Memo on Inventor Allocation

July 18, 2021

This memo consists on three parts, summarized below. In the email I discussed how I wanted to look at a fuller set of patents, and then figure out how to match it to Compustat. I did the former but none of the latter (perhaps it is not so necessary). On theory, you suggested working on a simple two-period model. I cannot seem to solve things analytically easily, but I can get some expressions that relate markups to research intensity, although not fully worked out. But I can get some early results in Matlab that make sense according to the story.

1. Data updates:

   (a) New sources:

      i. all patents (PATSTAT), classified by Zolas et al. (2016) at the Census into NAICS 6d, who kindly shared non-published data with me,

      ii. Xwalk between PATSTAT and PatentsView (my main source), I can match 30% of patents to classification above

      iii. Census measures of concentration at NAICS 4d, unfortunately just about 60 sectors, and 5-year intervals

      iv. Mercatus count of market restrictions at NAICS 4d to instrument for increase in concentration above (is the exclusion restriction satisfied?)

   (b) New data constructed:

      i. Inventor productivity along the line of the presentation, but on all patents assigned to companies, public or not, instead of just compustat-matched

      ii. Inventor productivity as the inventor fixed effect α_i from either of the regressions:

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

$$\#\text{Patents}_{cfit} = \alpha_i + \alpha_{cf} + \alpha_{ct} + \alpha_{ft} + \varepsilon_{cfit} \tag{2}$$

      at the level of CPC (Cooperative Patent Classification) class, c, assignee/company, f, inventor i and year t. I prefer 1, but clearly 2 allows the identification of more fixed effects. All these measures are strongly correlated, which is reassuring.

      iii. Concentration measured by Gini in different CPC classes and NAICS 4d as mapped from above

   (c) New findings:

      i. Inequality in effective inventors across CPC classes as well as NAICS 4d, as measured by the relevant Gini has increased substantially

      ii. Long-difference regression for 1997-2012 for share of effective inventors (fixed effects) over Census concentration measure, with knowledge markets fixed effects, report positive and significant coefficients, weighted by sector sales. Similar finding when using Mercatus product-specific restrictions as instruments.

   (d) To do:

      i. Re-compute knowledge markets on all patents, a lot more data than just Compustat.

      ii. Some accounting, how? Compute predicted share increase versus actual seems the natural place to start.

      iii. A lot of potential work with Compustat data: firm-specific markups; match with overall patent data and segment (4-digit) information, etc...

iv. Curious to see how markets are divided into upstream/downstream given the Bottlenecks paper (great presentation!)

2. Theory updates:

   (a) Done: simple two-period model based on a two- or one-stage Cournot game.

   (b) Findings similar to what I presented.

   (c) To do: Work on equilibrium.

   (d) Issues: I had to assume that there are full technological spillovers that are not accounted for by firms to avoid mechanical duplication. However this is also part of the story.

3. Steps forward:

   (a) Model remains an issue: prioritize clarity or quantification potential?

      i. Things get non-analytical real fast.

      ii. What is your take on existing literature to fit my analysis.

   (b) Empirical analysis: integration of new data from all patents with Compustat segments. Use higher-frequency concentration data from Compustat to interpolate the Census concentration? Still the size issue, model did not clarify much.

   (c) Establish a general flavor of the paper and potential.

1 Data updates

1.1 Description of new sources

My new data consists of five main elements. The first is the complete set of PatentsView utility patents that have as assignee a company (I could do all but this seemed most reasonable). The second is a mapping constructed by Zolas et al. (2016) between PATSTAT patents and NAICS 6-digit sectors through text analysis. The third is a crosswalk between PATSTAT application ID and patent ID in PatentsView built by Gianluca Tarasconi in 2019 (http://rawpatentdata.blogspot.com/2019/). Thus is a ready-made alternative to digging into PATSTAT. I could not find the documentation but it does allow to match about 30% of the PatentsView patents to those analyzed by Zolas.[1] The fourth is the set of concentration measures for the US Economic Census for 1997-2002-2007-2012. I can extend to 2017 but have not done it yet. The fifth is the Mercatus QuantGov database of counts of product restrictions from legislative sources from 1970-2020.

