Journal of Economic Perspectives—Volume 33, Number 3—Summer 2019—Pages 3–22

Are Price-Cost Markups Rising in the United States? A Discussion of the Evidence

Susanto Basu

Has the US economy entered a second Gilded Age? A pattern of increasing industrial concentration combined with rising inequality in income and wealth may seem to indicate as much. Yet industrial concentration can be interpreted as evidence either of increased market power or of greater competition, where more efficient firms are able to gain market share. Thus, economists have sought to estimate a less ambiguous measure of market power in order to see whether firms are able to exert greater control over the market prices of their outputs. The summary measure that has been the focus of much recent research is the markup of price over marginal cost.

Three main methods have been used to estimate markup trends in the US economy. The first method attempts to estimate economic profits using either aggregate or firm-level data and then, together with an assumption of constant returns to scale, generates an estimate of the size of markups. The second approach estimates a production function for various firms or sectors, based on a variety of inputs. Unlike the first approach, this allows for increasing returns to scale and recovers the markup by applying conditions for cost minimization to the estimated coefficients in the production function. The third method again estimates a production function, typically using firm-level data, but this time recovers the markup from the optimization condition for a single input. This approach again allows for increasing returns to scale, but unlike the first and second methods, it avoids the need to

■ *Susanto Basu is Professor of Economics, Boston College, Chestnut Hill, Massachusetts, and Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. His email address is susanto.basu@bc.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.33.3.3

doi=10.1257/jep.33.3.3

estimate the rate of economic profit (or to assume that it is zero). This method estimates the markup as the elasticity of output with respect to the single input, divided by the factor payment to the selected input as a share of the firm's revenue.

I begin with a conceptual overview of these three approaches, which shows how they are connected and provides an intuitive sense of how they allow estimation of markups for the economy as a whole. In the following sections, I examine empirical research that has implemented each of these approaches. I describe in more detail how the research was done and characterize some strengths and weaknesses of all three approaches. I show why estimates of large or steeply rising markups are implausible; for example, several of the prominent estimates suggest that the markup increased far more than would be necessary to explain the decline in labor's share.

The article offers some suggestions for future research on markups. The conclusion, in particular, asks researchers to link microeconomic estimates of markups to aggregate trends in the economy. The recent interest in market power stems in substantial part from the realization that higher markups may tend to depress the demand for factors of production, and thus the prices (incomes) those factors receive. For example, connections have been proposed between higher markups and a lower labor share of income and a lower investment rate. Yet, higher markups should also reduce hiring of workers and, ceteris paribus, raise the inflation rate. Several of the theoretical predictions of what patterns should accompany a substantial rise in markups are not easily verified in recent US data.

Three Methods of Measuring Markups

The markup of price over marginal cost is a basic measure of market power. With perfect competition in the goods market, a profit-maximizing firm will set price equal to marginal cost, and the markup will be equal to one. With imperfect competition, the firm produces at the quantity where marginal revenue equals marginal cost, and price will exceed marginal cost. In seeking to measure markups, an immediate hurdle is how to measure marginal cost—a variable that must be estimated or inferred rather than being directly observed in a market transaction like prices or revenues. Economists specializing in industrial organization have developed ways of estimating markups for particular firms and industries. But the challenge here is to develop measures of average markups for the economy as a whole.

This section describes three theoretical approaches that researchers have used. Those who attempt to estimate markups in a comprehensive manner, for most or all of the US economy, typically use a version of the cost-minimization framework described here.[1] The following three sections then describe in more detail how

[1] De Loecker and Eeckhout (2017) discuss the relationship between this approach and the traditional industrial organization approach to estimating markups using estimates of demand for individual goods or markets.

each approach has been implemented in recent empirical research and discuss some strengths and weaknesses of each approach.

Consider a firm producing output, Y, with a production function, F, that uses capital and labor as inputs, as well as freely available technology, Z: $Y = F(K, L, Z)$. (Researchers using firm- or industry-level data typically add intermediate inputs as a factor of production.) Suppose furthermore that while the firm may have market power in the goods market, it takes the prices of the two factors, the wage, W, and the required return to capital, R, as set in markets outside its control. Then, a profit-maximizing firm will make a cost-minimizing use of labor, which requires that it hire labor until the marginal product of labor equals the markup times the wage:[2]

$$PF_L = \mu W.$$

Here F_L is the marginal product of labor, P is the price of output, and μ is the markup of price over marginal cost. Perfect competition in the goods market corresponds to $\mu = 1$, which yields the familiar condition that an optimizing firm must equate the marginal product of labor to the real wage. However, a firm with market power hires less labor (and thus has a higher marginal product of labor for a given real wage), because it maximizes profit by producing less than the competitive level of output. More market power—a larger markup—corresponds to a lower level of desired output. Naturally, a similar condition applies to the equality between the value marginal product of capital and its rental rate, R, multiplied by μ.[3]

As we will see, the robust cost-minimization conditions alone allow us to *measure* the extent of a firm's market power, even though they do not by themselves answer the interesting question of *why* the firm has market power. Thus, the firm optimality condition above holds regardless of the form of imperfect competition that generates the markup. The firm can be a monopolist or an oligopolist, and it may follow either static or dynamic pricing policies. The firm's optimal choice of the markup is determined by its larger profit-maximization problem, of which cost minimization is only a part. But because we do not need to take a stand on the rest of the firm problem, which can be very complicated, we are able to measure the size of the markup with minimal assumptions.

One conclusion follows from multiplying the condition above by labor input and dividing by output, which has the effect of expressing this relationship in elasticity form:

$$\frac{F_L L}{Y} = \mu \frac{WL}{PY}.$$

[2]The fundamental condition arising from profit maximization is that the firm equates the marginal product of labor valued at marginal cost to the wage. The equation in the text follows from observing that marginal cost by definition equals the ratio of the output price to the markup, which is itself price over marginal cost.

[3]This cost-minimization condition holds even if it is costly for the firm to adjust its capital stock. However, in this case, the rental rate, R, must be redefined to include the marginal adjustment cost of capital and its expected rate of change (for discussion, see Basu, Fernald, and Shapiro 2001).

The left-hand side is the elasticity of output with respect to labor input.[4] The right-hand side is labor's share in revenue, multiplied by the markup. For a given output elasticity—and certainly in the Cobb–Douglas case, where the elasticity is constant—an increase in the markup depresses the share of revenue or national income going to labor.

