THE SERVICE ECONOMY AND THE RISE IN MARKUPS

RICARDO MARTO
Federal Reserve Bank of St. Louis

This paper examines the role of structural transformation toward services in explaining the evolution of markups. I document that the expansion and markup dynamics of the services sector account for most of the changes in market power. To interpret the evidence, I develop a multi-sector general equilibrium model with imperfect competition and novel non-homothetic preferences, in which household income and sectoral prices jointly determine the income and price elasticities of demand. As a result, markups evolve endogenously in response to income growth and structural transformation. Calibrated to U.S. data, the model replicates key macroeconomic trends between 1955 and 2020—including the rise in markups, the decline in the labor share, the shift toward services, and changes in relative prices—without requiring increased concentration or reduced competition. These findings suggest that rising markups can also reflect broader forces of economic development.

KEYWORDS: Endogenous markups, entry costs, income inequality, non-homothetic preferences, price elasticity of demand, services, structural change, structural transformation, technological progress.

JEL CLASSIFICATION: D11, D12, D22, D43, E21, E23, L11, L16, O41, O47.

Ricardo Marto: ricardo.marto@stls.frb.org

I am deeply indebted to my advisors Jeremy Greenwood, Aviv Nevo, and Joachim Hubmer for their guidance and continued support. I also thank Harun Alp, Yoshiki Ando, Gadi Barlevy, Harold Cole, Alessandro Dovis, Nezih Guner, Dohan Kim, Dirk Krueger, Agnieszka Markiewicz, Guillermo Ordoñez, Jordan Peeples, Michael Peters, Víctor Ríos-Rull, Raul Santaeulalia-Llopis, Ina Simonovska, Akos Valentinyi, and Jon Willis for their insightful comments and suggestions. I am also grateful to seminar participants at Arizona State University, Asian and European Meetings of the Econometric Society, Bundesbank's Structural Change conference, Católica Lisbon, Computing in Economics and Finance conference, ECB, ESADE, Federal Reserve Bank of Atlanta Firm Dynamics and the Macroeconomy conference, Federal Reserve Bank of Chicago Rookie conference, Federal Reserve Bank of St. Louis, Federal Reserve Board, FGV/EPGE, Insper, IMF, Inter-American Development Bank, Lisbon Macro conference, Penn State, PUC-Rio, SED, University of Oxford, University of Pennsylvania, University of Toronto, and Virtual Family Macro seminar for helpful suggestions. I thank the EEA for the 2023 Young Economist Award. The views expressed here are those of the author and do not necessarily reflect the views of the Federal Reserve Bank of St. Louis or the Federal Reserve System.

1. Introduction

The rise of firm markups in the U.S. economy has become one of the most debated macroeconomic trends in recent decades. Estimates of markups suggest they have risen markedly over the past several decades. This trend has sparked intense debate in both academic and policy circles. Are rising markups a symptom of declining competition and growing monopoly power, or could they reflect deeper structural changes in the economy?

Despite the attention this phenomenon has received, the underlying causes of rising markups remain an open question. Three broad explanations have emerged. The first attributes the rise in markups to increasing concentration and the emergence of superstar firms. Enabled by technological advantages and economies of scale, these firms charge higher markups than the average firm and have captured a growing share of output. The second view emphasizes entry barriers and declining business dynamism, suggesting that weakening competitive forces have allowed incumbent firms to exert greater pricing power. The third cautions that rising markups may reflect mismeasurement, due to assumptions about production technologies and costs.

This paper offers a complementary perspective. I argue that the long-run increase in aggregate markups is closely tied to structural change—specifically, the expansion and markup dynamics of the services sector. I begin by documenting four empirical patterns from 1955 to 2020 (Section 2): (i) the services share of value added rose from 53% to nearly 79% and of variable costs increased from 35% to 70%; (ii) the relative price of services increased by about 40%; (iii) the aggregate markup rose by 12 percentage points, a trend driven almost entirely by the services sector; and (iv) real economic profits grew substantially in services but remained flat in goods-producing industries. The importance of the services sector is robust to alternative measurements (Section A of the Supplemental Appendix discusses these in detail).

Models of structural change rely on two ingredients to generate the transition toward services and the rise of the relative price of services: differential rates of technological progress across sectors and non-homothetic preferences. I argue that these are sufficient to also engender a rise in markups when markets are imperfectly competitive. To do so, I develop a multi-sector general equilibrium model with imperfect competition in the final goods and services and a novel class of non-homothetic preferences (Section 3).

The key innovation of the model is to allow the price elasticity of demand to vary endogenously with household income and sectoral prices, linking rising incomes to firm-level markups. The latter captures the intuition that as consumers become wealthier, they are less sensitive to price changes. This is consistent with Harrod's (1936) *Law of Diminishing Elasticity of Demand* and recent empirical evidence from retail scanner data (Döpper et al. (2022), Sangani (2023)) Firms internalize these elasticities when setting prices, enabling them to raise markups even in the absence of changes in market structure or entry barriers. Section B of the

Supplemental Appendix provides the theoretical underpinnings for this avenue and Section C.1 explains why alternative preferences (e.g., Kimball or non-homothetic CES) are not suitable to address this problem.[1,2]

Models with monopolistic and oligopolistic competition are built and matched to U.S. data. Both successfully replicate key macroeconomic trends over the past 65 years, including the rise in aggregate markups, the decline of the labor share, the transition toward services, and the increase in the relative price of services (Section 4).

Two counterfactual experiments are performed: either (i) changes in total factor productivities, or (ii) changes in entry costs, are switched off over time (Section 5). In the monopolistically competitive economy, rising incomes fully account for the increase in markups. That is a natural consequence of firms' markups depending solely on consumers' price elasticity of demand. As productivity grows, consumers get richer and their demand for goods and services becomes less elastic. Under oligopolistic competition, a firm's markup also depends on its sales share. Changes in entry costs now affect the number of firms operating in the market and as a result their markups. Entry frictions matter quantitatively in that economy. Yet, rising incomes from technological change remains the dominant driver of the increase in markups (explaining 65% of the increase between 1955 and 2020).

Attributing the rise in markups exclusively to technological progress is misleading, however. Although it is a necessary condition for markups to grow, technological progress is not sufficient (Section 6). Without income effects, it would simply translate into lower prices.

Another important trend of the past 65 years is the increase in income inequality. How much did it impact the rise in markups? I disentangle the role of income inequality from the rise in incomes by allowing consumers to be heterogeneous in skills (Section 7). Both models with monopolistic and oligopolistic competition are simulated to also match the rise in income in-

[1]The appendix presents two key propositions to establish the effect of household income and product prices on a consumer's price elasticity of demand and therefore markups. The first proposition provides conditions for the price elasticity of demand to be decreasing in the consumer's income. This relates to Harrod's (1936) *Law of Diminishing Elasticity of Demand*, which Bretherton (1937) summarizes as follows *"...as people's incomes become larger, the ratio between the trouble involved in finding the cheapest market, and the real gain in utility which will result in so doing, increases."* This statement is interpreted here as implying that a consumer's price elasticity of demand for goods and services is lower the wealthier they are. The second proposition establishes the conditions for the price elasticity of demand to be increasing in the product's price. This relates to Marshall's (1890) *Second Law of Demand* and implies that demand is more elastic at higher prices.

[2]The two forces of structural change work as follows. As productivity grows faster in the non-services sector, these firms' marginal costs decline at a faster rate, allowing them to reduce the price of manufactured goods. As consumers' price elasticity of demand falls with lower prices, the cost pass-through is less than one. Hence, some of the efficiency gains will be retained by firms in the form of higher markups. This leads to an increase in the average markup of non-services firms and relative price of services, but also a decline of the services share as consumers buy more goods at lower prices. Income effects play the countervailing role. As household income grows, commodities that were luxuries become more accessible and consumption starts flowing toward the sector providing them, i.e., the services sector. As household income increases, their price elasticity of demand falls. As a result, firms are able to command higher markups, explaining jointly the rise of the services share and the average markup of services.

equality. Two additional counterfactual experiments are conducted. One in which inequality stays constant at its 1955 level, but incomes grow; and another in which inequality grows as in the data but the aggregate income stays constant over time. Although markups would be lower with less income inequality starting in 2010, rising living standards are across the board the major driver of the increase in markups.

The model calibrated to the U.S. economy can also offer predictions about the evolution of markups across many other countries over long time horizons (Section 8).

Section 9 concludes, pointing that these results do not imply that concerns about competition policy are misplaced. Rising markups may still reflect increased concentration in specific markets. But they caution against interpreting markup increases as evidence of declining competition. In an economy experiencing income growth and structural transformation, changes in consumer behavior can lead to higher markups even under monopolistic competition.

The paper makes several contributions. First, it offers an alternative to explanations focused solely on supply-side distortions. Second, it introduces a tractable general equilibrium framework that links endogenous markups to patterns of consumption over time, yielding closed-form expressions for firm-level markups. Third, it combines macro and micro data to ground the theoretical mechanism. Finally, it connects the literatures on structural change and imperfect competition, showing how shifts in sectoral composition can shape markups and prices.

Related literature. This paper connects and extends multiple strands of the literature on markups, structural change, and the evolution of firm behavior in general equilibrium settings.

First, a growing body of work documents a long-run increase in markups in the U.S., raising concerns about declining competition (De Loecker et al. (2020), Hall (2018)). Autor et al. (2020) linked these trends to the emergence of superstar firms. Related work emphasizes declining entry and weakened business dynamism (Decker et al. (2016), Akcigit and Ates (2021)) or the increasing use of intangible inputs (De Ridder (2021)) as additional forces driving the rise in market power. This paper complements these supply-side perspectives by highlighting the role of demand through which rising income and structural change can raise markups even in the absence of changes in competition.

The issue of measurement is tackled by Raval (2022) that shows that using other variable inputs to recover firms' markups can deliver a different distribution of markups. Traina (2018) also argues that including administrative expenses would display a smaller increase in markups in the United States. Bond et al. (2021) show that relying on firms' revenue to estimate output elasticities might distort the level of markups. De Ridder et al. (2022) show that biases from markups estimates are nonetheless highly correlated with true markups. This paper highlights that the relevance of the services sector persists despite measurement issues and is consistent with the increase in corporate profits of the services sector over the past 65 years.

Second, this paper contributes to a growing literature on demand-side explanations for firm pricing power. Recent work has highlighted how consumer heterogeneity and differences in price elasticities of demand can affect markups (Döpper et al. (2022), Bornstein (2025), Afrouzi et al. (2021), Sangani (2023)). I extend this logic in a general equilibrium setting, showing that non-homothetic preferences can generate endogenously rising markups. The mechanism is further validated through survey evidence linking household income to perceived price sensitivity.

Third, this work contributes to the literature on structural transformation. Foundational models emphasize how differential productivity growth and income effects shift consumption from goods to services over time (Kongsamut et al. (2001), Ngai and Pissarides (2007), Buera and Kaboski (2012), Herrendorf et al. (2013, 2014)). Recent work has focused on developing novel non-homothetic preferences. Boppart (2014) proposes preferences suited to analyze the joint role of changes in relative prices and income as drivers of structural change, and Comin et al. (2021) and Matsuyama (2019) propose non-homothetic CES preferences to study structural transformation. Bridgman and Herrendorf (2024) propose a model of structural change with input-output linkages to study the decline of the labor share and Moreira (2022) analyzes the role of market power and capital-biased technical change on the labor share.

Finally, the model contributes to work on imperfect competition in macroeconomic models. Recent work has developed models with variable markups. Atkeson and Burstein (2008) study differences in international relative prices with a nested production structure allowing for different elasticities of substitution. Edmond, Midrigan, and Xu (2021) introduce Kimball (1995) preferences relying on Klenow and Willis (2016) specification to study the welfare costs of markup distortions. Bertoletti, Etro, and Simonovska (2018) use indirectly additive preferences with monopolistically competitive firms to study the gains from trade liberalization. Matsuyama and Ushchev (2017) propose flexible homothetic preferences that allow the price of elasticity of demand to vary with product prices but not household income. This paper differs by embedding non-homothetic preferences into a tractable multi-sector framework, allowing markups to respond endogenously to the joint evolution of income and prices.

2. Empirical Motivation

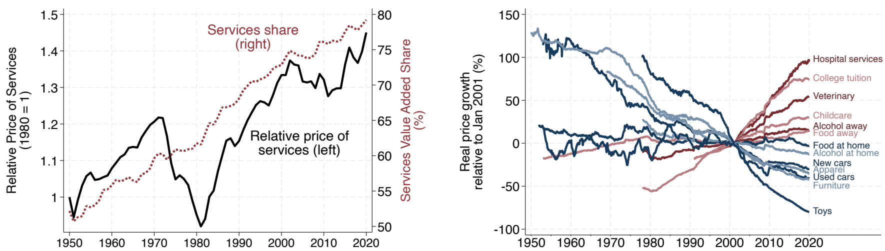
This section documents that (1) the rise of the services share was accompanied by an increase of the relative price of services, and (2) the rise of markups was driven by the services sector, which is consistent with the increase of the relative price of services and corporate profits. The relevance of the services sector in the rise of markups persists across several robustness checks. Data and key variables used in the analysis are described in Appendix A.

2.1. *The rise of the services share and the relative price of services*

The reallocation of economic activity and employment from agriculture and manufacturing toward the services sector—structural change—is accompanied in the United States and several other advanced countries by an increase in the relative price of services. Figure 2.1a depicts the services share and the relative price of service industries (over non-service industries) in terms of value added.[3] The services share increased by 27p.p. between 1950 and 2020, hovering now around 79% of the economy.[4] Over the same period, the relative price of services grew 45%, with a noticeable increase starting in 1980. Figure 2.1b displays the evolution of real prices of selected final goods (in blue) and services (in red) relative to January 2001. There was a rapid increase in the real prices of hospital services, college tuition, dental services, food and alcoholic beverages consumed away from home over the last seventy years, while the real prices of goods have risen at a much slower pace or declined.

Why are these trends important? The increase of the relative price of services is intimately related with the evolution of markups across sectors, hinting that the services sector had potentially larger markups than the non-services sector. On the other hand, the rise of the services share help explain the reallocation of economic activity towards higher markup firms. The next subsection provides evidence supporting that hypothesis.

FIGURE 2.1.—Structural change in the U.S., 1950-2020



(a) Rel. price of services and services share (b) Real price of goods and services

Note: Panel (a) shows the relative price of service industries (black), measured as the chain-weighted Fisher price index of the value added price indices of individual industries, and the the value added share of the services sector (red), using data from the BEA. Panel (b) shows the evolution of the real price of selected goods (blue) and services (red) relative to January 2001, using data from the BLS's Consumer Price Index. See Appendices A.1 and A.2 for details.

2.2. *Markups and the role of services*

A firm's markup is defined as the ratio of its output price to its marginal cost. The aggregate markup is then a weighted average of markups of all firms in the economy, with firms' variable

[3]Supplemental Appendix A.1 shows the services share and relative price of services for other advanced economies.
[4]The increase in the services share when measured with gross output or as the sum of labor compensation and intermediate inputs is commensurate with the increase in its value added share.

cost share used as their weight. As firms are assigned to sectors, the aggregate markup can also be written as the sum of the product of sectoral variable cost shares in the aggregate economy, $\boldsymbol{\omega}_{j_t}^{\text{costs}}$, and the average markup within that sector, $\overline{m}_{j_t}$.[5] For two sectors, services S and non-services G, the aggregate markup, M_t, can be expressed as

$$M_t = \left(1 - \boldsymbol{\omega}_{S_t}^{\text{costs}}\right) \overline{m}_{G_t} + \boldsymbol{\omega}_{S_t}^{\text{costs}} \overline{m}_{S_t}. \tag{2.1}$$

The services cost share, $\boldsymbol{\omega}_{S_t}^{\text{costs}}$, is measured using industry-level data from the BEA, which accounts for the entire industrial production of the U.S. economy starting in the late 1940s, and refers to the sum of *compensation of employees* and *intermediate inputs*.[6,7]

In the baseline results below, the average markup within each sector, $\overline{m}_{j_t}$, is based on firm-level data from Compustat, with the underlying assumption that the estimated markup of listed firms is a good proxy for the markups of nonlisted firms. It is computed as the weighted average of firm-level markups, where a firm's weight is the ratio of its variable costs to total variable costs (here *cost of goods sold* or `Cogs`). From a firm's cost minimization problem, a firm's markup can be shown to be equal to the ratio of its output elasticity to a variable input and its sales share. The numerator is usually obtained by estimating firms' production functions.[8] The denominator can be read off directly from balance sheet data, corresponding to the ratio of `Cogs` to `Sales` in Compustat. Alternative measures are discussed in the robustness checks.

Figure 2.2a shows the aggregate markup, M_t, over time. The aggregate markup increased 12p.p. between 1955 and 2020 and 19p.p. between 1980 and 2020 (with constant output elasticities). With time-varying output elasticities at the 2-digit NAICS level, the increase is more muted, about 4p.p. between 1955 and 2015 and 13p.p. between 1980 and 2015.[9]

Figure 2.2b displays the contribution of each sector to the aggregate markup, $\left(\boldsymbol{\omega}_{j_t}^{\text{costs}} \overline{m}_{j_t} / M_t\right)$, for the cases where the estimated output elasticities are constant and time-varying. The increase

[5]Appendix A.3 shows how to derive this result. It is also possible to derive the sales/gross output-weighted harmonic mean of firm-level markups in a similar fashion.

[6]This strategy differs from the usual markup literature relying solely on Compustat data. In Compustat, the services cost share (using `Cogs`) was 18% in 1950 and 57% in 2020. The BEA data shows that the services cost share was 35% in 1950 and above 70% in 2020. Note that this definition of variable costs from the BEA is consistent with the markup being estimated with data on cost of goods sold. See Appendix A.1 for details on the data.

[7]As the average markup within each sector is computed following the production function approach with the cost of goods sold as the variable input, the relevant variable cost shares are based on the sum of labor and intermediate inputs (the closest to what is in the cost of goods sold; it excludes the cost of capital). See Appendix A.4 for details.

[8]The same sample restrictions as De Loecker, Eeckhout, and Unger (2020) are applied when using Compustat data. The output elasticities are estimated at the two digit NAICS code level from a Cobb-Douglas production function of capital and `Cogs`. Results with constant and time-varying (five-year rolling window) output elasticities are discussed in the text. Using a translog production function yields similar results for the output elasticities and a very similar contribution of the services sector to the rise in markups. See Appendix A.4 for details.

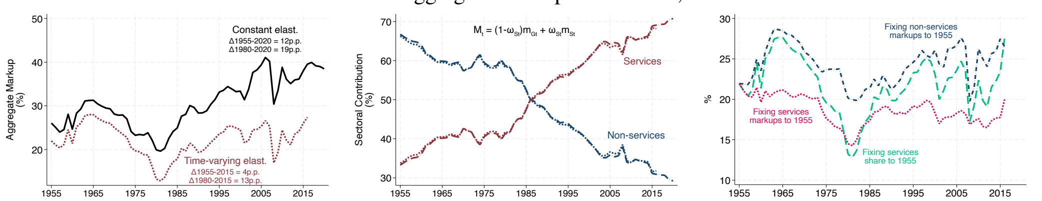
[9]The increase is significantly larger when gross output shares are used to measure both sectoral shares and average markups. When output elasticities are constant and gross output shares are used, the aggregate markup increases 53p.p. between 1955 and 2020, a value closer to De Loecker, Eeckhout, and Unger (2020).

in the aggregate markup is mostly driven by the services sector in both cases. Between 1955 and 2020, the contribution of services grew by more than 37p.p., from 33% to 71% of the aggregate markup when output elasticities are constant (from 34% to 68% between 1955 and 2015 when output elasticities vary over time).

Figure 2.2c depicts counterfactual markups; that is, what the aggregate markup would be if the sectoral shares or the average markup within the services or the non-services sector were kept constant over the past 65 years (when output elasticities are time-varying).[10] The increase in the average markup of services played a significant role in the overall rise in markups as the red line is fairly constant between 1955 and 2020.

The relevance of the services sector merits some additional discussion. As the average markup in non-service industries has been growing over time, albeit at a slower pace than in services, the reallocation of economic activity by itself is not sufficient to drive the entirety of the increase in the aggregate markup. Although De Loecker et al. (2020) find that the rise in the aggregate markup is mostly the result of within sector increases in markups, this paper stresses that increases in markups *within the services sector* have been the main engine of this rise. The aggregate markup trails closely the baseline if the markup within the non-services were held constant at its 1955 value (see blue line in panel (c)).

FIGURE 2.2.—Aggregate markups in the U.S., 1955-2020



(a) Aggregate markup (b) Sectoral contribution (c) Counterfactual markups

Note: Panel (a) shows the aggregate markup, M_t, measured as the cost-weighted average of markups when output elasticities with respect to `Cogs` are constant (black) and time-varying (red). The services variable cost share, $\boldsymbol{\omega}_{S_t}^{\text{costs}}$, uses BEA data and the average markups within sectors, $\overline{m}_{j_t}$, rely on Compustat. Panel (b) shows the sectoral contribution to the aggregate markup (non-services in blue, services in red), $\boldsymbol{\omega}_{j_t}^{\text{costs}}\,\overline{m}_{j_t}/M_t$, with constant (solid line) and time-varying (dotted line) output elasticities. Panel (c) shows the aggregate markup (with time-varying output elasticities), M_t, when the average markup within each sector, $\overline{m}_{j_t}$, is fixed at its 1955 level (non-services in blue, services in red) and when the services cost share, $\boldsymbol{\omega}_{S_t}^{\text{costs}}$, is fixed at its 1955 level (green).

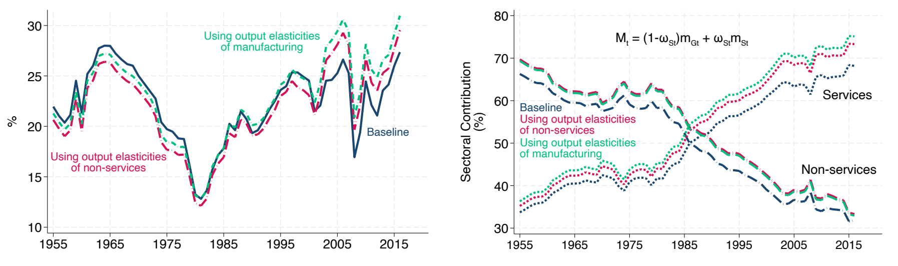
2.3. *Elasticities and other robustness checks*

How relevant are the output elasticities with respect to the variable input (`Cogs`) in driving the rise in markups? To assess their importance, I set the elasticities of service industries to be the

[10] As Figure A.2a in the Appendix A.2 shows, similar patterns emerge when the counterfactual markups are calculated with output elasticities held constant over time.

same as the average elasticity in non-service industries and in manufacturing (NAICS 31-33).[11] These changes only affect the evolution of the average markup of services, $\overline{m}_{S_t}$. Figure 2.3a shows the aggregate markup, M_t, when using these alternative time-varying output elasticities. In a nutshell, very little would differ from the baseline. The aggregate markup across these scenarios would still trail the original increase in markups, with the bulk of this increase driven by the services sector as Figure 2.3b displays.[12]

FIGURE 2.3.—The role of output elasticities in the aggregate markup and the services' contribution



(a) Aggregate markup

(b) Sectoral contribution

Note: Panel (a) shows the aggregate markup (with time-varying output elasticities), M_t, when the average markup of services, $\overline{m}_{S_t}$, is computed using the average output elasticity with respect to `Cogs` of non-service industries (red) and manufacturing (green). The baseline aggregate markup is in blue. Panel (b) shows the sectoral contribution to the aggregate markup (with time-varying output elasticities), $\boldsymbol{\omega}_{jt}^{\text{costs}}\, \overline{m}_{j_t}/M_t$, using the average output elasticity of non-service industries (red) and manufacturing (green). The sectoral contribution in the baseline data is in blue.

Additional robustness checks are presented in Supplemental Appendices A.3, A.4, A.5, A.6, and A.7. Using BEA data on industrial production for the entire economy, Appendix A.3 shows that computing average markups within sectors as the ratio of *gross output* to *cost of goods sold* or as the ratio of *value added* to *labor compensation* does not alter the role of the services sector.[13] Although the level of the aggregate markup differs between these measures and the baseline, their trends were similar. The services sector still accounts for more than 70% of the aggregate markup, in line with what was presented above.

Firms' income statements also report selling, general, and administrative expenses (`SGA`), which tend to be non-production costs that include shipping, rent, utilities, and marketing expenses. Traina (2018) and Basu (2019) highlight that not including them biases the overall increase in the aggregate markup. Appendix A.4 shows that indeed the aggregate markup would be lower. The relevance of the services sector persists, however. Keeping the average markup

[11]The average output elasticities with respect to `Cogs` for the non-service industries and the manufacturing sector are 0.8959107 and 0.9283032, respectively.

[12]The upshot of these experiments does not change when output elasticities are constant over time as shown in the Supplemental Appendix A.2.

[13]These are often referred to as accounting markups and they rely on the assumption that marginal costs equate average variable costs. Although the assumption is strong, the results point to the same direction as with the production function approach. Here, cost of goods sold refer to the sum of labor compensation and intermediate inputs.

of services or the services share constant at their 1955 values implies that the aggregate markup would have been even lower.