1.1.1 Data limitations

There are two main issues with the data, which emerge primarily in the regression analysis. First, the time frame is strongly limited. Zolas et al. classified patents only to 2016, the closest available economic census is 2017, and most importantly, there are no NAICS 4d concentration measures before 1997 outside manufacturing, and even the ones available afterwards cover only a subset of around 80 sectors, which include predominantly manufacturing. Second, the QuantGov database covers a non overlapping subset of 4-digit sectors, so that I can do IV only on 39 observations, but that is perhaps ok. I tried using regulations at 3-digits, but that is a very poor instrument for concentration changes at 4-digit. I could recover 5 observations by re-building the knowledge markets on all patents, rather than the Compustat set.

[1] I suspect that the small matching rate is due to the fact that PATSTAT has all applications, even those that do not necessarily result in a patent, or that do so with substantial lag. Further the two dataset do not overlap fully in general. I shall contact the author and dig more into the details perhaps.

1.2 New computations of effective scientists

As discussed above, I compute inventor productivity as the inventor fixed effect α_i from either of the regressions:

$$\#\text{Patents}_{cfit} = \alpha_i + \alpha_{cft} + \varepsilon_{cfit} \tag{3}$$

$$\#\text{Patents}_{cfit} = \alpha_i + \alpha_{cf} + \alpha_{ct} + \alpha_{ft} + \varepsilon_{cfit} \tag{4}$$

at the level of CPC (Cooperative Patent Classification) class, c, assignee/company, f, inventor i and year t. I prefer 3, but clearly 4 allows the identification of more fixed effects. All these measures are strongly correlated, which is reassuring. I compute it for various levels of the CPC classification (1, 3, 4 digits). I also compute an alternative measure as in the presentation, that is the number of patents per capita by each inventor in each year. I compute both an average productivity as well as the total number of effective patents. The correlation between all these measures is quite high, which is reassuring. My preferred measure is the α_i from 3, at the CPC level 1, the most aggregate level for patents. The reason is that I still fully saturate, but I can identify a lot more fixed effects than narrower classification. From my checks the measures are pretty close when they are both computed.

1.3 Increase in concentration across patent classes and Naics sectors

Given the above construction, I look at two things. First, I compute Gini coefficients of effective inventors, α_i (shifted to be nonnegative) across patent classes. This is reported in Figure 1. When looking outside Compustat only, the levels of concentration are more reasonable, and so is the increase.

Second, I use the subset of patents classified by Zolas (which ends in 2016) that I can match to PatentsView. This is shown in Figure 2. In both cases the coefficient increased by about 10% from 1978.

Figure 1: Effective inventor Gini coefficients at CPC-4 level

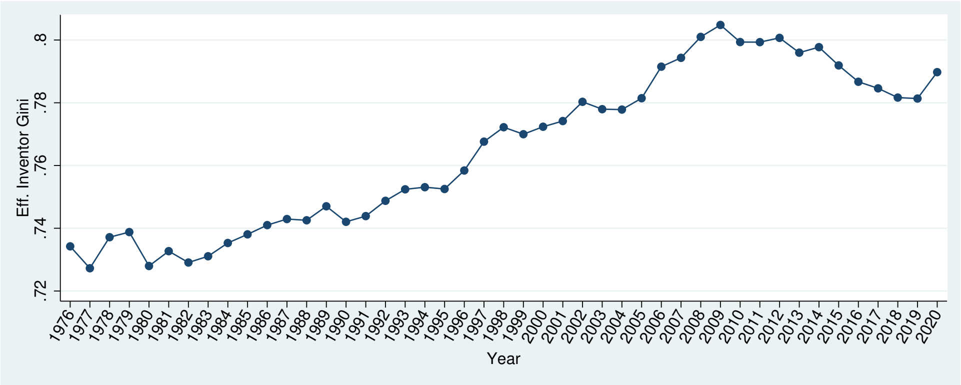
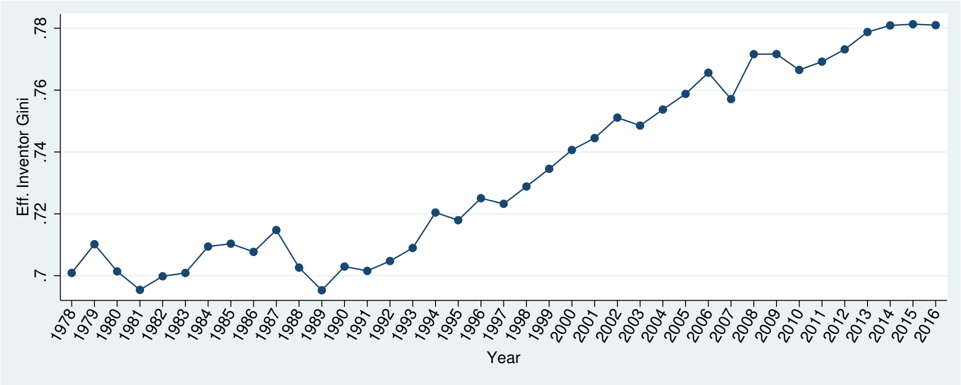