This implication of higher markups has drawn much interest in recent years, because the labor share of income in the United States has fallen sharply over the decades since 1980. After averaging about 0.64 in pre-1980 data, observations of labor's share have most recently been around 0.58, a sharp decline for what was sometimes called one of the "great ratios" of economic growth. Elsby, Hobijn, and Şahin (2013) attribute around one-third of the measured decline in the share to incorrect measurement of self-employment income, which still leaves an actual decline of four percentage points to be explained.

How much would the markup have to rise to explain this decline in labor's share? If the output elasticity of labor were constant, then the markup would have to increase by a factor of 1.07. That is, if perfect competition ($\mu = 1$) prevailed in 1980, we would require $\mu = 1.07$ now, so price would now be 7 percent higher than marginal cost. As we will see, this implied increase in the markup is modest relative to many of the estimates in the literature. If the markup is to have risen by more, the output elasticity of labor must have *risen* substantially in order to be consistent with the observed change in the labor share.[5] Of course, this back-of-the-envelope exercise attributes all of the decline in the labor share to changing market power. Elsby, Hobijn, and Şahin (2013) attribute much of the change to increased trade competition from globalization, while Karabarbounis and Neiman (2014) attribute it to changes in technology (their explanation implies that the output elasticity of labor should have *fallen* instead of rising).

As this quick calculation demonstrates, inferring changes in markups from changes in observed factor shares requires either an assumption about or, preferably, an estimate of the output elasticity in question. This is the method followed by De Loecker and Eeckhout (2017), which is one of the three approaches to markup estimation discussed at greater length below. They estimate an output elasticity with respect to input econometrically and divide it by the observed revenue share. By the equation above, the ratio gives an estimate of the markup. Repeating this exercise over time, they can also compute a trend in the markup.

The second method I review writes down an equation for efficient use of capital that is parallel to the condition for labor above, and adds the two equations to yield:

$$\frac{F_L L}{Y} + \frac{F_K K}{Y} = \mu\left[\frac{WL}{PY} + \frac{RK}{PY}\right] = \mu\frac{\text{Total cost}}{\text{Revenue}} = \mu(1 - s_\pi).$$

[4] By definition, the elasticity is $\frac{\partial \ln Y}{\partial \ln L}$. Note that $\frac{\partial \ln Y}{\partial \ln L} = \frac{\partial Y/\partial L}{Y/L} = \frac{F_L}{Y/L} = \frac{F_L L}{Y}$.

[5] This observation is due to Brent Neiman.

The left-hand side is the sum of the output elasticities of the production function, which is also the degree of returns to scale. The right-hand side is the markup times the ratio of total cost (including the rental cost of capital) to revenue. Since cost equals revenue minus economic profit, the right-hand side can also be written as the markup times one minus the profit rate s_π, the ratio of profit to revenue.[6]

This equation implies a different method of computing the markup, which is followed by another major strand of the literature discussed below. Suppose one estimates the degree of returns to scale in production, or simply assumes that returns to scale are constant (so the left-hand side equals one). Then one can compute the profit rate and thereby estimate the markup. As the equation above shows, the key to computing the profit rate is to impute a required return to capital, R. Under constant returns and competition, required payments to capital are just total revenue less payments to labor. But with imperfect competition, required payments to capital need to be estimated separately. Once capital payments are estimated and returns to scale are known, one can back out the markup.

This core relationship between markups, economic profit, and economies of scale also clarifies how a firm might have both a markup in excess of one and a near-zero rate of economic profit. This outcome is possible if the firm is producing in the area of increasing returns to scale, where average cost exceeds marginal cost. For example, this would be the case in the classic Chamberlinian model of monopolistic competition, in which long-run profits are zero due to free entry, but firms have market power to set price above marginal cost. Conversely, firms operating with increasing returns to scale—a situation that can arise when marginal costs are low compared with fixed costs—will find that they need to charge a markup above marginal cost to cover their fixed costs, or else they will make losses and go out of business. Of course, in both situations there is a welfare loss arising from the markup even with zero profits, as there is from any wedge (such as a tax) between price and marginal cost. Thus, contrary to suggestions in some papers, the markup is generally a better measure of market power than the profit rate.

Finally, one can derive a third method of estimating the markup by applying the same condition for cost minimization in a different context. This method, due to Hall (1988, 1990), begins by taking a first-order approximation in logs to the production function and taking differences over time of the resulting expression. Letting a lowercase letter represent the natural log of its uppercase counterpart (for example, $y \equiv \ln Y$) and letting a Δ represent a change over time, one gets:

$$\Delta y \simeq \frac{F_L L}{Y}\Delta l + \frac{F_K K}{Y}\Delta k + \Delta z,$$

[6]One can also derive this equation by manipulating the definition of the markup as the ratio of price to marginal cost. Multiply and divide by average cost, then recognize that the ratio of average to marginal cost is the degree of returns to scale for a cost-minimizing firm, while the ratio of price to average cost is also the ratio of revenue to total cost.

where Δz has the interpretation of technical change. Applying the conditions for cost minimization noted above, this equation becomes:

$$\Delta y \simeq \mu\left[\frac{WL}{PY}\Delta l + \frac{RK}{PY}\Delta k\right] + \Delta z.$$

As Hall (1990) emphasized, this approach generalizes Solow's (1957) classic method for calculating the growth rate of technology. If $\mu = 1$, as Solow assumed, then one can obtain a time series for technical change, Δz, as a residual by subtracting share-weighted input growth from output growth. (In Solow's case, required payments to capital are also easily observed as revenue minus labor payments, since there are no profits.) In the case where μ is allowed to exceed one but is unknown, it must be estimated econometrically, using the equation above as an estimating equation, with the unobserved Δz treated as the error term. Since one expects that changes in input usage, the composite right-hand-side variable, would be correlated with the change in technology, Hall uses an instrumental variables technique, where valid instruments must be correlated with input choice (the weighted average of Δk and Δl) but uncorrelated with technical change—loosely speaking, any type of "demand shock." Hall (2018) uses a small modification of this method to estimate markups using recent data.