The role of the right tail of the markup distribution is discussed in Appendix A.5. Compustat is composed of publicly listed firms that tend to be larger and more established, and hence might skew the average markup within sectors. To address this concern, firms in the top 1%, 5%, and 10% of the markup distribution within each sector and year are dropped. Although the aggregate markup is lower, the contribution of the services sector stays practically unchanged.

Markups are tightly connected to economic profits. The BEA's national income and product accounts (NIPA) provide data on corporate profits from *current production* and before taxes on corporate income for all financial and nonfinancial firms filing federal corporate tax returns. Appendix A.6 discusses the evolution of profits in the aggregate and across sectors from 1955 to 2020. In consonance with the rise in markups, real corporate profits have markedly increased. As the figures show, the services sector was the main driver of the increase in economic profits.

Appendix A.7 discusses the relationship between markups and the labor share of income. It highlights why a decline of the labor share does not necessarily imply an increase in markups.

Models of structural change have been able to explain the long-run trends depicted in Figure 2.1a thanks to differential rates of technological progress across sectors and non-homothetic preferences. Can these models also explain the trends in markups presented in Figure 2.2? Section 3 develops a model where these ingredients also engender a rise in markups when markets are imperfectly competitive.

3. A Model of Rising Markups and Services

The empirical patterns documented in Section 2 present a puzzle for standard models. In CES frameworks, markups are constant. In Kimball preferences, markups vary with relative prices but not income. Neither can explain why services—which face steadily rising relative prices—simultaneously experience rising markups and expanding expenditure shares. This section presents a multi-sector general equilibrium model with imperfect competition and non-homothetic preferences that reconciles these patterns through income-dependent demand elasticities. In particular, the central mechanism links household income and sectoral prices to endogenous price elasticities of demand, which in turn shape markups.[14]

Environment. Time is discrete and indexed by t. The economy is populated by a unit mass of identical households, endowed with one unit of productive time supplied inelastically in the labor market in exchange for the wage w. Households also receive nonlabor earnings Λ

[14] The reader is invited to immerse in Section B of the Supplemental Appendix, which presents the theoretical underpinning for this avenue. In a nutshell, when preferences are non-homothetic, a consumer's price elasticity of demand can vary endogenously along a variety's price and the consumer's income.

from owning firms. There are three sectors in this economy, one that produces consumption goods, another that produces services, and another that produces intermediate inputs used by firms in the other two sectors. Firms within the goods and services sectors are retailers, selling directly to consumers. Within each of these sectors, there is a continuum of monopolistically competitive firms producing a differentiated variety of goods or services. A variety within a sector differs in terms of its price and quality. The intermediate inputs producers are perfectly competitive. Labor is freely mobile across the three sectors and firms take factor prices as given.

3.1. *Households*

Preferences. Households have preferences over the consumption of different varieties of goods and services, denoted $\boldsymbol{c}_{G_t}$ and $\boldsymbol{c}_{S_t}$, and their respective quality, $\boldsymbol{q}_{G_t}$ and $\boldsymbol{q}_{S_t}$, where the bold variables correspond to vectors of the different varieties of goods and services. Preferences are represented by the direct utility function $u(\boldsymbol{c_{G_t}}, \boldsymbol{c_{S_t}}, \boldsymbol{q_{G_t}}, \boldsymbol{q_{S_t}})$. Each variety ω of goods and services is indexed by its price $p_{j_t}(\omega)$ and quality $q_{j_t}(\omega)$ taken as given by the household.

I start by defining the indirect utility function, i.e., the household's maximal attainable utility given her income, e_t, the vector of prices of goods and services, $\boldsymbol{p}_{G_t}$ and $\boldsymbol{p}_{S_t}$, and their respective quality, $\boldsymbol{q}_{G_t}$ and $\boldsymbol{q}_{S_t}$. This avenue allows me to highlight why the price elasticity of demand depends on the variety's price and the consumer's income.

Let the indirect utility be a composite of two sectoral indirect utilities, one for goods and another for services, aggregated in a Cobb-Douglas fashion according to

$$v(e_t, \boldsymbol{p}_{G_t}, \boldsymbol{p}_{S_t}, \boldsymbol{q}_{G_t}, \boldsymbol{q}_{S_t}) = v_G(e_t, \boldsymbol{p}_{G_t}, \boldsymbol{q}_{G_t})^{\lambda}\, v_S(e_t, \boldsymbol{p}_{S_t}, \boldsymbol{q}_{S_t})^{1-\lambda}, \tag{3.1}$$

where $\lambda \in (0,1)$ is the weight on the indirect utility from goods. Each sectoral indirect utility is in turn additively separable across the differentiated varieties of commodity j, implying

$$v_j(e_t, \boldsymbol{p}_{j_t}, \boldsymbol{q}_{j_t}) = \int_{\mathcal{N}_{j_t}} \widehat{v}_j\left(e_t, p_{j_t}(\omega), q_{j_t}(\omega)\right) d\omega, \tag{3.2}$$

where the sector-specific indirect subutility satisfies the standard properties of indirect utility functions as defined in Assumption B.1 in Appendix B. The sectoral indirect subutility for each variety ω of commodity j is taken to be

$$\widehat{v}_j\left(e_t, p_{j_t}(\omega), q_{j_t}(\omega)\right) = \frac{1}{1+\gamma}\left[\frac{\left(\phi_j e_t - p_{j_t}(\omega)\right) q_{j_t}(\omega)^{\delta}}{e_t}\right]^{1+\gamma} \quad \text{for } p_{j_t}(\omega) \leq \phi_j e_t \tag{3.3}$$

and zero otherwise. Here, $\phi_j e_t > 0$ is the sectoral choke price of any variety of commodity $j \in \{G, S\}$, i.e., the maximum price the household is willing to pay in order to consume a positive amount of that variety. A price above the consumer's choke price is not purchased and therefore yields a utility of zero. The higher the value of $\phi_j > 0$, the higher is the consumer's choke price. Similarly, the higher the household's income e_t, the higher is her choke price. Each variety is weighted by its quality $q_{j_t}(\omega)$. Varieties of higher quality are valued more than low-quality varieties. The parameter $\delta > 0$ is a quality-specific weight and $\gamma > 0$ ensures demand satisfies the law of demand. These parameters are common for both goods and services.

Proposition 3.1 shows that there is an analytic representation of the direct utility when the indirect utility has the above form (equations (3.1), (3.2), and (3.3)). Proposition 3.2 further demonstrates that the indirect utility collapses to the well-known two-sector CES utility.

PROPOSITION 3.1: (DIRECT UTILITY) The indirect utility (equation (3.1)) admits an analytic representation of the direct utility given by

$$u(\boldsymbol{c}_{G_t}, \boldsymbol{c}_{S_t}, \boldsymbol{q}_{G_t}, \boldsymbol{q}_{S_t}) = \psi \left[\frac{\widehat{C}_t - 1}{\widetilde{C}_{G_t}^{\lambda} \widetilde{C}_{S_t}^{(1-\lambda)}} \right]^{(1+\gamma)},$$

where $\widehat{C}_t = \phi_G \int_{\mathcal{N}_{G_t}} c_{G_t}(\omega) d\omega + \phi_S \int_{\mathcal{N}_{S_t}} c_{S_t}(\omega) d\omega$ denotes an aggregator of total consumption, and $\widetilde{C}_{j_t} = \left(\int_{\mathcal{N}_{j_t}} \left[\frac{c_{j_t}(\omega)}{q_{j_t}(\omega)^{\delta}} \right]^{\frac{1+\gamma}{\gamma}} d\omega \right)^{\frac{\gamma}{1+\gamma}}$ denotes a quality-adjusted composite of the different varieties of commodity $j = \{G, S\}$, and $\psi = (1+\gamma)^{-1} \lambda^{\lambda(1+\gamma)} (1-\lambda)^{(1-\lambda)(1+\gamma)} > 0$ is a parameter.

PROOF: See Appendix B.2. *Q.E.D.*

PROPOSITION 3.2: (TWO-SECTOR CES) Assume $\phi_j = 0$ for $j = \{G, S\}$, $\gamma < -1$, and $\delta < 0$. Then, these preferences collapse to a two-sector CES preferences with quality and $(-\gamma)$ as the elasticity of substitution within each sector.

PROOF: See Appendix B.3. *Q.E.D.*

Budget constraint. The budget constraint the household faces requires that total spending on goods and services, e_t, be paid for with labor income, w_t, and nonlabor earnings, Λ_t, according to

$$e_t \equiv \sum_{j=G,S} \int_{\mathcal{N}_{j_t}} p_{j_t}(\omega) c_{j_t}(\omega) \mathrm{d}\omega = w_t + \Lambda_t. \tag{3.4}$$

Demand for varieties. The household's demand for each variety of goods and services can be recovered using Roy's identity (see Appendix B.4 for details on how to derive it). Demand for variety ω of commodity $j \in \{G, S\}$ can then be expressed as

$$c_{j_t}(\omega) = \left[\underbrace{\phi_j\, e_t}_{\text{choke price}} - p_{j_t}(\omega) \right]^{\gamma} \underbrace{q_{j_t}(\omega)^{\delta(1+\gamma)}}_{\text{variety quality}} \underbrace{A_{j_t}}_{\text{sectoral composite}} \quad \text{for } p_{j_t}(\omega) \leq \phi_j e_t \tag{3.5}$$

and zero otherwise. So, if a firm sets a price above households' choke prices, then demand for its variety will be zero. Here, A_{j_t} is a household-specific sectoral composite.[15]

As $\gamma > 0$, consumer demand satisfies the law of demand, i.e., the quantity demanded varies inversely with the variety's price. In particular, it increases with the distance of the variety's price to the maximum amount the household is willing to pay to consume it (i.e., the commodity's choke price). Hence, all else equal, lower-priced varieties are associated with more consumption. Similarly, the higher the quality of the variety, the larger is the household's demand for that variety. The consumption demand for different varieties of the same commodity $j = \{G, S\}$ only varies as a result of differences in prices and quality.

Price elasticity of demand. The household's consumption demand yields a direct price elasticity of demand that depends on her income and the price of the particular variety demanded. Let $\xi_{j_t}(\omega)$ denote the (negative of the) percentage change in quantity demanded of variety ω of commodity j in response to a percentage change in its own price, or $\xi_{j_t}(\omega) \equiv -\frac{\partial c_{j_t}(\omega)}{\partial p_{j_t}(\omega)} \frac{p_{j_t}(\omega)}{c_{j_t}(\omega)}$. The household's price elasticity of demand is

$$\xi_{j_t}(\omega) = \frac{\gamma\, p_{j_t}(\omega)}{\phi_j e_t - p_{j_t}(\omega)}. \tag{3.6}$$

This expression satisfies the propositions B.3 and B.4 that allow the price elasticity of demand to vary with the variety's price and consumers' income.[16] Demand becomes *less* elastic when the household's income goes up, i.e., the price elasticity of demand is decreasing in the household's income. Figure 3.1 illustrates this case. Note that in this particular example the price elasticity of demand for services is lower than for goods, leading to a stronger increase in services firms' markups and a shift in expenditure shares toward services.

In addition, demand becomes *more* elastic when the price of the variety goes up, i.e., the price elasticity of demand is increasing in the variety's price. As a result, firms selling cheaper

---

[15]The household-specific sectoral composite is given by $A_{j_t} \equiv \widetilde{C}_{j_t}^{(1+\gamma)} \left[\psi_j\, e_t \left(\widehat{C}_t - 1 \right) \right]^{-\gamma}$, where $\psi_G = (1+\gamma)^{-\lambda}\, \lambda^{\lambda(1+\gamma)}$ and $\psi_S = (1+\gamma)^{-(1-\lambda)}\, (1-\lambda)^{(1-\lambda)(1+\gamma)}$.

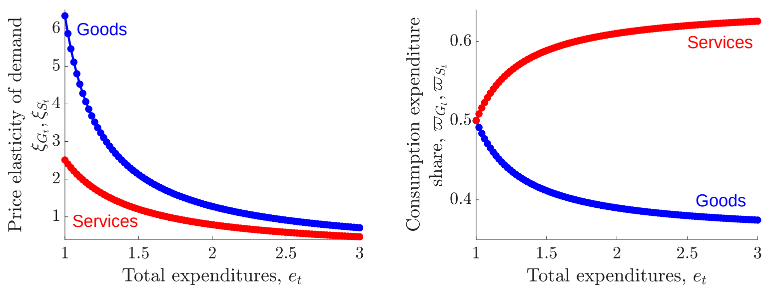
[16]The price super-elasticities of demand with respect to price and income are respectively given by

$$\frac{\partial \xi_{j_t}(\omega)}{\partial p_{j_t}(\omega)} \frac{p_{j_t}(\omega)}{\xi_{j_t}(\omega)} = 1 + \frac{\xi_{j_t}(\omega)}{\gamma} \quad \text{and} \quad \frac{\partial \xi_{j_t}(\omega)}{\partial e_t} \frac{e_t}{\xi_{j_t}(\omega)} = -\phi_j e_t.$$

varieties will be able to charge higher markups as $\xi_{j_t}(\widehat{\omega}) < \xi_{j_t}(\widetilde{\omega})$ for $p_{j_t}(\widehat{\omega}) < p_{j_t}(\widetilde{\omega})$.[17] Figure 3.2 illustrates this case. In this example the increase in the services expenditure share is accompanied by a decline in services' markups as the price of elasticity of demand for services goes up. Note, however, that in this case the price of services relative to goods would fall (conflicting with the observed trends in the data).

REMARK: The elasticity of substitution across varieties, i.e. how demand for variety $\widetilde{\omega}$ of commodity j changes in response to a change in the consumption of variety $\widehat{\omega}$ of commodity κ, is equal to the consumer's price elasticity of demand of the variety in the numerator.[18]

FIGURE 3.1.—Expenditure shares and price elasticity of demand when income rises



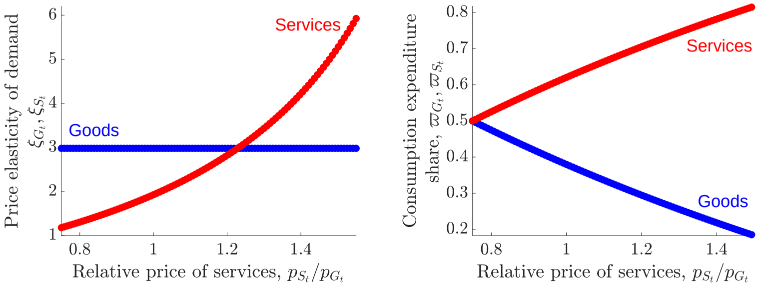
(a) Price elasticity of demand (b) Consumption expenditure shares

Note: Panel (a) shows the price elasticity of demand for goods (blue) and services (red) when total expenditures increase. Here, the prices of goods and services are assumed to be equal and constant. The choke price parameters were chosen so that both expenditure shares start at 50%, with $\phi_S > \phi_G$. Panel (b) shows the corresponding consumption expenditure shares on goods (blue) and services (red) when total expenditures increase.

[17]This property also holds for preferences that satisfy Marshall's second law of demand (e.g., Kimball preferences).

[18]Let the elasticity of substitution be $E(\widetilde{\omega}_{j_t}, \widehat{\omega}_{\kappa_t}) = -\frac{\partial\left(\frac{c_{j_t}(\widetilde{\omega})}{c_{\kappa_t}(\widehat{\omega})}\right)}{\partial\left(\frac{p_{j_t}(\widetilde{\omega})}{p_{\kappa_t}(\widehat{\omega})}\right)} \frac{\frac{p_{j_t}(\widetilde{\omega})}{p_{\kappa_t}(\widehat{\omega})}}{\frac{c_{j_t}(\widetilde{\omega})}{c_{\kappa_t}(\widehat{\omega})}}$. Then, $E(\widetilde{\omega}_{j_t}, \widehat{\omega}_{\kappa_t}) = \xi_{j_t}(\widetilde{\omega})$. Note that as the price elasticity of demand is greater than one, the elasticity of substitution is also greater than one, which implies that varieties are gross substitutes. Hence, the reduction in the relative quantity demanded of a variety exceeds the increase in its relative price. This leads to a decline of the relative expenditure on that variety. For CES preferences, the elasticity of substitution is given by $E(\widetilde{\omega}_{j_t}, \widehat{\omega}_{\kappa_t}) = -\gamma$.

FIGURE 3.2.—Expenditure shares and price elasticity of demand when the price of services rise



(a) Price elasticity of demand (b) Consumption expenditure shares

Note: Panel (a) shows the price elasticity of demand for goods (blue) and services (red) when the relative price of services increase. Here, the price of goods and total expenditures are assumed to be constant. As both the price of goods and total expenditures are constant, the price elasticity of demand for goods is also constant. The choke price parameters were chosen so that both expenditure shares start at 50%, with $\phi_S > \phi_G$. Panel (b) shows the corresponding consumption expenditure shares on goods (blue) and services (red) when the relative price of services increase.

Quality elasticity of demand. Let $\sigma_{j_t}(\omega)$ denote the percentage change in quantity demanded of variety ω of commodity j in response to a percentage change in its own quality, or $\sigma_{j_t}(\omega) \equiv \frac{\partial c_{j_t}(\omega)}{\partial q_{j_t}(\omega)} \frac{q_{j_t}(\omega)}{c_{j_t}(\omega)}$. As preferences are homothetic in quality, the quality elasticity of demand does not depend on the household's income, nor on the variety's price or quality, i.e.,

$$\sigma_{j_t}(\omega) = \delta\,(1+\gamma)\,, \tag{3.7}$$

for $j \in \{G, S\}$. As $\delta, \gamma > 0$, an increase in a variety's quality makes households increase their demand for that variety.

3.2. *Incumbent retailers*

Technology. Varieties of goods and services are produced by firms that differ in terms of their total factor productivity (TFP), z_{j_t}. The output of a firm is produced via a constant returns to scale Cobb-Douglas production function that combines labor n_j and intermediate inputs i_j according to

$$y_{j_t} = z_{j_t} n_{j_t}^{\theta} i_{j_t}^{1-\theta}. \tag{3.8}$$

Introducing intermediate inputs allows to decouple markups from the (inverse of the) labor share and is tightly linked to how markups were computed in the empirical exercise.

Costs. Firms' total costs comprise the wage bill, spending on intermediate inputs, expenses on quality, and other fixed costs, or

$$tc_{j_t} = w_t n_{j_t} + p_{I_t} i_{j_t} + \kappa q_{j_t}^{\vartheta} + f_{j_t}, \tag{3.9}$$

where $\kappa > 0$ and $\vartheta > 1$ are parameters common across sectors. I will refer to f_{j_t} as entry costs, which can vary over time. To draw a parallel with the empirical analysis, the wage bill and spending on intermediate inputs would correspond to `Cogs` and expenses on quality to `SGA`.

The firm solves its cost minimization problem (3.9) by optimally choosing the quantity of labor and intermediate inputs it needs subject to the technological constraint (3.8). By replacing these factor demands in the firm's variable cost function, an expression for its marginal cost is obtained as

$$mc_{j_t} = \frac{1}{z_{j_t}} \left(\frac{w_t}{\theta}\right)^{\theta} \left(\frac{p_{I_t}}{1-\theta}\right)^{1-\theta}. \tag{3.10}$$

The firm's marginal cost is decreasing in the firm's productivity and increasing in the aggregate wage and price of intermediate inputs.

Profit maximization. A firm sets a price, p_{j_t}, and a level of quality, q_{j_t}, to maximize its profits taking the aggregate demand for its variety as given. Since the firm's production technology is constant returns to scale, its marginal cost is equal to its average variable cost. A firm producing variety ω in sector j maximizes profits by solving the following problem

$$\pi_{j_t} = \max_{p_{j_t}, q_{j_t}} (p_{j_t} - mc_{j_t}) \, y_{j_t} - \kappa q_{j_t}^{\vartheta} - f_{j_t} \tag{3.11}$$

subject to $y_{j_t} = c_j(p_{j_t}, q_{j_t})$, where consumers' demand depends on both the price and quality chosen by the firm.

Price and markup. The solution to the firm's profit maximization problem yields its variety's price as a markup, m_j, over the marginal cost according to

$$p_{j_t} = m_{j_t} \, mc_{j_t}. \tag{3.12}$$

In turn, the firm's markup is a function of consumers' *endogenous* price elasticity of demand (equation (3.6)) in line with

$$m_{j_t} = \frac{\xi_j(e_t, p_{j_t})}{\xi_j(e_t, p_{j_t}) - 1}. \tag{3.13}$$

A decline in the price elasticity of demand leads to an increase in the firm's markup. The firm's optimal price is a fixed point as the consumers' demand elasticity depends on it. Solving for it yields the following analytic expression for the firm's markup

$$m_{j_t} = \underbrace{\frac{\gamma}{\gamma+1}}_{\text{CES terms}} + \underbrace{\frac{z_{j_t}}{\left(\frac{w_t}{\theta}\right)^{\theta}\left(\frac{p_{I_t}}{1-\theta}\right)^{1-\theta}}}_{\text{supply effect}} \times \underbrace{\frac{\phi_j e_t}{\gamma+1}}_{\text{demand effect}} . \tag{3.14}$$

The first term is constant and relates closely to the elasticity of substitution in the standard CES utility. The second term varies over time, highlighting the demand and supply channels driving a rise in markups when the non-homothetic preferences are introduced.

First, sector-specific technological progress (through an increase in z_{j_t}) reduces marginal costs, allowing firms to reduce their prices. At lower prices, households are more willing to buy a firm's product. This translates into a reduction in consumers' price elasticity of demand, which in turn allows firms to increase their markups. Similarly, an increase in input prices leads to a rise in the firm's marginal cost, which in turn puts downward pressure on the firm's markup.

Second, an increase in household income (through an increase in $e_t \equiv w_t + \Lambda_t$) makes consumers more willing to buy goods and services. This translates into a reduction in consumers' price elasticities of demand, leading firms to respond by increasing their markups. Note that the aggregate wage enters both as a supply and demand force because it leads to higher labor costs for the firm but also to heightened demand for its variety. The demand effect from wages dominates as long as $w_t/(w_t + \Lambda_t) > \theta^2$.

Quality. The firm's optimal choice of quality is tightly linked to its markup. A firm faces a tradeoff when choosing its price: improvements in quality require higher markups, as it weights its consumers' quality elasticity of demand and its markup. In particular, the firm equates the share of quality-related costs as a fraction of its sales to

$$\frac{\kappa q_{j_t}^{\vartheta}}{p_{j_t} y_{j_t}} = \frac{\sigma_{j_t}}{\vartheta} \frac{(m_{j_t} - 1)}{m_{j_t}}, \tag{3.15}$$

where σ_{j_t} corresponds to consumers' quality elasticity of demand defined in equation (3.7).

3.3. *Entrants in the retail market*

Potential entrants consider entering the market for goods or services as long as they can make profits. If a firm chooses to enter and produce a variety ω in sector $j = \{G, S\}$, it receives the expected profit $\left(\int_0^{N_{j_t}} \pi_{j_t}(\omega) d\omega\right)/N_{j_t}$. If instead the firm chooses to not enter the market, it gets a payoff of zero. Firms will thus keep entering the market driving down profits to zero. The free-entry condition determines the aggregate number of operating firms in each sector, which is denoted by N_{j_t}. The total number of operating retail firms in the economy is $N_t = N_{G_t} + N_{S_t}$.

3.4. *Retail oligopolists*

To give a fair chance to entry costs as a driver of the rise in markups, I recast the retail market structure as an oligopoly. As shown above, monopolistic competition delivers a markup that is solely a function of consumers' price elasticities of demand. In contrast, oligopolistic competition allows the evolution of the number of firms in the market to directly impact a firm's markup. Next, I highlight how changing the market structure affects the definition of markups.[19]

Assume now that firms compete *à la* Cournot in each retail market. A firm in sector $j = \{G, S\}$ chooses its level of output and quality to maximize profits taking the output of its competitors as given. Each firm now understands how its choices will affect the price of goods or services, as the equilibrium price of commodity j is now a function of all output produced in that sector. The optimal price is still a markup over marginal costs as in equation (3.12). However, the markup is now not only a function of consumers' price elasticity of demand, but also depends on the firm's share of total sales of commodity j, $s_j \in [0, 1]$. Now, equation (3.14) can be expressed as

$$m_{j_t} = \underbrace{\frac{\gamma}{\gamma + s_{j_t}}}_{\text{CES terms}} + \underbrace{\frac{z_{j_t}}{\left(\frac{w_t}{\theta}\right)^{\theta}\left(\frac{p_{I_t}}{1-\theta}\right)^{1-\theta}}}_{\text{supply effect}} \times \underbrace{\phi_j e_t}_{\text{demand effect}} \times \underbrace{\frac{s_{j_t}}{\gamma + s_{j_t}}}_{\text{competition effect}} . \tag{3.16}$$

As before, an increase in the firm's productivity or its customers' income results in an increase in markups. A new channel is now present. All else equal, a decline in the number of firms operating in sector j increases the sales share of a firm. That increase now also raises the firm's markup (recall that consumers purchase a variety as long as $\phi_j e_t \geq p_{j_t}(\omega) \geq mc_{j_t}(\omega)$). As entry costs directly affect the number of firms entering a sector, their increase will push firms' markups up.

3.5. *Intermediate input producers*

Intermediate input firms use a linear technology in labor, with productivity z_{I_t}, to produce inputs used by retailers in the goods and services sectors. The market for intermediate inputs is perfectly competitive. So, firms make zero profits and have a markup of one.

[19] Refer to Appendix B.5 for the details of this framework.

