NBER WORKING PAPER SERIES

DATA AND MARKUPS: A MACRO-FINANCE PERSPECTIVE

Jan Eeckhout
Laura Veldkamp

Working Paper 30022
http://www.nber.org/papers/w30022

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2022, Revised March 2024

We thank Vasco Carvalho, Joel David, Maryam Farboodi, Basile Grassi, Hugo Hopenhayn, Ariel Pakes, Bruno Pellegrino and Karthik Sastry as well as seminar audiences for insightful discussions. We benefited from superb research assistance by Renjie Bao, Ankit Bhutani, Yihao Li, Zihao Li, Huasheng Nie, Xiaobo Yu, and Judy Yue. Eeckhout gratefully acknowledges support from the ERC, Advanced grant 882499, and from PID2022-138443NB-I00. This paper originally circulated under the title, "Data and Market Power." The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Data and Markups: A Macro-Finance Perspective
Jan Eeckhout and Laura Veldkamp
NBER Working Paper No. 30022
May 2022, Revised March 2024
JEL No. D8,E3,L0

ABSTRACT

What does market power look like in a data economy? Economists typically use markups to measure market power. However, we use a simple model to show how firms' growing stocks of data can change markups, without changing their power to affect prices. Data's effects depend on how markups are aggregated. Growing data can produce differences in markup measures that match empirical facts. Markup aggregation wedges can measure data stocks and offer a way to purge markups of data's effects to reveal market power.

Jan Eeckhout
Department of Economics
Universitat Pompeu Fabra
Trias Fargas 25
Barcelona
08005
Spain
and CEPR
jan.eeckhout@upf.edu

Laura Veldkamp
Columbia Business School
725 Kravis Hall
665 W 130th St
New York, NY 10027
and NBER
lv2405@columbia.edu

I Introduction

Changes in firms' market power and the sources of those changes have become the focus of intense debate. Economists point to rising markups, economies of scale in information and the dominance of large, data-intensive firms as evidence that the unequal accumulation of data is responsible for a decline in competition (Jarsulic 2019). To assess such a claim, economists need to determine how to measure market power in data-rich firms. Our model is a tool to uncover potential problems with the interpretation of existing measures of market power and offer new techniques for measuring data's effects.

Besides the power for firms to affect their market prices, there are three additional forces that show up in product markups: cost advantages, compensation for risk and data barter. For firm and industry markups, additional aggregation effects arise. Data interacts with all of these. We first use a model with exogenous data to explain how data affects markups through cost, risk and aggregation. Later, we endogenize data to show how data barter depresses markups as well. Our work does not prove that each assumption is necessary, although the model is consistent with many empirical findings. Rather, the model shines a light on a number of potential issues and proposes a new data measurement technique. The goal is to explain relevant data-economy forces and enable a variety of new approaches to measurement.

The model in Section II draws on tools from multiple fields, each of which introduces a different relationship between firm data and markups. As in macro theory, data is modeled as information; as in corporate finance, firms price risk; as in industrial organization theory, firms exploit market power. Data is digitized information. The essence of information is that it reduces uncertainty. Just like a larger data set reduces the econometrician's standard error, more data for firms reduces their prediction errors. Making future events more predictable allows the firm to make better decisions that raise expected profits, and also means the firm faces less uncertainty and less risk. While abstracting from risk is appropriate to study many questions, abstracting from risk and the price of risk when studying data removes its essential character. Although the assumption that firms respond to risk is unusual in the firm competition literature, it is a bedrock principle of the field of corporate finance (Brealey, Myers, and Allen 2003; Eckbo 2008) and therefore worth considering as one of many possible links between firm data and markups. Section III shows that if firms use data to improve their forecasts and make their revenues more predictable and less risky, they produce more, which lowers price. So, holding costs fixed, the "risk channel" implies that more

data reduces markups.

However, data also makes it more profitable to grow large. In the model, firms choose an up-front investment, which lowers their future marginal cost of production. When data lowers the risk of investment by predicting future demand, firms invest more. This is *investment-data complementarity*. More investment means the firm grows larger, produces at lower marginal cost, and earns higher markups. What we call "investment channel" might be described in the language of Sutton (1991, 2001) as: our firms strategically use data to differentiate themselves and create a dominant position. While this logic is not novel or unique to data, it is prevalent in the data debate and worth including for consideration, alongside the other data-related effects.

A key theme of the paper is that, to disentangle data from markups, we should look to covariances. What distinguishes data from other assets and forces is that data facilitates prediction. Better predictions alter the composition of products and firms. Data-rich firms produce more of goods that their data predicts are likely to be profitable (high-markup goods). Thus, even if two firms sell identical goods at identical prices with identical marginal costs, the firm with more data will be measured as a higher-markup firm because that firm uses data to skew the composition of its goods production toward higher-markup goods. The covariance that data allows firms to achieve between production and markups is exactly the covariance that creates aggregation effects. This is useful because it means that aggregation wedges, in this case, the difference between the average product markup and the firm's markup, can be used to infer firms' data. Not only can data explain composition effects in markups, but some sort of information is necessary to explain the change in composition. Every firm would like to produce more of the more profitable goods and less of the less profitable ones. This is only a feasible (measurable) strategy if the firm can predict what will be profitable and what will not. Good prediction requires good data.

Our results are consistent with recent evidence by Galdon-Sanchez, Gil, and Uriz-Uharte (2023) on the behavior of small and medium size enterprises. They exploit an experiment by a bank that provides the firms with information about sales of their competitors. They show that the firms that access the information increase their revenue between 4.5% and 9%, and show that data allows the firms to identify the more profitable sales opportunities from among the existing ones and exploit them.

Data also causes the markup of an average firm and its industry to diverge. Data-investment complementarity ensures that high-data firms invest more and sell more. But if these high-data firms also have high firm-level markups because they skew the composition of their goods to-

ward high-markup goods, then high-markup firms are also larger firms. This effect is stronger in recessions where demand is most uncertain.

Trend and cyclical markup wedges are real. The model explains a curious feature of markup measurement that has been at the heart of a debate. From one perspective, markets are just as competitive today as in the past because good-level markups are stable (see for example Anderson, Rebelo, and Wong [2018]). Instead, growing firm-level and industry markups are evidence of declining competition (see Philippon [2019]; Furman and Orszag [2015]; Grullon, Larkin, and Michaely [2016]; De Loecker, Eeckhout, and Unger [2020]; Hall [2018]). Section V shows that the level of aggregation also changes the measured cyclicality of markups, in the way the model predicts (Nekarda and Ramey 2020; Bils 1985, 1987). These facts validate our model. In turn, the model lends an economic interpretation to these facts beyond the explanation that composition effects must be at work. The differences no longer represent a measurement or aggregation mistake made by one group or the other, but rather an interesting and useful measure of the quantity of firms' data.

Ultimately, most researchers are interested in markups because they are concerned about consumer welfare. Section VI discusses the relationship between markups, competitive outcomes, and welfare. Rising amounts of data can be good for consumers. After all, firms use data to produce more of goods that consumers want most. However, welfare may suffer when firms' data stocks become asymmetric.

The static model took data to be exogenous. That assumption matters. When data is a by-product of economic activity, a new effect on markups emerges. Section VII points to data barter that should push markups down. Firms that value data want to do more transactions to generate more data. To do more transactions, a firm must lower its price and thus its markup. The optimal production decision for a firm reveals that price and the marginal value of data enter as substitutes. Firms are effectively paid either with money or with their customers' data. This finding relates to the study of free digital goods. The idea that price does not fully capture the value of a transaction to a firm also provides one more reason that markups fail to capture market power in a data economy.

Since our contribution is to use theory to reinterpret existing facts and provide tools to indentify new ones, we leave measurement as a future agenda for multiple papers to tackle. Section VIII offers guidance for how this framework might be used to measure data or the market power arising from that data. Our model teaches us that the difference between firm and product markups is a

sufficient statistic for the amount of relevant data a profit-maximizing firm has about consumer demand. This would enable a sufficient statistics approach to measuring firms' data. The model could also be used for structural estimation. We discuss techniques to estimate firms' price of risk, approaches to measuring product characteristics and elasticities, and the map from the markup measures in our model to different empirical approaches in the markup literature.

RELATED LITERATURE Because we model data as digitized information, our tools are most similar to those in the information frictions literature in macroeconomics. Work by Lorenzoni (2009), Angeletos and La'O (2013), Asriyan, Laeven, and Martin (2022), David and Venkateswaran (2019), Nimark (2014) and Maćkowiak and Wiederholt (2009) feature similar information frictions, used to explain features of business cycles. Work by Rostek and Weretka (2012) explore a reverse question: the effect of market size and market power on price informativeness. Similar tools are used in models of banking competition as well (Vives and Ye 2021), where banks use information for forecasting and pricing risk. However, banks differ from firms: while goods-producing firms choose freely how many units of a good to produce, lenders typically cannot lend twice the requested amount to a promising borrower. The ability to scale production is central to market power. Finally, firms use data to forecast price, which depends on others' actions. Capturing this strategic use of data requires new solution tools to layer a forecasting-the-forecasts-of-others problem, on top of a imperfect information portfolio problem.

Existing work on the digital economy explores whether data can be a source of market power. Kwon, Ma, and Zimmermann (2022) argue that the timing and degree of rising concentration in an industry correlate closely with the industry's investment in information technology. Jones and Tonetti (2020) explore what data ownership rules facilitates economic growth. In Kirpalani and Philippon (2020), data enables directed two-sided search. Acemoglu et al. (2022) and Bergemann and Bonatti (2019) model data as information and explore whether data markets are efficient. Ichihashi (2020) shows how firms can use consumer data to price discriminate, while Liang and Madsen (2021) explore the use of data in labor markets. In De Ridder (2021) information technology raises fixed costs and reduces marginal costs. We do not dispute that data can be used for all of these purposes. However, we introduce uncertainty, aggregation effects and data barter and show how these affect markup levels, trends and cyclicality.

Our work obviously speaks to the large literature on markup measurement and complements it by providing new interpretations of results about trends and fluctuations in markups. Some new

papers model the mechanisms that give rise to trending markups (see for example De Loecker, Eeckhout, and Mongey [2021]). Those models and Edmond, Midrigan, and Xu (2019) evaluate the welfare consequences of markups. Our approach differs because we focus on firms' use of data.

Empirical work on the data economy often necessarily focuses on specific markets.[1] Lambrecht and Tucker (2015) take a strategy perspective on whether data has the necessary features to confer market power. Similarly, Goldfarb and Tucker (2017) discuss the many ways in which this digital economy is transformative.

Recent work by Burstein, Carvalho, and Grassi (2020) analyzes the business cycle properties of markups. They show analytically how the sign of markup cyclicality varies with aggregation and they establish the economic importance of these markups. Our results complement these insights by proposing a specific mechanism that causes markups to fluctuate, one that is rooted in how firms use data to gain a competitive advantage.

II Model

To explore the how data interacts with measures of market power, we build a model with multiple possible interactions. This model is not the simplest explanation for any given set of facts. Rather it is like a checklist of possible interactions to look for, meant to guide a measurement agenda to determine which are more relevant. They key assumptions are as follows. First, firms face uncertainty about consumer demand. It is not essential that uncertainty is about demand, rather than advertising, hiring, product placement or costs. We simply need a variable that is profit-relevant and uncertain. Second, data is used to resolve this uncertainty. Data is used to predict the profitability of various actions. Third, firms face a cost of bearing risk. This price of risk is what governs the magnitude of the link between data, uncertainty, and investment. Fourth, to explore the relationship between data and the composition of the goods a firm produces, we model firms that choose quantities of multiple goods. Allowing those goods to have correlated attributes, as in Pellegrino (2020), makes data relevant to multiple goods. Finally, since the data competition hypothesis is about high-data firms growing large, we allow firms to choose an initial investment, which reduces their marginal cost of production. This allows us to explore if high-data firms invest to operate at a larger scale and thus grow to have more market power.

[1] Athey, Mobius, and Pal (2017) and Athey and Gans (2010) examine media competition; Brynjolfsson, Hu, and Smith (2003) study booksellers; and Rajgopal, Srivastava, and Zhao (2021) measure digital technology firms. de Cornière and Taylor (2020) categorize uses of data as pro- or anticompetitive.

We first explore these features in a static model. Since our question is about what effects data has on competition measures, we take data to be exogenous and move it around in the model to observe its effect. Later, Section VII introduces dynamics and endogenizes data as a by-product of economic activity and something that can be purchased or sold. The forces we describe here will survive in that dynamic setting.