All three of these methods begin from the assumption that firms minimize costs taking input prices as given. This hypothesis is powerful, but it does not cover all important cases. For example, it assumes that individual firms do not have the power to set wages for their workers. A Council of Economic Advisers (2016) issue brief discussed evidence suggesting, to the contrary, that often firms do indeed have some power to set wages. Qualitatively, the implications of market power discussed above do not change much if firms also have power to set some factor prices. In most cases, such factor market power would also create a wedge between marginal products and factor prices, thus reinforcing the conclusions discussed above for the case of market power in the goods market alone. However, Hall (2018) shows via an insightful example that the quantitative conclusions drawn from applying cost minimization to data to estimate goods-market markups will typically give incorrect results if firms have market power in factor markets as well. Morlacco (2019) presents conditions under which one can reinterpret the evidence for *goods* market power obtained using the method of De Loecker and Eeckhout (2017) noted above as evidence of market power in the *factor* market instead.

Markup Estimates Based on Economic Profits and Constant Returns to Scale

As discussed in the previous section, an assumption of cost minimization makes it possible to derive a relationship between three parameters: returns to scale, the markup, and the rate of economic profit. If returns to scale are assumed to be constant, then calculations of economic profit will allow an estimate of markups.

Barkai (2016) applies this method to US national accounts data and obtains an estimate of the aggregate profit rate, which implies an average economy-wide markup. However, since aggregate time-series data are sparse and explanations for their behavior are typically abundant, Gutiérrez and Philippon (2017a, b) study cross-sectional data at the firm level from Compustat, which provides balance-sheet data on publicly listed US firms. In either case, because the profit rate is typically calculated period by period, this method produces a time series for the implied markup.

Perhaps the main advantage of this approach to markup estimation is that it avoids the need for econometric estimation of production functions, with the attendant difficulties of identification, which are used in the methods that follow. Conversely, the main problem with this approach is that economic profits are notoriously hard to calculate. A typical assumption in this approach is that profits are paid only to owners of capital. This assumption simplifies the computation, because observed payments to labor (and for intermediate inputs in firm- or industry-level data) can be treated as true factor costs, without any profit component. But one still faces the daunting challenge of separating required payments to capital (what the capital would earn on a competitive market) from economic profits, which are really a return to ownership of the firm but are bundled in the data with the implicit rental payments to capital. Efforts to separate the two require the researcher to impute a required return to capital, which when multiplied by the value of the capital stock yields the implicit rental payments.

The required rate of return includes the risk-free real rate, which can be observed from market interest rates on government debt: since 1997, inflation-indexed US government bond yields are available; prior to that, one needs to use nominal yields and subtract an expected inflation rate. It also includes the expected risk premium in excess of the safe rate, which typically must be imputed using an asset-pricing model. Barkai (2016) uses the AAA bond yield as the required return, which includes some compensation for risk. Gutiérrez (2017) explicitly imputes a risk premium, which he adds to a risk-free rate.

Another important component of the rental rate is the economic depreciation rate of the capital stock. Depreciation rates vary widely by type of capital; thus, required returns do as well. For example, Fraumeni (1997, table 3) reports annual depreciation rates of 2–3 percent for business structures, 10–20 percent for most types of business machinery, and 31 percent for office computers. The rate of economic depreciation includes the loss to the owner of capital from physical depreciation—the capital wearing out—as well as the expected capital gain or loss from the change in the resale price of the capital good relative to its purchase price. Most of the large depreciation rate for computers, for example, comes from the decline in the price of a computer over time due to technological progress in the manufacture of new computers, and not from the machine wearing out with use.

The rental cost of capital is then calculated as the sum of the required interest rate and the depreciation rate, multiplied by the market value of the capital stock for each type of capital, summed over all capital types.

While disaggregated stocks of capital are tracked at the level of large industries and the economy as a whole, they generally are not available at a firm level, where firm balance sheets typically report only the book (not market) value of the total capital stock. Furthermore, the national income accounts seek to estimate the rate of economic depreciation, while firm statements report only accounting depreciation. Thus, somewhat counterintuitively, the profit rate might be calculated with less error at an aggregate level, as in Barkai (2016), than at the firm level, as in Gutiérrez and Philippon (2017a, b).

Barkai (2016) calculates the profit rate on value added over the period 1984–2014 with US national accounts data. He finds a much lower profit rate at the start of his sample, 2.2 percent in 1984, than at the end, when it rises to 15.7 percent in 2014. The implied markup ratio μ thus rises from 1.02 to 1.19 over this period.

Gutiérrez and Philippon (2017a) calculate two measures of the profit rate. The first, which they term the "net operating margin," does not subtract the full required return to capital: their implicit rental payment includes depreciation, but not an interest rate. They also compute another markup estimate, based on a full user-cost measure subtracting the required interest payment to capital as well as depreciation. This measure rises by 0.05–0.10 over the period 1980–2015. Interestingly, it is relatively flat until about 2000 and then rises, which matches the timing of the change in labor's share, which is also fairly flat from 1980 to 2000 before declining sharply. Their estimated markup by the end of the sample is about 1.1.

While this estimate appears smaller than Barkai's, it is important to keep in mind that Gutiérrez and Philippon (2017a) are reporting a markup on firm *sales*, which is roughly equal to firm-level gross output, while Barkai reports a markup on *value added*, which summed over firms or industries equals GDP. This important distinction will be discussed further in the next section. For now, it suffices to note that on a common value-added basis, the two end-of-sample estimates are almost identical.

Note that the rise in both markup estimates is noticeably larger than what is implied by the decline in labor's share, as discussed in the previous section. If the markup did indeed rise to about 1.2, then to be consistent with the decline in labor's share corrected for mismeasurement, the output elasticity of labor would have to have risen by about 10 percent over the same period. It is not clear what could have caused such an increase.

A number of refinements to such calculations may be required to calculate the rental cost of capital, and therefore markups, with greater accuracy. Three refinements are worth particular mention. First and most straightforward is to correct the required return for taxes, following the classic method of Hall and Jorgenson (1967).