Figure 2: Effective inventor Gini coefficients at NAICS 4-digit level



1.4 Scatters and Regressions of Competition and Inventor share

The next step is computing the inventor shares of various sectors from above. To do so, I match the data above with the knowledge markets computed on NAICS 3-digits from Compustat. I see this as a first pass, while I construct more precise markets from the classified company patents, which will hopefully be at 4-digit. I then compute the share of effective inventors for each 4-digit NAICS, and run a long-difference specification between 1997-2012, weighted by Census-reported sales, and residualized by knowledge market:

$$\Delta\text{Share}_{pk} = f_k + \beta\Delta\text{Concentration}_p + \varepsilon_{pk}.$$

In a second analysis, I instrument the concentration change by the change in the number of NAICS 4-digit specific restrictions from Mercatus. I think the exclusion restriction is reasonable, although increase in product regulations might also increase the need to develop appropriate technologies and thus drive up the inventor share. However, I do not think that this effect ought to be major. Table 1 reports the results of the OLS ad IV regression. As I noted above, the Mercatus regulation measure is only available for a subset of sectors, so the sample is greatly reduced in the IV estimation. Two results stand out. First, the OLS specification shows a significant and positive effect, an increase of 1pp in the top 4 share in a NAICS 4-digit market results in an increase of 0.03pp in that sector's share of the relevant inventors.[2] Moving to the IV, the coefficient is reduced by about a third. I think this is reasonable in light of the reverse causality that logically exists between changes in concentration and inventors. The F-statistic is rather low, so I might want to look at Anderson-Rubin confidence intervals in the future. Graphically, Figure 3 displays a binscatter of changes residualized by knowledge market fixed effects (I made sure that the estimated slope holds when winsorizing at 1% as well).

Table 1: Long-differences specification, 1997-2012

	(1)	(2)
	OLS	IV
	Change Share Inventors	Change Share Inventors
Change Top-4 Sale Share	0.0338**	0.0201*
	(0.0164)	(0.0104)
Constant	-0.351*	
	(0.199)	
Observations	80	34
R-squared	0.212	0.032
K-market FE	Yes	Yes
First-stage F		12.75
AR Wald p-val.		0.0192

Regression weighted by total sales in 2012. Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