II.A Setup

PRODUCTS AND ATTRIBUTES The product space has N attributes, indexed by $j \in \{1, 2, \ldots, N\}$ and N goods, indexed by k, that are combinations of attributes. Each good $k \in \{1, 2, \ldots, N\}$ can be represented as an $N \times 1$ vector $\boldsymbol{a}_k$ of weights that good places on each attribute. The jth entry of vector $\boldsymbol{a}_k$ describes how much of attribute j the kth good requires. This collection of weights describes a good's location in the product space. Let the collection of $\boldsymbol{a}_k$'s for each good k be an $N \times N$, full-rank matrix A, such that

$$\mathbf{q}_i = A\tilde{\mathbf{q}}_i. \tag{1}$$

Conversely, the quantity of attributes that a firm i produces is a vector $\tilde{\mathbf{q}}_i$, with jth element $\tilde{q}_{ij}$. The attribute vector is the vector of firm i's product quantities, q_i, times the inverse attribute matrix: $\tilde{\mathbf{q}}_i = A^{-1}\mathbf{q}_i$. Similarly, we represent the price and the marginal cost of production of goods as the linear combinations of the vector of prices and costs of the attributes: $\mathbf{p}_i = A\tilde{\mathbf{p}}_i$ and $\mathbf{c}_i = A\tilde{\mathbf{c}}_i$.

For now, the mapping between attributes and products is fixed. Initially, we can equate goods and attributes. However, this structure facilitates the measurement of elasticities (Pellegrino 2020), which is essential for applying these tools. Later, we allow firms to choose how to position themselves in the product space by choosing A's.

FIRMS There are n_F firms, indexed by i: $i \in \{1, 2, \ldots, n_F\}$. Firm production profit π_i depends on quantities of each good, which are entries in the $N \times 1$ vector $\mathbf{q}_i$, the market price of each good, $\mathbf{p}$ of dimension $(N \times 1)$, and the marginal cost of production, $\boldsymbol{c}_i$ (also $N \times 1$):

$$\pi_i = q_i' \left(\mathbf{p} - \boldsymbol{c}_i\right). \tag{2}$$

Each firm chooses the number of units of each good they want to produce, an $N \times 1$ vector $\mathbf{q}_i$, to maximize risk-adjusted profit, where the price of risk is ρ_i.

$$U_i = \mathbf{E}\left[\pi_i | \mathcal{I}_i\right] - \frac{\rho_i}{2}\mathbf{Var}\left[\pi_i | \mathcal{I}_i\right] - g(\chi_c, \widetilde{\boldsymbol{c}}_i). \tag{3}$$

This mean-variance objective is consistent with empirical corporate finance evidence on firms' decisions (Eckbo 2008) and with macro evidence that firms increase investment and sales in response to a randomized data treatment (Kumar, Gorodnichenko, and Coibion 2023).

The last term in (3) is each firm's up-front investment. Let $\widetilde{\boldsymbol{c}}_i$ be the $(N \times 1)$ vector of marginal production costs for a unit of each attribute. The up-front investment choice is modeled as a choice of $\widetilde{\boldsymbol{c}}_i$ at an investment cost $g(\chi_c, \widetilde{\boldsymbol{c}}_i)$ to maximize $E[U_i]$. Assume that $g(\chi_c, \widetilde{\boldsymbol{c}}_i)$ is additively separable in attributes and strictly decreasing in $\widetilde{\boldsymbol{c}}_i$. Since lower choices of $\widetilde{\boldsymbol{c}}_i$ require a greater up-front investment, we interpret this as choosing a larger firm. Since we want to interpret χ_c as a parameter that governs the marginal cost of investment, we impose $\partial^2 g / \partial \chi_c \partial c_i < 0$. To guarantee non-negative interior marginal cost choices, $g\left(\chi_c, \widetilde{c}_i\right)$ is convex over $\widetilde{c}_i$, with $g\left(\chi_c, \overline{p}\right) = 0$, where $\bar{p}$ is the highest possible price, and $\lim_{\widetilde{c} \to 0} g\left(\chi_c, \widetilde{c}\right) = +\infty$.

PRICE Our demand system embodies the idea that goods with similar attributes are partial substitutes for each other. Therefore, the price of good i can depend on the amount every firm produces of every good.

Each attribute j has an average market price that depends on an attribute-specific constant and on the total quantity of that attribute that all firms produce:

$$\tilde{p}_j^M = \bar{p}_j - \frac{1}{\phi}\sum_{i=1}^{n_F} \tilde{q}_{ij}. \tag{4}$$

This demand system has the advantage that it holds the power to affect prices dp^M/dq_i fixed. The assumption is surly not true. But it ensures that when markups change, it is from effects other than the power to affect price.

Each firm does not receive the market price, but rather faces an uncertain price that depends on a demand shock $\mathbf{b}_i$. The demand shock $\mathbf{b}_i$ is a vector with jth element b_{ij}. This vector is random and unknown to the firm: $\mathbf{b}_i \sim N(0, I)$.[2] Demand shocks can covary across firms: $\xi = \boldsymbol{Cov}(b_i, b_l)\ \forall i \neq$

[2]The fact that the variance matrix is diagonal is without loss. We simply define attributes to be an orthogonal decomposition of the demand variance-covariance space. We investigate the effect of changing $Var(b_i)$ in Section V.

l. The price a firm receives for a unit of attribute j is thus $\tilde{p}_j + b_{ij}$. We can express firm i's price in vector form as

$$\tilde{\mathbf{p}}_i = \left[\tilde{p}_1^M, \tilde{p}_2^M, \ldots, \tilde{p}_N^M\right]' + b_i. \tag{5}$$

The firms's price of each good depends on its attributes. The price of good k is the units of each attribute a_{jk} times the price of each attribute $\tilde{p}_j$, summed over all the attributes:

$$p_k = \sum_{j=1}^{N} a_{jk}\tilde{p}_j. \tag{6}$$

INFORMATION Each firm generates n_{di} data points. Each data point is a signal about the demands for each attribute: $\tilde{\boldsymbol{s}}_{i,z} = \boldsymbol{b}_i + \tilde{\boldsymbol{\varepsilon}}_{i,z}$, where $\tilde{\boldsymbol{\varepsilon}}_{i,z} \sim N(\mathbf{0}, I)$ is an $N \times 1$ vector. Signal noises are uncorrelated across attributes and across firms.[3]

Because we are interested in how data affects competition, we take data (n_{di} and $\tilde{\Sigma}_e$) as given and exogenously change the amount of firms' data, n_{di}. Section VII explores what aspects of the results change when data is endogenously generated as a by-product of economic transactions.

We consider two possible information structures. In one, firms observe only their own data: $\mathcal{I}_i = \{\tilde{\boldsymbol{s}}_{i,z}\}_{z=1}^{ndi}$. In the other, data is public: all firms can observe every piece of data about every firm ($\mathcal{I}_i = \{\{\tilde{\boldsymbol{s}}_{i',z}\}_{z=1}^{ndi}\}_{i'=1}^{nF}$). This structure simplifies results and clarifies that the mechanism does not rely on asymmetric information. When more relevant data exists about a firm i, similar results arise, regardless of who else observes that data.

EQUILIBRIUM

1. Each firm chooses a vector of marginal costs $\tilde{\boldsymbol{c}}_i$, taking as given other firms' cost choices. Since the data realizations are unknown in this ex ante investment stage, the objective is the unconditional expectation of the utility in (3).

2. After observing the realized data, each firm updates beliefs with Bayes' law and then chooses the vector $\mathbf{q}_i$ of quantities to maximize conditional expected utility in (3), taking as given other firms' best responses.

3. Prices clear the market for each good.

[3]For now, we assume each firm has n_{di} data points about every attribute. Section VII relaxes this and allows data to differ by attribute. The assumption of signal variance I is then without loss. By Bayes' law, two independent data points function in the same way as one data point with twice the precision. By assuming $V(\tilde{\boldsymbol{\varepsilon}}_{i,z}) = I$, we are simply letting n_{di} govern data precision.

II.B Discussion of Assumptions

FIRMS THAT PRICE RISK. Pricing risk ($\rho_i > 0$) is not essential for the main aggregation results. The results cover the case where risk is not priced ($\rho = 0$). However, when firms face uncertain profits, evidence suggests we should take the effect of data on risk into account. In a randomized control trial, Kumar, Gorodnichenko, and Coibion (2023) provide some firm managers with predictive data and show that these firms invest more and produce more. This is risk pricing. It means that people do less activities that are perceived as risky and more of those activities when the risk is resolved.

While risk pricing is novel in the markups literature, it is a bedrock principle of corporate finance and is well supported by numerous empirical studies in many domains.[4] Because firms that take on risky projects will face a higher cost of capital, the price-of-risk term could be interpreted as an adjustment to their expected profit.[5] Our ρ_i term in (3) could capture both the price of risk and the covariance of the firm shock with market risk (the firm's beta).

In addition, our firms may price firm-specific risk because we are exploring a market with large players where firm-specific risk is not diversifiable. There is growing evidence that even idiosyncratic risk is priced, especially when firms face financial constraints (Hennesy and Whited 2007). Finally, it is also possible to interpret ρ_i as the absolute risk aversion of a firm manager who is compensated with firm equity.

DATA ABOUT CONSUMER DEMAND. One might question whether data is used to forecast demand or marginal cost. Conceptually, it shouldn't matter. Firms that face risk from their cost structure should also face a higher cost of becoming more productive. If data helps firms reduce profit risk, whether from the cost or the revenue side, it should embolden them to invest more and produce more at a lower market price. The same forces operate. Why then choose to model demand uncertainty? Markups are price divided by marginal cost. Having the random variable

[4]The Handbook of Empirical Corporate Finance fills chapter 18 with the evidence that firms price risk (Eckbo 2008). Most of the other chapters contain evidence in support of theories that are premised on risk pricing. For a textbook treatment of the topic, see Welch (2009), chapter 9. More recent work on this topic explores whether male and female CFOs are equally risk averse (Doan and Iskandar-Datta 2020). Management and psychology scholars (Lovallo et al. 2020) find that firms place too high a price on risk. In international economics, David, Ranciere, and Zeke (2022) document that multinational firms facing more risk hire less and compensate their capital owners with a greater share of income. Specifically, Brealey, Myers, and Allen (2003) argues for a price of risk ρ that matches the risk premium on the S&P 500. If a firm gets less return per unit of risk than this, the firm would be better off not investing in production and instead returning the cash to investors to invest in a market portfolio of equity.

[5]While a risk price affects a firm's cost of borrowing, simply including a risky interest rate in marginal cost does not suffice. The risk of debt is far less than the risk to the enterprise value of the whole firm. Markups need to compensate the firm for risk to its debt and its equity.

in the denominator makes it nearly impossible to characterize the average value of markups. If empiricists typically studied inverse markups, then it would be more practical to study cost uncertainty.

LINEAR DEMAND. We assume a linear demand system, which is common in the information aggregation literature.[6] Not only does the linearity assumption permit transparent results, recent work also shows that the linear setup fits the data well. Our model builds on Pellegrino (2020)'s Generalized Hedonic-Linear demand system, used to study market power in a network economy (see also Galeotti et al. [2022]). A feature of this model is the declining demand elasticity in firm size. This generates realistic higher markups for larger firms. Using the nonparametric estimates of Baqaee, Farhi, and Sangani (2021) for the demand, Ederer and Pellegrino (2022) show that the linear demand system fits the data better than the iso-elastic demand system.

However, if one wanted to change the relationship between elasticity and firm size, introducing a function $\phi(c_i)$ would only change the magnitudes of our results. The same trade-offs arise.

COMPETITION IN QUANTITIES. We model conduct by firms as competition in quantities because it holds market power ($\partial p/\partial q_i$) fixed. That assumption is surely wrong. But it holds fixed the non-data sources of markups, in order to see the effect of data more clearly. Of course, there are other conduct assumptions and frictions, including Bertrand competition, information frictions, transaction cost and barriers to entry. Appendix D.2. explores Bertrand competition. As is well-known, markups are lower under Bertrand than Cournot. Furthermore, while the risk effects may differ, our main results are robust.

NO PRICE DISCRIMINATION. We assume a uniform price for all consumers, at least for a given firm. The consumer demand uncertainty could represent uncertainty about how to best price-discriminate. The solution would be different. However, the idea that data reduces uncertainty and encourages firm growth, as well as the later results about covariances that information makes possible, would all still make sense. Since this agenda is still in its beginning stages, it makes sense to first understand uniform pricing, which corresponds ot the behavior of the vast majority of firms. We leave te price discrimination version of the model for future work.

[6] See Chapters 7–8 in Veldkamp (2011) for a textbook treatment.

GOODS AS BUNDLES OF ATTRIBUTES. We treat goods as collections of attributes, but this is not essential for our theoretical results. All results hold if $A = I$, in which case goods are attributes. However, the attribute structure creates correlation in demand across assets. That affects composition results and is important for measurement. To use this framework to measure a firm's data, it is crucial to recognize that information about one product can be informative about another. The correlated demand created by our attribute structure is what makes data relevant for multiple products.

Also, attribute-based demand is historically used in industrial organization (IO) economics because it allows researchers to predict what would happen if a new good was introduced.