Second, and much more difficult, is to correct for adjustment costs of capital. This refinement could make use of Hayashi's (1982) neoclassical interpretation of Tobin's (1969) q ratio, the value of installed capital relative to the purchase price of new investment goods. The valuation of both the existing capital stock and its expected rate of change should be done using the shadow value of installed capital,

marginal *q*, which can differ from the observed purchase price of capital due to adjustment costs. However, these marginal adjustment costs cannot be observed directly and must be estimated econometrically.[7]

Third, the measure of capital could be expanded to include "intangible capital," which seems to be growing in importance in the US economy (for discussion, see Corrado, Hulten, and Sichel 2009). While some forms of intangible capital, primarily software and research and development, are included in the US national accounts, most are not. It is possible that the imputed rentals to capital are too low, because tangible capital is substantially smaller than the true quantity of tangible plus intangible capital. It should be noted that Gutiérrez and Philippon (2017a, b) do incorporate intangible capital into their analysis, using methods in the literature following Corrado, Hulten, and Sichel (2009). However, it is possible that intangible investments grew even faster than the traditional measurements imply, as suggested by McGrattan and Prescott (2010).

In a closely related approach, Karabarbounis and Neiman (2018) also emphasize that the efforts to measure the rental rate of capital may require significant adjustments. They focus on the gap between revenue and imputed total costs, which at the national level is the sum of labor payments and imputed capital rents. They term this gap *factorless income*, because it cannot be attributed easily to either labor or capital, and examine its time-series behavior in aggregate US data from 1960 to 2016. (The method of measuring the markup discussed in this section assumes that "factorless income" represents economic profit resulting from market power, and that it is indeed a return to firm ownership, rather than a required payment to either factor of production.)

It turns out that their measure of factorless income was quite high in the 1960s and early 1970s, then declined, and has been high again since the 1990s. If factorless income is interpreted as profits, then markups must also have been high before 1980. But most of the hypotheses advanced to explain high market power—such as the rise of "superstar firms" (Autor et al. 2017) and a high industrial concentration ratio—fit the recent period but not the pre-1980 period. Similarly, while growth of intangible capital appears able to help explain high levels of factorless income in recent years (if it is interpreted as a return to unmeasured intangibles rather than economic profits), estimates of the quantity of intangible capital typically find that it has been rising steadily over time and was not large before 1970.

Somewhat by default, Karabarbounis and Neiman (2018) suggest that researchers are failing to measure the required rate of return to capital properly. Yet they do not demonstrate that including one or more of the variables typically omitted from the construction of the rental rate can actually account for a substantial

[7] In principle, the value of Tobin's *q* can be inferred from market prices of a firm's debt and equity. This method has two drawbacks. First, it applies only to publicly listed firms. Second, it requires the researcher to assume that asset market valuations reflect only fundamentals at all points in time, a "no-bubble" assumption that may be difficult to justify after the asset market run-ups and crashes observed around the world historically and in the past three decades.

share of the mysterious factorless income. Thus, Karabarbounis and Neiman's (2018) main contribution is to show that plausible changes in market power alone are unlikely to explain the full post–World War II time series of imputed profits/factorless income in aggregate US data.

Estimates Based on Econometric Estimation Using All Inputs

The methods discussed in this section and the next drop the assumption of constant returns to scale. Instead, they rely on estimates of production functions. These production functions may be estimated while imposing the cost-minimization conditions, as in Hall's approach, to give a one-step estimate of the markup. Or the estimated output elasticity for one of the inputs can be compared with that input's revenue share to obtain a two-step estimate, as in the approach of De Loecker and Eeckhout. Both applications also use firm- or industry-level data, where the output concept is gross output and intermediate inputs are an additional factor of production. These changes make no difference to the theory sketched above, but they do change the interpretation of the resulting estimates in an important way.

As noted in the methods section above, the approach pioneered by Hall (1988, 1990) and used recently by Hall (2018)—as well as a large intervening literature!—naturally estimates the markup as a single parameter over the entire sample period. By itself, this method would not provide an estimate of the change in the markup over time, which is the primary focus of the recent literature. Hall (2018) parameterizes each industry-level markup as the sum of a constant and a time trend, and he reports estimates for the weighted average markup at the beginning and end of his sample period, 1988–2015. (As noted above, the method is based on a first-order approximation to the production function, which implies that the output elasticities should be constant over time. To be consistent with the method while allowing a smooth rise in the markup, each input's share should be trending downward at the rate the markup is increasing—an implication that could be checked against the data.)

Hall uses an instrumental variables technique to address the concern that the error term is endogenous. Specifically, Hall uses four categories of military expenditures and the price of West Texas intermediate crude oil as instrumental variables that are arguably uncorrelated with technical change. Hall (2018) applies this technique to US data for 60 industries. Most of these industries are at the North American Industry Classification System (NAICS) two- or three-digit level of aggregation; some examples of large industries in the dataset include retail trade, wholesale trade, and construction.

A significant advantage of Hall's (2018) method is that it does not constrain returns to scale to be constant. The disadvantage is that it requires strictly more information, as well as econometric estimation with instrumental variables.

Notice that Hall's (2018) method still requires its user to compute the quantity of profits, because in his one-step production-function approach, the shares are

the *cost* of each input divided by revenue. The typical assumption is that profits are received only by capital, so one needs to impute the rental rate of capital, as in the papers discussed in the previous section. Hall acknowledges this issue. But in practice, he follows the construction of the KLEMS dataset from the US Bureau of Labor Statistics, which offers sectoral data on output, as well as inputs and shares of capital (K), labor (L), energy (E), materials (M), and purchased business services (S). The shares are constructed assuming that total cost equals total revenue, which of course is correct only if economic profits are zero. If profit rates are zero, then the estimates that Hall presents as markups are also estimates of the degree of returns to scale, as shown above.

Hall (2018) estimates that the weighted average industry markup is about 1.3 in 2015, and that industry markups fall in the range of 1.0–1.8. (Some 30 percent of the point estimates are below one but are constrained to equal one on economic grounds, because firms would never systematically price output below marginal production cost.) The time trend is estimated to be positive, implying that markups have been rising over time, although the estimate is not statistically larger than zero at conventional levels of significance.

