1 Data description

In this section, I detail source material and the construction of the dataset used in my empirical analysis. Subsection 1.1 lists the sources of the raw data. Subsection 1.2 focuses on the definition of inventor productivity measures and knowledge markets, which I identify through realized inventor flows across sectors. Subsection 1.3 briefly describes other data construction steps that are discussed in more detail in Appendix **??**.

1.1 Data Sources

My empirical analysis relies on the variation of concentration across product markets, as defined by 4-digit NAICS sectors, the impact of these shifts on the allocation of inventors with specific competences across these sectors, and the subsequent effect on inventor productivity. I use USPTO patent data to measure inventor productivity and establish the set of product markets that share similar inventors, and US Economic Census data to obtain concentration and productivity growth measures. Finally, I use a dataset of product market regulations, Mercatus RegData 4.0, to conduct an instrumental variable analysis, as well as NBER-CES to estimate the Lerner Index for the calibration of my theoretical model.

My primary source is USPTO patent data from PatentsView. This dataset contains disambiguated patent, inventor, and assignee identifiers, as well as Cooperative Patent Classification (CPC) classes for each of the patents registered from 1975 to 2021. I identify inventor flows across different sectors employing the ALP classification of 1976-2016 patents into NAICS sectors of application developed by **?**. Since this classification is constructed using the PATSTAT dataset, I rely on the crosswalk built by Gianluca Tarasconi to match these two sources.[1] This leaves me with one third of all the patents registered between 1975 and 2021, due to the restriction of the time frame to 1976-2016 and incomplete matching between PATSTAT and PatentsView. I comb patent records for self-citations, truncation-corrected forward citations, and patent generality, following **?** and **?**. I restrict my attention to utility patents, as I am interested in patents with a technological content and not just design improvements. **?** provide data on patents enforced in litigation cases between 2003 and 2016. This dataset allows me to analyze the evolution in patent litigation across NAICS 4-digit sectors for the sub-period 2002-2012.

My main source for concentration and sales data is the US Economic Census (EC), which reports sales shares for the top 4, 8, 20, and 50 firms; the Herfindal-Hirschman Index; sales and number of companies in various NAICS 4-digit sectors at a 5-year frequency. I focus on the period between 1997 and 2012 for three reasons. First, as I show below, this period saw substantial increase in the concentration of inventors in specific technology classes. Second, the start of this period coincides with an acceleration of market concentration and markups (see, e.g., **?**). Third, 1997 saw the adoption

[1] See https://patentsview.org/forum/7/topic/143, https://rawpatentdata.blogspot.com/2019/07/patstat-patentsview-concordance-update.html

of the NAICS classification, ensuring a consistent definition of product markets throughout the period I analyze. I select the HHI lower bound (**?**) as a measure of concentration, instead of the Economic Census HHI. This measure allows me to analyze a larger sample of sectors, since it requires only top sales shares in the Economic Census, which are available for a much wider set of industries than the census-computed HHI.[2] The Economic Census also provides sector-level growth in output per worker, which constitutes my main measure of productivity growth. I choose this measure instead of multi-factor productivity since the latter is available for selected sectors, mostly in manufacturing. I deflate sales using NAICS-specific price indices from the Bureau of Labor Statistics. All told, out of a total of 304 NAICS 4-digit sectors, I have assembled 157 business sectors for which I can measure the interrelation between concentration and knowledge markets.

I employ two additional data sources. First, I obtain sector-specific counts of regulations for various NAICS 4-digit sector from the Mercatus RegData 4.0 dataset. I employ them to conduct an instrumental variable analysis, strengthening the causal interpretation of my results.[3] Second, I use NBER-CES data to produce estimates of the Lerner Index following **?**.

1.2 Effective Inventors and Knowledge Markets

The main aim of this section is grouping product markets that share the same *required knowledge to innovate* and therefore compete for the same R&D inputs, namely inventors. I identify sectors that routinely exchange researchers through the Louvain community-detection algorithm (**?**).