NO ATTRIBUTE CHOICE. Appendix D.1. explores a model where a firm can choose the attributes of its good. The same forces are at work in that model. We choose to work with a simpler model without attribute choice to elucidate the main ideas more clearly.

NO VARIABLE CAPITAL COST. We made the investment in technology an up-front fixed cost. That means that the cost of capital is not part of the marginal cost that enters the markup calculation. One might object to that assumption on the grounds that the cost of capital is what captures the price of risk. Including a capital cost with a risk premium in marginal cost arguably absorbs the effect of risk on markups. This objection is tenuous. First, the capital cost is typically a borrowing cost. The risk premium on debt is not the same as the risk premium on equity. The firm cares about the variance of its cash flows, which is an equity claim. Second, the long-horizon risk that lenders care about is not the same as the short-term demand or cost fluctuations that data helps firms to forecast. These are substantially different risks. While including a variable capital cost with a risk premium in markup calculations probably improves their accuracy, this risk compensation has very little interaction with the way in which data helps to reduce operational uncertainties.

EXOGENOUS DATA. Section VII endogenizes data. The static forces are still present in that model and one new force emerges.

NO ENTRY OR EXIT. Adding entry would undoubtedly bring new insights. But that would also require a dynamic framework and a different paper. Since the static problem is not well understood, we start there. However, recent work by Baqaee and Farhi (2021) suggests that the aggregate distortions from market power are even larger once there is entry.

II.C General Solution

We solve the model by backwards induction, starting with the quantity choices and then working backwards to determine optimal firm investments in lowering marginal costs c_i.

OPTIMAL PRODUCTION The first-order condition with respect to goods production q_i is $\partial U_i / \partial \boldsymbol{q}_i$: $\mathbf{E}\left[\boldsymbol{p}_i|\mathcal{I}_i\right] - \boldsymbol{c}_i + \frac{\partial \mathbf{E}[\boldsymbol{p}_i|\mathcal{I}_i]}{\partial \boldsymbol{q}_i}\boldsymbol{q}_i - \rho_i \mathbf{Var}\left[\boldsymbol{p}_i|\mathcal{I}_i\right]\boldsymbol{q}_i \ = 0$. Rearranging delivers optimal production:

$$\boldsymbol{q}_i = \left(\rho_i \mathbf{Var}\left[\boldsymbol{p}_i|\mathcal{I}_i\right] - \frac{\partial \mathbf{E}\left[\boldsymbol{p}_i|\mathcal{I}_i\right]}{\partial \boldsymbol{q}_i}\right)^{-1}\left(\mathbf{E}\left[\boldsymbol{p}_i|\mathcal{I}_i\right] - \boldsymbol{c}_i\right). \tag{7}$$

The second term tells us that firms produce more of goods that have high expected prices, relative to their marginal costs. The first term tells us that uncertainty (conditional variance) or market power cause the firm to scale back their production response to changes in expected profit. By improving forecasts, data reduces uncertainty.

A key reason one should think about priced risk in this context is that risk mimics market power. Because market power enters only through the sensitivity term H in (7), more market power is mathematically equivalent to increasing the conditional variance $\mathbf{Var}\left[\boldsymbol{p}_i|\mathcal{I}_i\right]$. Both risk and market power restrain production. Both make firms less sensitive to expected changes in price or cost. In one case, it is because a risk-averse firm makes more conservative production decisions to manage its risk. In the other case, the firm makes more conservative decisions to minimize its price impact. This is one reason that data and market power are difficult to disentangle.

From differentiating the attribute pricing function (4), we find that the price impact of one additional unit of attribute output is

$$\frac{\partial \mathbf{E}\left[\tilde{\boldsymbol{p}}_i|\mathcal{I}_i\right]}{\partial \tilde{\boldsymbol{q}}_i} = -\frac{1}{\phi} I_N. \tag{8}$$

Define the sensitivity of production to a change in expected profit as

$$\hat{\boldsymbol{H}}_i \equiv \left(\rho_i \mathbf{Var}\left[\boldsymbol{p}_i|\mathcal{I}_i\right] + \frac{\boldsymbol{I}_N}{\phi}\right)^{-1}. \tag{9}$$

To simplify the problem, we can change the choice variable and have firms choose the optimal vector of attribute production $\tilde{\boldsymbol{q}}_i$. If we rewrite (7), replacing q, p, and c with attribute quantities, prices, and costs $\tilde{\boldsymbol{q}}_i$, $\tilde{\boldsymbol{p}}_i$, and $\tilde{\boldsymbol{c}}_i$, then we can substitute in the price impact (8) and conditional

expectation (10) to get $\tilde{\boldsymbol{q}}_i = \hat{\boldsymbol{H}}_i \left(\overline{\boldsymbol{p}} + \boldsymbol{K}_i \boldsymbol{s}_i - \frac{1}{\phi} \sum_{i'} \tilde{\boldsymbol{q}}_{i'} - \tilde{\boldsymbol{c}}_i \right)$.

II.D Solution with Public Data and Firm-Specific Shocks

While not as realistic, the special case of firm-specific shocks ($\xi = 0$) and public data allows us to more clearly illustrate the model's mechanics and the main economic forces of data.

BAYESIAN UPDATING According to Bayes' law for normal variables, observing n_{di} signals, each with signal noise variance $\tilde{\Sigma}_e$, is the same as observing the average signal $\boldsymbol{s}_i = (1/n_{di}) \sum_{z=1}^{n_{di}} s_{iz} = \boldsymbol{b}_i + \boldsymbol{\varepsilon}_i$, where the variance of $\boldsymbol{\varepsilon}_i$ is $\Sigma_{\epsilon_i} = \tilde{\Sigma}_e / n_{di}$. Therefore, do a change of variable, replacing $\tilde{\Sigma}_e / n_{di}$ with Σ_{ϵ_i}. In this representation, more data points (higher n_{di}) shows up as a lower composite signal noise Σ_{ϵ_i}.

Define $\boldsymbol{K}_i$ to be the sensitivity of price beliefs to the signal s_i: $\boldsymbol{K}_i := (I_N + \Sigma_{\epsilon_i})^{-1}$.[7] Then, firm i's expected value of the shock $\boldsymbol{b}_i$ can be expressed simply as $\mathbf{E}[\boldsymbol{b}_i | \mathcal{I}_i] = \boldsymbol{K}_i \boldsymbol{s}_i$. The expectation and variance of the pricing function (4) are

$$\begin{aligned} \mathbf{E}\left[\tilde{\boldsymbol{p}}_i | \mathcal{I}_i\right] &= \bar{\boldsymbol{p}} + \boldsymbol{K}_i \boldsymbol{s}_i - \frac{1}{\phi} \sum_{i'=1}^{n_F} q_{i'}, \\ \mathbf{Var}\left[\tilde{\boldsymbol{p}}_i | \mathcal{I}_i\right] &= \mathbf{Var}\left[\boldsymbol{b}_i | \mathcal{I}_i\right] = \left(I_N + \Sigma_{\epsilon_i}\right)^{-1} \Sigma_{\epsilon_i}. \end{aligned} \tag{10}$$

OPTIMAL PRODUCTION Next, sum production $\tilde{\boldsymbol{q}}_i$ over all firms i to get total production of each attribute $\sum_{i'} \tilde{\boldsymbol{q}}_{i'}$. This sum has a $\sum_{i'} \tilde{\boldsymbol{q}}_{i'}$ on both the left- and right-hand sides. Collect these terms and rearrange to get $\sum_{i'} \tilde{\boldsymbol{q}}_{i'} = \left(I + \frac{1}{\phi} \sum_i \hat{\boldsymbol{H}}_i \right)^{-1} \left[\sum_i \hat{\boldsymbol{H}}_i \left(\overline{\boldsymbol{p}} + \boldsymbol{K}_i \boldsymbol{s}_i - \widetilde{\boldsymbol{c}}_i \right) \right]$. Substituting this total production expression for $\sum_{i'=1}^{n_F} \tilde{\boldsymbol{q}}_{i'}$ in firm i's optimal production ($\tilde{\boldsymbol{q}}_i^*$) yields the optimal production of each attribute by each firm i:[8]

$$\widetilde{\boldsymbol{q}}_i^* = \hat{\boldsymbol{H}}_i \left(\overline{\boldsymbol{p}} + \boldsymbol{K}_i \boldsymbol{s}_i - \widetilde{\boldsymbol{c}}_i - \left(I_N + \frac{1}{\phi} \sum_{i'} \hat{\boldsymbol{H}}_{i'} \right)^{-1} \left[\sum_{i'} \hat{\boldsymbol{H}}_{i'} \left(\overline{\boldsymbol{p}} + \boldsymbol{K}_{i'} \boldsymbol{s}_{i'} - \widetilde{\boldsymbol{c}}_{i'} \right) \right] \right).$$

Finally, the product-level optimal production function is the attribute weighting matrix A times the optimal attribute production: $\boldsymbol{q}_i^* = A \tilde{\boldsymbol{q}}_i$.

[7] In a dynamic model, K_i would be called the Kalman gain.

[8] Since all signals are normally distributed, this formula does tell us that production can potentially be negative. We could bound choices to be non-negative, but this would make analytical solutions for covariances impossible. If parameters are such that all firms want negative production of a good or attribute, then the solution is simply to redefine the product as its opposite. In the numerical results, we simply choose parameters that make negative production extremely unlikely.

EQUILIBRIUM PRICE Substituting this aggregate quantity in the pricing function (4) yields an equilibrium average price of each attribute:

$$\tilde{\mathbf{p}}^M = \bar{p} - \left(I_N + \frac{1}{\phi} \sum_i \hat{\boldsymbol{H}}_i \right)^{-1} \left[\sum_i \hat{\boldsymbol{H}}_i \left(\overline{\boldsymbol{p}} + \boldsymbol{K}_i \boldsymbol{s}_i - \widetilde{\boldsymbol{c}}_i \right) \right]. \tag{11}$$

The average price of a good k with attribute vector a_k is then simply $p_k^M = \mathbf{a}_k' \tilde{\mathbf{p}}$, and firm i price of good k is $\mathbf{a}_k'(\tilde{\mathbf{p}}^M + \mathbf{b}_i)$.

OPTIMAL INVESTMENT CHOICES Firm i chooses cost $\boldsymbol{c}_i$ to maximize its unconditional expected utility $\mathbf{E}[U_i]$, taking all other firms' investment choices as given.

The optimal cost $\boldsymbol{c}_i$ for an interior solution satisfies (see Appendix A. for derivation):

$$\frac{\partial \mathbf{E}[U_i]}{\partial \tilde{\boldsymbol{c}}_i} = \frac{1}{2} \frac{\partial \mathbf{E}[\tilde{\boldsymbol{q}}_i]' \left(\frac{2}{\phi} I_N + \rho_i \mathbf{Var}\left[\boldsymbol{b}_i | \mathcal{I}_i \right] \right)^{-1} \mathbf{E}[\tilde{\boldsymbol{q}}]}{\partial \tilde{\boldsymbol{c}}_i} - \frac{\partial g(\chi_c, \tilde{\boldsymbol{c}}_i)}{\partial \tilde{\boldsymbol{c}}_i} = 0, \tag{12}$$

The first term is the marginal benefit. Lower production costs enable production at a greater scale and higher profit per unit. The second term is the marginal cost of the up-front investment.

II.E Solution with Private Data and Common Shocks

The optimal production takes the same form as before. The difference is in the expectation $\mathbf{E}\left[\boldsymbol{p}_i | \mathcal{I}_i\right]$ and the conditional variance.

The solution to this model is complicated by firms' need to forecast what other firms know, as in Angeletos and Pavan (2007). Because firms do not know other firms' data, they face strategic uncertainty. They use their own data to forecast what other firms will do. Data thus reduces risk in two ways—by predicting demand for the firm's products and by predicting the production decisions of other firms. This strengthens the risk channel because data reduces both demand uncertainty and strategic uncertainty.

Lemma 1. *With private data and common shocks, the equilibrium price takes the form*

$$\begin{aligned} \boldsymbol{p} &= \bar{\boldsymbol{p}}^M + \boldsymbol{F}\boldsymbol{b} - \frac{1}{\phi} \sum_{i=1}^{n_F} \hat{\boldsymbol{H}}_i \boldsymbol{K}_i \boldsymbol{\varepsilon}_i \qquad \textit{where} \\ \bar{\boldsymbol{p}}^M &= \left(\boldsymbol{I}_N + \frac{1}{\phi} \sum_{i=1}^{n_F} \hat{\boldsymbol{H}}_i \right)^{-1} \left(\bar{\boldsymbol{p}} + \frac{1}{\phi} \sum_{i=1}^{n_F} \hat{\boldsymbol{H}}_i \boldsymbol{c}_i \right) \qquad \textit{and } \boldsymbol{F}\ \boldsymbol{K}_i \textit{ and} \hat{\boldsymbol{H}}_i \textit{ are reported in Appendix B.} \end{aligned} \tag{13}$$

Solving this problem requires a technical innovation. Bayesian updating weights always depend on the covariance between the observed data s_i and the price the firm needs to forecast p_i. But, in this model, that covariance is endogenous. It depends on firms' production choices. To solve the fixed point of beliefs and output requires a state-space updating approach for informationally copmlex environments, as in Lambert, Ostrovsky, and Panov (2018). A state-space approach defines all the relevant model objects in terms of exogenous, orthogonal shocks and weights on those shocks. We extend this approach to the case of endogenous weights on each shock. To do that, choose Bayesian weights that maximize firms' objectives, given the beliefs that result.