3.6. *Equilibrium*

Definition (EQUILIBRIUM)**.** A symmetric equilibrium consists of a solution for: (1) consumers' demand for goods and services, c_G and c_S; (2) goods and services firms' price, p_G and p_S, quality, q_G and q_S, labor demand, n_G and n_S, and intermediate input demand, i_G and i_S; (3) the number of operating firms in each sector, N_G and N_S; (4) the intermediate input firms' price, p_I, and labor demand, n_I; (5) the economy's wage, w; and (6) nonlabor earnings transfered to consumers, Λ. These are determined such that

1. Given prices, p_G and p_S, quality, q_G and q_S, labor and nonlabor earnings, w and Λ, consumers' indirect utility satisfies (3.1), (3.2), and (3.3). The solution yields the allocations c_G and c_S.
2. Given consumers' demand and factor prices, w and p_I, incumbent firms in the retail sector $j = \{G, S\}$ maximize their profits according to (3.11), which determines a solution for prices, p_j, and quality, q_j. The labor and intermediate inputs demanded by incumbent firms, n_j and i_j, solve their cost minimization problem.
3. The free-entry condition holds in each sector.
4. Intermediate input producers maximize their profits such that $p_{I_t} z_{I_t} = w_t$ holds.
5. The labor supplied by households must equate the labor demanded by firms according to

$$\int_0^{N_{G_t}} n_{G_t}(\omega)\, d\omega + \int_0^{N_{S_t}} n_{S_t}(\omega)\, d\omega + n_{I_t} = 1.$$

6. The market for intermediate inputs must clear, so that the aggregate demand for intermediate inputs equates its aggregate supply according to

$$\int_0^{N_{G_t}} i_{G_t}(\omega)\, d\omega + \int_0^{N_{S_t}} i_{S_t}(\omega)\, d\omega = z_{I_t} n_{I_t}.$$

7. Expenses with quality and entry costs are rebated to households according to

$$\kappa \left[\int_0^{N_{G_t}} q_{G_t}(\omega)^{\vartheta}\, d\omega + \int_0^{N_{S_t}} q_{S_t}(\omega)^{\vartheta}\, d\omega \right] + N_{G_t} f_{G_t} + N_{S_t} f_{S_t} = \Lambda_t.$$

3.7. *Other extensions*

The parsimonious model here presented can be extended in various directions. Appendix C.1 in the Supplemental Appendix discusses the reasons why alternative preferences frequently used in the literature are not suitable for this class of problems. Appendix C.2 makes the model dynamic. Although having capital now changes the marginal cost of a firm, the savings decision does not alter the definition of a consumer's price elasticity of demand. Section 7 shows how consumer heterogeneity alters the model's predictions. As consumers have different price elasticities of demand, the firm's markup now depends both on the composition of customers and each consumer's price elasticity of demand.

4. Matching the Model to the U.S.

In this section the model is matched to U.S. data to be consistent with the key macroeconomic trends documented in Section 2. The calibration proceeds in two steps. In the first step, the parameters governing preferences, technology, and costs are estimated to match the main outcomes at two different points in time (namely, 1980 and 2020).[20] In particular, a set of parameter values can be backed out from the theory to match a set of data targets exactly. The remaining parameters are then chosen to minimize the model's prediction error relative to other targets. In the second step, the model is simulated to achieve the transition between 1955 and 2020. Only total factor productivities and entry costs vary over time. These time series are backed out to match the trends in the aggregate labor share, the services variable costs share, and firm entry rates across sectors, given the parameter values estimated in the first step. To validate the model, I contrast the time series of variables not targeted in the transition with their data counterparts. The model accounts well the untargeted trends of the past 65 years.

4.1. *Data targets*

The targeted moments used in the first step of the calibration are described below.[21]

Services share. The services share corresponds to the variable cost share of service industries as explained in Section 2 using data from the BEA. The targeted services variable cost shares for 1980 and 2020 are $\boldsymbol{\omega}_{S_t}^{\text{costs}} = \{0.417, 0.704\}$.

Relative price of services. The relative price of services is computed using BEA data and follows the methodology described in Section 2. In particular, sectoral prices are chain-weighted Fisher price indices of the price indices of individual industries, in which the relative price of services is normalized to one in 1950. The targeted relative prices of services for 1980 and 2020 are $\overline{p}_{S_t}/\overline{p}_{GI_t} = \{0.951, 1.450\}$.

Markups. The aggregate markup is measured using the sectoral variable cost shares and the average markups within each sector, as computed in Section 2, with data from the BEA and listed firms in Compustat. Only values for 2020 are targeted, with the aggregate markup given

[20] The choice of 1980 provides overall a slightly better fit for the economy with monopolistic competition. See Supplemental Appendix C.4 for the parameter values and fit when the model is confronted to 1955 and 2020 data.

[21] When mapping the model to the data, intermediate input firms are included in the non-services sector. To be precise, the non-services sector is a weighted average of final consumption goods and intermediate inputs, where the weight is the variable cost share of each in the non-services sector.

by $M_{2020} = 1.385$ and the sectoral average markups by $\overline{m}_{GI_{2020}} = 1.369$ and $\overline{m}_{S_{2020}} = 1.391$, where the average markup of the non-services sector corresponds to the average markup of consumption goods and intermediate inputs (weighted by their variable costs).

Labor share. The labor share corresponds to the labor compensation of private businesses as a share of aggregate output and is taken from the BLS. The targeted labor shares for 1980 and 2020 are $w_t/PY_t = \{0.678, 0.607\}$.

Quality expenses. The cost associated with quality as a share of sales in the services sector is targeted in the calibration. This is mapped to the `Cogs`-weighted average of the ratio of the sum of selling, general, and administrative expenses (`SGA`), and research and development expenses (`R&D`) to sales using Compustat data.[22] The targeted quality cost share in 2020 is $\kappa Q_{S_{2020}}/PY_{S_{2020}} = 0.080$.

Entry. The numbers of firms operating in each sector are taken from the Census' Business Dynamism Statistics and are rescaled by total population using BEA data. The number of firms operating in the non-services sector is normalized to one in 1955. The targeted number of firms in the non-services sector for 1980 and 2020 are $N_{G_t} = \{1.001, 0.818\}$, and the relative number of firms in the services sector are $N_{S_t}/N_{G_t} = \{4.054, 5.031\}$.

4.2. *Estimated parameters*

There are five preference parameters, $\{\gamma, \lambda, \phi_G, \phi_S, \delta\}$, seven technology parameters, $\{\theta, z_{G_t}, z_{I_t}, z_{S_t}\}$, and six cost parameters $\{\kappa, \vartheta, f_{G_t}, f_{S_t}\}$ for t in 1980 and 2020. Four of these parameters are exogenously set, while the remaining 14 are matched to the targets discussed above. In particular, technology and cost parameters are identified from the theory using first-order and equilibrium conditions. Preference parameters are then recovered from minimizing the model's prediction error. The aggregate wage is the numeraire, i.e., $w_t = \{1.0, 1.0\}$.

Technology and cost parameters. The vector of parameters $\Theta^\star \equiv \{\theta, z_{G_t}, z_{S_t}, \kappa, f_{G_t}, f_{S_t}\}$ is calibrated to match exactly ten data targets. This procedure uses the model's first-order and equilibrium conditions evaluated at the data targets to back out the parameter values. The solution to this system of nonlinear equations takes as given the values for the preference parameters, $\widetilde{\Theta}$, discussed below. The exponent on labor in firms' technology, θ, is identified by the labor share in 2020. The sectoral total factor productivity, z_{G_t} and z_{S_t}, help discipline the relative prices of services in 1980 and 2020, the labor share in 1980, and the aggregate markup in 2020. The quality cost parameter, κ, is used to match the fixed costs as a share of sales in the services sector in 2020. The entry costs, f_{G_t} and f_{S_t}, help match the number of firms operating in the non-services and services sectors in 1980 and 2020. The exponent on quality is exoge-

[22] To be consistent with the computation of the average markups in Compustat, the same sample restrictions as De Loecker, Eeckhout, and Unger (2020) are made.

nously set at $\vartheta = 2$ and total factor productivities of the intermediate input sector, z_{I_t}, are set so that the intermediate input prices equate the prices of consumption goods in 1980 and 2020.

Preference parameters. The vector of preference parameters $\widetilde{\Theta} \equiv \{\gamma, \lambda, \phi_G, \phi_S\}$ minimizes the model's prediction error with respect to the services shares in 1980 and 2020 as well as the average markups of goods and services in 2020.[23] This procedure internalizes how the choice of preference parameters in the outer loop affects the solution for the technology and cost parameters in the inner loop. In a nutshell, the indirect utility weight on goods, λ, and the exponent on the subutility, γ, help discipline the services share, while the choke price parameters, ϕ_G and ϕ_S, are estimated to match the average markup within each sector. Finally, δ is used to normalize the average quality of goods in 1980 to one.

TABLE 4.1

PARAMETER VALUES

Parameter	Description	Mon. comp.	Cournot	Identification
Preferences				
λ	Indirect utility's weight on goods	0.283	0.073	Services variable cost share
γ	Exponent in indirect subutility	2.134	2.436	Services variable cost share
ϕ_G	Choke price of goods	6.948	3.664	Average goods markups
ϕ_S	Choke price of services	8.278	10.766	Average services markups
δ	Exponent related with quality	0.183	0.167	Normalization ($\overline{q}_{G_{1980}} = 1$)
Technology				
θ	Exponent on labor	0.810	0.810	Labor share
z_{G_t}	TFP in goods sector in 1955, 2020	0.394, 0.769	0.648, 1.259	Labor share, aggregate markup
z_{I_t}	TFP in intermediate input sector in 1955, 2020	0.145, 0.192	0.275, 0.353	Normalization ($\overline{p}_{I_t} = \overline{p}_{G_t}$)
z_{S_t}	TFP in services sector in 1955, 2020	0.504, 0.410	0.861, 0.671	Relative price of services
Costs				
κ	Linear term related with quality	0.023	0.020	Quality expenses/sales in services
ϑ	Exponent related with quality	2.000	2.000	Exogenous
f_{G_t}	Entry costs in goods sector in 1955, 2020	0.058, 0.113	0.049, 0.113	Number of goods firms
f_{S_t}	Entry costs in services sector in 1955, 2020	0.036, 0.057	0.038, 0.057	Number of service firms

Results. Table 4.1 presents the parameter values used in the baseline exercise for the models with monopolistic and oligopolistic competition. Both models target exactly the same moments in the data. Of note, the services choke price parameter is larger than the one for goods, which reflects consumers' higher willingness to pay for services than for goods. The difference is more noticeable for the model with oligopolistic competition. The differential growth rate of productivities across sectors helps explain the strong decline of the price of manufactured goods over time. Both models predict strong productivity growth in the non-services sector. Entry costs in

[23]Specifically, denote the i'th data target by $\mathfrak{d}_i$ and the model's solution for this target by $\mathfrak{m}_i(\widetilde{\Theta}, \Theta^\star)$. Weighting each observation uniformly, the preference parameters solve the following minimization problem

$$\min_{\widetilde{\Theta}} \quad \sum_i \left[\frac{\mathfrak{d}_i - \mathfrak{m}_i(\widetilde{\Theta}, \Theta^\star)}{\mathfrak{d}_i} \right]^2 .$$

the goods sector doubled between 1980 and 2020, which helps sustain the strong decline in the number of goods-producing firms observed in the data.

Table 4.2 displays the results of the calibration exercise. The resulting fit is very good. The targeted moments are the same (up to machine precision) for the models with monopolistic and oligopolistic competition. The two models match perfectly the rise of the services share, the relative price of services, the labor share, and the number of firms in each sector in 1980 and 2020 as well as the aggregate markup, the average markups of goods and services, and the quality expenditures as a fraction of sales for the services sector in 2020.

TABLE 4.2

TARGETED MOMENTS: DATA VS. MODEL

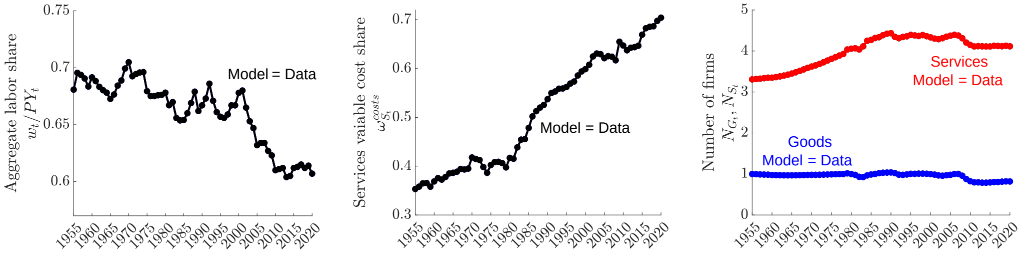
Moment	Description	Model 1980, 2020	Data 1980, 2020	Source
$\omega_{S_t}^{\text{costs}}$	Services variable cost shares	0.417, 0.704	0.417, 0.704	BEA
$\overline{p}_{S_t}/\overline{p}_{GI_t}$	Relative price of services	0.951, 1.450	0.951, 1.450	BEA
w_t/PY_t	Labor share	0.678, 0.607	0.678, 0.607	BLS
M_t	Aggregate markups	1.385	1.385	BEA, Compustat
$\overline{m}_{GI_t}$	Average non-services markups	1.369	1.369	Compustat
$\overline{m}_{S_t}$	Average services markups	1.391	1.391	Compustat
$\kappa Q_{S_t}/PY_{S_t}$	Sales share of quality expenses in services	0.080	0.080	Compustat
N_{G_t}	Number of non-services firms	1.001, 0.818	1.001, 0.818	BDS, BEA
N_{S_t}/N_{G_t}	Relative number of services firms	4.054, 5.031	4.054, 5.031	BDS, BEA

4.3. *Targeted trends*

Once all the parameters are estimated, the model is simulated yearly from 1955 to 2020 by solving for the values of the sectoral productivities, z_{G_t} and z_{S_t}, and entry costs, f_{G_t} and f_{S_t}, that match the time series of the aggregate labor share, the services variable cost share, and the number of firms in each sector. Both models with monopolistic and oligopolistic competition match these four aggregate trends perfectly (see Figure 4.1).[24] All other parameters are constant over time. The underlying productivity and entry costs are presented in Figure C.1 of the Supplemental Appendix. Technological progress is driven by the manufacturing firms that saw their productivity accelerate starting in the 1980s. In contrast, entry costs have increased sharply starting in the 2000s across both sectors, which help explain the slowdown in business dynamism observed in the data.

[24]The productivity of the intermediate input firms, z_{I_t}, is set so that the price of intermediate inputs equates the evolution of the price of consumption goods.

FIGURE 4.1.—Targeted trends, 1955-2020



(a) Labor share (b) Services variable cost share (c) Number firms

Note: Panel (a) shows the aggregate labor share of income (w_t/PY_t) in the data and baseline simulations. Panel (b) shows the services variable costs (labor and intermediate inputs) share ($\omega_{S_t}^{\text{costs}}$). Panel (c) shows the (normalized) number of firms (N_{G_t}, N_{S_t}). The trends are the same for the data and models with monopolistic and oligopolistic competition.

4.4. *Untargeted trends*

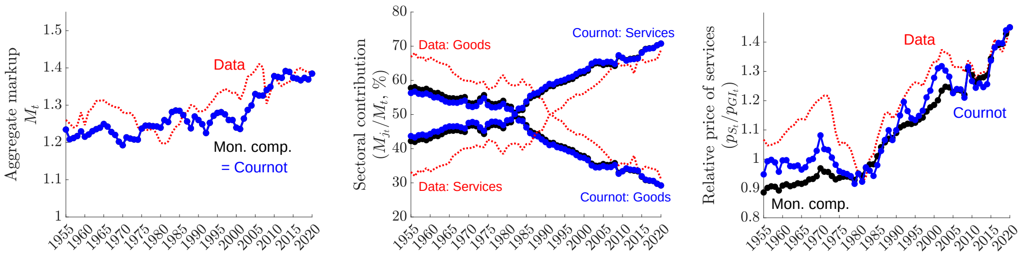
To validate the model out of sample, time series statistics not directly targeted in the simulation are now compared with the data. Figure 4.2 displays the aggregate markup (panel (a)) and the contribution of the services and non-services sectors to the aggregate markup (panel (b)). The model tracks the evolution of the aggregate markup and of the sectoral contribution remarkably well despite only targeting one data point in the entire simulation (year 2020). The model nonetheless underestimates the average markup of goods and overestimates the average markup of services at the beginning of the sample, which lowers the contribution of the goods sector in the 1950s and 60s. One explanation for the discrepancy relates to the assumption that the intermediate input sector is perfectly competitive.[25] It lowers the average markup of goods and indirectly distorts the estimated preference parameters, further pushing up the average markup of services.[26] Panel (c) shows that the model captures fairly well the rise of the relative price of services, in particular starting in 1980, and the model with oligopolistic competition does a better job overall.

The model with oligopolistic competition also produces more realistic price and income elasticities of demand. For instance, the income elasticity of demand for goods was 0.38, while that of services was 1.15. These values are consistent with goods being necessities and services being luxuries, and are in line with the estimates of Aguiar and Bills (2015) for different categories of goods and services. For instance, the income elasticity for goods is close to the one estimated for food at home (0.37), while several services (such as food away from home, entertainment, education, childcare) have income elasticities well above 1.15.

[25]The intermediate input sector accounts for 16% of the aggregate variable costs.

[26]An alternative explanation could be related to the quality of Compustat data prior to 1970. As fewer firms are covered prior to 1970, the average estimated markup might not be an accurate depiction of the true markup.

FIGURE 4.2.—Untargeted trends, 1955-2020



(a) Aggregate markup (b) Sectoral contribution (c) Relative price of services

Note: Panel (a) shows the aggregate markup (M_t) in the data (red) and in the models with monopolistic competition (black) and oligopolistic competition (blue). Panel (b) shows the sectoral contribution to the aggregate markup ($\omega_{jt}^{\text{costs}}\,\overline{m}_{jt}/M_t$). Panel (c) shows the relative price of services (p_{S_t}/p_{GI_t}). Only markups in 2020 and relative prices of services in 1980 and 2020 were targeted in the calibration.

5. DRIVERS OF THE RISE IN MARKUPS

5.1. *Synopsis*

The model is now used to decompose the forces driving the rise in markups over time. To do so, I simulate counterfactual economies in which technological progress or entry costs are held constant at their 1955 values and compute the marginal effects of shutting down their changes on the rise of the aggregate markup between 1955 and 2020.[27] Figure 5.1 shows the effect of shutting down changes in (i) productivity, z_{G_t}, z_{I_t}, and z_{S_t}, letting entry costs evolve as in the baseline (in red), and (ii) entry costs, f_{G_t} and f_{S_t}, letting productivity evolve as in the baseline (blue). The figure also shows the contribution of their interaction (orange).

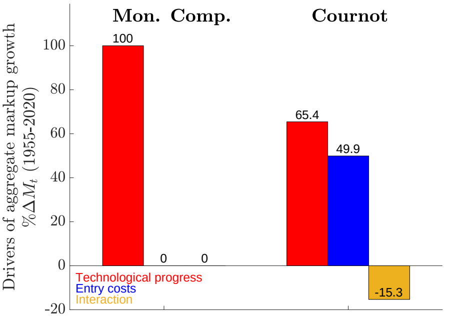
In the model with monopolistic competition, the rise in markups is entirely driven by technological progress. That is a natural consequence of firms' markups depending solely on consumers' price elasticity of demand. Technological progress is still the main driver of the rise in markups in the model with oligopolistic competition. However, it now accounts for 65% of the increase in markups between 1955 and 2020. In contrast, barriers to entry that explain the evolution of the number of firms account now for 50% of the increase in markups. This is the result of having firms' markups depend on their market share. The increase in productivity and entry costs generate interaction effects that help push the aggregate markup down.

[27] Specifically, the contribution of each experiment corresponds to the growth rate of markups in the counterfactual economy relative to the growth rate in the baseline economy, or

$$\text{Contribution} = 100 \times \left(\frac{M_{2020}^{\text{baseline}} - M_{2020}^{\text{experiment}}}{M_{2020}^{\text{baseline}} - M_{1955}^{\text{baseline}}} \right),$$

where M^{baseline} is the aggregate markup in the baseline and $M^{\text{experiment}}$ the markup in the counterfactual economy.

FIGURE 5.1.—Decomposing the rise of markups, 1955-2020



Note: The figure shows the contribution of each exogenous force on the rise of the aggregate markup between 1955 and 2020 (technological progress in red and entry costs in blue).

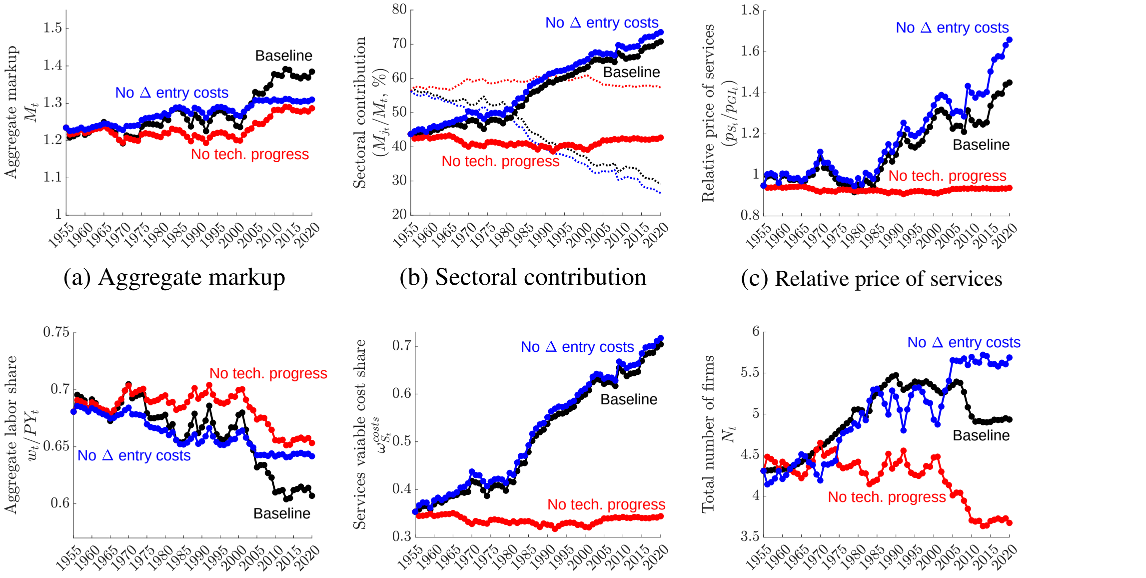
5.2. *Technological progress vs. barriers to entry*

Figure 5.2 contrasts the evolution of macroeconomic aggregates in the baseline economy (in black) with the economy without technological progress (i.e., productivities are constant and only entry costs rise; in red) and without changes in entry costs (i.e., technological progress is the only exogenous force active; in blue). The farther the counterfactual paths are to the baseline, the stronger the exogenous forces are at driving changes over time. The paths are for the economy with oligopolistic competition.

Panel (a) shows the aggregate markup. Technological progress was the main driver of the increase in markups. Without it, the aggregate markup would have been stable between 1955 and 2000, reaching 1.29 in 2020 (almost 10 p.p. lower than in the baseline). The rapid increase in manufacturing productivity relative to services explains the rise of the services share and of the relative price of services as well as the importance of the services sector in the aggregate markup (panels (b), (c), and (e)). Without technological progress, the services share would have settled at 34%, the services contribution would amount to 43% of the aggregate markup, and the price of services would equate that of manufactured goods.

Technological progress was behind the small drop in the labor share between 1955 and 2000, while the steep decline starting in the 2000s was the result of a rapid increase in entry costs (panel (d)). With higher entry costs, the number of operating firms takes a nosedive starting in the 2000s (panel (f)). This fall helps drive the increase in markups between 2000 and 2010 (panel (a)).

FIGURE 5.2.—Baseline economy vs. counterfactual economies, 1955-2020



(a) Aggregate markup (b) Sectoral contribution (c) Relative price of services

(d) Labor share (e) Services share (f) Number firms

Note: Panel (a) shows the the aggregate markup (M_t) in the baseline simulation (black) and in the counterfactual economies without technological progress (red) and changes in entry costs (blue) in the model with oligopolistic competition. Panel (b) shows the sectoral contribution to the aggregate markup ($\boldsymbol{\omega}_{jt}^{\text{costs}}\,\overline{m}_{jt}/M_t$). Panel (c) shows the relative price of services (p_{S_t}/p_{GI_t}). Panel (d) shows the aggregate labor share of income (w_t/PY_t). Panel (e) shows the services variable costs (labor and intermediate inputs) share ($\omega_{S_t}^{\text{costs}}$). Panel (f) shows the total (normalized) number of firms (N_t).

