag 1)	-0.352***	-0.220*	-0.238*
	(0.103)	(0.126)	(0.133)
Leverage Ratio (lag 2)		-0.109	0.135
		(0.124)	(0.151)
Leverage Ratio (lag 3)			-0.254*
			(0.142)
Net Income (lag 1)	0.006**	0.006**	0.008**
	(0.003)	(0.003)	(0.003)
Net Income (lag 2)		-0.0002	-0.001
		(0.003)	(0.003)
Net Income (lag 3)			0.003
			(0.003)
Herfindahl-Hirschman Index (lag 1)	-0.001	0.008	0.006
	(0.015)	(0.017)	(0.017)
Herfindahl-Hirschman Index (lag 2)		-0.013	-0.016
		(0.016)	(0,019)
Herfindahl-Hirschman Index (lag 3)			-0.001
			(0.020)

Notes. This table presents separate robustness checks for the impact of each accounting variable on the 4-digit NAICS complexity score. All the variables are in nature logged terms. Each column in this table shows one regression – and note that horizontal lines separate the robustness checks for each variable. For each variable, three models are estimated: one that includes only the one-year lag, a second that adds the two-year lag, and a third that includes both the two-year and three-year lags. ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

Table 4: Robustness Check for Patent Activity Indicators

	Excess Complexity Score		
Patent Numbers (lag 1)	-0.087***	0.276***	0.310***
	(0.034)	(0.096)	(0.095)
Patent Numbers (lag 2)		-0.389***	0.124
		(0.098)	(0.162)
Patent Numbers (lag 3)			-0.592***
			(0.142)
Growth of Patent Numbers (lag 1)	0.245***	0.323***	0.353***
	(0.070)	(0.078)	(0.080)
Growth of Patent Numbers (lag 2)		0.428***	0.518***
		(0.099)	(0.108)
Growth of Patent Numbers (lag 3)			0.395***
			(0.086)
Three-year Patent Growth Rate (lag 3)	0.509***		
	(0.069)		
One-year Citation Count (lag 1)	-0.120***	-0.033	-0.007
	(0.021)	(0.026)	(0.027)
One-year Citation Count (lag 2)		-0.145***	-0.095***
		(0.026)	(0.031)
One-year Citation Count (lag 3)			-0.128***
			(0.028)
Growth of One-year Citation (lag 1)	0.038**	0.048**	0.064***
	(0.018)	(0.020)	(0.021)
Growth of One-year Citation (lag 2)		0.028	0.060**
		(0.021)	(0.023)
Growth of One-year Citation (lag 3)			0.056***
			(0.022)
Two-year Citation Count (lag 2)	-0.163***	-0.013	
	(0.030)	(0.047)	
Two-year Citation Count (lag 3)		-0.203***	
		(0.048)	
Growth of Two-year Citation (lag 2)	0.085**	0.082**	
	(0.039)	(0.042)	
Growth of Two-year Citation (lag 3)		-0.044	
		(0.045)	
Three-year Citation Count(lag 3)	-0.109***		
	(0.035)		
Growth of Three-year Citation (lag 3)	-0.040		
	(0.047)		

Notes. This table presents separate robustness checks for the impact of each patent activity indicator on the 4-digit NAICS complexity score. Each column in this table shows one regression – and note that horizontal lines separate the robustness checks for each variable. For single-year measures (e.g., one-year patent counts or citations), three models are run: (1) the one-year lag only, (2) adding the two-year lag, and (3) adding the two-year and three-year lag. For multi-year measures—such as three-year patent growth or cumulative two- and three-year citation counts and their growth rates—we drop extra lags since these indicators already cover multiple years. ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

Table 5: Correlation Between the Accounting Variables and Patent Activity Indicators

	Emp	Sales	Asset	Net Income	Leverage Ratio	Patent Count	1-y Cit	2-y Cit	3-y Cit	g_ patent	3-y g_ patent	g_1-y Cit	g_2-y Cit	g_3-y Cit
Emp	1.0000													
Sales	0.9048	1.0000												
Asset	0.8587	0.9515	1.0000											
Net Income	0.4430	0.4832	0.4727	1.0000										
Leverage	-0.0381	-0.0614	-0.0265	-0.1467	1.0000									
Patent Num	0.3092	0.3481	0.3407	0.2256	-0.0494	1.0000								
1-y Citation	0.2821	0.2641	0.2485	0.1834	-0.0245	0.9040	1.0000							
2-y Citation	0.2667	0.2285	0.2099	0.1628	-0.0336	0.8660	0.9688	1.0000						
3-y Citation	0.2687	0.2298	0.2112	0.1634	-0.0466	0.8567	0.9492	0.9942	1.0000					
g_Patent	-0.0108	-0.0732	-0.0830	-0.0134	-0.0429	0.0579	0.2137	0.3119	0.3471	1.0000				
3-y g_Patent	0.0003	-0.0726	-0.0758	-0.0205	-0.0376	0.0606	0.2377	0.3422	0.3780	0.8866	1.0000			
g_1-y Cit	-0.0149	-0.0449	-0.0494	-0.0134	-0.0219	0.0247	0.1958	0.1800	0.1904	0.4854	0.4492	1.0000		
g_2-y Cit	-0.0280	-0.0869	-0.0996	-0.0283	-0.0331	0.0332	0.2250	0.3137	0.3341	0.7747	0.7350	0.6660	1.0000	
g_3-y Cit	-0.0317	-0.1010	-0.1182	-0.0338	-0.0352	0.0279	0.2314	0.3340	0.3688	0.8294	0.7817	0.5893	0.9291	1.0000