It may appear that Hall's (2018) estimate of average markup is only slightly larger than Barkai's (2016) estimate. This conclusion would be incorrect. Hall is using industry data and estimating a markup on gross output, while Barkai is estimating a markup on value added. A markup on gross output leads to a larger markup on value added when one takes into account the fact that firms use intermediate goods in production (Rotemberg and Woodford 1995; Basu and Fernald 2002). The intuition is that there is a "double-marginalization" phenomenon—firms sell some of their output for use as intermediate goods, which are bought by other firms that levy an additional markup on top of the markup they paid for their intermediates, and so on. Assuming an intermediate input share of 0.50, approximately the average value for the US economy over a long period of time, a markup of 1.3 on gross production translates to a markup on value added of 1.9, far larger than Barkai's estimate.[8]

One way to interpret this estimate, consistent with Hall's (2018) implicit assumption of zero economic profit, is that the production function for GDP using just capital and labor inputs must have returns to scale equal in size to the value-added markup, namely 1.9. To understand the implications of such a large degree of increasing returns in the aggregate production function, consider aggregate US data for 2015 as reported by the Bureau of Labor Statistics (2019). The BLS reports that private nonfarm business sector value-added output grew 3.5 percent in 2015, while the weighted average of capital and labor input in that sector grew 2.7 percent in the same year. For this growth in output and inputs to be consistent with returns to scale of 1.9 requires that true technological progress must have been *negative*

[8]The relationship between the two markup concepts is $\mu = \frac{\mu^G(1 - s^M)}{1 - \mu^G s^M}$, where μ is the markup on value added, μ^G is the markup on gross output, and s^M is the intermediate input share of revenue.

1.6 percent! Such a high rate of technological regress for the US economy as a whole seems quite implausible, casting doubt on the high returns to scale implied by Hall's estimate of the average markup in 2015.

Why might Hall (2018) be estimating markups/returns to scale that are too large? The key concerns regarding his procedure are those that arise whenever one attempts to estimate production functions in differences—that is, looking at change in output and changes in inputs—which is the core of Hall's method. First, consistent estimation of scale economies requires that we measure the real quantities of all the inputs correctly, as opposed to just their nominal payments, which is all that is required when computing profit rates. (Strictly speaking, the estimation is consistent if any measurement error in the inputs is uncorrelated with the instruments.) Second, when variables are measured in growth rates, the resulting correlation tends to emphasize high-frequency variation in output and inputs. A substantial macro literature has emphasized that actual capital and labor inputs vary at high frequencies in ways that are not recorded in the conventional production data that Hall uses (Bils and Cho 1994; Burnside, Eichenbaum, and Rebelo 1996; Basu, Fernald, and Kimball 2006). For example, firms may vary the workweek of capital by changing the number of shifts used to produce output, thus changing the true capital service input but without a change in the observed capital stock. Firms appear to vary capital's workweek and labor effort as they change their rate of production, in response to both demand and technology shocks. This unmeasured variation in utilization will probably lead to an upward bias in the estimated markup, and using demand instruments will not solve this problem.

As an example of how such considerations can affect the results, Basu, Fernald, and Kimball (2006) also use annual industry-level production data and a procedure similar to Hall's (2018), but with an additional control for variations in the intensity of factor usage. Their data cover the period 1949–1996, so they do not examine the past two decades (although the sample does cover the early post–World War II period, when there is also evidence of high profit rates/markups). Controlling for variable factor utilization, Basu, Fernald, and Kimball find few industry markup estimates that are greater than one at conventional levels of statistical significance. The clearest evidence of positive markups is in durables manufacturing, but even there the median industry markup is just 1.07 (on gross output).[9] Outside of durables manufacturing, only one industry (chemicals) is estimated to have a markup significantly larger than one. On the other hand, the estimated controls for utilization are positive and highly significant for both durables and nondurables manufacturing industries, and large although statistically insignificant outside manufacturing.

Thus, future research using Hall's (2018) method might proceed in several ways. First, when using this approach, it might be useful to apply similar utilization controls to see whether, by reducing the effect of short-term variations in

[9] Basu, Fernald, and Kimball (2006) present their results as estimates of returns to scale, but since they operate under the same zero-profit assumption as Hall, their estimates can be interpreted equally well as markups.

unmeasured inputs, one also reduces the markup estimates obtained. Second, it would be useful to investigate further the interpretation of the estimated μ, and the extent to which it is capturing economies of scale, economic profits, or some mixture of the two. Finally, because this approach requires a calculation of profit, all of the questions raised in the previous section about measuring the rental cost of capital apply here as well.

Econometric Estimation Using a Single Input

The theory presented in the first main section of this paper established a framework in which the markup must be the same for each input (because marginal cost must be the same along every margin). Thus, it should be possible to compute the markup on the basis of only a single input to production, not many inputs. Moreover, if one chooses an input that does not receive pure profits, then the issues of measuring required returns to capital and the profit rate do not arise. The single-input method would be an ideal one to apply to data on intermediate inputs, which probably do not share in pure profits and are measured with the least error due to utilization.[10]

De Loecker and Eeckhout (2017) take this single-input approach using balance-sheet data on publicly listed firms from Compustat. In a later version of this paper, De Loecker, Eeckhout, and Unger (2018) also use firm-level data from the US Census, which is a better source for production data and provides information on firms that are privately held as well. As of this writing, their results using these data had not been cleared for disclosure by the Census Bureau, and hence the discussion in this paper is based on the results using only Compustat data.

These authors use firms' expenditures in their accounting reports on a composite input termed cost of goods sold (COGS), which consists of most intermediate goods and a subset of labor input. They take COGS and the fixed capital stock, K, as their two inputs to production at the firm level. By hypothesis no profits are paid to COGS, so they can construct the share of this factor's cost to total revenue simply from reported data. Like many authors in the industrial organization literature, they use a variant of the technique introduced by Olley and Pakes (1996) to estimate a Cobb–Douglas production function without imposing constant returns to scale and obtain the relevant coefficient estimate. Then, they can divide the estimated output elasticity of COGS by its observed revenue share to calculate the markup. Using the cross-section dimension of their data, they are able to estimate

[10] Indeed, Dobbelaere and Mairesse (2013) apply a similar idea to firm-level data from France to allow for the possibility that both labor and capital bargain over the profits generated by markups. They estimate the markup from the intermediate input margin only, and estimate the bargaining power of capital and labor from the other margins. Their technique also suggests a possible method to allow for monopsony and monopoly power in the same estimation framework. Unfortunately, the Compustat data do not allow for this attractive approach, because most firms do not report separate expenditures on labor and intermediate inputs.

a different output elasticity for each factor for all years, and thus a time series for the markup. (Note that while the output elasticity is constrained to be equal across firms at a point in time, the revenue share and hence the markup estimate vary across firms and over time.)