When data is firm-specific, more data always reduces uncertainty. But when data is aggregate, it is possible for data to increase uncertainty about the price.[9] Since a firm would never choose data that raises uncertainty, going forward, we assume that the risk price of all firms is sufficiently low to ensure that $\partial \hat{\boldsymbol{H}}_i(j,j)/n_{di}(j) \geq 0 \ \forall j$. The parameter bound also ensures that when firm i acquires more data about attribute j, it does not make other firms less uncertain about the price of attribute j: $\partial \hat{\boldsymbol{H}}_{i'}(j,j)/n_{di}(j) \ \forall j$.

III Data and Product Markups

This section does not contain the main results. Higher or lower markups can be explained by many factors. However, these are a stepping stone to the aggregation results, which are more specific to data. We begin by exploring just two of the ways in which data affects markups, through cost and risk. The goal in this section is simply to establish the foundation of ideas upon which we build later.

By reducing the uncertainty a firm faces about consumer demand, data encourages the firm to produce more for a given level of investment. Reducing uncertainty also emboldens the firm to invest more in infrastructure that enables them to produce at a lower marginal cost. These two forces have opposite effects on markups. More production lowers prices, which in turn lowers markups. More initial investment lowers marginal cost, which raises markups. This section explores that tension.

[9] The reason is that if one firm is better informed, their beliefs might be less noise and more predictable to a competitor. Because the competitor can predict his rival's action, the competitor's uncertainty declines, causing the competitor to produce more aggressively, in response to his data. This aggressive response of the competitor to data unknown to the original firm could, in theory, overcome the original decline in uncertainty about the demand shock and cause a rise in uncertainty.

Definition 1 (Product markup). *The product-level markup for product k produced by firm i is* $M_{ik}^p := \mathbb{E}[\boldsymbol{p}_i(k)]/\boldsymbol{c}_i(k)$. *The average product-level markup is* $\overline{M}^p := \frac{1}{Nn_F}\sum_{i=1}^{n_F}\sum_{k=1}^{N} M_{ik}^p$.

To derive an expression for the product markup in the model, we simply divide each expected product price, using (11) and $E[\boldsymbol{b}_i] = 0$, by the marginal cost of that product, $\boldsymbol{c}_i = a_k'\widetilde{\boldsymbol{c}}_i$:

$$M_{ik}^p = \frac{1}{a_k'\widetilde{\boldsymbol{c}}_i} a_k' \left(\overline{\boldsymbol{p}} - \left(I_N + \frac{1}{\phi}\sum_{i=1}^{n_F} \hat{\boldsymbol{H}}_i \right)^{-1} \left(\sum_i \hat{\boldsymbol{H}}_i \boldsymbol{A} \left(\overline{\boldsymbol{p}} + \boldsymbol{K}_i \boldsymbol{s}_i - \widetilde{\boldsymbol{c}}_i \right) \right) \right). \tag{14}$$

Similarly, the average product markup for firm i is $\overline{M}_i^p = (1/N)\sum_{k=1}^{N} M_{ik}^p$.

Some of these causes of high markups in equation (14) are not surprising. For example, having lots of valuable attributes (with high expected value $\bar{p}$), fewer firms (low n_F) or low price elasticity (low ϕ) raise markups. Two forces show up in the markup formula that are affected by how much data a firm has. Those forces are risk and cost. We state and explain each in turn.

DATA, INVESTMENT, OUTPUT, AND MARKUPS The first two results encapsulate the standard logic about data and competition: Data enables firms to grow larger (invest more). These larger firms charge higher markups.

Lemma 2. ***Data-investment complementarity.*** *A firm with more data chooses a lower marginal cost* c_i*, which entails a higher cost investment and higher profitability* Π_i.

The proofs of this and all further results are in Appendix B.. The role of investment in data is to reduce the conditional variance of the firm's stochastic demand, which encourages the firm to produce more. Data increases the expected revenue of a firm by allowing it to produce more in states in which the price will be high. It also reduces the uncertainty around that investment and lowers the risk of the firm. Both of these effects increase the marginal benefit of production and the marginal benefit of investment. What this means is that high-data firms invest more and grow larger. As the next result shows, higher investment is also a channel through which data increases product markups.

Lemma 3. ***Higher investment raises product markups.*** *More investment (lower* c_i *choice) in any attribute j of good k, s.t.* $a_{jk} > 0$*, increases the markup of attribute j. If the markup on attribute j is less than the markup on product k (*$M_{i,k} \geq M^{\tilde{p}_{i,j}}$*), then this also raises the markup on good k.*

A firm that invests in producing an attribute can produce that attribute at a lower cost. If a good j does not load at all on that attribute ($a_{jk} = 0$), then the lower cost has no bearing on the

cost or markup of that good. But if good j contains some of that attribute ($a_{jk} > 0$), this investment lowers the cost of producing the good. Since markups are price divided by marginal cost, a lower cost raises the markup. Of course, a lower cost also lowers the equilibrium price of the good. However, the proof shows that price does not fall as much as cost. Therefore, the markup rises.[10]

The model teaches us that there is a second channel through which data affects markups: data reduces the risk of production, induces more production, and thereby lowers prices and markups.

Lemma 4. ***Data reduces product markups (risk premium channel).*** *Holding all firms' investments fixed (χ_c sufficiently high), an increase in any firm's data about any attribute of good k reduces the markup of good k.*

Data reduces markups because it reduces the risk in production. This induces firms to produce more. This effect can be seen in the firm's first-order condition (7) where the conditional variance in the denominator represents risk. When this variance declines, optimal production rises. More production lowers price and lowers markups. This force shows up in the markup (14) as high ρ makes H_i and $\bar{H}$ low. When we reduce risk with data, firms do not need as much markup compensation to be willing to produce.

The restriction on χ_c is there to shut down then investment channel, to isolate the risk effect. When χ_c or ρ is low, this risk premium channel is still present. But it may be overpowered by the investment channel working in the opposite direction.

Proposition 1. ***Data in(de)creases product markups when risk price or marginal cost of investment is sufficiently low (high).*** *If the price of risk ρ is sufficiently low or the investment cost χ_c is sufficiently low, then an increase in any firm's data about any attribute of good k increases the markup of good k, which loads positively on that attribute. Otherwise, an increase in any firm's data about any attribute of good k reduces the markup of good k.*

When firm investments greatly decrease marginal cost (low χ_c), then the cost channel is dominant and more data primarily increases investment, lowers costs, and raises markups. When the cost-reduction investment is inefficient (high χ_c), then data still prompts more investment, but this

[10] To see why not every attribute markup increase raises the product markup, consider a numerical example where a product uses 99% of an attribute with a price 101 and cost 100 and 1% of an attribute that costs 1 and has a price 5. The product markup is $\frac{99\%\cdot 101+1\%\cdot 5}{99\%\cdot 100+1\%\cdot 1} \approx 1.10403$. Now suppose we decrease the cost of the second attribute from 1 to 0.9 and reduce its price from 5 to 4.6. This implies that the second attribute's markup increases from 5 to 5.11. The new product markup is $\frac{99\%\cdot 101+1\%\cdot 4.6}{99\%\cdot 100+1\%\cdot 0.9} \approx 1.10373$, which is less than the original markup of 1.10403. Thus lowering the cost of a low-markup attribute can increase the product markup. This example is carefully contrived and not likely to be relevant. But it explains the reason for the parameter restriction in the Lemma.

has little effect on marginal cost. Instead, the dominant force is risk reduction. Similarly, if the price of risk is high, risk reduction is also the dominant force. A data-rich firm faces less cost from taking on more risk with a large production plan. By producing more, data-rich firms drive prices down and lower markups. Which scenario prevails depends on the strength of each force in a particular industry.

The main point is not whether data increases or decreases the markup. It is that, despite holding market power fixed, markups are contaminated by firms' use of data. To solve this problem, we need to know how much data firms have.

IV Measuring Markups and Measuring Data

In empirical work, markups are often measured at the firm or industry level. Measuring markups at these more aggregated levels often yields different answers about how competition is evolving. The next set of results show why aggregate markups differ from product-level markups in ways that vary systematically with the amount of data firms have. The difference between a firm's product- and firm-level markups turns out to be a good bound for the amount or quality of data that a firm must have.

These composition effects are quantitatively important. De Loecker, Eeckhout, and Unger (2020) find that two-thirds of the increase in measured industry markups comes from such composition effects. Crouzet and Eberly (2018) link the trend increase in markups to intangible assets, a broader category that includes data assets. They find that intangible-abundant firms have higher markups and that intangible-abundant industries have even higher markups. The results that follow contribute to this discussion by explaining why firms' use of predictive data can generate such statistical patterns and providing tools to measure firms' data.

IV.A Firm Markups

Economists have long known that difference in markup measurement at different levels of aggregation represent composition effects. What is less well understood is why such composition effects might change. We show how firms' data accumulation naturally gives rise to changes in the composition of firms' products. Data is what makes it possible for the firm to skew the composition of their products in the direction of high-markup goods. So, data strengthens the composition effect and makes firm markups larger and larger relative to that firm's average product markup.

Definition 2 (Firm Markup). *The firm markup for firm i is the firm's revenue divided by the firm's total variable costs:*

$$M_i^f := \frac{\mathbf{E}[\boldsymbol{q}_i' \boldsymbol{p}_i]}{\mathbf{E}[\boldsymbol{q}_i' \boldsymbol{c}_i]}. \tag{15}$$

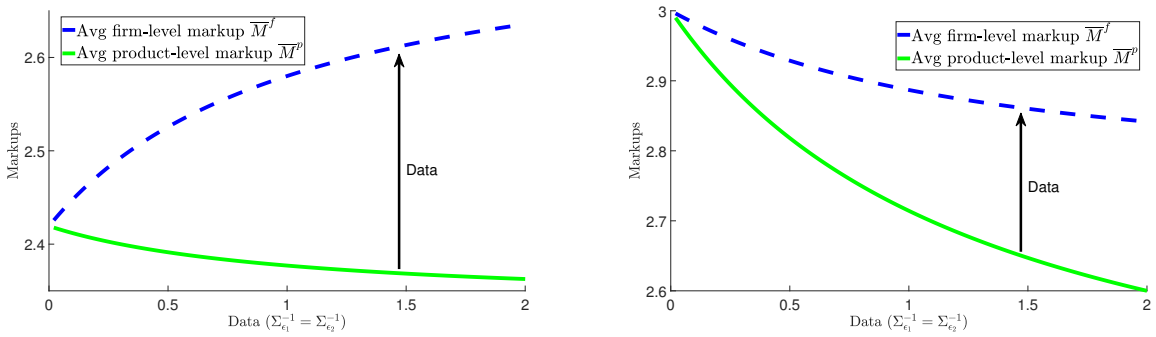
We can rewrite the expectation of the product of price and quantity as the product of expectations, plus a covariance term (trace of matrix):

$$M_i^f = \frac{\mathbf{E}[\boldsymbol{q}_i]' \mathbf{E}[\boldsymbol{p}_i] + \mathbf{tr}\left[\mathbf{Cov}(\boldsymbol{p}_i, \boldsymbol{q}_i)\right]}{\mathbf{E}[\boldsymbol{q}_i' \boldsymbol{c}_i]} = \underbrace{\frac{\sum_{j=1}^N M_{ij}^p \boldsymbol{c}_i(j) \mathbf{E}[\boldsymbol{q}_i(j)]}{\sum_{j=1}^N \boldsymbol{c}_i(j) \mathbf{E}[\boldsymbol{q}_i(j)]}}_{\text{Cost-weighted product markups}} + \frac{\mathbf{tr}[\mathbf{Cov}\left(\boldsymbol{p}_i, \boldsymbol{q}_i\right)]}{\sum_{j=1}^N \boldsymbol{c}_i(j) \mathbf{E}[\boldsymbol{q}_i(j)]}. \tag{16}$$

The second equality just comes from using the definition of the product markup to substitute: $\mathbf{E}[\boldsymbol{p}_i] = M_i^{p\prime} \boldsymbol{c}_i$ and then rewriting the vector products as sums. We learn that the firm markup is a cost-weighted sum of product markups, plus a term that depends on the covariance of prices and quantities. Firm data acts on this covariance term. It allows firms to produce more of goods that turn out to have high demand and thus high price.