Figure 1 exemplifies how I construct knowledge markets. Each node in the Figure represents a different NAICS 4-digit sector, and the black lines designate inventor flows, with thickness proportional to the size of the flow. I describe below how flows are computed. I employ a community detection algorithm to group together sectors most closely connected. Figure 1 depicts strong flows among grain, vegetable farming and animal food manufacturing, all of which involve knowledge related to agriculture and nutrition, and separately between footwear and tanning, which both require knowledge of leather crafting. In this illustrative example, my algorithm would identify two knowledge markets, one given by the agriculture and food manufacturing sectors, and the other by leather crafting sectors, leaving the laundry services sector isolated. Based on the strength of these connections, we would expect market conditions in footwear manufacturing to affect the distribution of inventors across the

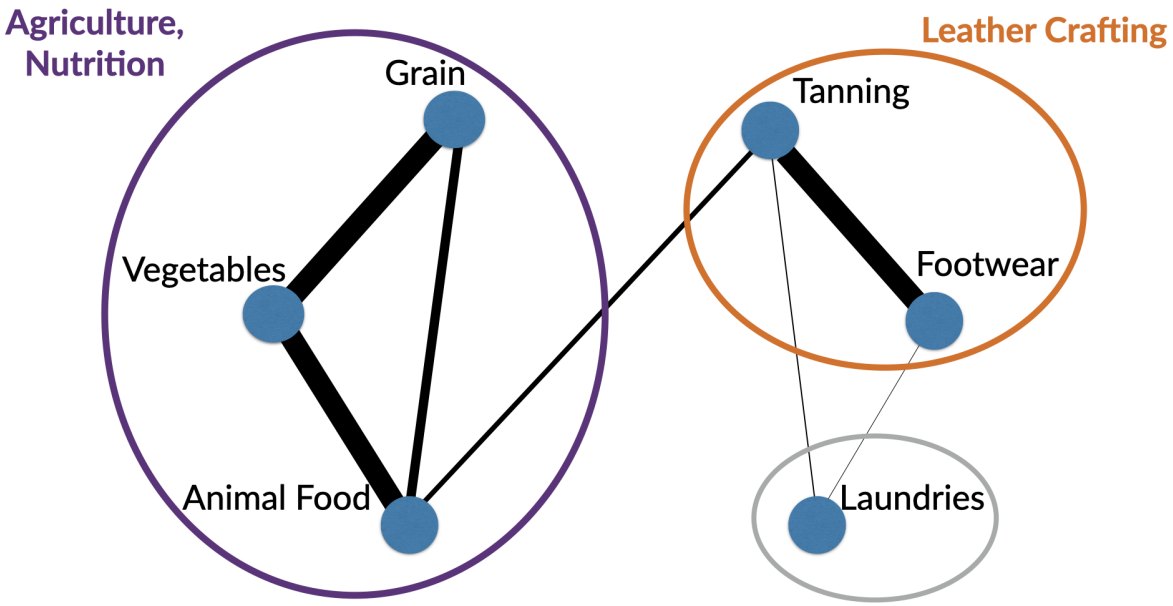
[2] The expression used to obtain this measure is:

$$\underline{\text{HHI}}_{p,t} = 4\left[\frac{\text{CR4}_{p,t}}{4}\right]^2 + 4\left[\frac{\text{CR8}_{p,t}-\text{CR4}_{p,t}}{4}\right]^2 + 12\left[\frac{\text{CR20}_{p,t}-\text{CR8}_{p,t}}{12}\right]^2 + 30\left[\frac{\text{CR50}_{p,t}-\text{CR20}_{p,t}}{30}\right]^2,$$

where "CR{X}" denotes the concentration ratio, that is the share of sales, of the top X firms. This measure is a lower bound, and coincides with the actual HHI if the sector has less than 50 firms, and sales share are distributed equally in each of the top 0-4, 5-8, 9-20, 21-50 brackets. **?** reports a correlation of $\underline{\text{HHI}}$ with the actual index of 0.93. Available at https://sites.google.com/site/drjankeil/data.

[3] Available at https://www.quantgov.org/bulk-download.