5.3. *Welfare*

In contrast to models where markups are driven solely by supply forces (e.g., Edmond, Midrigan, and Xu (2021) and De Loecker, Eeckhout, and Mongey (2021)), this framework predicts that welfare increased between 1995 and 2020 along with the rise of markups. Using equations (3.1), (3.2), and (3.3), I can now decompose the changes in the log of the household's indirect utility as stemming from: (i) changes in the difference between the varieties' prices and the consumer's choke price net of changes in her income (what I henceforth call *love for bargains*); (ii) changes in the varieties' quality (capturing the *love for quality*); and (iii) changes in the number of varieties (capturing the usual *love for variety*). In particular, the change in welfare between two points in time can be written as

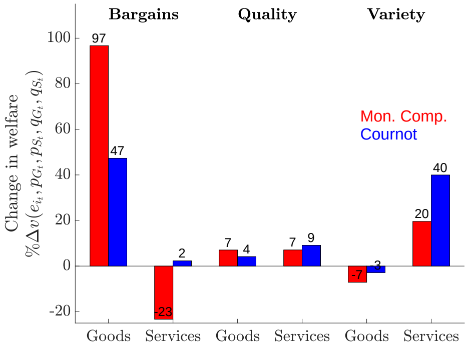
$$\Delta v(e_{i_t}, \boldsymbol{p}_{G_t}, \boldsymbol{p}_{S_t}, \boldsymbol{q}_{G_t}, \boldsymbol{q}_{S_t}) = \underbrace{(1+\gamma)\left[\lambda\Delta\left(\phi_G e_{i_t} - p_{G_t}\right) + (1-\lambda)\Delta\left(\phi_S e_{i_t} - p_{S_t}\right) - \Delta e_{i_t}\right]}_{\text{love for bargains (net of income changes)}}$$

$$+ \underbrace{(1+\gamma)\,\delta\left[\lambda\Delta q_{G_t} + (1-\lambda)\Delta q_{S_t}\right]}_{\text{love for quality}}$$

$$+ \underbrace{\lambda \Delta N_{G_t} + (1-\lambda) \Delta N_{S_t}}_{\text{love for variety}}. \tag{5.1}$$

Figure 5.3 shows this decomposition as a fraction of the total change in utility between 1955 and 2020 for both models with monopolistic competition and oligopolistic competition. The largest contributor to the increase in welfare is the *love for bargains* term. As consumers become richer and manufactured goods become cheaper, they have more disposable income to spend on goods and services. As a result, households' welfare increases noticeably. Although both models deliver the same number of firms, households in the model with oligopolistic competition put more weight on the utility derived from services (λ is larger than in the model with monopolistic competition), which makes the contribution of the *love for variety* of services more significant in the former. As households value less improvements in quality than the other two terms, the contribution of *love for quality* to welfare is less visible despite the increase in the quality of goods and services over time.

FIGURE 5.3.—Decomposing welfare gains



Note: The figure shows the contribution of each term in equation (5.1) to the change in consumers' indirect utility.

The equivalent variation measures the adjustment in income in 1955 that would make a consumer's utility equal to the level achieved in 2020.[28] Table 5.1 shows that consumers' income would have to increase by 39% in 1955 so that they would enjoy the same utility as they did in 2020, when markups were higher. Despite the higher markups, households enjoy goods and services of higher quality and more varieties to choose from in 2020 relative to 1955.

[28]This corresponds to the value of $\varepsilon_i^{\text{ev}}$ that solves the following equation

$$v(e_{i_{1955}}(1+\varepsilon_i^{\text{ev}}), \boldsymbol{p}_{G_{1955}}, \boldsymbol{p}_{S_{1955}}, \boldsymbol{q}_{G_{1955}}, \boldsymbol{q}_{S_{1955}}) = v(e_{i_{2020}}, \boldsymbol{p}_{G_{2020}}, \boldsymbol{p}_{S_{2020}}, \boldsymbol{q}_{G_{2020}}, \boldsymbol{q}_{S_{2020}}).$$

Consumers are better off in the counterfactual economies in which productivity or entry costs are held constant at their 1955 values.[29] Although productivity in the consumption goods and intermediate input sectors would be lower in that economy, the productivity of service firms would be higher, allowing consumers to enjoy services at a lower price. Households would be willing to pay between 37 and 39% of their 2020 baseline income to enjoy that level of utility. Lower entry costs in the goods and services sector also make households better off as they would imply a much higher number of varieties to choose from. In that economy, consumers would be willing to pay between 34 and 37% of their 2020 baseline income to enjoy that level of utility.

TABLE 5.1

EQUIVALENT VARIATIONS

ε_i^{ev}, %	Mon. Comp.	Cournot
Baseline economy, 1955 vs. 2020	39.1	39.1
Baseline economy vs. Counterfactual economy, 2020		
Productivities constant at 1955 values	39.1	37.3
Entry costs constant at 1955 values	34.3	36.5

6. IS TECHNOLOGICAL PROGRESS SUFFICIENT?

The previous section shows that technological progress is a necessary condition for markups to grow as it increases consumer income, which in turn reduces their price elasticity of demand. This section now studies whether technological progress alone—i.e., without income effects—is sufficient to engender a generalized increase in markups. The short answer is no.

The details follow. Proposition 3.2 shows that income effects in consumers' price elasticity of demand can be shut down by setting choke price parameters, ϕ_j, to zero. This delivers the usual CES demand with $(-\gamma)$ as the elasticity of substitution. The resulting firm's markup in the economy with oligopolistic competition is

$$m_{j_t} = \frac{(-\gamma_j)}{(-\gamma_j) - s_{j_t}},$$

where γ is allowed to differ across sectors. It is straightforward to see that in such economy markups, the labor share, and sectoral shares are constant over time. The only long-run trend this model can match is the increase in the relative price of services. The reason for that is that

[29]The equivalent variation now delivers a value of $\varepsilon_i^{\text{ev}}$ that solves the following

$$v\left(e_{i_{2020}}^{\text{baseline}}(1+\varepsilon_i^{\text{ev}}), \boldsymbol{p}_{G_{2020}}^{\text{baseline}}, \boldsymbol{p}_{S_{2020}}^{\text{baseline}}, \boldsymbol{q}_{G_{2020}}^{\text{baseline}}, \boldsymbol{q}_{S_{2020}}^{\text{baseline}}\right) = v\left(e_{i_{2020}}^{\text{exp}}, \boldsymbol{p}_{G_{2020}}^{\text{exp}}, \boldsymbol{p}_{S_{2020}}^{\text{exp}}, \boldsymbol{q}_{G_{2020}}^{\text{exp}}, \boldsymbol{q}_{S_{2020}}^{\text{exp}}\right).$$

technological progress that reduces marginal costs are translated one to one into a reduction in prices. Hence, technological progress alone cannot generate an increase in markups, a fall in the labor share, or the rise of the services sector. The Supplemental Appendix C.6 provides details about the calibration and simulation of this counterfactual economy.

7. Rising living standards or income inequality?

Thus far I have shown that income effects play a crucial role in driving the rise in markups as they allow consumers' price elasticity of demand to fall as their incomes grow. Income inequality also widened markedly over the past 70 years. Can it explain the rise in markups? To some extent, but not as much as the rise in incomes. In this section I show that although markups would be lower with less income inequality, rising living standards across the board was the major driver of the increase in markups.

Consider now an economy populated by consumers who differ in terms of skills. A fraction μ_t is high-skilled and $(1-\mu_t)$ is low-skilled. High and low-skilled labor, h_j and ℓ_j, are imperfect substitutes in production, with $1/(1-\iota)$ as the elasticity of substitution. Let x_t denote an aggregate skill-biased productivity.[30] Firm labor is now given by the following CES function

$$n_{j_t} = \left[\alpha\, x_t h_{j_t}^{\iota} + (1-\alpha)\, \ell_{j_t}^{\iota}\right]^{1/\iota}.$$

High-skilled workers receive a skill premium in the labor market, i.e., $w_{H_t}/w_{L_t} > 1$, and changes in the skill premium are driven by skill-biased productivity. Finally, a rule is needed to split aggregate expenses with entry and quality between households. The share rebated to high-skilled households, $\phi_{\Lambda_t} \equiv \mu_t \Lambda_{H_t}/\Lambda_t$, is governed by data on relative earnings of individuals with at least a bachelor's degree. Supplemental Appendix C.7 provides details about the data.

Assume firms cannot price discriminate consumers. Consumers' price elasticity of demand is still given by equation (3.6). Since consumers face the same price and high-skilled households have higher incomes ($e_{H_t} \geq e_{L_t}$), their price elasticity of demand is lower than that of low-skilled households, i.e., $\xi_{H,j_t} \leq \xi_{L,j_t}$. As consumers differ, the firm's markup is now a function of the *average* price elasticity of demand according to

$$m_{j_t} = \frac{\overline{\xi}_j(e_{H_t}, e_{L_t}, \boldsymbol{p}_{G_t}, \boldsymbol{p}_{S_t})}{\overline{\xi}_j(e_{H_t}, e_{L_t}, \boldsymbol{p}_{G_t}, \boldsymbol{p}_{S_t}) - s_{j_t}},$$

where the average price elasticity of demand, $\overline{\xi}_{j_t}$, is a weighted average of each consumer's own price elasticity of demand and their demand share (see Appendix C.3 for details).

[30] Buera et al. (2022) allow skill-biased productivity to differ across sectors. Constraining it to be the same across sectors is consistent with a time trend in the estimation of output elasticities.

Both models with monopolistic and oligopolistic competition are taken to the data. Their fit is overall very good. They both match the evolution of the aggregates discussed in Section 4. Calibration and simulations are discussed in the Supplemental Appendix C.7.

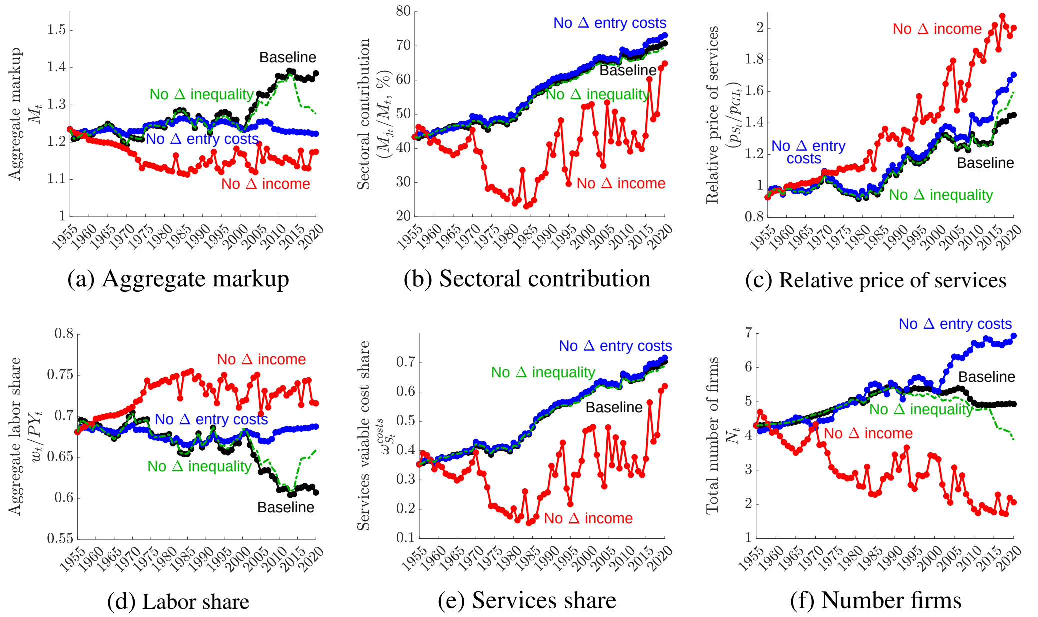
To disentangle the role of rising income inequality from rising incomes, I compute two counterfactual economies: one in which income inequality is held constant at its level in 1955, but aggregate household income grows as in the baseline simulation; and another in which income inequality evolves as in the baseline simulation, but aggregate household income is constant at its 1955 level. This is achieved by solving for the values of x_t and ϕ_{Λ_t} that minimize the distance between $\left\{e_t^{\text{exp}}, e_{H_t}^{\text{exp}}/e_{L_t}^{\text{exp}}\right\}$ from $\left\{e_t^{\text{baseline}}, e_{H_t}^{\text{baseline}}/e_{L_t}^{\text{baseline}}\right\}$, while letting neutral productivities and entry costs evolve as in the baseline simulation.

Figure 7.1 depicts the evolution of macroeconomic aggregates in the baseline economy with heterogenous households (in black) and the counterfactual economies where aggregate income (red), income inequality (green), and entry costs (blue) are constant in the model with oligopolistic competition.

If the aggregate income was constant over time—but inequality grew—the aggregate markup would have fallen relative to 1955 (panel (a)). In 2020, the aggregate markup would be 1.17 (about 20 p.p. lower than in the baseline). In contrast, keeping income inequality at its level in 1955 has little effect on the aggregate markup. Noticeable differences start only in the last decade, with the aggregate markup reaching 1.28 in 2020. It would require large differences in price elasticities of demand across consumers for income inequality to be quantitatively meaningful. Note also that rising barriers to entry is a stronger driver of the increase in markups than income inequality as the aggregate markup settles at a lower level in 2020 (about 1.22).

Despite the smaller increase of the services share in the economy with constant aggregate income (panel (e)), the services sector would still be a major contributor to the rise in markups (panel (b)) and services would be even pricier than goods (panel (c)). The labor share would have increased (panel (d)) as the denominator is constant but the aggregate wage in the numerator grows thanks to skill-biased technological progress. As consumers would be poorer, the number of firms entering the market would shrink substantially (panel (f)).

FIGURE 7.1.—Baseline economy vs. counterfactual economies with heterogeneous consumers, 1955-2020



(a) Aggregate markup (b) Sectoral contribution (c) Relative price of services

(d) Labor share (e) Services share (f) Number firms

Note: Panel (a) shows the the aggregate markup (M_t) in the baseline simulation (black) and in the counterfactual economies without changes in aggregate income (red), inequality (green), and entry costs (blue) in the model with oligopolistic competition. Panel (b) shows the sectoral contribution to the aggregate markup ($\omega^{\text{costs}}_{jt}\,\overline{m}_{jt}/M_t$). Panel (c) shows the relative price of services (p_{S_t}/p_{GI_t}). Panel (d) shows the aggregate labor share of income (w_t/PY_t). Panel (e) shows the services variable costs (labor and intermediate inputs) share ($\omega^{\text{costs}}_{S_t}$). Panel (f) shows the total (normalized) number of firms (N_t).

8. CROSS-COUNTRY EXPERIMENTS

In this section the model calibrated to the U.S. economy is confronted with data from other countries. Using only time series data on the labor and services shares of output, the model is simulated to trace the evolution of markups across many countries.[31] I focus on countries with at least 20 years of data that experienced a sustained increase in their services shares. Details and caveats about data and simulations are available in the Supplemental Appendix C.8.

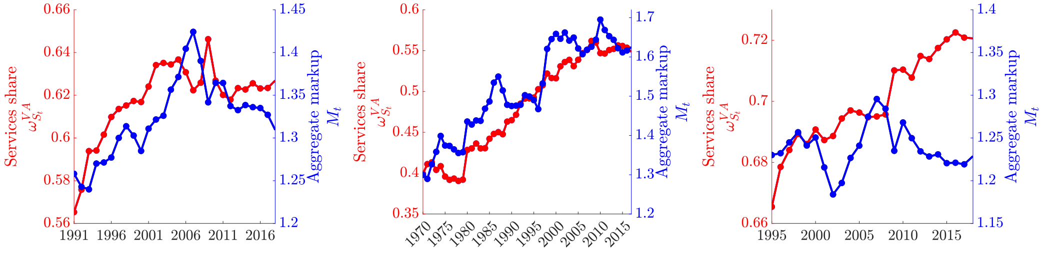
Díez, Fan, and Villegas-Sánchez (2021) document a generalized increase in markups in advanced and emerging economies between 2000 and 2015. Although there is limited historical data on the evolution of markups for most countries, the model fit to U.S. data can be used to shed light on their evolution over longer time horizons. In line with the evidence, the model delivers an increase in markups across most—yet, not all—countries that saw their services shares rise and productivity grow, with the services sector driving the increase in markups.

Figure 8.1 shows a few examples. Start with South Korea (panel (b)), which has the longest time series available. The services sector grew from 40% of value added in 1970 to 55% in

[31] Ideally, I would need the services variable cost shares to simulate the model. Since these shares are not widely available, the services value added shares are used instead.

2017. The model implies that the aggregate markup increased from 1.30 to 1.63 over the past 50 years, with services accounting for most of the increase (64% of the aggregate markup in 2017 vs. 49% in 1970). Faster productivity growth in manufacturing relative to the services sector triggered both the rise of the services share and the aggregate markup.[32]

FIGURE 8.1.—Simulated markups across countries



(a) Germany (b) Korea, Rep. (c) Switzerland

Note: The figure shows the targeted services value added share (black) together with the simulated aggregate markup (blue). Figure C.4 in the Supplemental Appendix shows the services shares and aggregate markups for other advanced economies.

Germany is another example (panel (a)). Its services share corresponded to 56% of value added in 1991. It grew until the Great Financial Crisis and then declined to values around 62% over the past decade. The aggregate markup followed a similar trend: from 1.26 in 1991 to 1.31 in 2018, peaking at 1.42 in 2007. The slowdown in markups was driven by a productivity slowdown after the crisis.[33] There are a few countries that experienced an increase in their services share without an increase in markups. Switzerland is one such example (panel (c)). Its services share grew 5 p.p. between 1995 and 2018, but its aggregate markup was flat as the underlying productivity growth was somewhat stagnant over the past 25 years.[34] Additional examples are presented in the Supplemental Appendix C.8.

9. CONCLUSION

This paper argues that the long-run rise in aggregate markups in the U.S. reflects not only changes in market structure but also deeper forces of structural transformation. I document that the expansion and markup dynamics of the services sector account for nearly all of the increase in markups since 1955. This evidence is robust to different measures of markups and

[32] Aghion et al. (2021) estimate markups for Korean manufacturing firms and find an average markup around 2.5 between 1992 and 2003.

[33] Ganglmair et al. (2020) find that markups averaged 1.35 between 2007 and 2016, with services having higher markups than manufacturing. Mertens and Mottironi (2025) find that markups averaged 1.1 between 1995 and 2016.

[34] Switzerland's productivity growth averaged less than 1% annually over the past 20 years. Steiner and Stucki (2025) find that average markups in manufacturing were stable between 2012 and 2017 at 1.26.

cuts of the data. I develop a multi-sector general equilibrium model with imperfect competition and a novel class of non-homothetic preferences to study the drivers of the rise in markups. Technological progress coupled with income effects drove most of the increase between 1955 and 2020. Increased barriers to entry played some role in the economy with oligopolistic competition. When comparing the role of rising incomes with income inequality, the former is quantitatively a lot more important in explaining the increase in markups.

These findings do not imply that concerns about competition policy are misplaced. They call, however, for a broader understanding of markup dynamics in modern economies, where income growth and structural change play a significant role. In addition, the increasing importance of services poses new challenges that have yet to be quantified. Service-producing firms tend to offer more targeted and specialized products to consumers and have a higher ability to price discriminate them. The advent of digital advertising and big data may have facilitated this. Those considerations are left for future research.

REFERENCES

D. Acemoglu and D. Autor. Skills, tasks and technologies: Implications for employment and earnings. *Handbook of Labor Economics*, 4(12):1043–1171, 2011. [20]

H. Afrouzi, A. Drenik, and R. Kim. Growing by the masses: Revisiting the link between firm size and market power. *Working Paper*, 2021. [4]

P. Aghion, S. Guriev, and K. Jo. Chaebols and firm dynamics in korea. *Economic Policy*, 36(108):593–626, 2021. [32]

M. Aguiar and M. Bills. Has consumption inequality mirrored income inequality? *American Economic Review*, 105 (9):2725–2756, 2015. [23]

U. Akcigit and S. T. Ates. Ten facts on declining business dynamism and lessons from endogenous growth theory. *American Economic Journal: Macroeconomics*, 13(1):257–298, 2021. [3]

A. Atkeson and A. Burstein. Pricing-to-market, trade costs, and international relative prices. *American Economic Review*, 98(5):1998–2031, 2008. [4]

D. Autor, D. Dorn, L. F. Katz, C. Patterson, and J. Van Reenen. The fall of the labor share and the rise of superstar firms. *The Quarterly Journal of Economics*, 135(2):645–709, 2020. [3]

S. Basu. Are price-cost markups rising in the united states? a discussion of the evidence. *Journal of Economic Perspectives*, 33(3):3–22, 2019. [8, 3]

P. Bertoletti, F. Etro, and I. Simonovska. International trade with indirect additivity. *American Economic Journal: Microeconomics*, 10(2):1–57, 2018. [4]

S. Bond, A. Hashemi, G. Kaplan, and P. Zoch. Some unpleasant markup arithmetic: Production function elasticities and their estimation from production data. *Journal of Monetary Economics*, 121:1–14, 2021. [3, 4]

T. Boppart. Structural change and the kaldor facts in a growth model with relative price effects and non-gorman preferences. *Econometrica*, 82(6):2167–2196, 2014. [4]

G. Bornstein. Entry and profits in an aging economy: The role of consumer inertia. 2025. [4]

R. F. Bretherton. A note on the law of diminishing elasticity of demand. *The Economic Journal*, 47(187):574–577, 1937. [2]

B. Bridgman and B. Herrendorf. Labor share, markups, and input-output linkages: Evidence from the national accounts. *Working Paper*, 2024. [4]

F. J. Buera and J. P. Kaboski. The rise of the service economy. *American Economic Review*, 102(6):2540–2569, 2012. [4]

F. J. Buera, J. P. Kaboski, R. Rogerson, and J. I. Vizcaino. Skill-biased structural change. *Review of Economic Studies*, 89(2):592–625, 2022. [29, 20]

D. Comin, D. Lashkari, and M. Mestieri. Structural change with long-run income and price effects. *Econometrica*, 89(1):311–374, 2021. [4, 14]

J. De Loecker and P. F. Warzynski. Markups and firm-level export status. *American Economic Review*, 102:2437–2471, 2012. [3]

J. De Loecker, J. Eeckhout, and G. Unger. The rise of market power and the macroeconomic implications. *The Quarterly Journal of Economics*, 135(2):561–644, 2020. [3, 6, 7, 20, 1]

J. De Loecker, J. Eeckhout, and S. Mongey. Quantifying market power and business dynamism in the macroeconomy. *mimeo*, 2021. [26]

M. De Ridder. Market power and innovation in the intangible economy. *mimeo*, 2021. [3]

M. De Ridder, B. Grassi, and G. Morzenti. The hitchhiker's guide to markup estimation. *mimeo*, 2022. [3]

R. Decker, J. Haltiwanger, R. Jarmin, and J. Miranda. Declining business dynamism: What we know and the way forward. *American Economic Review*, 106(5):203–07, 2016. [3]

F. J. Díez, J. Fan, and C. Villegas-Sánchez. Global declining competition? *Journal of International Economics*, 132, 2021. [31]

H. Döpper, A. MacKay, N. H. Miller, and J. Stiebale. Rising markups and the role of consumer preferences. *mimeo*, 2022. [1, 4]

C. Edmond, V. Midrigan, and D. Xu. How costly are markups? *Journal of Political Economy*, 2021. [4, 26]

N. Fabra and M. Reguant. Pass-through of emissions costs in electricity markets. *American Economic Review*, 104 (9):2872–2899, 2014. []

R. I. Feenstra, Robert C. and M. P. Timmer. The next generation of the penn world table. *American Economic Review*, 105(10):3150–3182, 2015. [23]

B. Ganglmair, N. Hahn, M. Hellwig, A. Kann, B. Peters, and I. Tsanko. Price markups, innovation, and productivity: Evidence from germany. *Produktivität für Inklusives Wachstum, Bertelsmann Stiftung, Gütersloh*, (8), 2020. [32]

R. E. Hall. New evidence on the markup of prices over marginal costs and the role of mega-firms in the u.s economy. *NBER Working Paper*, (24574), 2018. [3]

R. Harrod. *The Trade Cycle: An Essay*. Clarendon Press, Oxford, 1936. [1, 2, 6, 11]

B. Herrendorf, R. Rogerson, and A. Valentinyi. Two perspectives on preferences and structural transformation. *American Economic Review*, 103:2752–2789, 2013. [4]

B. Herrendorf, R. Rogerson, and A. Valentinyi. Growth and structural transformation. *Handbook of Economic Growth*, 2:855–941, 2014. [4]

L. F. Katz and K. M. Murphy. Changes in relative wages, 1963-1987: Supply and demand factors. *The Quarterly Journal of Economics*, 107:35–78, 1992. [20]

M. S. Kimball. The quantitative analytics of the basic neomonetarist model. *Journal of Money, Credit and Banking*, pages 1241–1277, 1995. [4, 13]

P. J. Klenow and J. L. Willis. Real rigidities and nominal price changes. *Economica*, 83:443–472, 2016. [4, 13, 14]

P. Kongsamut, S. Rebelo, and D. Xie. Beyond balanced growth. *Review of Economic Studies*, 68(4):869–882, 2001. [4]

A. Marshall. *The principles of economics*. London: Macmillan, 1890. [2, 6, 11]

K. Matsuyama. Engel's law in the global economy: Demand-induced patterns of structural change, innovation, and trade. *Econometrica*, 87(2):497–528, 2019. [4]

K. Matsuyama and P. Ushchev. Beyond ces: Three alternative classes of flexible homothetic demand systems. *Working Paper*, 2017. [4]

M. Mertens and B. Mottironi. Do larger firms exert more market power? markups and markdowns along the size distribution. *Economics Letters*, 255(112500), 2025. [32]

S. F. Moreira. Inside the decline of the labor share: technical change, market power, and structural change. *Working Paper*, 2022. [4]

E. Nakamura and D. Zerom. Accounting for incomplete pass-through. *The Review of Economic Studies*, 77(3): 1192–1230, 2010. []

R. Ngai and C. Pissarides. Structural change in a multisector model of growth. *American Economic Review*, 97(1): 429–443, 2007. [4]

D. Raval. Testing the production approach to markup estimation. *Review of Economic Studies*, 2022. [3]

R. Roy. La distribution du revenu entre les divers biens. *Econometrica*, 15(3):205–225, 1947. [8]

K. Sangani. Markups across the income distribution: Measurement and implications. *Working Paper*, 2023. [1, 4]

E. Steiner and Y. Stucki. The impact of exchange rate fluctuations on markups – firm-level evidence for switzerland. *Journal of International Money and Finance*, 159(103422), 2025. [32]

J. Traina. Is aggregate market power increasing? production trends using financial statements. 2018. Stigler Center Working Paper No. 17. [3, 8]

ONLINE APPENDIX

A: EMPIRICS

A.1. *Industry-level data for services share and relative price of services*

Industry-level data is taken from the U.S. Bureau of Economic Analysis (BEA). The benefit of using the BEA data is that it covers the entire economy and not only listed firms as is the case with Compustat. The data are annual and cover all industries starting in 1947. *Value added* is taken from here for the 1947-1997 period and from here for the 1997-2020 period. *Gross output* is taken from here for the 1947-1997 period and from here for the 1997-2020 period. *Intermediate inputs* is taken from here for the 1947-1997 period and from here for the 1997-2020 period. *Compensation of employees* is taken from here for the 1947-1997 period and from here for the 1997-2020 period. *Chain-type price indexes for value added* are taken from here for the 1947-1997 period and from here for the 1997-2020 period.