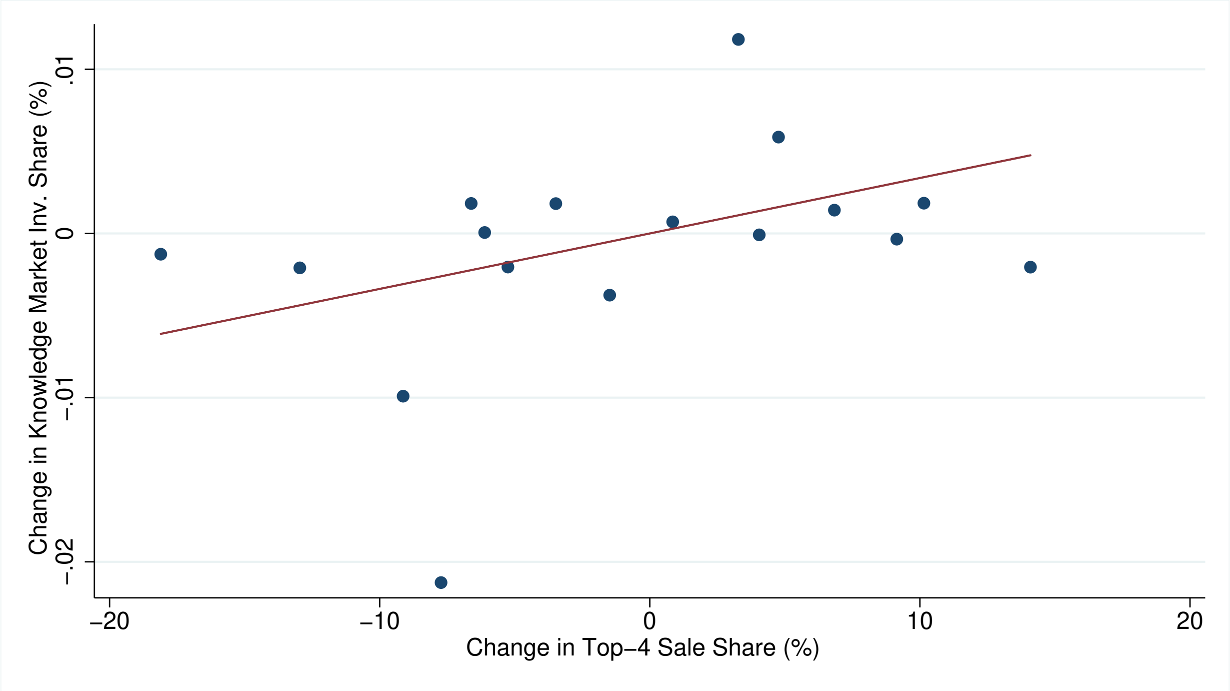
2 Theory updates: model sketch and simulations

The model is a very simple two Cournot-Nash game, where firms can choose quantity as well as hiring of R&D labor. In what follow, I solve the problem of a single sector with N firms that are Cournot competitors. I assume an isoelastic demand for the good produced by the sector:

$$Q = P^{-\varepsilon}, \quad \varepsilon > 1$$

[2] This effect might seem small, but it is important to consider that the largest inventor market can encompass as many as 40 sectors, in which case an even split of inventors across markets would result in a share of 2.5pp per sector. In this scenario a 1pp increase in the top-4 share would increase the inventor share by about 1%. Another way to get a sense of the magnitude is considering that the coefficient is 10% of the constant and that the (within) R-squared is around .21.

Figure 3: Binscatter of 1997-2020 change in inventor share over changes in NAICS 4-digit top-4 sale share



Note: Regression line is estimated weighting by total sales in 2020.

The production function is linear in labor, so the firm i's operating profits are:

$$\pi_i = Pq_i - \tilde{c}_i w q_i,$$

where P is the good's price, w is the production wage and $\tilde{c}_i$ is the firm's labor requirement per unit of output. This labor requirement is determined as:[3]

$$\tilde{c}_i = \frac{c}{(1+x_i)^{\gamma}} \quad \gamma(\varepsilon - 1) < 1,$$

where x_i is the firm's R&D labor. Each R&D worker commands a wage w^{RD} so total profits are given by:

$$\pi_i - w^{RD} x_i.$$

There are two ways of setting the problem. The first assumes that R&D labor and quantity are set in the same period. The second assumes that the game has two stages. First, firms choose R&D spending taking other firms' spending as given, and then they play a Cournot quantity game. The resulting demands for R&D are different but both setups give the desired result.

2.1 One-stage game

Firms choose R&D and quantities simultaneously, taking others' quantities and R&D as given:

$$\begin{aligned} \max_{q_i, x_i} \quad & q_i \left[P - \tilde{c}_i w\right] - w^{RD} x_i \\ \text{s.t.} \quad & \tilde{c}_i = \frac{c_i}{(1+x_i)^{\gamma}}, \; \gamma(\varepsilon - 1) < 1 \\ & P = \left(\sum q_i\right)^{-\frac{1}{\varepsilon}} \end{aligned}$$

[3] The condition on γ is required to have a downward sloping demand for R&D labor.