Proposition 2. ***Data accumulation widens the wedge between product and firm markups.*** *Holding all firms' investments fixed ($(c_1, ..., c_{nF})$ given), an increase in firm i's data about any attribute increases $E[M_i^f - \overline{M}_i^p]$.*

Firm markups rise when data increases the covariance of firm's production decision q_i with the price p_i in (16). Without any data to predict demand, this covariance is low because firms cannot know which markups would be high and which goods to produce more of. The positive effect of data on the price-quantity covariance shows up in the production first-order condition (7), where a reduction in the conditional variance of demand makes production decisions q_i more sensitive to expected changes in price p_i. That higher sensitivity is a higher covariance.



(a) Low price of risk.

(b) High price of risk.

Figure 1: More data may raise or lower markups but always causes product and firm markups to diverge. Parameters used: $\bar{p} = 5$, $\phi = 0.1$, and $A = I$. Firm marginal costs are not chosen here. They are fixed as $c_1 = c_2 = 1$. On the left, $\rho_1 = \rho_2 = 1$. On the right, $\rho_1 = \rho_2 = 10$.

AN ILLUSTRATIVE EXAMPLE OF THE PRODUCT-FIRM MARKUP WEDGE To illustrate the mechanisms at work, Figure 1 plots the competing effects data has on product and firm/industry markups in a specific example. When the price of risk is high, the product-level markup falls as both firms' data rises. The reason the product markup is falling is that data is resolving risk. It is allowing the firms to be less uncertain because data allows them to forecast demand more precisely. Firms that are less uncertain require a lower markup to compensate them for the lower risk. When the price of risk is low, more data may result in higher firm markups, as high-data firms invest, grow, and lower their marginal costs.

What the model teaches us so far is that increases or decreases in markups, at either the product level or the firm level, are not indicative of a firm that has a larger stock of data. As a firm accumulates more data, both product and firm markups may increase, both may decrease, or they may move in opposite directions. Instead, data governs the difference in markups. Data changes the composition of products and firms and makes various measures of markups diverge. This is a theme that will recur as we proceed to explore markups at the industry level.

IV.B Industry Markup Definitions

Typically, researchers are interested in the markup for an industry because the regulatory question of interest is whether that industry is a competitive one or not. However, there are multiple ways to aggregate markups into a single industry measure. We examine four of the most common measures inside the model. The model lends an interpretation to the differing trends arising from the different ways empirical researchers measure industry markups.

Definition 3. *The equally-weighted average firm markup in an industry is*

$$\bar{M}^f := (1/N) \sum_{i=1}^{n_F} M_i^f. \tag{17}$$

Definition 4. *The cost-weighted markup for an industry is*

$$M^c := \sum_{i=1}^{n_F} w_i^c M_i^f \quad \textit{where cost weights are} \quad w_i^c = \frac{\mathbf{E}\left[\boldsymbol{q}_i' \boldsymbol{c}_i\right]}{\sum_{i=1}^{n_F} \mathbf{E}\left[\boldsymbol{q}_i' \boldsymbol{c}_i\right]}. \tag{18}$$

In contrast to the unweighted $\bar{M}^f$, the cost-weighted markup M^c weights larger firms more. Larger firms that produce more have larger variable costs, not per unit, but in total. The next definition also weights larger firms more, where larger is based on gross revenues.

Definition 5. *The sales-weighted markup is*

$$M^s := \sum_{i=1}^{n_F} w_i^s M_i^f \quad \textit{where sales weights are} \quad w_i^s = \frac{\mathbf{E}\left[\boldsymbol{q}_i' \boldsymbol{p}_i\right]}{\sum_{i=1}^{n_F} \mathbf{E}\left[\boldsymbol{q}_i' \boldsymbol{p}_i\right]}. \tag{19}$$

Definition 6. *The industry- aggregates markup is*

$$M^{ind} := \frac{\mathbf{E}\left[\sum_{i=1}^{n_F} \boldsymbol{q}_i' \boldsymbol{p}_i\right]}{\mathbf{E}\left[\sum_{i=1}^{n_F} \boldsymbol{q}_i' \boldsymbol{c}_i\right]}. \tag{20}$$

The industry-aggregates markup uses data already aggregated at the industry level. It is the ratio of the total industry sales over the total industry variable cost.

IV.C Data's Effect on Industry Markup Measures

Industry aggregation effects from the covariance of data and firm size. Data-investment complementarity means that more data makes larger up-front investment (larger firms) optimal. So, high-data, large firms are weighted more, relative to the unweighted firm average. High-data firms use data to skew their production toward high-markup goods (Proposition 2). Thus, the measures that weight large, high-data firms more will also weight high-markup firms more, generating a higher predicted industry markup.

Proposition 3. ***Growing data increases the wedges between industry markup measures.*** *An increase in firm i's data about any attribute*

a. increases the difference between cost-weighted and unweighted firm markups $E[M^c - \overline{M}^f]$, if the price

of risk is not too high ($\rho_i < \bar{\rho}$);

If, in addition, all firms are initially symmetric, then an increase in firm i's data about any attribute

b. *increases the difference between sales-weighted and cost-weighted markups $E[M^s - M^c]$, and*

c. *increases the difference between the sales-weighted and industry-aggregates markup $E[M^s - M^{ind}]$.*

Mathematically, the key to each of these results is a covariance. In the first case (a), the covariance is between the firm markup and the total production of a firm. If the risk channel is not so strong that it overpowers both the cost channel and the firm markup aggregation effect, then high-data firms are high-markup firms and these firms get weighted more than small firms by the cost weights.

Economically, this effect arises because data has economies of scale. Firms get the most value from their data if they grow large. The way they get value from data is to use the data to forecast which goods are high-margin and produce more of them. Thus, more data increases the covariance between size and markups and makes the aggregate markup larger than the average firm markup.

In cases (b) and (c), the key covariance is between a firm's markup and the firm's revenue, relative to its costs. High-data firms are firms that are able to produce more of the products that have high price relative to their cost of production. Therefore, these high-data firms have higher sales-weighted markups relative to their cost-weighted markups.

Part (c) follows from (b) because industry-aggregates markups are identical to cost-weighted markups:

$$M^c := \sum_{i=1}^{N} \frac{\mathbf{E}\left[\boldsymbol{q}_i' \boldsymbol{c}_i\right]}{\sum_{i=1}^{N} \mathbf{E}\left[\boldsymbol{q}_i' \boldsymbol{c}_i\right]} M_I^f = \sum_{i=1}^{N} \frac{\mathbf{E}\left[\boldsymbol{q}_i' \boldsymbol{c}_i\right]}{\sum_{i=1}^{N} \mathbf{E}\left[\boldsymbol{q}_i' \boldsymbol{c}_i\right]} \frac{\mathbf{E}\left[\boldsymbol{q}_i' \boldsymbol{p}_i\right]}{\mathbf{E}\left[\boldsymbol{q}_i' \boldsymbol{c}_i\right]} = \frac{\mathbf{E}\left[\sum_{i=1}^{N} \boldsymbol{q}_i' \boldsymbol{p}_i\right]}{\mathbf{E}\left[\sum_{i=1}^{N} \boldsymbol{q}_i' \boldsymbol{c}_i\right]} := M^{ind}. \tag{21}$$

Identical holds in theory. In practice, with different sources of measurement error at the firm and aggregate level, they differ somewhat, but have similar trends.

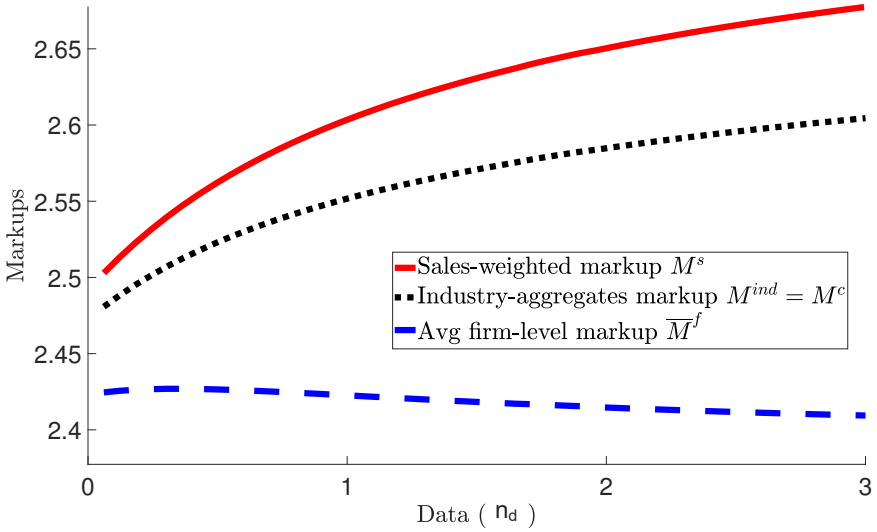


Figure 2: Data Accumulation Makes Industry Markup Measures Diverge. Investment cost function is $g(\chi_c, c_i) = \chi_c / c_i^2$, with $\chi_c = 1$. Parameters are $\bar{p} = 5$, $\rho_1 = 1$, $\rho_2 = 5$, $\phi = 0.8$, and $A = I$. Firm 1's data is measured on the x-axis. Firm 2's data is fixed at $\Sigma_{\epsilon_2}^{-1} = 1$.

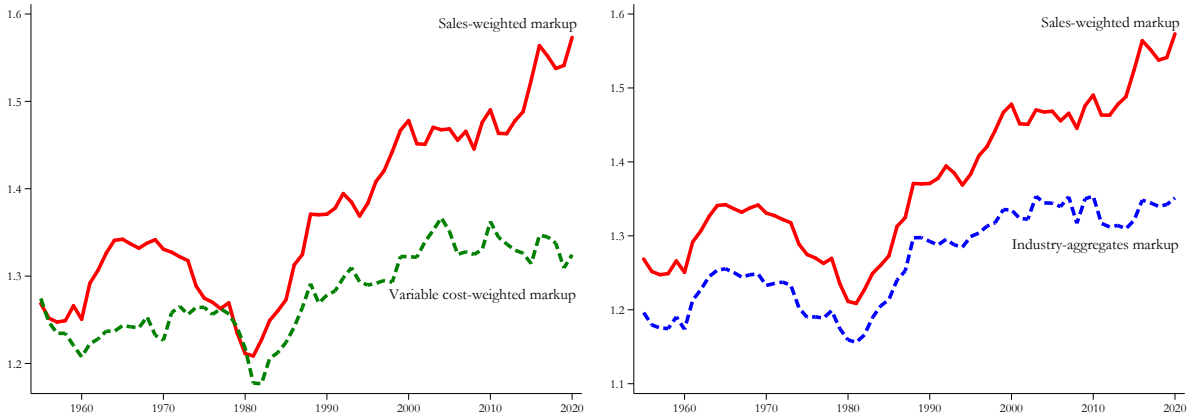
AN ILLUSTRATIVE EXAMPLE OF INDUSTRY MARKUP DIVERGENCE Figure 2 illustrates the divergence. The firm-level markup (dashed line at the bottom) rises less than the industry-aggregates markup (middle dotted line), which rises less than that sales-weighted markup (solid, top line). That result suggests that as firms process more data over time, the differences between markups measured at various levels of aggregation will continue to grow.

With aggregate markups, there are now four ways in which data affects markups, aside from true changes in market power. Data increases markups because of reduced cost, cross-product aggregation, and cross-firm aggregation. Data decreases markups because it induces firms to produce more (risk premium channel).

IV.D Empirical Evidence from Industry Markups

The empirical literature finds that there is a wedge between the sales-weighted markup and the cost-weighted markup, and that this wedge is growing over time since the early 1980s (see Figure 3a, from De Loecker, Eeckhout, and Unger (2020)). Firms that have market power sell at higher prices and therefore have higher revenue and relatively lower costs. This difference between sales and costs therefore drives a wedge between sales- and cost-weighted markup measures. This is consistent with what we find as firms that have market power boost their sales with fewer inputs since they have higher markups. In our model, firms who invest heavily in data do exactly that,

and the more important the role of data, the bigger the wedge between the input- and output-weighted aggregate markup. Our contribution is to propose a theory based on the role of data in creating these wedges, and how they grow as the data becomes more important.



(a) Sales-weighted markups, M^s, (solid line) vs. cost-weighted markups, M^c, (dashed line)

(b) Sales-weighted markups, M^s, (solid line) vs. industry-aggregates markups, M^{ind}, (dashed line)

Figure 3: Markups Measured and Aggregated in Different Ways Diverged Over Time.
Note. From De Loecker, Eeckhout, and Unger (2020) (updated), Figure XVI.A (left panel) and Figure V (right panel).