The services sector comprise all industries with NAICS code 42 and above. The non-services sector encompasses all the other remaining primary and secondary sectors (i.e., agriculture, forestry, fishing, and hunting; mining; utilities; construction; manufacturing). To be consistent with the sample restrictions applied by De Loecker, Eeckhout, and Unger (2020) with Compustat data, NAICS 55 "Management of companies and enterprises" is excluded from the analysis. That industry includes the activity of offices of bank and corporate holding companies. Public administration (NAICS 92) is also excluded from the analysis, so that the industry data covers only private sector activity.

The cost-based services share is computed using the sum of *Compensation of employees* and *Intermediate inputs*, which is the closest to the accounting definition of *Cost of goods sold* reported in Compustat. The sales-based services share is computed using *Gross output*, while the valued added services share is computed using *Value added*.

A.2. *Real price of goods and services*

The real price of goods and services is based on the Bureau of Labor Statistics (BLS) seasonally adjusted *Consumer Price Index for All Urban Consumers* (here), which covers the monthly evolution of consumer prices starting in 1947. The real prices are computed as the ratio of the good or service index over the index for all items, normalized to 0 in January 2001.

A.3. *Aggregate and average markups*

The aggregate markup is computed as the variable cost weighted average of markups of all firms in the economy, or

$$M_t = \sum_{i=1}^{N_t} \omega_{i_t} m_{i_t},$$

where $\omega_{i_t} = vc_{i_t}/vc_t$ is the variable cost share of firm i in the economy's aggregate variable costs, m_{i_t} is the firm's markup, and N_t is the aggregate number of firms in the economy. As firms are assigned to one of the broad sectors $\mathcal{J}$, the aggregate markup can also be written as

$$M_t = \sum_{j\in\mathcal{J}}\sum_{i=1}^{N_t} \omega_{i_t} m_{i_t} \mathbb{1}_{i\in j}.$$

With two sectors, it is easy to deduce that the aggregate markup can be written as

$$M_t = \boldsymbol{\omega}_{G_t}\overline{m}_{G_t} + \boldsymbol{\omega}_{S_t}\overline{m}_{S_t}.$$

where the sectoral cost share is denoted by $\boldsymbol{\omega}_{j_t} = vc_{j_t}/vc_t$, with $\boldsymbol{\omega}_{G_t} \equiv (1-\boldsymbol{\omega}_{S_t})$. The average markup within sector $j = \{G, S\}$ is given by

$$\overline{m}_{j_t} = \sum_{i=1}^{N_t} \widetilde{\omega}^j_{i_t} m_{i_t} \mathbb{1}_{i\in j}, \tag{A.1}$$

where firm i's cost share in sector j's variable costs is simply $\widetilde{\omega}^j_{i_t} = vc_{i_t}/vc_{j_t}$. Different measures of average markups are presented in the baseline results presented in Section 2 and in the robustness checks in the Supplemental Appendix.

A.4. *Firm-level data for sector-specific average markups*

Compustat North America Fundamentals Annual published by Standard and Poor's and retrieved through WRDS is used to calculate the average markups of the services and non-services sectors. The dataset provides financial information on listed firms in the U.S. starting in 1950, including measures of sales, costs of goods sold, and capital. It also includes firms' main industry classification, which allows me to group firms into non-services and services sectors. The dataset is widely used to compute measures of markups for the U.S. as it provides a long time series and covers a large proportion of economic activity. Relying on Compustat also ensures that results presented here are comparable to other studies analyzing the evolution of market power in the U.S.

Despite being widely used, the dataset also poses some limitations. First, it only includes publicly traded firms. Second, only sales are recorded and therefore prices cannot be distinguished from quantities. Third, variable costs cannot be split into labor and intermediate inputs.

Sample restrictions. I follow exactly the same sample restrictions imposed by De Loecker, Eeckhout, and Unger (2020). Firms whose data format is standard (i.e., $\texttt{datafmt} = \texttt{STD}$), population source is domestic (i.e., $\texttt{popsrc} = \texttt{D}$), and are consolidated (i.e., $\texttt{consol} = \texttt{C}$) are kept in the sample. Only data in the industrial format ($\texttt{indfmt} = \texttt{INDL}$) is kept for firms with data reported in both the industrial and financial services formats, which removes duplicates in particular for firms in the "Finance and insurance," and "Real estate and rental and leasing" industries (NAICS 52 and 53). Firms in NAICS 55 "Management of companies and enterprises" are also excluded from the sample. Firms with negative sales ($\texttt{sale}$), cost of goods sold ($\texttt{cogs}$), or selling and administrative expenses ($\texttt{xsga}$) are dropped. The sample is trimmed to exclude firms whose ratio of sales to costs of goods sold (i.e., $\texttt{sale}/\texttt{cogs}$) is above the 99th percentile or below the 1st percentile.

Average markups. The average markups within the services and non-services sectors, for $j = \{G, S\}$, are given by equation (A.1) above. A firm's variable cost share in sector $j = \{G, S\}$ variable costs are given by its share in the total *cost of goods sold* ($\texttt{cogs}$) according to $\widetilde{\omega}^j_{i_t} = \texttt{cogs}_{i_t} / \left(\sum_{\kappa=1}^{N_t} \texttt{cogs}_{\kappa_t} \mathbb{1}_{\kappa \in j} \right)$. Following De Loecker and Warzynski (2012), a firm's markup is retrieved from its cost minimization problem subject to its production function. As long as the the production function is homogeneous, the solution to the firm's problem implies that its markup is equivalent to

$$m_{i_t} = \frac{\alpha_{i_t}}{\texttt{cogs}_{i_t} / \texttt{sales}_{i_t}},$$

where $\alpha_{i_t} = \frac{\partial f_{i_t}(g_{i_t}(\ell_{i_t}, m_{i_t}))}{\partial g_{i_t}(\ell_{i_t}, m_{i_t})} \frac{g_{i_t}(\ell_{i_t}, m_{i_t})}{f_{i_t}(k_{i_t}, g_{i_t}(\ell_{i_t}, m_{i_t}))}$ is firm i's output elasticity with respect to the variable input (here assumed to be a composite of labor and materials, consistent $\texttt{cogs}$).

In the robustness checks in the Supplemental Appendix A, I also used accounting measures of markups, where the markup of an industry is simply given by the ratio of measures of output/value added and measures of marginal costs—without the need for an output elasticity.

Output elasticities. To recover the output elasticity with respect to the variable input, a Cobb-Douglas production function is estimated for each 2-digit NAICS industry j according to

$$y^j_{i_t} = \alpha^j_t \texttt{cogs}_{i_t} + \beta^j_t k_{i_t} + \omega^j_{i_t} + \varepsilon^j_{i_t},$$

where $y^j_{i_t}$ is a measure of output of firm i in industry j, $k^j_{i_t}$ is the firm's capital, $\omega^j_{i_t}$ is its productivity, and $\varepsilon^j_{i_t}$ are disturbances. The elasticity of interest is α^j_t, which differs across industries to account for differences in technology. The estimated output elasticities can also be constant or time-varying to account for technological change over time. Give the amount of data in Compustat, the time-varying elasticities are estimated in a five-year rolling window.

To deal with firms' unobserved productivity shocks, $\omega_{i_t}^j$, the control function approach is used. Here, the cost of goods sold is the control variable. Another concern with this method pertains to the units of output and variable inputs used. Compustat provides data on revenue and expenditures rather than production and input use. Bond, Hashemi, Kaplan, and Zoch (2021) discuss the issue of using revenues as opposed to quantities in greater detail.

B: MODEL

B.1. *Assumptions on the indirect subutility*

ASSUMPTION B.1: (SECTORAL INDIRECT SUBUTILITY) The sector-specific indirect subutility $v_j(e, p_j, q_j)$ satisfies the standard properties of indirect utilities, namely: $v_j(e, p_j, q_j)$ is continuous on $\mathbb{R}^3$; decreasing in prices, $\frac{\partial v_j}{\partial p_j(\omega)} \leq 0$; strictly increasing in income, $v'_{j_e} > 0$; homogeneous of degree 0 in $(e, p_j(\omega))$; convex, and hence quasiconvex, in $(e, p_j(\omega))$ up to a choke price, which is the the maximum willingness to pay for each variety of commodity j (common to all households and possibly infinite). For any price above that choke price, the indirect subutility is such that $v_j = v'_{j_p} = v'_{j_e} = 0$ (and it is thus assumed that $v_j > 0$ for any price below). It is further assumed that v_j is at least thrice differentiable, with $v''_{j_{p,p}} > 0$, $v''_{j_{p,w}} < 0$, $v'''_{j_{p,p,p}} < -\frac{v''_{j_{p,p}}}{p_j(\omega)} < 0$, and $v'''_{j_{p,p,w}} < 0$, which ensures that the price elasticity of demand is positive and that commodities consumed conform to the law of demand.

B.2. *Proof of Proposition 3.1 (From the indirect to the direct utility)*

Start from the household's consumption demand for some variety ω of commodity $j \in \{G, S\}$ using Roy's identity

$$c_{j_t}(\omega) = -\frac{\left(\partial \widehat{v}_j\left(e_t, p_{j_t}(\omega), q_{j_t}(\omega)\right) / \partial p_{j_t}(\omega)\right)\, e_{i_t}}{\left(v_j(e_t, \boldsymbol{p}_{j_t}, \boldsymbol{q}_{j_t}) / \lambda_j\right)\, \Phi_t(e_t, \boldsymbol{p}_{G_t}, \boldsymbol{p}_{S_t}, \boldsymbol{q}_{G_t}, \boldsymbol{q}_{S_t})},$$

where $\partial \widehat{v}_j\left(e_t, p_{j_t}(\omega), q_{j_t}(\omega)\right) / \partial p_{j_t}(\omega) = -\frac{1}{e_t}\left(\frac{\phi_j e_t - p_{j_t}(\omega)}{e_t}\right)^{\gamma_j} q_{j_t}(\omega)^{\delta_j(1+\gamma_j)}$. Rearrange this expression to write

$$\left(\frac{\phi_j e_t - p_{j_t}(\omega)}{e_t}\right)^{1+\gamma_j} = -\left[\frac{v_j(e_t, \boldsymbol{p}_{j_t}, \boldsymbol{q}_{j_t})\, \Phi_t(e_t, \boldsymbol{p}_{G_t}, \boldsymbol{p}_{S_t}, \boldsymbol{q}_{G_t}, \boldsymbol{q}_{S_t})\, c_{j_t}(\omega)}{\lambda_j}\right]^{\frac{1+\gamma_j}{\gamma_j}} q_{j_t}(\omega)^{-\frac{\delta_j(1+\gamma_j)^2}{\gamma_j}}.$$

Use this in the sectoral indirect utility (equation (3.3)) to write it as a function of consumption and quality. This results in the following sectoral *direct* utility u_j given by

$$u_j(\boldsymbol{c}_{G_t}, \boldsymbol{c}_{S_t}, \boldsymbol{q}_{G_t}, \boldsymbol{q}_{S_t})^{-\frac{1}{\gamma_j}} = -\frac{1}{1+\gamma_j}\left[\frac{\Phi_t(e_t, \boldsymbol{p}_{G_t}, \boldsymbol{p}_{S_t}, \boldsymbol{q}_{G_t}, \boldsymbol{q}_{S_t})\,\widetilde{C}_{j_t}}{\lambda_j}\right]^{\frac{1+\gamma_j}{\gamma_j}},$$

where $\widetilde{C}_{j_t} = \left(\int_0^{N_{j_t}}\left[\frac{c_{j_t}(\omega)}{q_{j_t}(\omega)^{\delta_j}}\right]^{\frac{1+\gamma_j}{\gamma_j}}\mathrm{d}\omega\right)^{\frac{\gamma_j}{1+\gamma_j}}$. To derive an expression of the income elasticity of the indirect utility, Φ, as a function of consumption, use the equations above to write the consumption spending share on commodity j as

$$\frac{\int_0^{N_{j_t}} p_{j_t}(\omega)c_{j_t}(\omega)\mathrm{d}\omega}{e_t} = \phi_j\int_0^{N_{j_t}} c_{j_t}(\omega)\mathrm{d}\omega + \left[\frac{u_j(\boldsymbol{c}_{G_t}, \boldsymbol{c}_{S_t}, \boldsymbol{q}_{G_t}, \boldsymbol{q}_{S_t})\,\Phi_t(e_t, \boldsymbol{p}_{G_t}, \boldsymbol{p}_{S_t}, \boldsymbol{q}_{G_t}, \boldsymbol{q}_{S_t})}{\lambda_j}\right]^{\frac{1}{\gamma_j}}\widetilde{C}_{j_t}^{\frac{1+\gamma_j}{\gamma_j}}.$$

Note that from the definition of the direct utility above, we have

$$\frac{\int_0^{N_{j_t}} p_{j_t}(\omega)c_{j_t}(\omega)\mathrm{d}\omega}{e_t} = \phi_j\int_0^{N_{j_t}} c_{j_t}(\omega)\mathrm{d}\omega - (1+\gamma_j)\frac{\lambda_j}{\Phi_t(e_t, \boldsymbol{p}_{G_t}, \boldsymbol{p}_{S_t}, \boldsymbol{q}_{G_t}, \boldsymbol{q}_{S_t})}.$$

Next, use that expression in the budget constraint to get

$$\widehat{C}_t - \frac{1}{\Phi_t(e_t, \boldsymbol{p}_{G_t}, \boldsymbol{p}_{S_t}, \boldsymbol{q}_{G_t}, \boldsymbol{q}_{S_t})}\left[(1+\gamma_G)\lambda_G + (1+\gamma_S)\lambda_S\right] = 1,$$

where $\widehat{C}_t = \phi_G\int_{\Omega_{G_t}} c_{G_t}(\omega)\mathrm{d}\omega + \phi_S\int_{\Omega_{S_t}} c_{S_t}(\omega)\mathrm{d}\omega$. Now that an expression for Φ was obtained as a function of consumption and parameters, we can replace it in the definition of the sectoral direct utility according to

$$u_j(\boldsymbol{c}_{G_t}, \boldsymbol{c}_{S_t}, \boldsymbol{q}_{G_t}, \boldsymbol{q}_{S_t}) = -\frac{1}{1+\gamma_j}\left[\frac{\lambda_j(1+\gamma_j)}{\lambda_G(1+\gamma_G)+\lambda_S(1+\gamma_S)}\right]^{1+\gamma_j}\left(\frac{\widehat{C}_t - 1}{\widetilde{C}_{j_t}}\right)^{1+\gamma_j}.$$

To get the direct utility, aggregate the two sectoral direct utilities using the Cobb-Douglas weights λ_G and λ_S according to

$$u(\boldsymbol{c}_{G_t}, \boldsymbol{c}_{S_t}, \boldsymbol{q}_{G_t}, \boldsymbol{q}_{S_t}) = \psi \left(\frac{\widehat{C}_t - 1}{\widetilde{C}_{G_t}} \right)^{\lambda_G(1+\gamma_G)} \left(\frac{\widehat{C}_t - 1}{\widetilde{C}_{S_t}} \right)^{\lambda_S(1+\gamma_S)},$$

where $\psi = (1+\gamma_G)^{-\lambda_G} (1+\gamma_S)^{-\lambda_S} \left[\frac{\lambda_G(1+\gamma_G)}{\lambda_G(1+\gamma_G)+\lambda_S(1+\gamma_S)} \right]^{\lambda_G(1+\gamma_G)}$
$\left[\frac{\lambda_S(1+\gamma_S)}{\lambda_G(1+\gamma_G)+\lambda_S(1+\gamma_S)} \right]^{\lambda_S(1+\gamma_S)}$.

B.3. *Proof of Proposition 3.2 (Two-sector CES)*

Assume $\phi_j = 0$, $\gamma_j < -1$, and $\delta_j < 0$ for $j = \{G, S\}$. Denote the elasticity of substitution across varieties by $\vartheta_j = -\gamma_j$ and let P_{j_t} denote the sectoral ideal price index given by

$$P_{j_t} = \left[\int_0^{N_{j_t}} p_{j_t}(\omega)^{1-\vartheta_j} q_{j_t}(\omega)^{\delta(1-\vartheta_j)} \mathrm{d}\omega \right]^{\frac{1}{1-\vartheta_j}}.$$

The direct utility is then given by

$$u(\boldsymbol{c}_{G_t}, \boldsymbol{c}_{S_t}, \boldsymbol{q}_{G_t}, \boldsymbol{q}_{S_t}) = \psi \widetilde{C}_{G_t}^{\lambda_G(\vartheta_G-1)} \widetilde{C}_{S_t}^{\lambda_S(\vartheta_S-1)}$$

and the indirect utility by

$$v(e_t, \boldsymbol{p}_{G_t}, \boldsymbol{p}_{S_t}, \boldsymbol{q}_{G_t}, \boldsymbol{q}_{S_t}) = \left[\int_0^{N_{G_t}} \frac{1}{\vartheta_G - 1} \left[\frac{e_t}{(\phi_G e_t - p_{G_t}(\omega))\, q_{G_t}(\omega)^{\delta_G}} \right]^{\vartheta_G - 1} \mathrm{d}\omega \right]^{\lambda_G} \left[\int_0^{N_{S_t}} \frac{1}{\vartheta_S - 1} \left[\frac{e_t}{(\phi_S e_t - p_{S_t}(\omega))\, q_{S_t}(\omega)^{\delta_S}} \right]^{\vartheta_S - 1} \mathrm{d}\omega \right]^{\lambda_S}.$$

The consumption demand for variety ω of commodity j is

$$c_{j_t}(\omega) = \left[\frac{p_{j_t}(\omega)}{P_{j_t}} \right]^{-\vartheta_j} q_{j_t}(\omega)^{\delta_j(1-\vartheta_j)} \widetilde{C}_{j_t}.$$

B.4. *Using Roy's identity to derive demand (3.5)*

Use Roy's identity to express the consumer's demand for variety ω of commodity j as

$$c_{j_t}(\omega) = -\frac{\partial v(e_t, \boldsymbol{p}_{G_t}, \boldsymbol{p}_{S_t}, \boldsymbol{q}_{G_t}, \boldsymbol{q}_{S_t}) / \partial p_{j_t}(\omega)}{\partial v(e_t, \boldsymbol{p}_{G_t}, \boldsymbol{p}_{S_t}, \boldsymbol{q}_{G_t}, \boldsymbol{q}_{S_t}) / \partial e_t},$$

which simplifies to

$$c_{j_t}(\omega) = -\frac{(\partial \widehat{v}_j(e_t, p_{j_t}(\omega), q_{j_t}(\omega))/\partial p_{j_t}(\omega))\ e_t}{\left(v_j(e_t, \boldsymbol{p}_{j_t}, \boldsymbol{q}_{j_t})/\lambda_j\right)\ \Phi_t},$$

where $\Phi_t = \left(\partial v(e_t, \boldsymbol{p}_{G_t}, \boldsymbol{p}_{S_t}, \boldsymbol{q}_{G_t}, \boldsymbol{q}_{S_t})/\partial e_t\right) \left(e_t/v(e_t, \boldsymbol{p}_{G_t}, \boldsymbol{p}_{S_t}, \boldsymbol{q}_{G_t}, \boldsymbol{q}_{S_t})\right) > 0$ is the total utility's income elasticity. The denominator corresponds to the sectoral composite in equation (3.5), with $A_{j_t} = \left(v_j(e_t, \boldsymbol{p}_{j_t}, \boldsymbol{q}_{j_t})/\lambda_j\right)\ \Phi_t$, while the numerator can be expressed as $-\left(\partial \widehat{v}_j(e_t, p_{j_t}(\omega), q_{j_t}(\omega))/\partial p_{j_t}(\omega)\right)\ e_t = [\phi_j\, e_t - p_{j_t}(\omega)]^\gamma\, q_{j_t}(\omega)^{\delta(1+\gamma)}$.

B.5. *Oligopoly model*

Let ω_j denote a variety of commodity $j \in \{G, S\}$. The firm's problem (3.11) can now be rewritten as

$$\pi_{\omega_{j_t}} = \max_{y_{\omega_{j_t}}, q_{\omega_{j_t}}} \left(p(y_{j_t}) - mc_{\omega_{j_t}}\right) y_{\omega_{j_t}} - \kappa q_{\omega_{j_t}}^{\vartheta} - f_{j_t}$$

subject to the demand constraint $y_{j_t} = c_j(p_{j_t}, q_{\omega_{j_t}})$ and the total market supply $y_{j_t} = \sum_{\widetilde{\omega}_j=1}^{N_{j_t}} y_{\widetilde{\omega}_{j_t}}$. The firm's optimal pricing decision is still a markup over marginal costs as in equation (3.12). The markup is now not only a function of consumers' price elasticity of demand, but also depends on the firm's share of total sales of commodity j, $s_{\omega_{j_t}} \equiv \frac{p(y_{j_t}) y_{\omega_{j_t}}}{p(y_{j_t}) y_{j_t}}$, according to

$$m_{\omega_{j_t}} = \frac{\xi_{\omega_{j_t}}(e_t, p_{j_t})}{\xi_{\omega_{j_t}}(e_t, p_{j_t}) - s_{\omega_{j_t}}}. \tag{B.1}$$

In a symmetric equilibrium, a firm's sales share of sectoral output is $s_{\omega_{j_t}} = 1/N_{j_t}$. Rearranging equation (3.12) and replacing the firm's marginal cost yields equation (3.16). Note that now the choice of the firm's output is a best response to all other competing firms' choices of output.

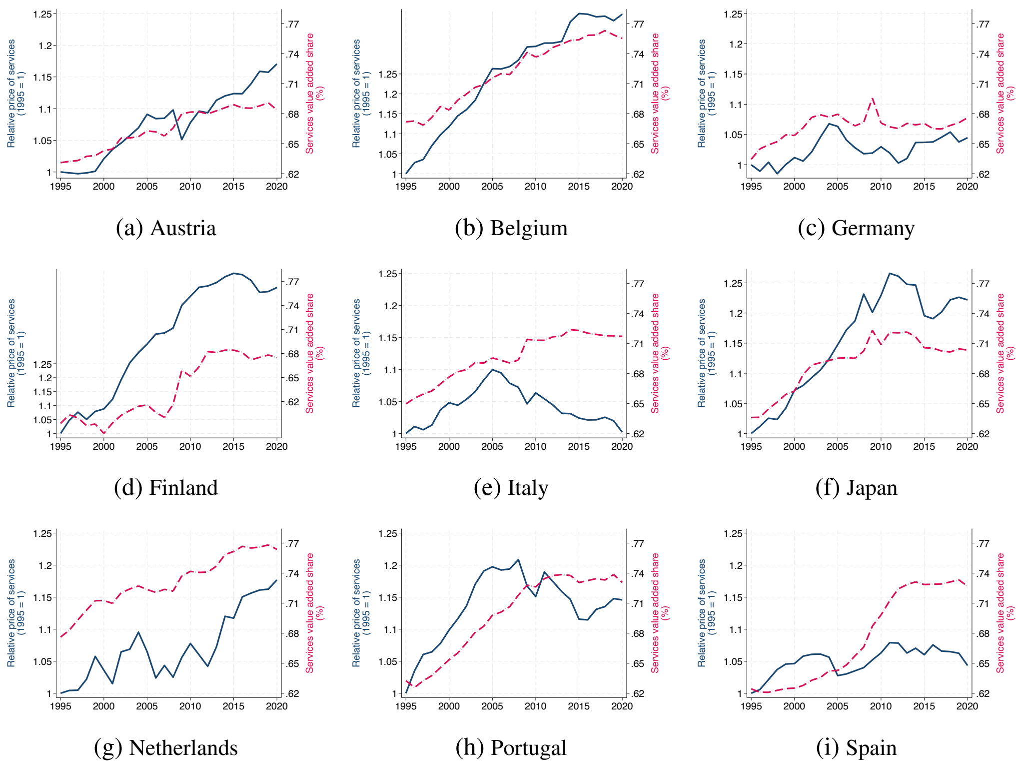
SUPPLEMENTAL APPENDIX

A: EMPIRICAL EXTENSIONS AND ROBUSTNESS CHECKS

A.1. *Services share and relative price of services across other advanced economies*

The rise of the services share was also accompanied by an increase in the relative price of services across several advanced economies. The data are taken from the EUKLEMS & INTANProd database, National Accounts, made available by the Luiss Lab of European Economics (here). The non-services sector corresponds to NACE codes A to F, and the services sector covers NACE codes G to S. The relative price of services corresponds to the chain-weighted Fisher price index of the value added price indices of individual industries.