This gives the system of equations:

$$P\left[1-\frac{s_i}{\varepsilon}\right] = \tilde{c}_i w, \quad \forall i, \tag{5}$$
$$\sum s_k = 1,$$
$$\gamma \frac{c_i}{(1+x_i)^{1+\gamma}} w q_i = w^{RD}, \quad \forall i, \tag{6}$$

with $s_i = q_i/Q$. The set of equations 5 implies:

$$P^{\star} = \frac{\varepsilon}{\varepsilon N - 1} \sum_{i=1}^{N} \tilde{c}_i w \tag{7}$$

and corresponding quantities:

$$q_i^{\star} = P^{\star - \varepsilon}\left[\varepsilon - (\varepsilon N - 1)\frac{\tilde{c}_i}{\sum_{j=1}^{N} \tilde{c}_j}\right] \tag{8}$$

This general solution can be explored further assuming that the N firms are symmetric, $c_i = c$ for all i. In this case the equilibrium quantities and R&D are jointly determined by he system:

$$q^{\star} = \frac{1}{N}\left[\frac{\varepsilon N}{\varepsilon N - 1}\frac{c_i}{(1+x^{\star})^{\gamma}} w\right]^{-\varepsilon}, \tag{9}$$

$$\frac{w}{w^{RD}} \gamma c_i q_i = \left(1 + x_i^{\star}\right)^{1+\gamma}. \tag{10}$$

Finally yielding:

$$x^{\star} = \left(\frac{\varepsilon N - 1}{\varepsilon N}\left[\frac{\varepsilon N}{\varepsilon N - 1} w c\right]^{1-\varepsilon} \frac{1}{N}\frac{\gamma}{w^{RD}}\right)^{\frac{1}{1-\gamma(\varepsilon-1)}} - 1 \tag{11}$$

In this model, equilibrium R&D labor is an increasing function of the quantity produced by the individual firm. Thus, ceteris paribus, it decreases with a larger markup, $\frac{\varepsilon N}{\varepsilon N - 1}$, and with the number of firms N. This inevitably gives rise to negative duplication effects in more competitive (larger N) markets, which mechanically results in R&D being more productive in less competitive environments. As a result, when doing simulations, I assume that aggregate R&D, X, spills over to individual firms, who take aggregate R&D as given :

$$\tilde{c}_i = \frac{c_i}{\left(1 + x_i + \alpha \frac{(N-1)^{\xi}}{N} X\right)^{\gamma}}. \tag{12}$$

In the simulations below I assume full spillovers, $\alpha = \xi = 1$, so that there is no waste from duplication. Solving the problem with this specification gives gives individual R&D:

$$x^{\star} = \frac{1}{N}\left[\left(\frac{\varepsilon N - 1}{\varepsilon N}\left[\frac{\varepsilon N}{\varepsilon N - 1} w c\right]^{1-\varepsilon} \frac{1}{N}\frac{\gamma}{w^{RD}}\right)^{\frac{1}{1-\gamma(\varepsilon-1)}} - 1\right],$$

so that aggregate R&D corresponds to individual R&D without spillovers.

2.2 Two-stage game

In this solution, firms play Cournot in period 2, and choose R&D strategically in period 1. The problem in period 1 is then:

$$\begin{aligned} \max_{x_i} \quad & q_i\left[P - \tilde{c}_i w\right] - w^{RD} x_i \\ \text{s.t.} \quad & q_i = q_i^{\star}(x_i, x_{-i}),\ P = P^{\star}(x_i, x_{-i}) \\ & \tilde{c}_i = \frac{c}{(1+x_i)^{\gamma}},\ \gamma(\varepsilon - 1) < 1, \end{aligned}$$

where $q_i^\star, P^\star$ denote the Cournot equilibrium price and quantities given R&D choices. Assuming symmetry and full spillovers we get individual R&D:

$$x^\star = \frac{1}{N}\left[\left(\left(\frac{\varepsilon N-1}{\varepsilon N}+\frac{N-1}{N}\left[1-\frac{1+\varepsilon}{\varepsilon N}\right]\right)\left[\frac{\varepsilon N}{\varepsilon N-1}wc\right]^{1-\varepsilon}\frac{1}{N}\frac{\gamma}{w^{RD}}\right)^{\frac{1}{1-\gamma(\varepsilon-1)}}-1\right] \tag{13}$$

Thus, firms conduct more R&D than in the one-stage game. This extra "strategic" R&D in equilibrium is governed by the term:

$$S \equiv \frac{N-1}{N}\left[1-\frac{1+\varepsilon}{\varepsilon N}\right],$$

which is unsurprisingly 0 when the firm is a monopolist, $N = 1$, and increases towards 1 as N grows.