Our theory predicts differences between product, firm, and industry markups. To date, there is still limited evidence comparing product versus firm markups using the same data source. However, there is consistent evidence comparing firm markups to industry markups. In fact, the seminal paper on markup measurement by means of the production approach by Hall (1988) uses industry, not firm-level, data to construct aggregate markup measures (see also Hall [2018] for recent industry estimates using KLEMS data). With firm-level data and industry classification codes, we can mimic the industry aggregates using exactly the same set of firms underlying the industry aggregates. Based on De Loecker, Eeckhout, and Unger (2020) using data on publicly traded firms, Figure 3b shows that industry markups (blue dashed line) have increased by half as much as sales-weighted firm markups (red line). In other words, they find that there is a wedge between the industry markup and the sales-weighted firm markup, and that wedge is increasing as investment in data increases. Note that industry markups (in Figure 3b) look remarkably similar to cost-weighted firm markups (in Figure 3a). This is due to the systematic relation between input-weighting and industry aggregates in equation (21).

V Cyclicality of Markups

A key question for mainstream New Keynesian models of the type often used by central banks is whether markups are countercyclical. This question has created stark disagreement. Researchers who measure markups at the firm or industry level find clear evidence of countercyclical markups (Bils 1985, 1987). In contrast, researchers who measure markups at the product level do not find evidence of countercyclicality (Nekarda and Ramey 2020). Our model offers a way to reconcile these facts.

Our explanation builds on the progress in Burstein, Carvalho, and Grassi (2020). They show analytically how composition changes can turn procyclical markups into countercyclical ones, depending on how markups are aggregated. Our model provides a specific economic mechanism for these composition changes. The cyclical markup evidence, in turn, supports the realism of the model's assumptions.

To use the model to explore the cyclicality of markups, we first need to define what is a boom or recession in the context of this model. There are two relevant features of a boom: 1) demand rises and 2) the variances of demand and of output fall. In contrast, recessions are volatile, uncertain times. To formalize this new assumption, we introduce a variable $Boom \in \{0, 1\}$ that makes the level of demand procyclical and the demand variance countercyclical:

$$\bar{p} = d_0 + d_1 * Boom \qquad \text{where } \; d_0, d_1 \geq 0, \tag{22}$$

$$\Sigma_b = d_2 - d_3 * Boom \qquad \text{where } \; d_2, d_3 \geq 0. \tag{23}$$

High demand in a boom (22) regulates how countercyclical or acyclical product markups are. Falling variance in a boom (23) is what makes the cyclical behavior of aggregate markups differ relative to product markups. The second statement is formalized in the next proposition.

Proposition 4. ***Product markups diverge from firm and industry markups when volatility rises.*** *Suppose the investment cost structure is such that firms choose identical investments ($c_i = c_j \;\; \forall \, i, j$).*

a. *The product-level markup is strictly increasing in demand variance, $\partial \mathbf{E}[M_{ij}^p] / \partial \Sigma_{b,j} > 0$, and converges to a constant as $\Sigma_{b,j} \to \infty$.*

b. *If demand variance is large enough, firm and industry markups are strictly increasing, $\partial \mathbf{E}[M_{ij}^f] / \partial \Sigma_{b,j} > 0$ and $\partial \mathbf{E}[M_{ij}^m] / \partial \Sigma_{b,j} > 0$, and asymptotic to a function increasing in variance,*

$\lim_{\Sigma_{b,k}\to\infty} \partial \mathbf{E}[M_{ij}^f]/\partial \Sigma_{b,j}, \partial \mathbf{E}[M_{ij}^m]/\partial \Sigma_{b,j} > 0.$

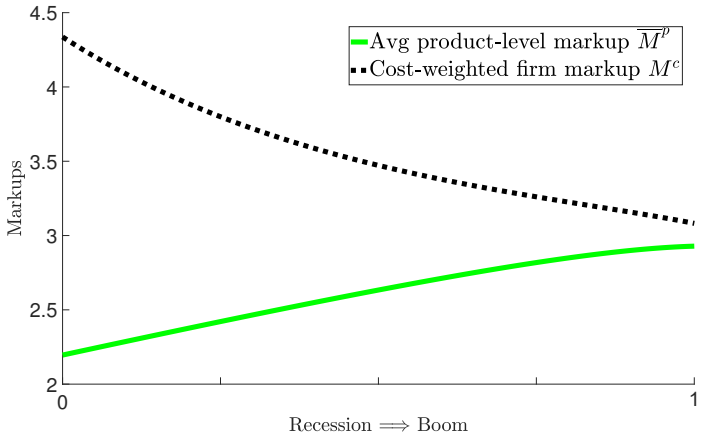


Figure 4: Procyclical product markups can coexist with countercyclical firm / industry markups. Left (right) on the x-axis represents recessions (booms), as described in (22) and (23), where $d_0 = 7/2$, $d_1 = 2$, $d_2 = 5$, and $d_3 = 4$. A decreasing line represents a countercyclical markup. Remaining parameters are $\rho_1 = \rho_2 = 1$, $c_1 = c_2 = 1$, $\phi = 1$, and $\Sigma_{\epsilon_1} = \Sigma_{\epsilon_2} = 1$.

The coexistence of a procyclical product markup and a countercyclical firm or industry markup is illustrated in Figure 4. The covariance of demand and a firm's output is what makes firm markups different from product markups. Firms have higher markups in more volatile environments because that volatility allows them to produce more of products that have extremely high markups. In other words, when variance of demand rises, covariance rises as well. This strengthens the composition effects that push firm markups up higher than product markups.

VI Welfare

Typically, economists are interested in markups because they indicate welfare loss. In this setting, markups perform a dual role—they are compensation for firm risk-taking and indicators of deadweight loss. This section shows that more data typically improves welfare, but it also makes distortions from market power more costly. When firms' stocks of data are asymmetric, the effect of exacerbating the data asymmetry depends on whether the risk or the investment effect dominates.

If firms are not compensated for the risk they bear, they will not produce. So a zero markup cannot be the efficient benchmark. Instead, we define prices to be efficient if they arise from production choices of firms that behave as if they were in a competitive market. Competitive firms

are those who take market prices as given, as if their price impact were zero: $\partial E_i[p]/\partial q_i = 0$. If we set price impact to zero in the firm's first-order condition, optimal production is $\boldsymbol{q}_i^{comp} = \frac{1}{\rho_i}\mathbf{Var}\left[\boldsymbol{p}_i|\mathcal{I}_i\right]^{-1}\left(\mathbf{E}\left[\boldsymbol{p}_i|\mathcal{I}_i\right] - \boldsymbol{c}_i\right)$, which implies a competitive markup $H_{ik}^{comp} > 0$.

The challenge this poses is that measuring H_{ik}^{comp} requires estimating a firm's data and price of risk. But using the markup wedges to measure data, as described in section VIII, makes this feasible.

WELFARE BENEFITS OF DATA When all firms get more data, this can be a Pareto improvement. Firm owners benefit because more information improves forecasts, which reduce risk that they are averse to. Also, firms with more data invest to be more efficient. On top of that, consumer surplus increases because lower production cost and more information both tighten competition among firms. The next result formalizes this logic.

Proposition 5. ***Data improves welfare.*** *When firms are symmetric, then more data for every firm increases social welfare.*

Recall that uncertainty mimics market power because both enter additively in $\hat{\mathbf{H}}_i$ in (7). Therefore, resolving uncertainty is like lowering market power. It raises welfare.

Figure 5 illustrates this force. The upward slope of the lines tells us that welfare is increasing in the amount of data. This is true even when there is perfect competition. Even when there is no risk aversion, the ability to produce more goods to meet demand still enhances welfare.

DATA AMPLIFIES MARKET POWER COSTS Figure 5 decomposes the welfare loss into risk aversion and market power. The loss due to market power is much higher on the right, where data is abundant.

The reason that data makes market power more powerful can be seen in the first-order condition (7) of the firm's choice of production quantities q. The right term is expected profit per unit. That expected profit is divided by the term $\rho_i\mathbf{Var}\left[\boldsymbol{p}_i|\mathcal{I}_i\right] - \frac{\partial\mathbf{E}[\boldsymbol{p}_i|\mathcal{I}_i]}{\partial\boldsymbol{q}_i}$, which captures risk price ρ_i times risk (the conditional variance), plus the expected price impact of a trade (market power). Imagine that the product of risk price and risk is large. Then, adding some market power to this large number does not change it by much. When we divide by a large number or a slightly larger number, the answer is almost the same. Thus, when data is scarce and variance is high, market power has little effect on production.

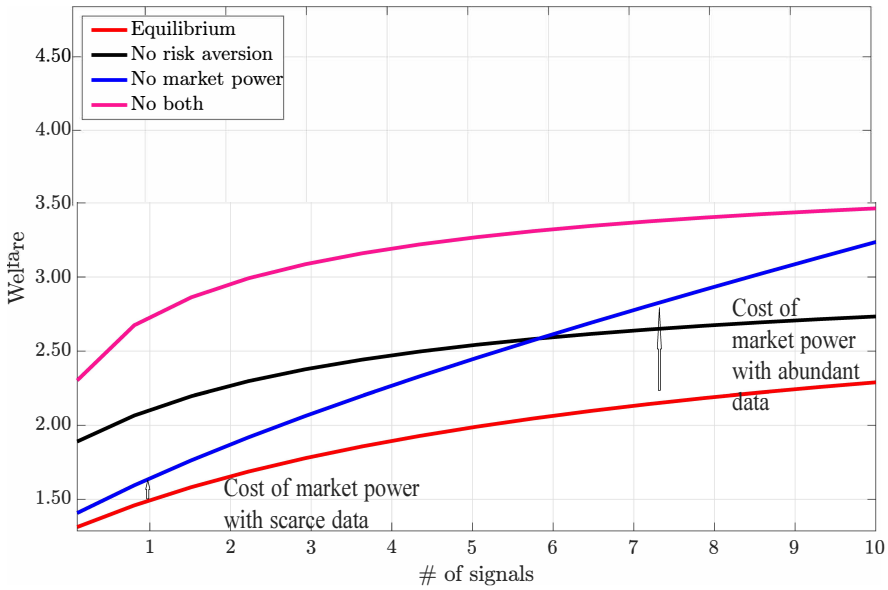


Figure 5: Welfare: Abundant data raises welfare, makes market power more costly

Notes: This counterfactual exercise is constructed over a single-good duopoly example. The x-axis is the number of data points that both firms have. The investment cost function is assumed as $g(\chi_c, c_i) = \chi_c (\bar{c} - c_i)^2 / 2$ with $\chi_c = 1$ and $\bar{c} = 3$. Other parameters are: $\overline{p} = 5$, $\phi = 1$, $\sigma_b = 1$, $\mu_b = 0$, $\sigma_e = 2$, and $\rho_1 = \rho_2 = 1$.

But when data is abundant, the conditional variance is low. Lots of data makes the firm less uncertain. If the first term is small, then adding market power to it makes a big difference. Dividing by a number close to zero or a number slightly less close to zero makes a big difference. Thus, when data is abundant and risk is low, market power has an outsized effect on production choices and thus on prices and markups.

DATA ASYMMETRY Increasing data asymmetry has two opposite welfare effects: (1) increasing market power and hence deadweight loss, and (2) lower disutility from risks because the firm with more information will produce more. (See Appendix C.2.) When the marginal cost c_i is relatively small, a difference in data precision creates a large difference in investment and thus firm size. This force can easily enable one firm dominate the market.

The idea that there are socially good and bad aspects to markups, and that the balance between the two may change over time, is consistent with the evidence of Covarrubias, Gutiérrez, and Philippon (2020).

While these results do not offer a simple answer or prediction about whether data is good or bad, they do provide an important input into the data policy debate. Data asymmetry and market dominance are not synonymous with welfare loss. Uncertainty is also a powerful drag on

economic efficiency and on welfare. Sound data policy needs to trade off traditional market power harms against the benefits of resolving risk. This model produces a simple tool to evaluate that trade-off.

VII Data Barter and Markups: A Dynamic Model

In a dynamic economy, there is one more effect of data on markups: a role for data as a means of payment. We call this data barter. The dynamic model teaches us that, when assessing competition, customers can pay with money or data. This is not just true for free digital apps. The price of every good should be affected. Despite this additional force, the original insights derived from the static model survive, even when data comes from firms' transactions with their customers.

Finally, this model shows how the static problem fits in a broader quest to understand how competition evolves across product space and over time, when data is a strategic asset that is produced and accumulated.

Dynamic model setup. Consider an economy where each firm i chooses an $n \times 1$ vector a_{it} that describes their location in the product space and a quantity q_{it} to produce. As before, firms maximize expected profit, with a price of risk adjustment as in equation (3). There are n attributes and demand shocks for each of those attributes.