FIGURE A.1.—Structural change across advanced economies



(a) Austria (b) Belgium (c) Germany

(d) Finland (e) Italy (f) Japan

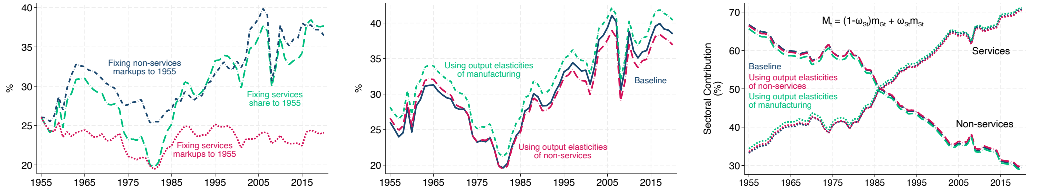
(g) Netherlands (h) Portugal (i) Spain

Note: The figure shows the services value added share (red) and the relative value added price of services (black) for selected advanced economies, using data from the EUKLEMS & INTANProd database made available by the Luiss Lab of European Economics (here).

A.2. *Aggregate markups with constant output elasticities*

The figures below reproduce Figures 2.2c, 2.3a, and 2.3b when the estimated output elasticities are constant over time. Although the level of the markup differs over time, the services sector still is the main driver of its increase. Figure A.2a shows that if the average markup of the services sector was held constant at its 1955 level, the aggregate markup would have been constant over time. Figure A.2b shows that using different output elasticities still delivers an increase in the aggregate markup and Figure A.2c confirms the importance of the services sector when the output elasticities of the manufacturing sector or goods sector are used.

FIGURE A.2.—Aggregate markups and services' contribution with constant output elasticities



(a) Markups fixing average markups and sectoral share

(b) Markups with different elasticities

(c) Sectoral contribution with different elasticities

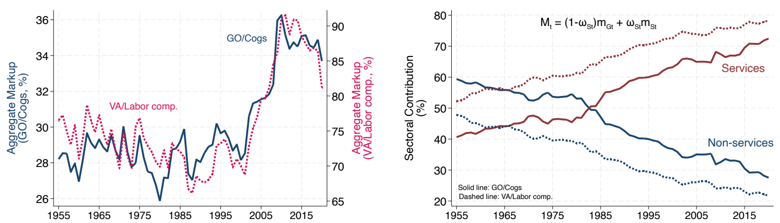
Note: Panel (a) shows the aggregate markup (with constant output elasticities), M_t, when the average markup within each sector, $\overline{m}_{j_t}$, is fixed at its 1955 level (non-services in blue, services in red) and when the services cost share, $\boldsymbol{\omega}_{S_t}^{\text{costs}}$, is fixed at its 1955 level (green). Panel (b) shows the aggregate markup (with constant output elasticities), M_t, when the average markup of services, $\overline{m}_{S_t}$, is computed using the average output elasticity with respect to `Cogs` of the non-services industries (red) and of manufacturing (green). The aggregate markup in the baseline data is in blue. Panel (c) shows the sectoral contribution to the aggregate markup (with constant output elasticities), $\boldsymbol{\omega}_{j_t}^{\text{costs}}\,\overline{m}_{j_t}/M_t$, using the average output elasticity with respect to `Cogs` of the non-service industries (red) and of manufacturing (green). The sectoral contribution in the baseline data is in blue.

A.3. *Aggregate markups with accounting markups from industry-level data*

An alternative measure of markups can be computed without estimating production functions. Instead of using firm-level data, I proceed with data from the BEA on the entire U.S. industrial production (see Appendix A.1 for details about the data). I now compute each industry's markup, m_{i_t}, as the ratio of gross output to costs of goods sold or as the ratio of value added to labor compensation. This measure implicitly assumes that the production function within each industry is constant returns to scale. Each sector's average markup, $\overline{m}_{j_t}$, is now a weighted average of all industries' markup within the sector, with their weight, $\widetilde{\omega}_{i_t}^{j}$, being the industry's variable cost share. To be precise, when markups are computed as the ratio of gross output to costs of goods sold, the appropriate industry weight uses the costs of goods sold. In contrast, when markups are computed as the ratio of value added to labor compensation, the industry weight uses labor compensation.

Figure A.3a depicts the aggregate markup when each sector's average markup is computed using the ratio of gross output to costs of goods sold and the ratio of value added to labor compensation. Although the level of the aggregate markup differs across both measures, their trends were similar. Both measures suggest that there was a trough in the early 1980s and that the aggregate markup increased rapidly after—with the most noticeable growth happening in the 2000s. The contribution of the services sector is similar to what was depicted in Section 2. The services sector corresponded to less than 50% of the aggregate markup in the 1950s under both measures as Figure A.3b displays. The services sector is now more than 70% of the aggregate markup.

FIGURE A.3.—Aggregate markups and services' contribution with accounting markups



(a) Aggregate markup (unit elasticities) (b) Sectoral contribution (unit elasticities)

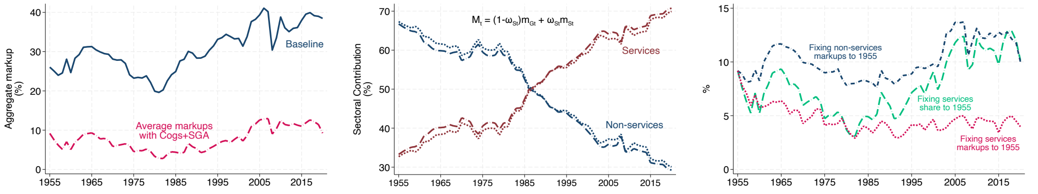
Note: Panel (a) shows the aggregate markup (with unit output elasticities), M_t, when the sectoral average markups are computed as the ratio of Gross output to Costs of goods sold (in blue) and as the ratio of Value Added to Labor compensation (in red) using industry-level data from the BEA. Panel (b) shows the sectoral contribution to the aggregate markup (with unit output elasticities), $\boldsymbol{\omega}_{jt}^{\text{costs}}\,\overline{m}_{jt}/M_t$, when the sectoral average markups are computed as the ratio of Gross output to Costs of goods sold and as the ratio of Value Added to Labor compensation (the services contribution is in red and the non-services contribution is in blue) using industry-level data from the BEA.

A.4. *Aggregate markups with selling, administrative, and general expenses*

The right measure of firms' variable costs might be debatable. In addition to the cost of goods sold, a firm's income statement reports selling, general, and administrative expenses (SGA), which tend to be non-production costs that include rent, utilities, and marketing. Traina (2018) and Basu (2019) highlight that not including them might bias the overall increase in the aggregate markup. Figure A.4a shows that this is the case. When average markups include SGA in both the firm-level markup, m_{i_t}, and in the firm's weight in each sector, $\widetilde{\omega}_{i_t}^{j}$, the aggregate markup would be lower. Instead of growing 12p.p. between 1955 and 2020 and 19p.p. between 1980 and 2020, the aggregate markup would have increased 1p.p. and 6p.p. respectively.

The sectoral contribution of the services sector is not affected by the inclusion of SGA as evidenced by Figure A.4b. In particular, shutting down changes in the average markup of services over the past 65 years would translate into a decline of the aggregate markup from 9% in 1955 to 4% in 2020 (viz. Figure A.4c). Fixing the services share to its value in 1955 would have also implied a lower aggregate markup for much of the 20th century.

FIGURE A.4.—Aggregate markups and services' contribution including SGA data



(a) Aggregate markup with SGA (b) Sectoral contribution with SGA (c) Counterf. markups with SGA

Note: Panel (a) shows the aggregate markup, M_t, in the baseline (blue) and when the average markups within each sector are computed using the sum of `Cogs` and `SGA` (red), using BEA data to measure the services share, $\omega_{S_t}^{\text{costs}}$, and Compustat data to measure average markups within sectors, $\overline{m}_{j_t}$. Panel (b) shows the sectoral contribution to the aggregate markup (non-services in blue, services in red), $\omega_{j_t}^{\text{costs}}\,\overline{m}_{j_t}/M_t$ in the baseline and when the average markups within each sector are computed using the sum of `Cogs` and `SGA` (dotted line). Panel (c) shows the aggregate markup, M_t, when the average markup within each sector, $\overline{m}_{j_t}$, is fixed at its 1955 level (non-services in blue, services in red) and when the services cost share, $\omega_{S_t}^{\text{costs}}$, is fixed at its 1955 level (green).

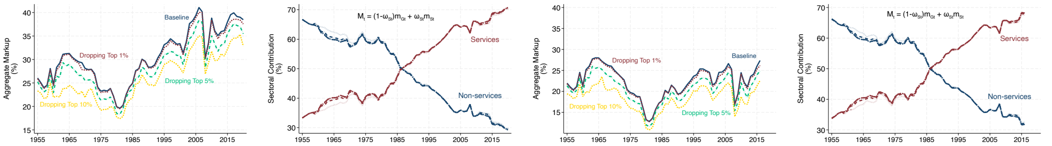
A.5. *Aggregate markups without the right tail of firms*

I now focus on the right tail of the markup distribution. Compustat is composed of publicly listed firms that tend to be larger and more established, and hence might skew the average markups. To address this potential concern, firms that are in the top 1%, 5%, and 10% of the markup distribution within each sector and year are dropped. Figure A.5a shows the aggregate markup when output elasticities are constant over time. When dropping firms that are in the top 1%, 5%, and 10% of the markup distribution within each sector and year, the increase in aggregate markups is still noticeable—albeit to a smaller extent. The services sector is still the largest contributor to the aggregate markup as Figure A.5b shows. When output elasticities vary over time, the aggregate markup is reduced, yet the importance of the services sector is not diminished as Figures A.5c and A.5d display.

A.6. *Aggregate profits*

I now use data on corporate profits from the BEA's national income and product accounts (Table 6.17). Corporate profits are the second-largest component of national income after employee compensation and consists of net dividends and undistributed profits from current production for all financial and nonfinancial firms required to file federal corporate tax returns as well as profits originating in the rest of the world that are received by U.S. residents (i.e., dividends from foreign corporations to U.S. investors and firms). I focus on corporate profits from

FIGURE A.5.—Dropping the right tail of the markup distribution



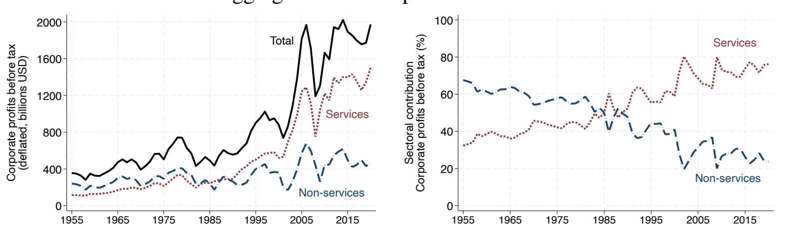
(a) Aggregate markup (constant elasticities) (b) Sectoral contribution (constant elasticities) (c) Aggregate markup (time-varying elasticities) (d) Sectoral contribution (time-varying elast.)

Note: Panel (a) shows the aggregate markup (with constant output elasticities), M_t, in the baseline (in blue) and when firms in the top 1%, 5%, and 10% of the markup distribution within each sector and year are removed (red, green, and yellow, respectively). Panel (b) shows the sectoral contribution to the aggregate markup (with constant output elasticities), $\boldsymbol{\omega}_{jt}^{\text{costs}}\,\overline{m}_{jt}/M_t$, in the baseline and when firms in the top 1%, 5%, and 10% of the markup distribution within each sector and year are removed (the curves are barely distinguishable among each other). Panels (c) and (d) replicate panels (a) and (b) using time-varying output elasticities at the 2-digit NAICS level.

domestic industries and before taxes on corporate income. Profits are deflated using the BEA's seasonally-adjusted GDP implicit price deflator (with 2017 as the base year).

Figure A.6a shows aggregate profits and profits from services and non-service industries over time. Real corporate profits have increased markedly, in particular from 1980 until the Great Financial Crisis in line with the rise in aggregate markups. The services sector was the main driver of this increase. Services accounted for 32% of aggregate profits in 1955; the sector represented 76% of aggregate profits in 2020 as Figure A.6b displays.

FIGURE A.6.—Aggregate domestic profits and sectoral contributions



(a) Aggregate profits (b) Sectoral contribution

Note: Panel (a) shows deflated aggregate profits from domestic industries (in black), profits from service industries (in red) and from non-service industries (in blue) in billions of US$. Panel (b) shows the sectoral contribution to aggregate profits.

A.7. *Markups and the labor share of income*

I now discuss the relationship between markups and the labor share of income. As previously defined, the markup of a firm i can be expressed as

$$m_{i_t} = \frac{\alpha_{i_t} p_{i_t} y_{i_t}}{w_t \ell_{i_t} + p_{m_t} m_{i_t}},$$

where α_{i_t} is the firm's output elasticity with respect to the costs of goods sold (the sum of labor and intermediate expenses). The average markup within sector $j = \{G, S\}$ is then given by

$$\overline{m}_{j_t} = \sum_{i=1}^{N_t} \widetilde{\omega}_{i_t}^{j} \left(\frac{\alpha_{i_t} w_t \ell_{i_t}}{w_t \ell_{i_t} + p_{m_t} m_{i_t}} \right) \frac{p_{i_t} y_{i_t}}{w_t \ell_{i_t}} \mathbb{1}_{i \in j},$$

where $\widetilde{\omega}_{i_t}^{j}$ is the firm's cost weight defined in Appendix A.3.The labor share in sector j is usually defined as the ratio of the sum of total labor compensation to aggregate output, or

$$\text{Labor share}_{j_t} = \left(\sum_{i=1}^{N_t} w_t \ell_{i_t} \mathbb{1}_{i \in j} \right) / \left(\sum_{i=1}^{N_t} p_{i_t} y_{i_t} \mathbb{1}_{i \in j} \right) \equiv \beta_{j_t}.$$

For simplicity, assume that all firms within sector j are symmetric. Hence, the sector's average markup can now be written as

$$\overline{m}_{j_t} = \left(\frac{w_t \ell_{j_t}}{w_t \ell_{j_t} + p_{m_t} m_{j_t}} \right) \frac{\alpha_{j_t}}{\beta_{j_t}}.$$

It transpires that a decline in the sector's labor share does not necessarily imply an increase in the sector's average markup as either the labor share of variable costs, $w_t \ell_{j_t} / (w_t \ell_{j_t} + p_{m_t} m_{j_t})$, or the output elasticity with respect to the variable cost, α_{j_t}, may have also fallen.

B: From Structural Change to Rising Markups: Theoretical Underpinnings

This section proposes the key ingredients needed for structural change to impact markups. It starts by offering a novel theorem linking the price elasticity of demand to the income elasticity of demand for a general class of preferences. In particular, it shows that non-homothetic preferences, which imply that for individuals of different income levels some commodities are luxuries and some are necessities, also mean that individuals of different income levels will have a different price elasticity of demand for the same commodity, affecting markups.

The theorem is used as a stepping stone for two key results. The first states the conditions for the price elasticity of demand to be increasing in a commodity's own price, i.e., individuals' price sensitivity is lower for cheaper products. This is often referred to as Marshall's (1890) *Second Law of Demand*. The second states the conditions for the price elasticity to be decreasing in a consumer's income, i.e. individuals' price sensitivity is lower the wealthier they are. This second result connects with Harrod's (1936) *Law of Diminishing Elasticity of Demand*.

B.1. *Why demand matters more than you think*

Pricing with market power. A firm's markup depends on the slope of the demand curve as profit maximizing firms set their prices by equating marginal revenues to marginal costs. A firm's marginal revenue depends on the price and quantity of the product it is selling, which in turn depend on its consumers' own price elasticities of demand. If aggregate demand is composed of different consumers, all facing the same price, the price elasticity of the total demand faced by the firm can be written as the average of each individual's own price elasticity of demand weighted by their consumption share. Proposition B.1 shows that in models in which firms have market power markups can be written as a weighted average of the firm's consumers' price elasticities of demand. Let $\xi_i(q^\star)$ denote individual i's own price elasticity of demand and $\varpi_i(q^\star)$ her consumption share, where $q^\star = \sum_j y_j = \sum_i c_i$ is the aggregate quantity traded in equilibrium, and $\epsilon_j(q^\star) \equiv \left(\frac{\partial y_j(q^\star)}{\partial q^\star}\right)^{-1} \frac{y_j(q^\star)}{q^\star}$ firm j's output elasticity of aggregate demand.

PROPOSITION B.1: (MARKUP) In models of imperfect competition, in which the market structure is composed of a monopolist, monopolistic competitors or oligopolists à la Cournot, firm j's markup, $m_j(q^\star)$, is given by

$$m_j(q^\star) = \frac{\sum\limits_i \varpi_i(q^\star)\,\xi_i(q^\star)}{\sum\limits_i \varpi_i(q^\star)\,\xi_i(q^\star) - \epsilon_j(q^\star)}.$$

If firm j is a monopolist or a monopolistic competitor, then $\epsilon_j(q^\star) = 1$.

PROOF: Assume firms have constant returns to scale technologies. If the firm is a monopolist or a monopolistic competitor, it solves the following profit maximization problem

$$\max_{c>0}\; p(c)\,c - mc\,c.$$

A solution to this problem must satisfy the first-order condition, which equates the marginal revenue to the marginal cost. Dividing both sides by $p(c)$, we have

$$\frac{\partial p(c)}{\partial c}\frac{c}{p(c)} + 1 = \frac{1}{m(c)},$$

where $m(c)$ is the firm's markup, and rearranging

$$m(c) = \frac{\xi(c)}{\xi(c) - 1},$$

where $\xi(c)$ is the price elasticity of aggregate demand or the weighted of each individual's own price elasticity of demand $\xi(c) = \sum_i \varpi_i(c)\,\xi_i(c)$.

If the firm is an oligopolist competing à la Cournot, it solves the following problem

$$\max_{c_j > 0}\; p(c)\, c_j - mc_j\, c_j \quad \text{s.t. } c_j + \sum_{\kappa=1}^{N} c_\kappa = c,$$

where N is the number of competitors. As before, a solution to this problem must satisfy the first-order condition, which equates the marginal revenue to the marginal cost according to

$$\frac{\partial p(c)}{\partial c}\frac{c}{p(c)}\frac{\partial c}{\partial c_j}\frac{c_j}{c} + 1 = \frac{1}{m_j(c)},$$

and rearranging

$$m_j(c) = \frac{\xi(c)}{\xi(c) - \epsilon_j(c)}.$$

The markup is increasing in $\epsilon_j(c)$ as $\frac{\partial m_j(c)}{\partial \epsilon_j(c)} = \frac{m_j(c)^2}{\xi(c)} > 0$. *Q.E.D.*

The roots of market power are thus intertwined with how preferences are defined as they determine in equilibrium consumers' price elasticities of demand, $\xi_i(q^\star)$, and their consumption shares, $\varpi_i(q^\star)$. To proceed I resort to the indirect utility and Roy's (1947) identity. Let e_i denote individual i's expenditures (or income), $p(\omega)$ the price of variety ω, and $\boldsymbol{p}$ a vector of all prices. The identity establishes that demand for a variety, $c(e_i, p(\omega), \boldsymbol{p})$, can be derived using an individual's indirect utility, $v(e_i, \boldsymbol{p})$, and its derivatives with respect to the variety's price as

$$c(e_i, p(\omega), \boldsymbol{p}) = -\frac{\partial v(e_i, \boldsymbol{p})/\partial p(\omega)}{\partial v(e_i, \boldsymbol{p})/\partial e_i}, \tag{B.1}$$

where the indirect utility satisfies the usual properties postulated in Assumption B.1 below. The results that follow require the additional Assumption B.2, which ensures all objects are well defined. In particular, Assumption B.2 (i) is needed to ensure both the price and income elasticities of demand are well defined, while (ii) ensures the pass-through between the price and income elasticities of demand is not degenerate (i.e. $\chi(e_i, p(\omega), \boldsymbol{p}) \neq 0$) and (iii) ensures the price elasticity of demand is positive (i.e. $\xi(e_i, p(\omega), \boldsymbol{p}) > 0$).

ASSUMPTION B.1: (INDIRECT UTILITY) The indirect utility $v(e_i, \boldsymbol{p})$ is: (i) continuous on $\mathbb{R}^N \times \mathbb{R}$; (ii) decreasing in prices, $\frac{\partial v(e_i,\boldsymbol{p})}{\partial p(\omega)} \leq 0$ for all $p(\omega)$; (iii) strictly increasing in income, $\frac{\partial v(e_i,\boldsymbol{p})}{\partial e_i} > 0$; (iv) homogeneous of degree 0 in $(e_i, \boldsymbol{p})$; (v) quasiconvex in $(e_i, \boldsymbol{p})$.

ASSUMPTION B.2: (DIFFERENTIABILITY) The indirect utility function $v(e_i, \boldsymbol{p})$ is at least twice continuously differentiable and satisfies (i) $\frac{\partial v(e_i,\boldsymbol{p})}{\partial p(\omega)} < 0$ for all $p(\omega)$; (ii) $\frac{\partial^2 v(e_i,\boldsymbol{p})}{\partial e_i \partial p(\omega)} \neq 0$ for all $p(\omega)$; and (iii) $\frac{\partial^2 v(e_i,\boldsymbol{p})}{\partial p(\omega)^2} > 0$.

Price elasticity of demand. Start with an inidividual's price elasticity of demand, $\xi(e_i, p(\omega), \boldsymbol{p}) \equiv -\frac{\partial c(e_i,p(\omega),\boldsymbol{p})}{\partial p(\omega)} \frac{p(\omega)}{c(e_i,p(\omega),\boldsymbol{p})}$. Using the indirect utility, the price elasticity of demand can be expressed as

$$\xi(e_i, p(\omega), \boldsymbol{p}) = -p(\omega) \left[\frac{\partial^2 v(e_i, \boldsymbol{p})/\partial p(\omega)^2}{\partial v(e_i, \boldsymbol{p})/\partial p(\omega)} - \frac{\partial^2 v(e_i, \boldsymbol{p})/\partial e_i \partial p(\omega)}{\partial v(e_i, \boldsymbol{p})/\partial e_i} \right]. \tag{B.2}$$

This expression highlights the different channels through which changes in the price elasticity of demand materialize. Notably, changes in the variety's price, and possibly all other prices, and in the household's income can alter a consumer's price elasticity of demand. In models without strategic interactions, the dependence on competitors' prices does not affect the price elasticity of demand directly. Likewise, in models with homothetic preferences the price elasticity of demand does not depend on the consumer's income. Finally, demand for a variety is said to be inelastic when the price elasticity is less than one (i.e. $\xi(e_i, p(\omega), \boldsymbol{p}) < 1$): that is, changes in price have a relatively small effect on the quantity demanded (perfectly inelastic if the elasticity is zero). Demand for a variety is said to be elastic when the elasticity is greater than one (i.e. $\xi(e_i, p(\omega), \boldsymbol{p}) > 1$; perfectly elastic if the elasticity is infinity). Varieties conform to the law of demand as long as $\xi(e_i, p(\omega), \boldsymbol{p}) \geq 0$.

Income elasticity of demand. The income elasticity of demand, $\eta(e_i, p(\omega), \boldsymbol{p}) \equiv \frac{\partial c(e_i,p(\omega),\boldsymbol{p})}{\partial e_i} \frac{e_i}{c(e_i,p(\omega),\boldsymbol{p})}$, measures how demand changes in response to changes in income. Using the consumer's indirect utility, the elasticity is given by

$$\eta(e_i, p(\omega), \boldsymbol{p}) = e_i \left[\frac{\partial^2 v(e_i, \boldsymbol{p})/\partial e_i \partial p(\omega)}{\partial v(e_i, \boldsymbol{p})/\partial p(\omega)} - \frac{\partial^2 v(e_i, \boldsymbol{p})/\partial e_i^2}{\partial v(e_i, \boldsymbol{p}) \partial e_i} \right]. \tag{B.3}$$

A variety is said to be a luxury for the consumer if the income elasticity is greater than one (i.e. $\eta(e_i, p(\omega), \boldsymbol{p}) > 1$), a necessity if the elasticity is positive but less than one (i.e. $0 < \eta(e_i, p(\omega), \boldsymbol{p}) < 1$), and an inferior good if the elasticity is negative (i.e. $\eta(e_i, p(\omega), \boldsymbol{p}) < 0$).