Notes: This table reports the correlation matrix for the natural logs of all industry-level variables, including both accounting data and patent activity measures. Here, "Emp" stands for the number of employees; "1-y cit," "2-y cit," and "3-y cit" refer to the citations received within one, two, and three years, respectively, by patents filed in that year; "g_patent" is the annual growth rate of patent counts, while "3-y g_Patent" is the three-year growth rate; and "g_1-y Cit," "g_2-y Cit," and "g_3-y Cit" denote the annual growth rates of the one-, two-, and three-year citation counts, respectively.

Table 6: Regression Results for Determinants of Industrial Complexity

	Industrial Complexity Level					
	Full Sample		*HHI<p50*		*HHI≥p50*	
	(1)	(2)	(3)	(4)	(5)	(6)
Accounting Variables:						
Total Assets	-0.029**	-0.035**	-0.034	-0.059	-0.019	-0.014
	(0.013)	(0.016)	(0.032)	(0.040)	(0.016)	(0.019)
Net Income	0.006**	0.006*	-0.002	-0.001	0.020***	0.020***
	(0.003)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)
Leverage Ratio	-0.263**	-0.219*	-0.033	0.071	-0.266**	-0.210
	(0.107)	(0.123)	(0.249)	(0.282)	(0.120)	(0.137)
Patent Activity Indicators:						
1-year Patent Growth Rate	0.242***		0.264***		0.127	
	(0.073)		(0.089)		(0.097)	
3-year Patent Growth Rate		0.612***		0.605***		0.475***
		(0.074)		(0.093)		(0.121)
1-year Citation Count	-0.114***		-0.199***		0.020	
	(0.023)		(0.031)		(0.032)	
3-year Citation Count		-0.239***		-0.394***		0.068
		(0.045)		(0.062)		(0.064)
Absorb Industry Effect	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Obs	6,840	5,858	3,476	2,971	3,364	2,887
Adj. R^2	48.90%	43.98%	49.14%	44.67%	55.44%	51.48%

Notes. This table presents the coefficients and robust standard errors for six regressions using Equations 8 and 9 absorbing at the industry level (4-digit NAICS) with year fixed effects. Columns (1), (3), and (5) show the results using Equation 8. Columns (2), (4), and (6) are for Equation 9. Columns (1) and (2) present results for the full sample. Columns (3) to (6) report results after dividing industries by their Herfindahl–Hirschman Index: columns (3) and (4) each covers industries with an HHI below the median (more competitive), and columns (5) and (6) each covers those with an HHI above the median (less competitive). $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 7: Robustness Test Results for Determinants of Industrial Complexity

	Industrial Complexity Level			
	HHI<p33		*HHI≥p67*	
	(1)	(2)	(3)	(4)
Accounting Variables:				
Total Assets	-0.026	-0.062	0.025	-0.010
	(0.043)	(0.055)	(0.018)	(0.022)
Net Income	-0.003	0.002	0.018**	0.018**
	(0.005)	(0.005)	(0.007)	(0.008)
Leverage Ratio	-0.179	-0.004	-0.322**	-0.211
	(0.304)	(0.337)	(0.139)	(0.158)
Patent Activity Indicators:				
1-year Patent Growth Rate	0.235**		0.060	
	(0.109)		(0.141)	
3-year Patent Growth Rate		0.573***		0278**
		(0.104)		(0.131)
1-year Citation Count	-0.247***		0.041	
	(0.040)		(0.041)	
3-year Citation Count		-0.406***		-0.123
		(0.074)		(0.079)
Absorb Industry Effect	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Obs	2,267	1,915	2,190	1,870
Adj. R^2	53.84%	48.78%	55.48%	50.95%

Notes. This table presents the coefficients and robust standard errors for four robustness check regressions using Equations 8 and 9 absorbing at the industry level (4-digit NAICS) with year fixed effects. Columns (1) and (3) show the results using Equation 8. Columns (2) and (4) are for Equation 9. Industries were split into three groups (tertiles) based on their Herfindahl–Hirschman Index: columns (1) and (2) each covers industries with an HHI below the first tertile (the more competitive), and columns (3) and (4) each covers those with an HHI above the second tertile (less competitive). ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

Online Appendix

A Conversion from CPC to NAICS

A.1 Conversion of Complexity Score

As mentioned in Broekel (2019), when calculating the NDS for the network of each 4-digit CPC code, 50 nodes are selected, and a subnetwork is drawn by a random walktrap of 150 steps starting from each of the node. These 50 subnetworks are used to calculate the iNDS and the final NDS is the average of the 50 iNDS. This process emphasizes the importance of topologies' diversity rather than the size of the network, so when the case is that multiple 4-digit CPC industries are mapped into one 6-digit NAICS industry, the scale complexity score should not be aggregated.