What makes data an asset that retains value over multiple periods is that those demand shocks are persistent. If they were not persistent, if demand were independent each period, then data about yesterday's demand would have no value in predicting today's demand. Data would have one-period value. It would not be a long-lived asset. Therefore, we assume a persistent demand process that is an AR(1):

$$b_t = \rho b_{t-1} + \eta_{bt}, \qquad \eta_{bt} \sim iid\ N(0, \sigma_\eta I). \tag{24}$$

At the same time, there needs to be some transitory noise in prices. If there were not, the price of a good would be a sufficient statistic for all past data. If prices revealed all the information in past data, then data would confer only a one-period advantage. It would also not be a long-lived asset. Therefore the demand shock for each attribute is the persistent process (24), plus some transitory noise:

$$\tilde{b}_t = b_t + \epsilon_{bt} \qquad \epsilon_{bt} \sim iid\ N(0, \sigma_\epsilon I) \tag{25}$$

The fact that these demand shocks are common to all firms makes data from one firm relevant for

another firm.

Data is produced as a by-product of economic activity. In other words, the more a firm produces and sells, the more it learns about its customers, its suppliers and its optimal choices. We can model this as a number of data points that depends on the amount produced q. Since data is about the demand for each attribute, the amount of data observed d_{it} is also proportional to the good's loading on the attribute:

$$d_{it} = q_{i,t-1} a_{i,t-1}. \tag{26}$$

This captures the idea that firms learn more about attributes they produce. If they produce cell phones, they learn about the demand for (or cost of) electronics, not about demand for food. If they want to learn about the demand for food, they need to produce something edible, or buy such data. Notice that this makes production a form of active experimentation in the product space. Firms are like gamblers in a classic bandit problem, learning about the profitability of each action by observing its result.

The amount of data that a firm has to inform their decisions depends on data production as data purchases or sales: $D_{it} = d_{it} + m_{it}\mathcal{P}_t$ where m_{it} is the amount of data purchased by firm i at date t and $\mathcal{P}_t$ is the time-t market price per unit of data. Firms also choose an amount of data to sell l_{it}. Since data is non-rival, data that is sold is not lost. However, selling data may not be optimal if better-informed rivals reduce a firm's own production and profits.

Each data point is a signal about the demand shock vector b_t, with precision Σ_e per signal. Firms update demand forecasts using Bayes law. Thus, when a firm obtains D_{it} units of data about each attribute, Bayes law tells the firm to average the signals to arrive at a composite signal that has precision $D'_{i,t}\Sigma_e^{-1}D_{i,t}$. Notice that Bayes' law allows us to incorporate non-integer numbers of signals. So we can proceed considering d_{it} and D_{it} to be any real, non-negative numbers.

Dynamic model solution Let Ω_t be the set of all firm's data precisions $\{\omega_{it}\}_{i=1}^N$. The firms' optimal production $\{q_{i,t},\, a_{i,t}\}$ and data purchases / sales $\{m_{i,t},\, l_{i,t}\}$ solve:

$$V(\Omega_t) = \max_{q_{i,t}, a_{i,t}, m_{i,t}, l_{i,t}} (P_t - c) q_{i,t} a_{i,t} + \mathcal{P}_t(l_{i,t} - m_{i,t}) + \left(\frac{1}{1+r}\right) V(\Omega_{t+1}), \tag{27}$$

where the law of motion for $\Omega_{i,t}$ is

$$\Omega_{i,t+1} = \left[\rho^2 \Omega_{i,t}^{-1} + \sigma_\epsilon\right]^{-1} + (n_{i,t} + m_{i,t})\sigma_\epsilon^{-2} \tag{28}$$

and the number of data points produced by the firm is $n_{i,t} = q_{i,t}a_{i,t}$.

The first-order condition for the quantity of production looks similar to (7) in the static model. Optimal production depends on risk and price impact, in the denominator, and expected profit $(p - c)$, in the numerator.

$$\boldsymbol{q}_i a_i = \left(\rho_i \mathbf{Var}\left[\tilde{\boldsymbol{p}}|\mathcal{I}_i\right] + \frac{\partial \mathbf{E}\left[\tilde{\boldsymbol{p}}|\mathcal{I}_i\right]}{\partial \boldsymbol{q}_i}\right)^{-1} \left(\mathbf{E}\left[\tilde{\boldsymbol{p}}|\mathcal{I}_i\right] + \frac{\partial V(\Omega_{t+1})}{\partial q_i} - \boldsymbol{c}_i\right) \tag{29}$$

However, there is one new term in dynamic model: $\partial V/\partial q_i$ is the increase in the future value of the firm, from producing data.

Notice that the future value of data enters additively with the price. This means that monetary payments and data payments are substitutes for the firm. In other words, customers pay for goods, in part with data. This is a partial barter trade where goods are partly paid for with data, as when you receive a loyalty card discount at a supermarket or pharmacy. These discounts are similar to those in customer acquisition models (Nakamura and Steinsson 2011).

Data barter changes the interpretation of markups. The solution (29) reveals that the price of a good is not the complete payment for the good. The relevant measure of income from selling a unit of a good is $p + \partial V/\partial q_i$. So markups underestimate market power because they fail to account for the data payment that accompanies the monetary payment from customers. Firms in areas of the product space where data is valuable should keep their measured markups low, in order to generate more transactions, to generate more valuable data.

While this dynamic extension introduced new ideas about the interaction of data and markups, it did not change the main conclusions of the static model. Data still complicates the interpretation of markups as measures of market power. In this model, there are three main forces at work in dynamic product markups: (i) the classic effect of market power, (ii) a risk premium, and (iii) data barter. In a data-intensive sector, markups reflect the value of data and its effect on risk as well. Data still shows up as a force that changes how markups are aggregated. Firms use data to predict which goods will have high demand and produce more of those goods. Firms that do this prediction well will have higher firm markups and will grow bigger and get higher weights in their industry markup. But this model suggests that simply correcting markups for a risk premium will not be enough to solve the problem of measuring competition in data-intensive industries.

VIII Mapping Theory to Data

One reason it is important to have models that describe the relationships between quantities like data and markups is that models inform measurement. In this case, the model teaches us how to measure the amount of data a firm has and how to determine what risks that data is about. While executing the measurement is a separate paper, this section is meant to aid others who might choose to use the model as a structure for empirical analysis.

The next result shows that we can measure the amount of data a firm has by looking at the gap between average product markups and firm markups. This is analogous to looking at the alpha of a fund manager to infer how much they know.

Corollary 1. ***Markup wedges are measures of data.*** *The production-aggregation wedge* $E[M_i^f - \overline{M}_i^p]$ *is a monotonic function of firm i's data.*

This result is a straightforward conclusion from Proposition 2. But it is key to measurement. For many measurement exercises, an econometrician may need to know how much data a firm or a collection of firms has. This suggests a measurement approach is to look at the markups at various degrees of aggregation and use the aggregation wedge to infer a corresponding level of data.

WHAT IS DATA ABOUT? MEASURING CHARACTERISTIC LOADINGS Measuring attributes is novel in finance, but more standard in IO. One way to gauge attribute loadings is by looking at demand variance-covariance across goods and extracting principal components. The eigenvectors are loadings. There are also other orthogonal decompositions one can use. But the eigen decomposition has a nice interpretation in terms of principal components.

Another way of measuring characteristic loadings is to use the Hoberg-Phillips measure of cosine similarity from textual analysis of firms' earnings reports. This measure determines how similarly different firms describe their products to their investors.

One might think of a characteristic of a good as being its location. Rossi-Hansberg, Sarte, and Trachter (2018) discovered a different divergence in measures of market power, one between local and national markets. That difference in market power is not expressed in markups but in concentration indices such as HHI (Herfindahl-Hirschman Index). Expressed in markups, there is no documented local-national divergence.[11]

[11] Benkard, Yurukoglu, and Zhang (2021) argue that HHI is defined over the market where consumers are located,

Our predictions are consistent with the superstar firm economy of Autor et al. (2020) and the increasing span of control in Aghion et al. (2019) and Lashkari, Bauer, and Boussard (2018). The rise in firm concentration, the rise in average markups that comes from high-markup firms growing larger, and the correlation between productivity and concentration are all features of U.S. and international markets and are features of our model. Similarly, Crouzet and Eberly (2018) argue that large modern firms have high levels of intangible investment, which is correlated with having high markups. What our work adds is a mechanism—an explanation for why the accumulation of customer data can explain these trends.

MEASURING THE PRICE OF RISK Measuring risk price is novel in IO, but standard in finance. A key parameter that governs the sign of many of the predictions is ρ, the price of risk. Finance has developed a whole battery of tools to determine this risk price in various ways. A common approach is to use the market prices of equities to estimate the compensation investors demand for risk in that domain and then carry the same price over to determine the price of risk that a firm faces. The argument for doing that is that the manager should be maximizing equity holders' interests. The firm's equity holders are the same agents who hold other market equities, with the same risk preferences.[12]

DISTINGUISHING DATA FROM COMPETITION Where data and market competition differ is in $cov(p, q_i)$. Data boosts the covariance between price and quantity by allowing firms to have better forecasts of demand and thereby price. Market competition also changes this covariance by making production decisions more sensitive to expected price changes. But data enhances that sensitivity and also makes expected price and actual price more highly correlated.

Data also enables more accurate forecasting, while market competition does not. Another approach to measuring and identifying firms' data would be to assess the accuracy of firms' forecasts.

Rossi-Hansberg, Sarte, and Trachter (2018) discovered a different divergence in measures of market power, one between local and national markets. That difference in market power is not expressed in markups but in concentration indices such as HHI (Herfindahl-Hirschman Index). Expressed in markups, there is no documented local-national divergence.[13]

whereas data used to measure HHI is based on the location of production, which leads to misleading and inconsistent findings when aggregating. Eeckhout (2020) argues that the discrepancy stems from a mechanical relation between population size and the market definition.

[12] See Brealey, Myers, and Allen (2003) for a more complete explanation of the rationale and execution.

[13] Benkard, Yurukoglu, and Zhang (2021) argue that HHI is defined over the market where consumers are located, whereas data used to measure HHI is based on the location of production, which leads to misleading and inconsistent

Our predictions are consistent with the superstar firm economy of Autor et al. (2020) and the increasing span of control in Aghion et al. (2019) and Lashkari, Bauer, and Boussard (2018). The rise in firm concentration, the rise in average markups that comes from high-markup firms growing larger, and the correlation between productivity and concentration are all features of U.S. and international markets and are features of our model. Similarly, Crouzet and Eberly (2018) argue that large modern firms have high levels of intangible investment, which is correlated with having high markups. If a firm is a data-abundant firm, they should have high levels of intangible assets. However, since there are other intangible assets, the reverse might not be true.

IX Conclusion

The hypothesis that data encourages large firms to grow larger and gain market power is both plausible and incomplete. Because data improves both prediction and firms' profitability, we need to consider competitive effects using a framework where firms compete and face uncertain outcomes that require prediction. In other words, wrestling with the competitive effects of data requires incorporating risk.

We used the a model to illustrate data's competing effects and guide new measurement to disentangle these effects from market power. We find that high-data firms do invest more, grow larger, and exert more impact on prices. However, if uncertain firms scale back production, then more data that resolves their uncertainty also pushes markups down. The effect of data may not be seen in markups.

Instead, the effects of data should show up in markup aggregation. Firms react to data about demand by shifting their production to high-demand goods. These are high-markup goods. So data changes the composition of production. This composition effect leads firms to shift production toward high-markup goods, which raises markups. The tug-of-war between risk reduction and the composition effects induced by data plays out differently for product, firm, and industry markups. A model designed to explore the logic of data and large firms turned out to explain why econometricians got different answers about what was happening to markups over time when they measured at different levels of aggregation. Our model suggests a new interpretation of existing facts. Constant product markups and rising firm and industry markups are not competing facts. They are consistent with an economy where firms are getting better and better at forecasting

findings when aggregating. Eeckhout (2020) argues that the discrepancy stems from a mechanical relation between population size and the market definition.

future demand. Both are helpful in the attempt to understand and measure firms' use of data.

References

Acemoglu, Daron, Ali Makhdoumi, Azarakhsh Malekian, and Asu Ozdaglar, "Too Much Data: Prices and Inefficiencies in Data Markets," *American Economic Journal: Microeconomics*, forthcoming (2022).

Aghion, Philippe, Antonin Bergeaud, Timo Boppart, Peter Klenow, and Huiyu Li, "A Theory of Falling Growth and Rising Rents," NBER Working Paper 26448, 2019.

Anderson, Eric, Sergio Rebelo, and Arlene Wong, "Markups across Space and Time," NBER Working Paper 24434, 2018.

Angeletos, George-Marios, and Jennifer La'O, "Sentiments," *Econometrica*, 81 (2013), 739–779.

Angeletos, George-Marios, and Alessandro Pavan, "Efficient Use of Information and Social Value of Information," *Econometrica*, 75 (2007), 1103–1142.

Asriyan, Vladimir, Luc Laeven, and Alberto Martin, "Collateral Booms and Information Depletion," *Review of Economic Studies*, 89 (2022), 517–555.