Income elasticity and super-elasticity of utility. The income elasticity of utility, $\Phi(e_i, \boldsymbol{p}) \equiv \frac{\partial v(e_i,\boldsymbol{p})}{\partial e_i} \frac{e_i}{v(e_i,\boldsymbol{p})}$, measures how the consumer's utility changes when income changes. This elasticity is common to all varieties and takes into account all the possible interactions across varieties when income changes. As households tend to enjoy more utility if their income grows, $\Phi(e_i, \boldsymbol{p})$ is usually positive. In turn, the income super-elasticity of utility, $\varphi(e_i, \boldsymbol{p}) \equiv \frac{\partial \Phi(e_i,\boldsymbol{p})}{\partial e_i} \frac{e_i}{\Phi(e_i,\boldsymbol{p})}$, measures how responsive the utility's income elasticity is to changes in household income. It can also be written as $\varphi(e_i, \boldsymbol{p}) = (1 - \Phi(e_i, \boldsymbol{p})) + e_i \frac{\partial v^2(e_i,\boldsymbol{p})/\partial e_i^2}{\partial v(e_i,\boldsymbol{p})/\partial e_i}$.

Pass-through. The variety's pass-through, $\chi(e_i, p(\omega), \boldsymbol{p})$, measures the relative strength of the income elasticity of demand and the price elasticity of demand and it is given by

$$\chi(e_i, p(\omega), \boldsymbol{p}) = -\frac{p(\omega)}{e_i} \frac{\partial^2 v(e_i, \boldsymbol{p})/\partial p(\omega)^2}{\partial^2 v(e_i, \boldsymbol{p})/\partial e_i \partial p(\omega)}. \tag{B.4}$$

Proposition B.2 below establishes the relationship between the price elasticity of demand and the income elasticity of demand, and is the fundamental mechanism behind the demand channel underlying markups. Often overlooked and obscured by simplifying assumptions, this relationship has important implications for the rise in markups observed in the data.

PROPOSITION B.2: (PRICE AND INCOME ELASTICITIES OF DEMAND) Given Assumptions B.1 and B.2, the price elasticity of demand of individual i for variety ω is related to their income elasticity of demand through the following expression

$$\underbrace{\xi(e_i, p(\omega), \boldsymbol{p})}_{\text{price elast of demand}} = \underbrace{\alpha(e_i, p(\omega), \boldsymbol{p})}_{\text{fixed effect}} + \underbrace{\chi(e_i, p(\omega), \boldsymbol{p})}_{\text{pass-through}} \left[\underbrace{\eta(e_i, p(\omega), \boldsymbol{p})}_{\text{income elast of demand}} + \underbrace{(\Phi(e_i, \boldsymbol{p}) + \varphi(e_i, \boldsymbol{p}))}_{\substack{\text{income elast of utility} \\ \text{and super-elast}}} - 1 \right],$$

where $\alpha(e_i, p(\omega), \boldsymbol{p}) = p(\omega) \frac{\partial^2 v(e_i, \boldsymbol{p})/\partial e_i \partial p(\omega)}{\partial v(e_i, \boldsymbol{p})/\partial e_i}$ is a variety-specific fixed effect.

PROOF: Rearrange equation (B.2) to have

$$\frac{1}{\partial v(e_i, \boldsymbol{p})/\partial p(\omega)} = \frac{1}{\partial^2 v(e_i, \boldsymbol{p})/\partial p(\omega)^2} \left[\frac{\partial^2 v(e_i, \boldsymbol{p})/\partial e_i \partial p(\omega)}{\partial v(e_i, \boldsymbol{p})/\partial e_i} - \frac{\xi(e_i, p(\omega), \boldsymbol{p})}{p(\omega)} \right]$$

and rewrite the income super-elasticity of utility $\varphi(e_i, \boldsymbol{p})$ as

$$\varphi(e_i, \boldsymbol{p}) = -\left[1 + \Phi(e_i, \boldsymbol{p}) + e_i \frac{\partial^2 v(e_i, \boldsymbol{p})/\partial {e_i}^2}{\partial v(e_i, \boldsymbol{p})/\partial e_i} \right].$$

Next, plug these in equation (B.3) to have the income elasticity of demand as

$$\eta(e_i, p(\omega), \boldsymbol{p}) = 1 + \Phi(e_i, \boldsymbol{p}) + \varphi(e_i, \boldsymbol{p}) + \frac{[\xi(e_i, p(\omega), \boldsymbol{p}) - \alpha(e_i, p(\omega), \boldsymbol{p})]}{\chi(e_i, p(\omega), \boldsymbol{p})}.$$

Rearranging this equation gives the result in the proposition, i.e.

$$\xi(e_i, p(\omega), \boldsymbol{p}) = \alpha(e_i, p(\omega), \boldsymbol{p}) + \chi(e_i, p(\omega), \boldsymbol{p}) \left[\eta(e_i, p(\omega), \boldsymbol{p}) + (\Phi(e_i, \boldsymbol{p}) + \varphi(e_i, \boldsymbol{p})) - 1 \right]$$

If the price elasticity of demand is instead defined as $\xi(e_i, p(\omega), \boldsymbol{p}) = -p(\omega)\frac{\partial^2 v(e_i,\boldsymbol{p})/\partial p(\omega)^2}{\partial v(e_i,\boldsymbol{p})/\partial p(\omega)}$, we have that

$$\alpha(e_i, p(\omega), \boldsymbol{p}) - p(\omega)\frac{\partial^2 v(e_i, \boldsymbol{p})/\partial e_i \partial p(\omega)}{\partial v(\partial p(\omega))/\partial e_i} = 0,$$

which defines the variety-specific fixed effect. In that case, the relationship between the price and income elasticities of demand is simply

$$\xi(e_i, p(\omega), \boldsymbol{p}) = \chi(e_i, p(\omega), \boldsymbol{p})\left[\eta(e_i, p(\omega), \boldsymbol{p}) + (\Phi(e_i, \boldsymbol{p}) + \varphi(e_i, \boldsymbol{p})) - 1\right].$$

Q.E.D.

REMARK 1: It is common to drop the variety-specific fixed effect term, $\alpha(e_i, p(\omega), \boldsymbol{p})$, and define the price elasticity of demand as $\xi(e_i, p(\omega), \boldsymbol{p}) = -p(\omega)\frac{\partial^2 v(e_i,\boldsymbol{p})/\partial p(\omega)^2}{\partial v(e_i,\boldsymbol{p})/\partial p(\omega)}$. In that case, the relationship still holds with only a minor change, i.e. the fixed effect is dropped and

$$\xi(e_i, p(\omega), \boldsymbol{p}) = \chi(e_i, p(\omega), \boldsymbol{p})\left[\eta(e_i, p(\omega), \boldsymbol{p}) + (\Phi(e_i, \boldsymbol{p}) + \varphi(e_i, \boldsymbol{p})) - 1\right]. \tag{B.5}$$

B.2. *Price elasticity of demand: Two key results*

Assumptions B.3 and B.4 provide additional conditions for the price elasticity to vary. Two results then follow. First, the price elasticity of demand must be decreasing in the consumer's income, which sustains Harrod's (1936) *Law of Diminishing Elasticity of Demand* (Proposition B.3). Second, the price elasticity of demand must be increasing in the variety's price in line with Marshall's (1890) *Second Law of Demand* (Proposition B.4).

ASSUMPTION B.3: (INDIRECT UTILITY AND INCOME) The indirect utility $v(e_i, \boldsymbol{p})$ is at least thrice continuously differentiable with $\frac{\partial^3 v(e_i,\boldsymbol{p})/\partial p(\omega)^2 \partial e_i}{\partial^2 v(e_i,\boldsymbol{p})/\partial p(\omega)^2} < \frac{\xi(e_i,p(\omega),\boldsymbol{p})}{\chi(e_i,p(\omega),\boldsymbol{p})\, e_i}$.

ASSUMPTION B.4: (INDIRECT UTILITY AND PRICE) The indirect utility $v(e_i, \boldsymbol{p})$ is at least thrice continuously differentiable with $\frac{\partial^3 v(e_i,\boldsymbol{p})\partial p(\omega)^3}{\partial^2 v(e_i,\boldsymbol{p})/\partial p(\omega)^2} > -\frac{(1+\xi(e_i,p(\omega),\boldsymbol{p}))}{p(\omega)}$.

PROPOSITION B.3: (PRICE ELASTICITY OF DEMAND ACROSS INCOME) Under Assumptions B.1, B.2, and B.3, the price elasticity of demand for a variety ω is decreasing in the consumer's income.

PROOF: The derivative of a consumer's price elasticity of demand for a variety ω with respect to her income is given by

$$\frac{\partial \xi(e_i, p(\omega), \boldsymbol{p})}{\partial e_i} = \xi(e_i, p(\omega), \boldsymbol{p}) \left[\frac{\partial^3 v(e_i, \boldsymbol{p})/\partial p(\omega)^2 \partial e_i}{\partial^2 v(e_i, \boldsymbol{p})/\partial p(\omega)^2} - \frac{\partial^2 v(e_i, \boldsymbol{p})/\partial p(\omega) \partial e_i}{\partial v(e_i, \boldsymbol{p})/\partial p(\omega)} \right].$$

The expression in the square brackets must be negative for the price elasticity of demand for a variety ω to be decreasing in the consumer's income, e_i. Recall that under Assumption B.2, the following holds: (i) $\frac{\partial v(e_i, \boldsymbol{p})}{\partial p(\omega)} < 0$ for all $p(\omega)$; (ii) $\frac{\partial^2 v(e_i, \boldsymbol{p})}{\partial e_i \partial p(\omega)} \neq 0$ for all $p(\omega)$; and (iii) $\frac{\partial^2 v(e_i, \boldsymbol{p})}{\partial p(\omega)^2} > 0$. Using the definition of the pass-through (equation (B.4)) and rearranging the term in the square brackets implies the result in the proposition, i.e.,

$$\frac{\partial^3 v(e_i, \boldsymbol{p})/\partial p(\omega)^2 \partial e_i}{\partial^2 v(e_i, \boldsymbol{p})/\partial p(\omega)^2} < \frac{\xi(e_i, p(\omega), \boldsymbol{p})}{\chi(e_i, p(\omega), \boldsymbol{p}) e_i}.$$

Q.E.D.

PROPOSITION B.4: (PRICE ELASTICITY OF DEMAND ACROSS PRICE) Under Assumptions B.1, B.2, and B.4, the price elasticity of demand for a variety ω is increasing in its price.

PROOF: The derivative of a consumer's price elasticity of demand for a variety ω with respect to its price is given by

$$\frac{\partial \xi(e_i, p(\omega), \boldsymbol{p})}{\partial p(\omega)} = \frac{\xi(e_i, p(\omega), \boldsymbol{p})}{p(\omega)} \left[1 + p(\omega) \frac{\partial^3 v(e_i, \boldsymbol{p})/\partial p(\omega)^3}{\partial^2 v(e_i, \boldsymbol{p})/\partial p(\omega)^2} + \xi(e_i, p(\omega), \boldsymbol{p}) \right].$$

The expression in the square brackets must be positive for the price elasticity of demand for a variety ω to be increasing in its price, $p(\omega)$. Recall that under Assumption B.2, $\frac{\partial^2 v(e_i, \boldsymbol{p})}{\partial p(\omega)^2} > 0$. Rearranging the term in the square brackets implies the result in the proposition, i.e.,

$$\frac{\partial^3 v(e_i, \boldsymbol{p})/\partial p(\omega)^3}{\partial^2 v(e_i, \boldsymbol{p})/\partial p(\omega)^2} < -\frac{(1 + \xi(e_i, p(\omega), \boldsymbol{p}))}{p(\omega)}.$$

Q.E.D.

Given Propositions B.3 and B.4, firms' markups will be higher the lower the price of the variety they sell and/or the wealthier their consumers are. This has implications both for the cross-sectional distribution of markups and for the distribution of markups over time. In addition, if Proposition B.3 holds and households are heterogeneous in terms of income, then changes in the composition of demand also have an effect on markups—even if each household's price elasticity of demand does not change.

C: Model Extensions and Robustness Checks

C.1. *Alternative preferences*

It's worth explaining why alternative preferences are not suitable to study this problem. First, Kimball and non-homothetic CES preferences do not deliver both an increase in the services share and sectoral markups. To be precise, when a discrete number of commodities (e.g. goods and services) with many varieties within them are aggregated, they imply that either a consumer spends a constant share of income on services or that markups are constant. To break that result, these preferences have to be nested, for instance, within a Stone-Geary utility function. Under some restrictions, these preferences allow the price elasticity of demand to be falling in the consumer's income and increasing in the variety's price. These issues are now explored.

Kimball. Kimball (1995) preferences have been extensively used to introduce markups that vary endogenously across firms. Although these preferences allow markups to vary over time, they do not allow the services share to increase as incomes rise. To see this, let a consumer's direct utility be represented by a Cobb-Douglas function of a goods bundle, C_G, and a services bundle, C_S. The consumer's problem is then to solve the following

$$\max_{\{c_{G_t}(\omega)\},\{c_{S_t}(\omega)\},C_{G_t},C_{S_t}} \quad C_{G_t}^{\lambda} C_{S_t}^{1-\lambda}, \tag{C.1}$$

subject to the budget constraint (3.4). Here the sector-specific consumption bundle is implicitly defined by the Kimball aggregator $\Upsilon_j(\cdot)$ for $j=\{G,S\}$ according to

$$\int_{\mathcal{N}_{j_t}} \Upsilon_j\left(\frac{c_{j_t}(\omega)}{C_{j_t}}\right) \mathrm{d}\omega = 1,$$

where $\Upsilon_j(\cdot)$ satisfies the constraints $\Upsilon_j(1)=1$, $\Upsilon_j'(\cdot)>0$, and $\Upsilon_j''(\cdot)<0$. The solution to the consumer's problem delivers a demand for a variety of commodity j given by

$$c_{j_t}(\omega) = \Psi_j\left(\frac{p_{j_t}(\omega)D_{j_t}}{P_{j_t}}\right) C_{j_t},$$

where D_{j_t} is a sector-specific demand index defined as $D_{j_t} = \int_{\mathcal{N}_{j_t}} \Upsilon_j'\left(\frac{c_{j_t}(\omega)}{C_{j_t}}\right)\frac{c_j(\omega)}{C_j}\,\mathrm{d}\omega$, the sector-specific price indices are $P_{G_t} = \frac{\lambda e_t}{C_{G_t}}$ and $P_{S_t} = \frac{(1-\lambda)e_t}{C_{S_t}}$, and $\Psi_j(\cdot) \equiv \Upsilon_j'^{-1}(\cdot)$ is the inverse of the derivative of the Kimball aggregator $\Upsilon_j(\cdot)$ such that $\Psi_j(\cdot)>0$ and $\Psi_j'(\cdot)<0$.[35]

[35]Following Klenow and Willis (2016), the Kimball aggregator is defined as $\Upsilon(x;\phi,\gamma) = 1 + (\gamma - 1)\ \mathrm{e}^{1/\phi}\phi^{\,\gamma/\phi-1}\left[\Gamma\left(\frac{\gamma}{\phi},\frac{1}{\phi}\right) - \Gamma\left(\frac{\gamma}{\phi},\frac{x^{\phi/\gamma}}{\phi}\right)\right]$, where $\Gamma(u,z)$ is the incomplete gamma function $\Gamma(u,z) = \int_z^\infty s^{u-1}\mathrm{e}^{-s}\mathrm{d}s$, γ is the steady-state elasticity of substitution across varieties and ϕ is a super-elasticity that controls

In turn, a consumer's price elasticity of demand is given by

$$\xi_{j_t}(\omega) = -\frac{\frac{p_{j_t}(\omega)D_{j_t}}{P_{j_t}}\Psi_j'\left(\frac{p_{j_t}(\omega)D_{j_t}}{P_{j_t}}\right)}{\Psi_j\left(\frac{p_{j_t}(\omega)D_{j_t}}{P_{j_t}}\right)}.$$

Under the Klenow and Willis (2016) specification, this expression simplifies to

$$\xi_{j_t}(\omega) = \frac{\gamma}{1+\phi_j \ln\left(\frac{\gamma-1}{\gamma}\frac{P_{j_t}}{p_{j_t}(\omega)D_{j_t}}\right)}.$$

As before, a consumer's price elasticity of demand is increasing in the price of the variety, i.e. $\frac{\partial \xi_{j_t}(\omega)}{\partial p_{j_t}(\omega)} > 0$.[36] This implies that firms can charge higher markups if they sell their products at lower prices. Note, however, that these preferences annihilate any role for changes in demand, in particular through rising incomes or shifts in demand shares. As a consequence, market power arises solely from the supply side as shifts in productivity trigger changes in prices. Finally, these preferences imply that services spending is a constant fraction of total income over time, with the income share of services spending given by

$$\int_{\mathcal{N}_{S_t}} p_{S_t}(\omega)c_{S_t}(\omega)\mathrm{d}\omega = (1-\lambda)e_t.$$

Non-homothetic CES. Another popular utility function is the one proposed by Comin, Lashkari, and Mestieri (2021). We can write the consumer's problem as the one described above (equation (C.1)), but replace how the sector-specific consumption bundle is defined. In this particular case, let C_j be given by

$$\int_{\mathcal{N}_{j_t}} \left(\frac{c_{j_t}(\omega)}{C_{j_t}^{\phi_j}}\right)^{\frac{\gamma-1}{\gamma}} \mathrm{d}\omega = 1,$$

the strength of the strategic complementarities between varieties. For $\phi \to 0$, the Kimball aggregator reduces to the constant elasticity of substitution (CES) aggregator with $\Upsilon(x) = x^{\frac{\gamma-1}{\gamma}}$.

[36]To be more precise, the price elasticity of demand is increasing in a variety's price as long as $\xi_j(\omega) > -1 - \left(\frac{p_j(\omega)D_j}{P_j}\right)\frac{\Psi_j''\left(\frac{p_j(\omega)D_j}{P_j}\right)}{\Psi_j'\left(\frac{p_j(\omega)D_j}{P_j}\right)}$.

where γ is the price elasticity of demand and ϕ_j controls the income elasticity of demand. The demand for a particular variety of commodity j is defined by

$$c_{j_t}(\omega) = \left(\frac{p_{j_t}(\omega)}{D_j e_t}\right)^{-\gamma} C_{j_t}^{\phi_j(1-\gamma)},$$

where D_j is a time-invariant sector-specific demand index such that $D_G = \frac{\lambda/\phi_G}{\lambda/\phi_G+(1-\lambda)/\phi_S}$ and $D_S = \frac{(1-\lambda)/\phi_S}{\lambda/\phi_G+(1-\lambda)/\phi_S}$. Now the price elasticity of demand is constant over time and the same for both goods and services (i.e., given by γ for both varieties of goods and services).

Stone-Geary. Stone-Geary preferences are particularly popular in the structural transformation literature because they generate non-unitary income elasticities of demand. These are characterized by a subsistence point in the direct utility and can be easily combined with (homothetic or non-homothetic) CES. Now define the consumer's problem as above (equation (C.1)), but let each consumption bundle be explicitly defined by

$$C_{j_t} = \left[\int_{\mathcal{N}_{j_t}} \left(c_{j_t}(\omega) + \bar{c}_j\right)^{\frac{\gamma-1}{\gamma}} \mathrm{d}\omega\right]^{\frac{\gamma}{\gamma-1}},$$

where $\bar{c}_j > 0$ is a sector-specific subsistence point. For the price elasticity of demand to vary for both goods and services, we need both $\bar{c}_G$ and $\bar{c}_S$ to be different from zero. In addition, for the price elasticity of demand to satisfy the properties defined in Section B in the Online Appendix B, we must impose that $\bar{c}_j$ is positive (more on this below). A consumer's demand for a variety of commodity j is given by

$$c_{j_t}(\omega) = \left(\frac{p_{j_t}(\omega)}{P_{j_t}}\right)^{-\gamma} C_{j_t} - \bar{c}_j,$$

where P_{j_t} is a sectoral price index such that

$$P_{j_t} C_{j_t} = \lambda_j \left[e_t + \bar{c}_G \int_{\mathcal{N}_{G_t}} p_{G_t}(\omega)\mathrm{d}\omega + \bar{c}_S \int_{\mathcal{N}_{S_t}} p_{S_t}(\omega)\mathrm{d}\omega\right],$$

where $\lambda_j = \lambda$ for goods and $\lambda_j = 1-\lambda$ for services. It is easy to see that the spending share of services can increase if the income share of the goods and services subsistence baskets, $\left[\bar{c}_G \int_{\mathcal{N}_{G_t}} p_{G_t}(\omega)\mathrm{d}\omega + \bar{c}_S \int_{\mathcal{N}_{S_t}} p_{S_t}(\omega)\mathrm{d}\omega\right]/e_t$, rises over time.

The price elasticity of demand can now vary both as a result of changes in price and income, as was the case in the baseline economy. A consumer's price elasticity of demand for a variety of commodity j can in turn be written as

$$\xi_{j_t}(\omega) = \gamma \frac{\left(\frac{p_{j_t}(\omega)}{P_{j_t}}\right)^{1-\gamma} P_{j_t} C_{j_t}}{\left(\frac{p_{j_t}(\omega)}{P_{j_t}}\right)^{1-\gamma} P_{j_t} C_{j_t} - p_{j_t}(\omega)\bar{c}_j}.$$

How does the price elasticity of demand varies with price and income? The super-elasticity of demand with respect to price is now given by

$$\frac{\partial \xi_{j_t}(\omega)}{\partial p_{j_t}(\omega)} \frac{p_{j_t}(\omega)}{\xi_{j_t}(\omega)} = \xi_{j_t}(\omega) - \gamma$$

and the super-elasticity of demand with respect to income is

$$\frac{\partial \xi_{j_t}(\omega)}{\partial e_t} \frac{e_t}{\xi_{j_t}(\omega)} = \frac{\lambda_j}{P_{j_t} C_{j_t}} e_t \left[1 - \frac{\xi_{j_t}(\omega)}{\gamma}\right],$$

where $\lambda_j = \lambda$ for goods and $\lambda_j = 1 - \lambda$ for services. Hence, a consumer's price elasticity of demand is increasing in the price of the variety and falling in income as long as $\xi_{j_t}(\omega) > \gamma$. For that condition to hold, it must be that $\bar{c}_j > 0$. Contrast these expressions with the ones derived in the baseline model (footnote 16). Although the super-elasticity of demand with respect to price is only a function of the level of the elasticity itself, the super-elasticity with respect to income now depends on both the income of the consumer and the elasticity itself.

C.2. *Dynamic model*

Assume now that households can save a fraction of their income in exchange for a return rate R_t. Discounting the future at rate $1/\beta$, a consumer has a lifetime indirect utility given by

$$\sum_{t=0}^{\infty} \beta^t \, v(e_t, \boldsymbol{p}_{G_t}, \boldsymbol{p}_{S_t}, \boldsymbol{q}_{G_t}, \boldsymbol{q}_{S_t}), \tag{C.2}$$

where $v(e_t, \boldsymbol{p}_{G_t}, \boldsymbol{p}_{S_t}, \boldsymbol{q}_{G_t}, \boldsymbol{q}_{S_t})$ is defined as in equations (3.1), (3.2), and (3.3). The budget constraint the consumer now faces is given by

$$e_t + a_{t+1} = w_t + R_t a_t + \Lambda_t, \tag{C.3}$$

where a_t is the amount of wealth owned by the household. A consumer now maximizes equation (C.2) subject to (C.3). The static problem of deciding how much to spend on goods and services is the same as in Section 3. Hence, the spending share on services is still the same and the price and quality elasticities of demand are still given by equations (3.6) and (3.7), respectively. The consumer's optimal savings decision is in turn given by the following Euler equation

$$\left(\frac{e_{t+1}}{e_t}\right)^{2+\gamma} = \beta R_{t+1}\left(\frac{A_{t+1}}{A_t}\right),$$

where $A_t = \lambda \int_0^{N_{G_t}} [\phi_G e_t - p_{G_t}(\omega)]^{\gamma}\, p_{G_t}(\omega)\, q_{G_t}(\omega)^{\delta(1+\gamma)}\, \mathrm{d}\omega + (1-\lambda)\int_0^{N_{S_t}} [\phi_S e_t - p_{S_t}(\omega)]^{\gamma}$ $\times\, p_{S_t}(\omega)\, q_{S_t}(\omega)^{\delta(1+\gamma)}\, \mathrm{d}\omega$.[37]

Firms now produce a variety of commodity $j \in \{G, S\}$ using capital and labor according to the following constant returns to scale technology

$$y_{j_t} = z_{j_t} k_{j_t}^{\alpha} n_{j_t}^{\theta} i_{j_t}^{1-\alpha-\theta}.$$

Capital is mobile across sectors and in order to use it firms have to pay the rental rate r_t. A firm's marginal cost can now be written as

$$mc_{j_t} = \frac{1}{z_{j_t}}\left(\frac{r_t}{\alpha}\right)^{\alpha}\left(\frac{w_t}{\theta}\right)^{\theta}\left(\frac{p_{I_t}}{1-\alpha-\theta}\right)^{1-\alpha-\theta}.$$

Introducing capital does not alter the optimal decision of a firm with respect to the price and quality of its variety. The firm's markup is still given by equation (3.12). The resulting markup in the model with monopolistic competition is now given by

$$m_{j_t} = \frac{\gamma}{\gamma+1} + \underbrace{\frac{z_{j_t}}{\left(\frac{r_t}{\alpha}\right)^{\alpha}\left(\frac{w_t}{\theta}\right)^{\theta}\left(\frac{p_{I_t}}{1-\alpha-\theta}\right)^{1-\alpha-\theta}}}_{\text{supply effect}} \times \underbrace{\frac{\phi_j e_t}{\gamma+1}}_{\text{demand effect}}.$$

Finally, the asset market clearing condition requires that all households' savings, A_t, equate the total capital demanded by firms, $K_t \equiv \left[\int_0^{N_{G_t}} k_{G_t}(\omega)\, d\omega + \int_0^{N_{S_t}} k_{S_t}(\omega)\, d\omega + k_{I_t}\right]$. This equilibrium condition determines the rental rate of capital, where $R_t = 1 + r_t - \delta$ and δ is the capital depreciation rate.[38]

---

[37] We can derive an *approximate* intertemporal elasticity of substitution of expenditure as $\frac{\partial(e_{t+1}/e_t)}{\partial R_{t+1}} \frac{R_{t+1}}{(e_{t+1}/e_t)} \approx \frac{1}{2+\gamma}$. As the interest rate rises, the growth rate of total consumption spending is approximately equal to $1/(2+\gamma)$.