Figure 1: Graphical Illustration of Knowledge Markets



Note: This figure displays examples of knowledge markets as sets of product markets with the same required knowledge to innovate. This illustration is based on a subset of transitions and classifications from my data. Additional sectors inside and outside these knowledge markets are excluded for ease of exposition. Nodes represent NAICS sectors 1111 (Oilseed and Grain Farming), 1112 (Vegetable and Melon Farming), 3111 (Animal Food Manufacturing), 3161 (Leather and Hide Tanning and Finishing), 3162 (Footwear Manufacturing), and 8123 (Drycleaning and Laundry Services). Edges represent inventor transitions, with width proportional to the size of undirected inventor flows.

two leather crafting sectors, leaving agriculture and nutrition sectors largely unaffected.

Measuring Inventor Transitions I employ the USPTO patent data classified into 4-digit NAICS sectors by **?** to construct knowledge markets. Table 1 depicts a hypothetical matching of the USPTO dataset with NAICS classifications. Inventors are each assigned a disambiguated ID corresponding to the serial number of their first patent. In this example, inventor 00001-1 and 00001-2 both cooperate on the development of patent US00001. The third column in Table 1 shows the **?** classification for NAICS 4-digit industries. This classification is not limited to a single sector per patent, and includes multiple sectors in almost all instances. For instance, patent US00001 relates to multiple sectors, while patent US00002 is applicable to just one sector. Importantly, this classification captures the *technological nature* of the patent and the sectors of application of the knowledge required to develop that patent. While other classifications, like the CPC or the USPC, also describe the technological nature of patents, they do not allow a direct match to sectors of application.

I define a transition in two ways. First, I consider transitions *within patents*. This transition occurs between two sectors if an inventor works on a patent that applies to both. The direction of flows does not matter for the definition of knowledge markets, since I am only interested in grouping sectors that exchange researchers. Table 1 depicts two transitions between sectors 1111 and 1112 in 1980. The

Table 1: USPTO Data Structure

Patent ID	Inventor ID	? NAICS	Year
US00001	00001-1	1111	1980
US00001	00001-1	1112	1980
US00001	00001-2	1111	1980
US00001	00001-2	1112	1980
US00002	00001-1	3111	1981

Note: This table displays a hypothetical example of the structure of my data. The first two columns report patent and inventor identifiers from PatentsView; the third column reports NAICS 4-digit classifications.

second type of transition that I consider is *across patents*. This transition occurs when an inventor applies his knowledge to patents in different product markets, such as between sector 1112 and 3111 by inventor 00001-1. Raw transition counts are the basis of my measure of inventor flows.

Weighting Inventor Flows: Effective Inventors I construct two alternative measures to assess the effective flow of inventors across sectors. The first measure weighs each transition equally, computing inventor flows as the raw count of researchers moving across NAICS. The second measure adjusts for the productivity of individual inventors, since raw counts might overstate or understate the importance of each transition, depending on the size of origin and destination sectors, their technological nature, as well as the proficiency of each inventor. I therefore define a measure of "effective inventors" to correct for these and other omitted factors. For each inventor, I estimate the fixed effect, α_i, in the fully-saturated regression

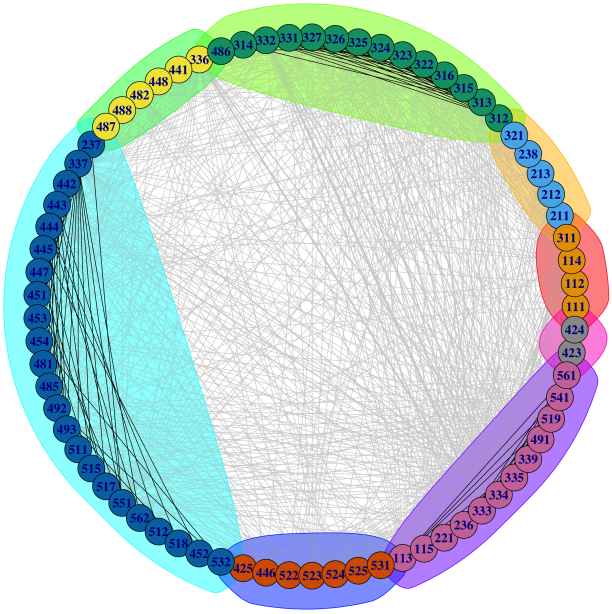
$$\#\text{Patents}_{cfit} = \alpha_i + \gamma_{cft} + \varepsilon_{cfit}, \tag{1}$$