In their headline estimates, De Loecker, Eeckhout, and Unger (2018) report that the weighted average of the markup ratio at the firm level rises from 1.21 in 1980 to 1.61 in 2016, with most of the increase taking place in the 1980s and 1990s.[11] (Note that the timing of the estimated markup change does not match particularly well the timing of the decline in labor's share of national income, which drops sharply starting in the 2000s but is fairly stable earlier.) The authors emphasize that the trend in the average markup is being driven by increased heterogeneity at the firm level. They plot the distribution of their estimated markups across firms in 1980 and 2016. Both distributions have a mode that is very similar, about 1.3, but the distribution in 2016 has a higher standard deviation and a thicker right tail. This increase in density in the right tail leads to the much higher estimate of the average markup by the end of their sample.

Markups as large as those reported by De Loecker, Eeckhout, and Unger (2018) at the end of their sample have some implausible implications. For example, because the authors report that average returns to scale are about 1.05 in 2016, but that the average markup is 1.61, the relationship given earlier between markups, returns to scale, and economic profits suggests that the average economic profit rate must be extremely high, on the order of 35 percent of firm sales. Since sales on average are about twice as large as value added, this calculation suggests that about 70 percent of GDP is pure economic profit! Profit rates of this size are too large to be credible. Also, because labor receives slightly less than 60 percent of GDP as compensation, and the authors assume that none of that payment is profit, economic profits of this size would mean that there is not enough output to pay both labor and profit, let alone any required return to capital.

Another implausible implication arises because, in keeping with the rest of the recent literature, De Loecker, Eeckhout, and Unger (2018) assume that profits are paid only to owners of capital. On average across US industries, firms spend about 50 percent of their revenues on intermediate goods and 30 percent on labor. With these shares and a markup of 1.61, the output elasticity of labor and intermediates must sum to about 1.3. But since they estimate average returns to scale of only 1.05, the implied output elasticity of capital must be on the order of −0.25.[12] It seems very

[11] These are the results from what De Loecker, Eeckhout, and Unger (2018) term the "traditional" production function, which they denote PF1. They also report results for an alternative specification, PF2, in which markups rise from about 1 in 1980 to 1.32 in 2016. However, the paper stresses the results from PF1, which are the only ones mentioned in the abstract and are presented first in the introduction. Thus, the discussion here focuses on the PF1 results, while noting the PF2 estimates when they differ significantly.

[12] On its face, the finding by De Loecker, Eeckhout, and Unger (2018) that returns to scale have not risen over time does not support one common hypothesis for explaining the rise of markups. A standard implication of a "knowledge economy" is that production is characterized by large fixed or sunk costs but

unlikely that firms spend billions of dollars on investment to accumulate capital that will reduce their output. Furthermore, the implied negative capital elasticity is inconsistent with De Loecker, Eeckhout, and Unger's own production-function estimate of the output elasticity of capital, which averages approximately 0.2. This inconsistency between the implied and the directly estimated capital elasticities is an indication that the estimation is producing problematic results.

Finally, if the average markup of 1.61 is put on a value-added basis as before, the implied markup on GDP is in excess of 4, enormously higher than standard estimates in the macro literature, which typically estimates values between 1.1 and 1.4 (for example, Basu and Fernald 1997; Christiano, Eichenbaum, and Evans 2005).

The preceding calculations have taken the average markup reported by De Loecker, Eeckhout, and Unger and examined its implications as if it were the markup of the representative firm in the economy. But given their finding of extreme heterogeneity in markups, these calculations should be performed at the firm level. The authors can easily perform the first two calculations, of profit rates and the implied output elasticity of capital, at the firm level using their existing estimates, with the results reported as distributions. The third calculation, of the value-added markup, is also easily done, but the data may not be available for every firm. The reason is that many firms in Compustat do not report the data necessary to construct an intermediate input share at the firm level, which is required for the conversion. Given the great interest that this paper has generated, many economists would be keen to see these additional results.

The method used by De Loecker, Eeckhout, and Unger (2018) seems excellent in principle, so why is it leading to implausible estimates of markup levels and trends? As a matter of accounting, the sharp rise in the markup the authors estimate could be driven by either a rising output elasticity of their key input measure, cost of goods sold, or a decline in its share over time. In fact, the output elasticity is estimated to be nearly constant over the sample period, so the rise in the estimated markup is being driven completely by the decline in the share.

This fact should motivate us to think harder about this measure of inputs. Cost of goods sold is an accounting concept used to value changes in inventory holdings for firms that produce to stock. But such industries—agriculture, mining, and manufacturing—produce only about 16 percent of private-sector output in the US economy (for data from 2016, see table 5 in Bureau of Economic Analysis 2018). The concept is much less meaningful when applied to the service industries that produce twice as large a share of US GDP, such as finance and insurance, health care, education, and professional and business services.[13] For some intuition behind this issue, one might try to describe how the reported COGS for a few large, publicly

low marginal cost. The classic example is software, which can take huge resources to develop but then can be replicated at essentially zero cost.

[13] In service industries, cost of goods sold is often renamed "cost of revenue," but the two are conceptually similar.

listed service companies (like Facebook, Goldman Sachs, or a for-profit hospital chain such as HCA Healthcare) is meaningful in an economic sense.

In Compustat, firms actually report two measures of operating (noncapital) expenses. One is cost of goods sold; the other is selling, general, and administrative expenses (SGA). If an increasingly larger share of the inputs that used to be classified as COGS is now recorded as being part of SGA, then such mismeasurement could explain why the COGS share of firm revenue is falling over time, and consequently why markups are estimated as rising over time. Traina (2018) redoes the procedure using the sum of COGS and SGA as the measure of noncapital input and finds no evidence that markups have increased over time.

Some of the controversy following Traina's (2018) paper has focused on the extent to which cost of goods sold can be interpreted as a variable cost and selling, general, and administrative expenses as a fixed cost. (Here the word "fixed" is used to mean overhead inputs, ones that do not vary with the amount of output produced, at least locally, as opposed to inputs that are quasi-fixed, meaning costly to adjust.) The controversy is misplaced, because the underlying theory does not require that *all* of the input on the examined margin be variable. It requires only that there be *some* variable inputs in the input bundle under consideration, and that the bundle be defined consistently over time. If overhead inputs are a higher share of total inputs, the estimated output elasticity of that input bundle will be larger—appropriately so, since overhead inputs are one important source of increasing returns to scale.