[38] In the model with oligopolistic competition, the firm's markup is $m_{j_t} = \frac{\gamma}{\gamma+s_{j_t}} + \frac{z_{j_t}}{\left(\frac{r_t}{\alpha}\right)^{\alpha}\left(\frac{w_t}{\theta}\right)^{\theta}\left(\frac{p_{I_t}}{1-\alpha-\theta}\right)^{1-\alpha-\theta}} \times \phi_j e_t \times \frac{s_{j_t}}{\gamma+s_{j_t}}$ and the aggregate capital demand is $K_t \equiv \left[\sum_{\omega_G=1}^{N_{G_t}} k_{\omega_{G_t}} + \sum_{\omega_S=1}^{N_{S_t}} k_{\omega_{S_t}} + k_{I_t}\right]$.

C.3. *Model with heterogeneous skills*

The average price elasticity of demand for a firm is the weighted average of each consumer's own price elasticity of demand, ξ_{i,j_t}, given by

$$\overline{\xi}_{j_t} = \omega_{H,j_t}\xi_{H,j_t} + (1-\omega_{H,j_t})\xi_{L,j_t},$$

where ω_{i,j_t} is the consumer's demand share in the firm's sales, with the consumption share of high-skilled consumers being $\omega_{H,j_t} = \mu_t c_{H,j_t}/y_{j_t}$, and that of low-skilled consumers $\omega_{L,j_t} = (1-\mu_t)c_{L,j_t}/y_{j_t}$. Using each consumer's price elasticity of demand and the firm's marginal cost, the firm's markup can now be written as

$$m_{j_t} = \frac{\gamma\left[\omega_{H,j_t} + (1-\omega_{H,j_t})\widehat{\phi}_{j_t}\right]}{\gamma\left[\omega_{H,j_t} + (1-\omega_{H,j_t})\widehat{\phi}_{j_t}\right]+1} + \underbrace{\frac{z_{j_t}}{\left\{\frac{w_{L_t}}{\theta}\left[(\alpha x_t)^{\frac{1}{1-\iota}}\left(\frac{w_{H_t}}{w_{L_t}}\right)^{\frac{\iota}{\iota-1}} + (1-\alpha)^{\frac{1}{1-\iota}}\right]^{\frac{\iota-1}{\iota}}\right\}^{\theta}\left(\frac{p_{I_t}}{1-\theta}\right)^{1-\theta}}}_{\text{supply effect}} \times \underbrace{\frac{\phi_j e_{H_t}}{\gamma\left[\omega_{H,j_t} + (1-\omega_{H,j_t})\widehat{\phi}_{j_t}\right]+1}}_{\text{demand effect}},$$

where $\widehat{\phi}_{j_t} = (\phi_j e_{H,j_t} - p_{j_t})/(\phi_j e_{L,j_t} - p_{j_t})$ denotes the relative willingness to pay for a variety for a high-skilled consumer.

Now the firm's markup also depends on the skill premium and the composition of its customers. An increase in the skill premium, raises the marginal cost that in turn reduces the firm's markup. Alternatively, an increase of the share of high-skilled consumers in the economy raises the probability the firm meets a wealthier shopper. As these consumers have a lower price elasticity of demand, the firm can now increase its markup.

C.4. *Parameter values and model fit when confronted to 1955 and 2020 data*

Table C.1 presents the parameter values for the models with monopolistic and oligopolistic competition when estimated to match 1955 and 2020 data. Table C.2 presents the model's fit.

TABLE C.1

PARAMETER VALUES

Parameter	Description	Mon. comp.	Cournot	Identification
Preferences				
λ	Indirect utility's weight on goods	0.369	0.073	Services variable cost share
γ	Exponent in indirect subutility	2.701	1.565	Services variable cost share
ϕ_G	Choke price of goods	9.813	5.934	Average goods markups
ϕ_S	Choke price of services	10.128	15.424	Average services markups
δ	Exponent related with quality	0.170	0.223	Normalization ($\overline{q}_{G_{1955}} = 1$)
Technology				
θ	Exponent on labor	0.810	0.810	Labor share
z_{G_t}	TFP in goods sector in 1955, 2020	0.402, 0.775	0.385, 0.734	Labor share, aggregate markup
z_{I_t}	TFP in intermediate input sector in 1955, 2020	0.129, 0.166	0.151, 0.181	Normalization ($\overline{p}_{I_t} = \overline{p}_{G_t}$)
z_{S_t}	TFP in services sector in 1955, 2020	0.368, 0.352	0.500, 0.391	Relative price of services
Costs				
κ	Linear term related with quality	0.032	0.016	Quality expenses/sales in services
ϑ	Exponent related with quality	2.000	2.000	Exogenous
f_{G_t}	Entry costs in goods sector in 1955, 2020	0.070, 0.142	0.041, 0.113	Number of goods firms
f_{S_t}	Entry costs in services sector in 1955, 2020	0.036, 0.048	0.047, 0.057	Number of service firms

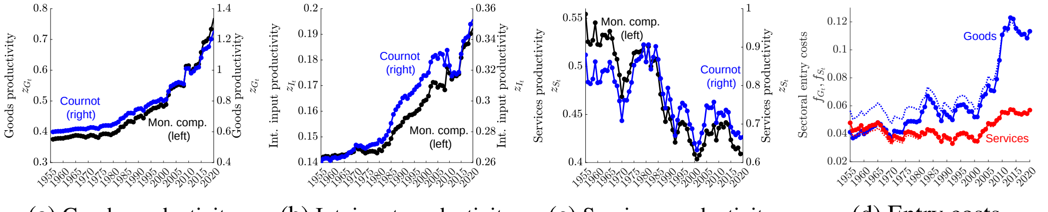
TABLE C.2

TARGETED MOMENTS: DATA VS. MODEL

Moment	Description	Model 1955, 2020	Data 1955, 2020	Source
$\omega^{\text{costs}}_{S_t}$	Services variable cost shares *(monopolistic comp.)*	0.553, 0.704	0.353, 0.704	BEA
	(Cournot)	0.353, 0.704	0.353, 0.704	BEA
$\overline{p}_{S_t}/\overline{p}_{GI_t}$	Relative price of services *(monopolistic comp.)*	1.067, 1.449	1.067, 1.450	BEA
	(Cournot)	1.067, 1.450	1.067, 1.450	BEA
w_t/PY_t	Labor share	0.682, 0.607	0.681, 0.607	BLS
M_t	Aggregate markups	1.385	1.385	BEA, Compustat
$\overline{m}_{GI_t}$	Average non-services markups	1.369	1.369	Compustat
$\overline{m}_{S_t}$	Average services markups	1.391	1.391	Compustat
$\kappa Q_{S_t}/PY_{S_t}$	Sales share of quality expenses in services	0.080	0.080	Compustat
N_{G_t}	Number of non-services firms	1.000, 0.818	1.000, 0.818	BDS, BEA
N_{S_t}/N_{G_t}	Relative number of services firms	3.308, 5.031	3.308, 5.031	BDS, BEA

C.5. *Additional figures from the baseline simulation*

FIGURE C.1.—Model-implied trends in productivity and entry costs, 1955-2020



(a) Goods productivity (b) Int. input productivity (c) Services productivity (d) Entry costs

Note: Panel (a) shows the evolution of productivity in the final consumption goods sector, z_{G_t}, in the models with monopolistic competition (black) and oligopolistic competition (blue). Panel (b) shows the evolution of productivity in the intermediate input sector, z_{I_t}. Panel (c) shows the evolution of productivity in the services sector, z_{S_t}. Panel (d) shows the evolution of entry costs in the goods (blue; f_{G_t}) and services (red; f_{S_t}) sectors in the models with monopolistic competition (solid line) and oligopolistic competition (dotted line).

C.6. *Quantitative analysis of the model with CES demand*

Calibration and simulation. The model with CES demand sets the choke price parameters, ϕ_G and ϕ_S, to zero. The price elasticity of demand is now allowed to vary across sectors with $(-\gamma_G) \neq (-\gamma_S) > 0$. As preferences are now homothetic, the sectoral shares, the labor share of output, and markups are constant over time. Table C.3 presents the parameter values for the model with monopolistic competition when consumers' demand is given by CES and Table C.4 contrasts the model's fit with the data. The model gets close to the 2020 targets, but it is unable to generate a change in markups, the labor share, or sectoral shares. Simulating the model over time delivers a constant time path these aggegates.

TABLE C.3

PARAMETER VALUES

Parameter	Description	Mon. comp.	Identification
Preferences			
λ	Indirect utility's weight on goods	0.300	Services variable cost share
γ_G	Price elast. of demand for goods	-2.377	Average goods markups
γ_S	Price elast. of demand for services	-3.466	Average services markups
δ	Exponent related with quality	-0.226	Normalization ($\overline{q}_{G_{1955}} = 1$)
Technology			
θ	Exponent on labor	0.810	Labor share
z_{G_t}	TFP in goods sector in 1980, 2020	2.806, 2.806	Labor share, aggregate markup
z_{I_t}	TFP in intermediate input sector in 1980, 2020	1.000, 1.000	Normalization ($\overline{p}_{I_t} = \overline{p}_{G_t}$)
z_{S_t}	TFP in services sector in 1980, 2020	2.402, 1.575	Relative price of services
Costs			
κ	Linear term related with quality	0.018	Quality expenses/sales in services
ϑ	Exponent related with quality	2.000	Exogenous
f_{G_t}	Entry costs in goods sector in 1955, 2020	0.100, 0.122	Number of goods firms
f_{S_t}	Entry costs in services sector in 1955, 2020	0.060, 0.059	Number of service firms

TABLE C.4

TARGETED MOMENTS: DATA VS. MODEL

Moment	Description	Model 1980, 2020	Data 1980, 2020	Source
$\omega^{\text{costs}}_{S_t}$	Services variable cost shares	0.704, 0.704	0.417, 0.704	BEA
$\overline{p}_{S_t}/\overline{p}_{GI_t}$	Relative price of services	0.951, 1.446	0.951, 1.450	BEA
w_t/PY_t	Labor share	0.607, 0.607	0.678, 0.607	BLS
M_t	Aggregate markups	1.385	1.385	BEA, Compustat
$\overline{m}_{GI_t}$	Average non-services markups	1.336	1.369	Compustat
$\overline{m}_{S_t}$	Average services markups	1.405	1.391	Compustat
$\kappa Q_{S_t}/PY_{S_t}$	Sales share of quality expenses in services	0.080	0.080	Compustat
N_{G_t}	Number of non-services firms	1.001, 0.818	1.001, 0.818	BDS, BEA
N_{S_t}/N_{G_t}	Relative number of services firms	4.054, 5.031	4.054, 5.031	BDS, BEA

C.7. *Quantitative analysis of the model with heterogeneous consumers*

Data, calibration, and simulation. The model with heterogeneous households features eight new parameters, $\{\alpha, \iota, x_t, \phi_{\Lambda_t}, \mu_t\}$, for t in 1980 and 2020 that have to be calibrated. Acemoglu and Autor (2011), Buera, Kaboski, Rogerson, and Vizcaino (2022), and Katz and Murphy (1992) estimate the elasticity of substitution between high and low-skilled labor for different periods and find values ranging from -2.9 to -1.4, which corresponds to a value of $\iota \in [0.291, 0.661]$. This range is consistent with skill-biased technological change decreasing marginal costs as high and low-skilled labor are substitutes. A value of $\iota = 0.4$ is chosen. Skill-biased productivity is normalized to 1 in 1980, i.e., $x_{1980} = 1$. The high-skilled labor parameter α and skill-biased productivity in 2020, x_{2020}, help match the skill premium, w_{H_t}/w_{L_t}, in 1980 and 2020. The skill premium corresponds to the ratio of the median income of individuals with at least a bachelor's degree vs. less than a four-year college degree taken from the Census' American Community Survey (ACS). The share of aggregate expenses with quality and entry costs rebated to high-skilled consumers, $\phi_{\Lambda_t} \equiv \mu_t \Lambda_{H_t}/\Lambda_t$, is used to match the relative total earnings of individuals with at least a bachelor's degree from the ACS, or e_{H_t}/e_{L_t} in the model. Finally, the share of high-skilled households, μ_t, is measured directly from the data as the share of individuals with at least a bachelor's degree, also taken from the ACS.

Table C.5 presents the parameter values for the models with monopolistic and oligopolistic competition with heterogeneous consumers. The parameter values are close to the ones estimated in Section 4, with the exception of productivity. Technological progress is now skill-biased. Table C.6 contrasts the model's fit with the data. The model with oligopolistic competition matches all targets perfectly, while the model with monopolistic competition slightly overstates the services share in 1980 and understates the relative price of services in 2020.

TABLE C.5

PARAMETER VALUES

Parameter	Description	Mon. comp.	Cournot	Identification
Preferences				
λ	Indirect utility's weight on goods	0.266	0.086	Services variable cost share
γ	Exponent in indirect subutility	2.526	2.194	Services variable cost share
ϕ_G	Choke price of goods	6.875	5.212	Average goods markups
ϕ_S	Choke price of services	8.410	13.430	Average services markups
δ	Exponent related with quality	0.163	0.179	Normalization ($\overline{q}_{G_{1955}} = 1$)
Technology				
θ	Exponent on labor	0.810	0.810	Labor share
z_{G_t}	TFP in goods sector in 1980, 2020	0.726, 0.685	0.882, 0.784	Labor share, aggregate markup
z_{I_t}	TFP in intermediate input sector in 1980, 2020	0.315, 0.164	0.403, 0.197	Normalization ($\overline{p}_{I_t} = \overline{p}_{G_t}$)
z_{S_t}	TFP in services sector in 1980, 2020	0.955, 0.361	1.178, 0.418	Relative price of services
α	High-skilled weight in labor	0.403	0.403	Skill premium
ι	Elasticity of sub. between high and low-skilled	0.400	0.400	Exogenous
x_t	Skill-biased prod. in 1980, 2020	1.000, 2.286	1.000, 2.286	Normalization, skill premium
Costs				
κ	Linear term related with quality	0.022	0.021	Quality expenses/sales in services
ϑ	Exponent related with quality	2.000	2.000	Exogenous
f_{G_t}	Entry costs in goods sector in 1955, 2020	0.054, 0.158	0.053, 0.154	Number of goods firms
f_{S_t}	Entry costs in services sector in 1955, 2020	0.042, 0.077	0.042, 0.078	Number of service firms
Other				
μ_t	Share of high-skilled in 1980, 2020	0.201, 0.423	0.201, 0.423	College-educated share
ϕ_{Λ_t}	Share of profits to high-skilled in 1980, 2020	0.276, 0.552	0.276, 0.552	Relative earnings of high-skilled

The simulation over time follows the strategy outlined in Section 4. I now also have to specify paths for $\{x_t, \mu_t, \phi_{\Lambda_t}\}$ between 1955 and 2020. These variables target the skill premium, the share of college-educated individuals, and their total earnings relative to non-college-educated individuals over time (i.e., the targets used in the calibration). Since the ACS is available every

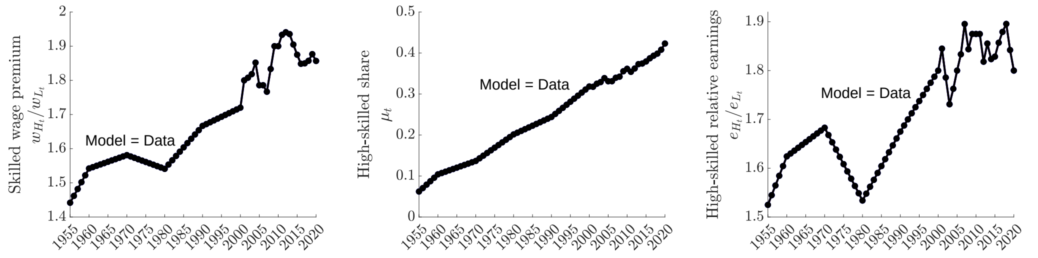
TABLE C.6

TARGETED MOMENTS: DATA VS. MODEL

Moment	Description	Model 1980, 2020	Data 1980, 2020	Source
$\omega_{S_t}^{\text{costs}}$	Services variable cost shares *(monopolistic comp.)*	0.426, 0.704	0.417, 0.704	BEA
	(Cournot)	0.417, 0.704	0.417, 0.704	BEA
$\overline{p}_{S_t}/\overline{p}_{GI_t}$	Relative price of services *(monopolistic comp.)*	0.951, 1.446	0.951, 1.450	BEA
	(Cournot)	0.951, 1.450	0.951, 1.450	BEA
w_t/PY_t	Labor share	0.678, 0.607	0.678, 0.607	BLS
M_t	Aggregate markups	1.385	1.385	BEA, Compustat
$\overline{m}_{GI_t}$	Average non-services markups	1.369	1.369	Compustat
$\overline{m}_{S_t}$	Average services markups	1.391	1.391	Compustat
$\kappa Q_{S_t}/PY_{S_t}$	Sales share of quality expenses in services	0.080	0.080	Compustat
N_{G_t}	Number of non-services firms	1.001, 0.818	1.001, 0.818	BDS, BEA
N_{S_t}/N_{G_t}	Relative number of services firms	4.054, 5.031	4.054, 5.031	BDS, BEA
w_{H_t}/w_{L_t}	Skilled wage premium	1.541, 1.857	1.541, 1.857	ACS
e_{H_t}/e_{L_t}	High-skilled earnings share	1.534, 1.800	1.534, 1.800	ACS

decade prior to 2000, these time series are interpolated within each decade to get continuous time series. In addition to the paths illustrated in Figure 4.1, this model also matches these three time series perfectly (see Figure C.2).

FIGURE C.2.—Matched trends related to income inequality, 1955-2020



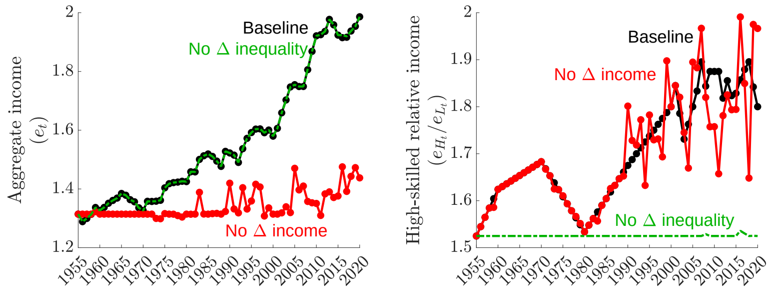
(a) Skilled wage premium (b) High-skilled share (c) Skilled relative earnings

Note: Panel (a) shows the skilled wage premium (w_{H_t}/w_{L_t}) in the data and baseline simulations. Panel (b) shows the share of high-skilled households in the economy (μ_t). Panel (c) shows the earnings of high-skilled households over low-skilled households (e_{H_t}/e_{L_t}). The trends are the same for the data and models with monopolistic and oligopolistic competition.

Experiments. In addition to the counterfactual economy in which entry costs, f_{G_t} and f_{S_t}, are held constant over time at their 1955 values, two additional economies are simulated. One in which the aggregate household income is constant and income inequality is as in the baseline; and another in which income inequality is constant and the aggregate income is as in the baseline. These are computed by finding the values of skill-biased productivity and high-skilled households' transfers share, x_t and ϕ_{Λ_t}, that minimize the distance between the aggregate income and relative income of the counterfactual economy, e_t^{exp} and $e_{H_t}^{\text{exp}}/e_{L_t}^{\text{exp}}$, and their corresponding baseline values, e_t^{baseline} and $e_{H_t}^{\text{baseline}}/e_{L_t}^{\text{baseline}}$. Neutral productivities, z_{G_t}, z_{I_t}, and z_{S_t}, entry costs, f_{G_t} and f_{S_t}, and all other parameters are at their baseline values.

Figure C.3 shows the evolution of the aggregate income (panel (a)) and high-skilled relative income (panel (b)) for the baseline and the two counterfactual economies. The economy with no changes in income inequality is precisely estimated, while the algorithm for the economy without changes in the aggregate income has a hard time matching the baseline values in income and income inequality in particular starting in the mid 1990s.

FIGURE C.3.—Aggregate income and high-skilled relative income



(a) Aggregate income (b) High-skilled relative income

Note: Panel (a) shows the aggregate income (e_t) in the baseline economy (black), the economy with constant income inequality (green) and with constant income (red) for the model with heterogeneous consumers and oligopolistic competition. Panel (b) shows the corresponding relative income of high-skilled households (e_{H_t}/e_{L_t}).

C.8. *Cross-country quantitative analysis*

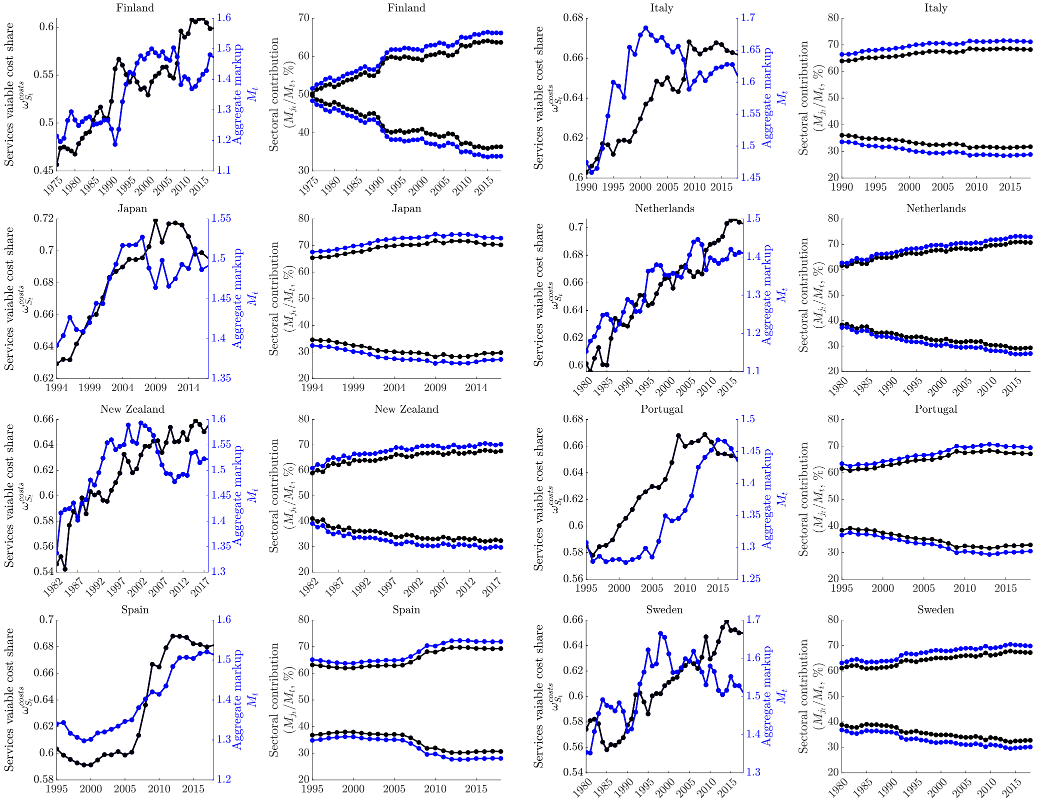
Data and simulation. The cross-country analysis requires data on the labor share and the services share to discipline the evolution of sectoral productivities, z_{G_t} and z_{S_t}.[39] The Penn World Table (PWT, Feenstra and Timmer (2015)) provides historical data on the labor shares of output and the World Bank on the services value added shares. I keep countries for which there is at least 20 years of data for both series and focus on countries for which the services share increased over time. This resulted in a sample of more than 40, mostly advanced, economies. Given the limited availability of historical data on the number of firms per industry to discipline the evolution of entry costs, f_{G_t} and f_{S_t} are kept constant throughout the simulation at the U.S. values in 1955. Hence, only differences in sectoral productivity drive results across countries. All other parameters are set to the U.S. values presented in Table 4.1 and both models with monopolistic and oligopolistic competition were simulated.

Cross-country markups. Figure C.4 shows the evolution of the services share and the aggregate markup for selected countries (first and third columns) as well as the contribution of the services and non-services sectors to the increase in the aggregate markup (second and fourth

[39]Ideally, I would need the services variable cost shares to simulate the model. Since these are not widely available, the services value added shares are used instead. As was the case for the U.S. simulation, the productivity of intermediate input producers is ensures the price of intermediate inputs equates the price of consumption goods.

columns). The simulations predict that most advanced economies that experienced an increase in their services shares also saw their aggregate markup increase.

FIGURE C.4.—Model-implied trends in markups and their sectoral decomposition across countries



Note: The figure shows the evolution of the services value added share (in black) together with the aggregate markup (in blue; first and third panels) as well as the sectoral contribution to the aggregate markup (second and fourth panels) in the models with monopolistic competition (black) and oligopolistic competition (blue).