2.3 Simulation

The following graphs display the properties of main aggregates in a set of alternative economies with different numbers of firms. Note that here I am setting a given wage for production and R&D workers and looking at sector demands for R&D and ensuing equilibrium quantities, growth and marginal product of inventors, computed as:

$$MP(X^*) = \frac{\partial Q(X^*)}{\partial X},$$

further note that in this model, sector growth is just:

$$g = \mathrm{d}\log Q(X^*) = \gamma\varepsilon\log(1+X^*).$$

Figure 4 displays the main quantities of interest relative to their value under monopoly, for both the one-stage and two-stage games. Both feature qualitatively similar features, although the two-stage game has a higher value of R&D due to the strategic interaction between firms. This strategic effect is reponsible for the hump in total R&D, output and growth in the respective panels. The concavity of sectoral R&D returns in turn implies that growth per inventor and marginal products increase with the number of firms, since demand for inventors falls with the number of firms. It is evident from the first-order condition 10 that R&D is decreasing in the number of firms due to a market-size effect. The more firms are active, the lower the fraction of the market captured by the individual firm, and the lower its incentive to conduct R&D. Figure 5 reports the case where I assume that there are no spillovers across firms. As is evident from the botton-middle panel here duplication kicks in after a point, reducing total growth per inventor. Of course all these results are for an illustrative calibration just to highlight the properties of the model and should not be taken too seriously.

2.4 Multiple sectors (in progress)

Next, I intend to endogenize the wage of production and research workers and put together the multiple sectors above. I could then match the distribution of the number of firms with the empirical distribution of top-4 sale shares in the data, and do some computations under reasonable parameters. Solving the model in equilibrium requires micro-founding the downward-sloping demand:

$$Y = \left[\sum_{k=1}^{S} Q_k^{\frac{\varepsilon-1}{\varepsilon}}\right]^{\frac{\varepsilon}{\varepsilon-1}},$$

where S is the number of sectors in the economy and each sector k behaves as described above. Taking the price of the final good as numeraire, equilibrium R&D is defined implicitly by a system of equations, that should be combined with market clearing for production and R&D labor:

$$\begin{aligned}
x_k^\star &= \left(\left(\frac{1}{\mu_k}+\frac{N_k-1}{N_k}\left[1-\frac{1+\varepsilon}{\varepsilon N_k}\right]\right)\left[wc_k\mu_k\right]^{1-\varepsilon}\frac{1}{N_k}\frac{\gamma\mathbf{Y}}{w^{RD}}\right)^{\frac{1}{1-\gamma(\varepsilon-1)}}-1,\ k=1,\dots S-1\\
Y &= \frac{\left(\sum_{k=1}^{N_s}\left[\mu_k\tilde{c}_k\right]^{1-\varepsilon}\right)^{\frac{\varepsilon}{\varepsilon-1}}}{\sum_{k=1}^{N_s}\tilde{c}_k^{1-\varepsilon}\mu_k^{-\varepsilon}},
\end{aligned}$$

where I define the markup, $\mu_k \equiv \frac{\varepsilon N_k}{\varepsilon N_k - 1}$. Clearly, everything goes numerical quite fast.