Given that there is no harm in deriving the markup using an input aggregate that includes some overhead inputs, it is safer to use a more comprehensive input measure. For example, by convention, payments to salaried workers are classified as selling, general, and administrative expenses, while variable (hourly and commission) labor payments are in cost of goods sold. If there has been a general change in compensation practices for workers fulfilling the same function, shifting them from being classified as part of COGS to being included in SGA, then COGS could not be used to compute the trend in the markup, but the sum of COGS and SGA could be used for this purpose.[14]

This single-input approach to estimating markups would be best executed using a single, distinct measure of physical input, such as production-worker labor hours or purchased energy. However, such data are not available for a large fraction of firms in the Compustat dataset. The next-best choice would be to use a comprehensive measure of composite inputs that includes most or all variable inputs and some overhead inputs, but where the measure of inputs is sufficiently comprehensive that we can be confident it is defined consistently over time.

[14] Outsourcing work would reduce the labor component of cost of goods sold but increase the intermediate input component, and it should not change the total appreciably. This hypothesis is consistent with the findings of Meixell, Kenyon, and Westfall (2014).

Directions for Future Research: Reconciling Micro Estimates and Macro Facts

Two of the three approaches to markup estimation reviewed above yield end-of-sample estimates of the average value-added markup that are too large to be credible. By the underlying logic of cost minimization, high markups must be matched either by equally high returns to scale or by large rates of pure economic profit. Yet returns to scale of the magnitude required would imply large rates of technological regress for the US economy, while the implied rates of economic profit would displace all of the required payments to capital, as well as some of the observed payments to labor. In both cases, the larger estimates of increases in the markup greatly overshoot what is required to explain the decline in labor's share: the true puzzle becomes why labor's share has not fallen far more! Only the approach based on constant returns and computed profit rates leads to estimates of the level and change in the markup that might be consistent with the observed decline in labor's share, and even these estimates are on the high side.

It is worth reviewing some other macro implications of higher markups to contrast them in an informal way with recent data for the US economy. According to standard models, higher markups should reduce the demand for inputs of labor and capital, leading for example to weak growth in jobs and wages, and should raise inflation relative to a welfare-theoretic measure of slack in the economy, because higher markups act as "cost-push" shocks to the Phillips curve. In classic endogenous-growth models, higher markups should also spur innovation, as firms compete more fiercely to displace incumbents from profitable markets, leading to higher rates of productivity growth.[15]

With one partial exception, none of these predictions is borne out in recent US data. The labor market is extremely tight, inflation is subdued, and productivity growth has been weak. (The second fact is particularly striking in light of the first and third.) Even wage growth has been stronger than one would conclude from standard data: Daly and Pedtke (2018) find that median weekly earnings adjusted for labor-force composition have grown 1.5 to 2 percentage points faster per year over the period 2013–2017 than the unadjusted data suggest. An adjustment of this magnitude nearly doubles the growth rate of median weekly earnings. Because rising markups should depress labor demand, it is difficult to reconcile the hypothesis of rising markups with strong growth in both the quantity and the price of labor input over the past several years.

One piece of macro evidence that does go in the direction that the rising markup hypothesis predicts is low business investment. Gutiérrez and Philippon (2017a, b) argue that total investment in both tangible and intangible capital has

[15] This prediction is more nuanced in recent models. Aghion and Griffith (2005) show that the relationship between productivity growth and the markup may have an inverted-U shape, with high markups reducing growth. Aghion et al. (2019) suggest that high markups may lead to first higher and then lower rates of total factor productivity growth.

been weak in recent years, particularly when conditioned on Tobin's q ratio, a common measure of fundamentals. Higher markups leading to a positive profit rate can indeed make the average (asset-market) q high, while making marginal q, which determines investment, low. (Markups by themselves are neither sufficient nor necessary to break the link between average and marginal q; a positive economic profit rate is the key wedge.)

Another piece of evidence that may be consistent with a rising markup is a low natural rate of interest. The link between the two is that a higher markup is supposed to create expectations of declining consumption growth due to low demand for capital and labor, which in turn should pull down the interest rate. With the two factor markets for capital and labor seemingly moving in opposite directions, it is not obvious that the markup channel is at work. Carvalho, Ferrero, and Nechio (2016) and Gagnon, Johannsen, and Lopez-Salido (2016) show that demographic forces likely explain much of the decline in long-term real rates.

Thus, future research needs to address two puzzles. The first is why most markup estimates based on micro data are implausibly large and grow too fast in relation to the macro facts to be explained. The second is why most macro data appear to indicate that markups are low and stable, but the investment rate is sending a different signal. A full understanding of markup trends and their economic effects requires an explanation of these two issues.

■ *I thank the editors for useful suggestions that have improved the paper, and Jan Eeckhout, John Fernald, Germán Gutiérrez, Robert Hall, Brent Neiman, Thomas Philippon, Valerie Ramey, Fabio Schiantarelli, Chad Syverson, and J. Christina Wang for helpful comments and discussions.*

References

Aghion, Philippe, Antonin Bergeaud, Timo Boppart, Peter J. Klenow, and Huiyu Li. 2019. "A Theory of Falling Growth and Rising Rents." Federal Reserve Bank of San Francisco Working Paper 2019-11.

Aghion, Philippe, and Rachel Griffith. 2005. *Competition and Growth: Reconciling Theory and Evidence*. Cambridge, MA: MIT Press.

Autor, David, David Dorn, Lawrence F. Katz, Christina Patterson, and John Van Reenen. 2017. "The Fall of the Labor Share and the Rise of Superstar Firms." NBER Working Paper 23396.

Barkai, Simcha. 2016. "Declining Labor and Capital Shares." Stigler Center New Working Paper Series 2.

Basu, Susanto, and John G. Fernald. 1997. "Returns to Scale in U.S. Production: Estimates and Implications." *Journal of Political Economy* 105(2): 249–83.

Basu, Susanto, and John G. Fernald. 2002. "Aggregate Productivity and Aggregate Technology." *European Economic Review* 46(6): 963–91.

Basu, Susanto, John G. Fernald, and Miles S. Kimball. 2006. "Are Technology Improvements

Contractionary?" *American Economic Review* 96(5): 1418–48.