where $\#\text{Patents}_{cfit}$ denotes the number of patents registered in CPC class c; firm (assignee) f; and year t, that include inventor i. In this regression γ_{cft} denotes a of CPC class by firm (assignee) by year fixed effect. I choose to include indicators for one-digit CPC classes, the broadest classification, to identify as many fixed effects as possible. The fixed effect γ_{cft} controls for specific technological features of the patented technology, the firm environment, as well as the year. Further, this specification produces an estimate of inventor productivity that accounts for the number of collaborators on each patent. Given this specification, I define an *effective inventor* as one unit of the resulting fixed effect α_i, rescaled to take nonnegative values. Since these fixed effects might be inconsistently estimated, I check the robustness of all my results, including the construction of knowledge markets, to the use of the raw count of inventors.

I define *effective inventor flows* between sector j and sector k at time t as:

$$flow_{j\to k,t} = \sum_i \#\{i\text{'s transitions } j \to k \text{ in } t\} \cdot \alpha_i,$$

Figure 2: Knowledge Markets Obtained from NAICS 3-digit Sectors



Note: This figure displays the network of inventor flows between NAICS 3-digit sectors and the knowledge markets resulting from the Louvain algorithm. Lines denote inventor transitions within (black) and between (gray) knowledge markets, with width proportional to effective undirected inventor flows.

that is, the sum of transition counts weighted by effective inventors. The *total undirected flow* between two sectors is the sum of inflows and outflows with ends in one of the two sectors over all sample years, t:

$$flow_{jk} = \sum_t \left(flow_{j \to k,t} + flow_{k \to j,t} \right).$$

This flow measure defines a network of inventor transitions across product markets, where the nodes, j,k, are given by 4-digit NAICS codes and edges are transitions across sectors. Edge weights are a rescaled version of $flow_{jk}$ and measure the strength of the connection between pairs of sectors in the network. Rescaling the flow measure is necessary to remove mechanical effects of sector sizes and to avoid double counting of inventors. More details are in Appendix **??**.

Community Detection and Resulting Knowledge Markets I use the rescaled undirected flow measure as a network edge weight to identify communities through the Louvain algorithm developed by **?**. This procedure maximizes the modularity of the network by choosing the number of communities (knowledge markets) and the assignment of nodes (NAICS sectors) to communities. Modularity, a commonly used measure of connectedness of networks, measures the distance between the density of links *within* communities versus *between*.

This procedures produces 10 sets of NAICS 4-digit sectors that share the same inventors and have

concentration measures. Applying the community detection algorithm results in knowledge markets that do not overlap: Each NAICS 4-digit sector belongs to one and only one knowledge market. Figure 2 displays the result of my procedure applied to NAICS 3-digit sectors. I report this exercise since the 4-digit equivalent would be too dense to depict. However, the knowledge markets identified by the two exercises are qualitatively similar although they are clearly more numerous in the 4-digit case. In this figure, lines denote inventor transitions, with width proportional to the effective undirected inventor flow between sectors. Black lines depict flows within knowledge markets, while gray lines represent transitions between communities.

Three features are worth emphasizing. First, the network is very dense, and transitions across 3-digit as well as 2-digit sectors are pervasive, differing largely in intensity. The approach I propose is therefore far more illuminating than grouping sectors based on broad product markets, which would neglect the linkages across disparate markets, or pooling all sectors together, which would neglect the difference in the strength of inventor flows. Second, the flows between communities appear more numerous than within communities, but this is solely a by-product of the circular layout of the network, whereby nodes mask flows within close communities on the circle. Less than a third of flows occur between communities, as expected since the community detection algorithm maximizes the density of within-community linkages. Third, the classification sensibly groups together sectors that we might expect to share similar knowledge to innovate. Starting from sector 111 and going counter-clockwise, the knowledge markets are as follows. The first market groups sectors involving agricultural production (111, 112 and 114) and food manufacturing (311). The second market, starting with 211, includes oil, gas, and mining. The green cluster at the top of the figure groups several heavy manufacturing industries, such as chemicals, plastics and petroleum products, and pipeline transportation (486). The market in yellow consists largely of transportation services and manufacturing as well as motor vehicle dealers. The large blue cluster captures many retail sectors, as well as data processing, telecom, and broadcasting services. The remaining three markets include insurance and finance (red cluster), computer, electronics, machinery manufacturing and professional services (violet), and wholesalers (gray).