Figure 4: Equilibrium Objects, Index Relative to Monopoly, No Duplication of R&D

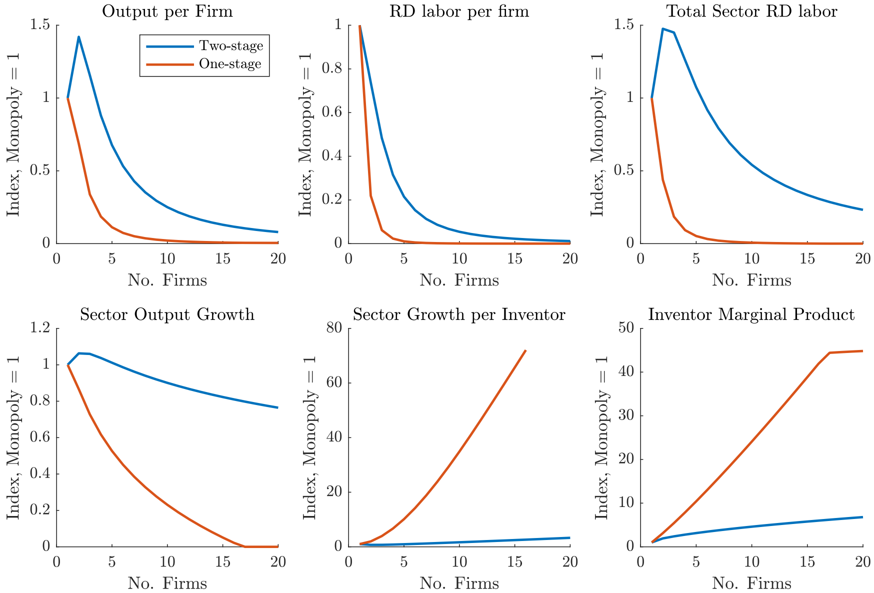
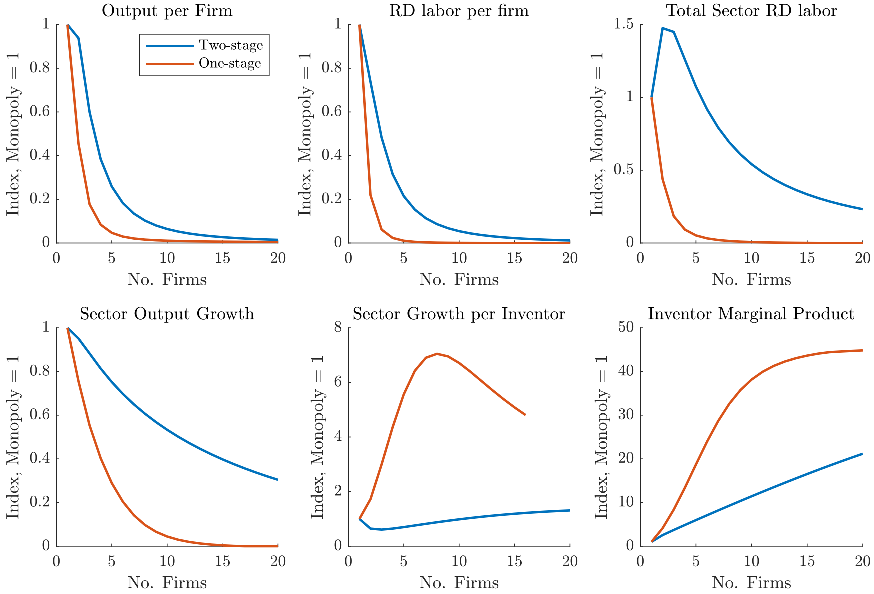


Figure 5: Equilibrium Objects, Index Relative to Monopoly, Duplication of R&D



3 Next Steps

To sum up, I think the empirics is getting stronger, and the theory a bit clearer. But there is a long way to go.

I think I need plenty of advice on the theory side and on the quantification of the mechanism. I think the above lays out the basic mechanism that I want to highlight, and I do not mind too much that the reason for misallocation is essentially that firms are larger in monopoly. It seems reasonable that firms would accept more incremental innovations that they can spread on many products, and this insight is consistent with e.g. Acemoglu and Linn and a number of papers by Ufuk on incremental innovation versus radical.

However, the model is highly mechanical, in the sense that misallocation arises somewhat trivially from the fact that there are decreasing returns to R&D, which makes the mechanism really not that novel. In addition, what this simple theory implies is that size cannot be really decoupled from competition when doing some empirical analysis.

Clearly, I need a more serious model for quantification, but I am wary of spending a lot of time learning a new model that might not work, so I would welcome your advice on where to direct my energies.

On the empirical side, I think the situation is much better. While the instrumental analysis has a lot of weaknesses, I think it conveys the suggestion very well, reinforcing the raw correlation. Perhaps I can use the Mercatus instrument on Compustat, where I could also build better measures of scientist share, and move to firm-specific markups. The chief weakness is that Compustat is awful to measure concentration. I know that there are methods to interpolate low-frequency data using higher frequency, so I could try to use Compustat concentration to predict overall 4d concentration at a frequency higher than the Census 5-years, and extend the analysis to include firm controls. I believe this would give quite some extra precision.

Regardless of next Compustat steps, the thing I wanted to do is quantify even just the correlation that I obtain in the regressions above. I am unsure how to measure the effects on TFP? I would still need a model to aggregate them...