The probability weight needed for the conversion comes from Nathan et al. (2019) which utilized the 'Algorithmic Links with Probabilities' method and provided direct probabilistic linkages between classification systems and supports joint analysis of patent and economic data. The files that map 6-digit NAICS to 4-digit CPC were used, so for each 6-digit NAICS, the sum of probability weight across all the CPC codes it is mapped to is 1. The original score was calculated under each 6-digit NAICS code i might be mapped to n 4-digit CPC code js with the respective probability weight and original score for each j as $Pr(j|i)$ and C^j_{CPC-4}. Then the 6-digit NAICS code's complexity score should be the sum of all the 4-digit CPC code's scores times the respective weight with the weight summing up to 1 as the equation 10.

$$C^i_{NAICS-6} = \sum_j C^j_{CPC-4} \cdot Pr(j|i) \quad \text{with} \sum_j Pr(j|i) = 1 \tag{10}$$

A.2 Conversion of Patent Count and Citation Data

After calculating the number of patents for each 4-digit CPC code in each year and the count of citations these patents receive in the subsequent one-, two-, and three- year periods, conversion to NAICS is needed for future analysis. The difference from the previous conversion is that the total

patent numbers and citation counts should be preserved after the crosswalk, so equation 10 cannot be applied to the patent and citation number.

As the complexity score, the original patent, or citation count is calculated under the 4-digit CPC level, a crosswalk to the 6-digit NAICS level is necessary for economic analysis. For this conversion, the files used mapped one or multiple NAICS industries to one CPC code, so the sum of weight assigned to these NAICS industries would be 1, and in consequence, the number of patents and citations does not change during the conversion.

For each 6-digit NAICS industry i, it may show up in the mapping into m 4-digit CPC code js with the respective probability weight for each i as $Pr(j|i)$ and the patent or citation number as N^j_{CPC-4}. Then, the 6-digit NAICS code's patent or citation number should be the sum of all the 4-digit CPC code's patent or citation number, times the respective weight, as in equation 11.

$$N^i_{NAICS-6} = \sum_j N^j_{CPC-4} \cdot Pr(j|i) \tag{11}$$

B Conversion between difference versions of NAICS

The crosswalks from CPC to NAICS are only feasible for the first three NAICS versions, which are the NAICS 1997, NAICS 2002, and NAICS 2007, so the conversion between different versions of NAICS is necessary between the recent 4 NAICS versions, which are NAICS 2007, NAICS 2012, NAICS 2017, and NAICS 2022. Once the conversion between different versions of NAICS are available, the conversion from CPC to the later versions of NAICS can be achieved.

B.1 Conversion of Complexity Score

To convert complexity scores between NAICS versions, we use the Census Bureau's concordance files to assign probability weights. For instance, when mapping from 2007 to 2012 codes, all corresponding 6-digit NAICS 2007 entry is equally weighted across the 6-digit NAICS 2012 codes they map to. As a result, the weights for each 2012 code — summed over all its linked 2007 codes

— add up to 1. This ensures that when multiple 2007 codes merge into a single 2012 code, the aggregated complexity score remains on the same scale.

The complexity scores are first calculated or converted to one version. To move from one version to the next, we use equal probability weights from the Census concordance. Specifically, for each 6-digit code i in the new version that maps to m industries j in the old version, we give each link a weight of $\frac{1}{m}$ with the respective complexity score for each j as $C^j_{NAICS_n}$. Then the new version's 6-digit NAICS complexity score is the average weighted of the old version's 6-digit NAICS complexity scores as the equation 12 .

$$C^i_{NAICS_{n+1}} = \frac{1}{m} \sum_j C^j_{NAICS_n} \tag{12}$$

B.2 Conversion of Patent Count and Citation Data

As mentioned in Appendix A.2, we need to keep the total patent and citation counts unchanged during any version conversion, so we can't use Equation 12 for these variables.

The probability weights are also assigned by using the the Census Bureau's concordance files between two adjacent versions. To convert from old version to the new one, equal probability weights are assigned. Specifically, for each 6-digit code j in the old version, it may show up in the concordance with m industries i of the new version, we give each link a weight of $Pr(j) = \frac{1}{m}$ with the respective patent count or citation count for each j as $N^j_{NAICS_n}$. The count for each new code i is then the sum of all linked old counts multiplied by their respective weights. This approach, shown in Equation 13, guarantees that the aggregate number of patents or citations stays the same through every step of the conversion process.

$$N^i_{NAICS_{n+1}} = \sum_j C^j_{NAICS_n} \cdot Pr(j) \tag{13}$$