I identify knowledge markets using effective inventors, but I obtain nearly identical results using raw inventor counts; more than 97% of 4-digit NAICS sectors are classified in the same manner using the two measures. That is, 97% of sector pairs belong to the same knowledge market according to both measures.

1.3 Other Constructed Measures and Aggregation at Census Frequency

Patent Citation Measures For each patent classified by **?**, I compute forward citations, a measure of patent generality, and self-citations. I describe self-citation measures together with the related

estimation results in the Appendix.

I count forward citations and measure patent generality following **?**. The forward citation measures compute the average number of citations received by each firm's patents, giving an indication of the importance of each patent for future technological developments. As in **?**, I correct for truncation by estimating the empirical CDF of the forward citations lag distribution of patents in the relevant CPC 2-digit technology class, and dividing the overall number of forward citations at the latest available date by the inverse of this CDF. The procedure suggested by **?** uses only information pertaining to the CPC 2-digit technology class of the cited patent. I also compute an alternative correction that estimates a separate distribution for each citing CPC 2-digit class and sums the corrected forward citations across all citing classes. Patent generality also measures the scope of application of the patent by computing the dispersion of citations received across different CPC classes. The higher the dispersion, the wider the technological applicability of the patent.[4]

Regulation Data Mercatus RegData provides a count of restrictions imposed on a number of NAICS 4-digit product markets, obtained by matching a set of keywords in NAICS descriptions to regulatory texts, and then taking the best match for each document. However, the available data does not include a set of codes due to data quality reasons. Therefore, I process the description of NAICS 4-digit codes and compute the cosine-similarity between all pairs of sectors. I build an estimate of sector-relevant restrictions for missing sectors by taking an average weighted by cosine similarity of sectors included in RegData. I include in the average the five most similar NAICS codes if similarity is larger than .2, and I attribute the regulations of the most similar sector otherwise. I chose this threshold by inspecting the similarity associated to various NAICS pairs, and the assignment of regulations to sectors is not highly sensitive to this choice.

Inventor Distribution Measures I employ the measure of effective inventors constructed as detailed above to compute measures of researchers' concentration within sectors for each year in my sample. Specifically, I use the PatentsView assignee ID to identify firms that employ specific inventors in each sector, and then compute several measures of the concentration of inventors within sectors. I also compute the HHI of inventors across NAICS to document increasing concentration of inventors in specific sectors.

Patent Litigation Cases I match the data on litigation cases compiled by **?** with the data on inventors by NAICS 4-digit. For each sector, I compute the number of litigation cases per patent. These data are available only for the sub-period 2003-2016, which does not allow me to reliably estimate an empirical CDF to correct for truncation. I therefore choose to keep only the litigation cases occurring in the same

[4] The interested reader should consult **?** for a detailed discussion, and the related appendix for details on the construction of these measures.

year as the patent registration, which amounts to assuming that the time profile of cases is constant over time and across sectors.[5] I then average the litigations per patent over the years 2003-2006 and 2013-2016 for the Economic Census waves 2002 and 2012.

Aggregation at Census Frequency Data from the Economic Census are available at five-year intervals for the years 1997-2017, which requires aggregating the other data at the same frequency. Since I am interested in the effect of concentration on the allocation of inventors, I average all variables related to inventors and productivity using the five-year window *starting* in the census year (e.g., 1997- 2001 for 1997), while I use concentration measures for the corresponding census year. In the IV regression I use product restrictions as an instrument for concentration, which is why I average restrictions in the five-year window *ending* in the census year (e.g., 1993-1997 for 1997). Since **?**'s matching only covers the period up to 2016, I run all specifications in long-differences over the time frame 1997-2012. The only exception is the patent litigation regression, which uses the period 2002-2012.

[5] More precisely, this would be the same as using only the contemporaneous patent litigation cases and correcting for truncation dividing by the inverse CDF at period 0, which would scale the estimated coefficient upwards.