Basu, Susanto, John G. Fernald, and Matthew D. Shapiro. 2001. "Productivity Growth in the 1990s: Technology, Utilization, or Adjustment?" *Carnegie-Rochester Conference Series on Public Policy* 55(1): 117–65.

Bils, Mark, and Jang-Ok Cho. 1994. "Cyclical Factor Utilization." *Journal of Monetary Economics* 33(2): 319–54.

Bureau of Economic Analysis. 2018. "Gross Domestic Product by Industry: First Quarter 2018." Bureau of Economic Analysis News Release, July 20, 2018. https://www.bea.gov/system/files/2018-07/gdpind118_3.pdf.

Bureau of Labor Statistics. 2019. "Private Business and Private Nonfarm Business Multifactor Productivity Tables." March 20, 2019. https://www.bls.gov/mfp/special_requests/mfptable.xlsx.

Burnside, Craig, Martin Eichenbaum, and Sergio Rebelo. 1996. "Capital Utilization and Returns to Scale," in *NBER Macroeconomics Annual 1995*, vol. 10, edited by Ben S. Bernanke and Julio J. Rotemberg, 67–124. Cambridge, MA: MIT Press.

Carvalho, Carlos, Andrea Ferrero, and Fernanda Nechio. 2016. "Demographics and Real Interest Rates: Inspecting the Mechanism." *European Economic Review* 88(1): 208–26.

Christiano, Lawrence J., Martin Eichenbaum, and Charles L. Evans. 2005. "Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy." *Journal of Political Economy* 113(1): 1–45.

Corrado, Carol, Charles Hulten, and Daniel Sichel. 2009. "Intangible Capital and U.S. Economic Growth." *Review of Income and Wealth* 55(3): 661–85.

Council of Economic Advisers. 2016. "Labor Market Monopsony: Trends, Consequences and Policy Responses." Council of Economic Advisers Issue Brief, October 2016.

Daly, Mary C., and Joseph H. Pedtke. 2018. "Revisiting Wage Growth." Federal Reserve Bank of San Francisco Economic Letter.

De Loecker, Jan, and Jan Eeckhout. 2017. "The Rise of Market Power and the Macroeconomic Implications." NBER Working Paper 23687.

De Loecker, Jan, Jan Eeckhout, and Gabriel Unger. 2018. "The Rise of Market Power and the Macroeconomic Implications." Unpublished, September 14, 2018.

Dobbelaere, Sabien, and Jacques Mairesse. 2013. "Panel Data Estimates of the Production Function and Product and Labor Market Imperfections." *Journal of Applied Econometrics* 28(1): 1–46.

Elsby, Michael W. L., Bart Hobijn, and Ayşegül Şahin. 2013. "The Decline of the U.S. Labor Share." *Brookings Papers on Economic Activity*, Fall, 1–52.

Fraumeni, Barbara M. 1997. "The Measurement of Depreciation in the U.S. National Income and Product Accounts." *Survey of Current Business* 77(7): 7–23. https://apps.bea.gov/scb/account_articles/national/0797fr/maintext.htm.

Gagnon, Etienne, Benjamin K. Johannsen, and David Lopez-Salido. 2016. "Understanding the New Normal: The Role of Demographics." Finance and Economics Discussion Series 2016-080.

Gutiérrez, Germán. 2017. "Investigating Global Labor and Profit Shares." Unpublished, October 2017.

Gutiérrez, Germán, and Thomas Philippon. 2017a. "Declining Competition and Investment in the U.S." NBER Working Paper 23583.

Gutiérrez, Germán, and Thomas Philippon. 2017b. "Investmentless Growth: An Empirical Investigation." *Brookings Papers on Economic Activity*, Fall, 89–190.

Hall, Robert E. 1988. "The Relation between Price and Marginal Cost in U.S. Industry." *Journal of Political Economy* 96(5): 921–47.

Hall, Robert E. 1990. "Invariance Properties of Solow's Productivity Residual," in *Growth/Productivity/Unemployment: Essays to Celebrate Bob Solow's Birthday*, edited by Peter A. Diamond, 71–112. Cambridge, MA: MIT Press.

Hall, Robert E. 2018. "New Evidence on Market Power, Profit, Concentration, and the Role of Mega-Firms in the US Economy." Unpublished, September 23, 2018.

Hall, Robert E., and Dale W. Jorgenson. 1967. "Tax Policy and Investment Behavior." *American Economic Review* 57(3): 391–414.

Hayashi, Fumio. 1982. "Tobin's Marginal *q* and Average *q*: A Neoclassical Interpretation." *Econometrica* 50(1): 213–24.

Karabarbounis, Loukas, and Brent Neiman. 2014. "The Global Decline of the Labor Share." *Quarterly Journal of Economics* 129(1): 61–103.

Karabarbounis, Loukas, and Brent Neiman. 2018. "Accounting for Factorless Income." NBER Working Paper 24404.

McGrattan, Ellen R., and Edward C. Prescott. 2010. "Unmeasured Investment and the Puzzling US Boom in the 1990s." *American Economic Journal: Macroeconomics* 2(4): 88–123.

Meixell, Mary J., George N. Kenyon, and Peter H. Westfall. 2014. "The Effects of Production Outsourcing on Factory Cost Performance: An Empirical Study." *Journal of Manufacturing Technology Management* 25(6): 750–74.

Morlacco, Monica. 2019. "Market Power in Input Markets: Theory and Evidence from French Manufacturing." Unpublished, March 20, 2019.

Olley, G. Steven, and Ariel Pakes. 1996. "The Dynamics of Productivity in the Telecommunications Equipment Industry." *Econometrica* 64(6): 1263–97.

Rotemberg, Julio J., and Michael Woodford. 1995. "Dynamic General Equilibrium Models with Imperfectly Competitive Product Markets," in *Frontiers of Business Cycle Research*, edited by Thomas F. Cooley, 243–93. Princeton, NJ: Princeton University Press.

Solow, Robert M. 1957. "Technical Change and the Aggregate Production Function." *Review of Economics and Statistics* 39(3): 312–20.

Tobin, James. 1969. "A General Equilibrium Approach to Monetary Theory." *Journal of Money, Credit and Banking* 1(1): 15–29.

Traina, James. 2018. "Is Aggregate Market Power Increasing? Production Trends Using Financial Statements." Stigler Center New Working Paper Series 17.