Athey, Susan, and Joshua S. Gans, "The Impact of Targeting Technology on Advertising Markets and Media Competition," *American Economic Review*, 100 (2010), 608–613.

Athey, Susan, Mark Mobius, and Jeno Pal, "The Impact of News Aggregators on Internet News Consumption: the Case of Localization," Stanford Technical Report 3353, 2017.

Autor, David, David Dorn, Lawrence F Katz, Christina Patterson, and John Van Reenen, "The Fall of the Labor Share and the Rise of Superstar Firms," *The Quarterly Journal of Economics*, 135 (2020), 645–709.

Baqaee, David, and Emmanuel Farhi, "Entry vs. Rents: Aggregation with Economies of Scale," NBER Working Paper 27140, 2021.

Baqaee, David Rezza, Emmanuel Farhi, and Kunal Sangani, "The Darwinian Returns to Scale," NBER Working Paper 27139, 2021.

Benkard, Lanier, Ali Yurukoglu, and Anthony Lee Zhang, "Concentration in Product Markets," NBER Working Paper 28745, 2021.

Bergemann, Dirk, and Alessandro Bonatti, "The Economics of Social Data: An Introduction," Cowles Foundation Discussion Paper 2171R, 2019.

Bils, Mark, "The Cyclical Behavior of Marginal Cost and Price," *American Economic Review*, 77 (1987), 838–855.

Bils, Mark J, "Real Wages over the Business Cycle: Evidence from Panel Data," *Journal of Political Economy*, 93 (1985), 666–689.

Brealey, RA, SC Myers, and F. Allen, *Principles of Corporate Finance*, 7th ed. (New York, NY: McGraw-Hill, 2003).

Brynjolfsson, Eric, Y Hu, and M Smith, "Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers," *Management Science*, 49 (2003), 1580–1596.

Burstein, Ariel, Vasco M Carvalho, and Basile Grassi, "Bottom-Up Markup Fluctuations," NBER Working Paper 27958, 2020.

de Cornière, Alexandre, and Greg Taylor, "Data and Competition: a General Framework with Applications to Mergers, Market Structure, and Privacy Policy," TSE Working Paper 1076, 2020.

Covarrubias, Matias, Germán Gutiérrez, and Thomas Philippon, "From Good to Bad Concentration? US Industries over the Past 30 Years," *NBER Macroeconomics Annual*, 34 (2020), 1–46.

Crouzet, Nicolas, and Janice Eberly, "Understanding Weak Capital Investment: The Role of Market Concentration and Intangibles," in *Jackson Hole Economic Policy Symposium*, 87–149, 2018.

David, Joel M, Romain Ranciere, and David Zeke, "International Diversification, Reallocation, and the Labor Share," Chicago Fed Mimeo, 2022.

David, Joel, and Venky Venkateswaran, "The Sources of Capital Misallocation," *American Economic Review*, 109 (2019), 2531–2567.

De Loecker, Jan, Jan Eeckhout, and Simon Mongey, "Quantifying Market Power and Business Dynamism in the Macroeconomy," NBER Working Paper 28761, 2021.

De Loecker, Jan, Jan Eeckhout, and Gabriel Unger, "The Rise of Market Power and the Macroeconomic Implications," *Quarterly Journal of Economics*, 135 (2020), 561–644.

De Ridder, Maarten, "Market Power and Innovation in the Intangible Economy," Cambridge Working Paper 1931, 2021.

Doan, Trang, and Mai Iskandar-Datta, "Are Female Top Executives More Risk-Averse or More Ethical? Evidence from Corporate Cash Holdings Policy," *Journal of Empirical Finance*, 55 (2020), 161–176.

Eckbo, B. Espen, ed. *Handbook of Empirical Corporate Finance Set* (Amsterdam: Elsevier, 2008).

Ederer, Florian, and Bruno Pellegrino, "A Tale of Two Networks: Common Ownership and Product Market Rivalry," NBER Working Paper 30004, 2022.

Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu, "How Costly Are Markups?", NBER Working Paper 24800, 2019.

Eeckhout, Jan, "Comment on: Diverging Trends in National and Local Concentration," in *NBER Macroeconomics Annual 2020*, 35, 151–166: NBER), 2020.

Furman, Jason, and Peter Orszag, "A Firm-Level Perspective on the Role of Rents in the Rise in Inequality," Speech, 2015.

Galdon-Sanchez, Jose Enrique, Ricard Gil, and Guillermo Uriz-Uharte, "The Value of Information in Competitive Markets: Evidence from Small and Medium Enterprises," 2023, Queens University mimeo.

Galeotti, Andrea, Benjamin Golub, Sanjeev Goyal, Eduard Talamàs, and Omer Tamuz, "Taxes and Market Power: A Principal Components Approach," Northwestern University Mimeo, 2022.

Goldfarb, Avi, and Catherine E. Tucker, "Digital Economics," NBER Working Paper 23684, 2017.

Grullon, Gustavo, Yelena Larkin, and Roni Michaely, "Are US Industries Becoming More Concentrated?", mimeoTechnical report, 2016.

Hall, Robert E., "The Relation between Price and Marginal Cost in U.S. Industry," *Journal of Political Economy*, 96 (1988), 921–947.

Hall, Robert E, "New Evidence on the Markup of Prices over Marginal Costs and the Role of Mega-Firms in the US Economy," NBER Working Paper 24574, 2018.

Hennesy, Christopher, and Toni Whited, "How Costly Is External Financing? Evidence from a Structural Estimation," *Journal of Finance*, 62 (2007), 1705–1745.

Ichihashi, Shota, "Online Privacy and Information Disclosure by Consumers," *American Economic Review*, 110 (2020), 569–595.

Jarsulic, Marc, "Antitrust Enforcement for the 21st Century," *The Antitrust Bulletin*, 64 (2019), 514–530.

Jones, Charles, and Chris Tonetti, "Nonrivalry and the Economics of Data," *American Economic Review*, 110 (2020), 2819–58.

Kirpalani, Rishabh, and Thomas Philippon, "Data Sharing and Market Power with Two-Sided Platforms," NBER Working Paper 28023, 2020.

Kumar, Saten, Yuriy Gorodnichenko, and Olivier Coibion, "The Effect of Macroeconomic Uncertainty on Firm Decisions," *Econometrica*, forthcoming (2023).

Kwon, Spencer Yongwook, Yueran Ma, and Kaspar Zimmermann, "100 Years of Rising Corporate Concentration," SAFE Working Paper 359, 2022.

Lambert, Nicolas S., Michael Ostrovsky, and Mikhail Panov, "Strategic Trading in Informationally Complex Environments," *Econometrica*, 86 (2018), 1119–1157.

Lambrecht, Anja, and Catherine E. Tucker, "Can Big Data Protect a Firm from Competition?", mimeoTechnical report, 2015.

Lashkari, Danial, Arthur Bauer, and Jocelyn Boussard, "Information Technology and Returns to Scale," Boston College Mimeo, 2018.

Liang, Annie, and Erik Madsen, "Data and Incentives," mimeoTechnical report, 2021.

Lorenzoni, Guido, "A Theory of Demand Shocks," *American Economic Review*, 99 (2009), 2050–2084.

Lovallo, Dan, Tim Koller, Robert Uhlaner, and Daniel Kahneman, "Your Company Is Too Risk-Averse," *Harvard Business Review*.

Maćkowiak, Bartosz, and Mirko Wiederholt, "Optimal Sticky Prices Under Rational Inattention," *American Economic Review*, 99 (3) (2009), 769–803.

Nakamura, Emi, and Jón Steinsson, "Price setting in forward-looking customer markets," *Journal of Monetary Economics*, 58 (2011), 220–233.

Nekarda, Christopher, and Valerie Ramey, "The Cyclical Behavior of the Price-Cost Markup," *Journal of Money, Credit, and Banking*, 52 (2020), 319–353.

Nimark, Kristoffer P., "Man-Bites-Dog Business Cycles," *American Economic Review*, 104 (2014), 2320–2367.

Pellegrino, Bruno, "Product Differentiation and Oligopoly: a Network Approach," WRDS Research Paper, 2020.

Philippon, Thomas, *The great reversal: How America gave up on free markets*: Harvard University Press, 2019).

Rajgopal, Shivaram, Anup Srivastava, and Rong Zhao, "Do Digital Technology Firms Earn Excess Profits? An Alternative Perspective," mimeoTechnical report, 2021.

Rossi-Hansberg, Esteban, Pierre-Daniel Sarte, and Nicholas Trachter, "Diverging Trends in National and Local Concentration," NBER Working Paper 25066, 2018.

Rostek, Marzena, and Marek Weretka, "Price Inference in Small Markets," *Econometrica*, 80 (2012), 687–711.

Sutton, John, *Sunk Costs and Market Structure: Price Competition, Advertising, and the Evolution of Concentration* (Cambridge, MA: MIT Press, 1991).

——— *Technology and Market Structure: Theory and History* (Cambridge, MA: MIT Press, 2001).

Veldkamp, Laura, *Information Choice in Macroeconomics and Finance* (Princeton, NJ: Princeton University Press, 2011).

Vives, Xavier, and Zhiqiang Ye, "Information Technology and Bank Competition," IESE Working Paper, 2021.

Welch, Ivo, *Corporate Finance An Introduction* (Boston, MA: Pearson, 2009).

Online Appendix

A. Appendix: Solution Details

We start by solving the model with firm-specific shocks and public information. Then we solve the model where shocks are aggregate, and derive analogous expressions for two key objects in the model that govern the sensitivity of beliefs to signals and the sensitivity of production to changes in expected price. Then, in Appendix B., we use these solutions to prove the propositions and show that the same properties hold for both models. Appendix C has the most technical lemmas that are inputs into the proposition proofs.

A.1. Preliminaries

PRODUCTS AND ATTRIBUTES We consider n_F firms, indexed by i. Each firm potentially produces N goods indexed by k. The product space has N independent attributes indexed by $j \in \{1, 2, ..., N\}$. Therefore, each good $k \in 1, 2, ..., N$ can be represented by an $N \times 1$ vector $\mathbf{a}_k$ of weights that good k places each attribute. The collection of weights describes a good's location in the product space. Let the collection of weights $(\mathbf{a}_k{}'s)$ be an $N \times N$ full-rank matrix $\boldsymbol{A}$, such that

$$\boldsymbol{q}_i = \boldsymbol{A}\tilde{\boldsymbol{q}}_i$$

The linear mapping $\boldsymbol{A}$ between good and attribute spaces allows us to transform the original model into attribute-competition model in which n_F firms choose upfront investments and attributes to maximize their mean-variance utility. As the attributes are assumed to be orthogonal, the model can be solved by considering one attribute at a time.

PRICES AND COSTS Similarly, we can represent the price and marginal cost of production of goods as the linear combinations of the vector of prices and costs of the attributes:

$$\boldsymbol{p}_i = \boldsymbol{A}\tilde{\boldsymbol{p}}_i \tag{30}$$

$$\boldsymbol{c}_i = \boldsymbol{A}\tilde{\boldsymbol{c}}_i. \tag{31}$$

Each attribute j has an an average market price $\bar{p}_j^M$ that depends on an attribute-specific constant $(\bar{p}_j)$ and on the total quantity of that attribute that all firms produce. It is given by the inverse demand function, which holds for each attribute j:

$$\tilde{p}_j^M = \bar{p}_j - \frac{1}{\phi}\sum_{i=1}^{n_F}\tilde{q}_{ij} \tag{32}$$

INFORMATION Each firm i has n_{di} data points, each of which is a signal of the attribute demand shock $\boldsymbol{s}_{i,z} = \boldsymbol{b_i} + \boldsymbol{\varepsilon}_{i,z}$ where $z = 1, \ldots, n_{di}$. We assume signal noises are uncorrelated and normally distributed with zero mean and precision 1. This is equivalent to a compound signal $\boldsymbol{s}_i$ with total data precision $n_{di} = \sum_{z=1}^{n_{di}} 1$, the sum of the precisions of all of data points about firm i's demand. Therefore, we define a composite signal for each firm-attribute pair, which is the average of the n_{di} data points, each of which is the jth entry of the vector $\boldsymbol{s}_{i,z}$:

$$s_{i,j} = \frac{1}{n_{di}}\sum_{z=1}^{n_{di}}\boldsymbol{s}_{i,z}(j) \tag{33}$$

This signal is a sufficient statistic for all data observed about firm i, attribute j. The posterior variance of the demand shock, conditional on the composite signal $s_{i,j}$, is $\mathbb{V}\left(b_{i,j}|s_{i,j}\right) = \frac{1}{1+n_{di}}$. The posterior mean (the prediction) of demand is $\mathbb{E}[b_{i,j}|s_{i,j}] = (n_{di}/(1+n_{di}))s_{i,j}$, because the prior mean of $b_{i,j}$ is zero, with precision 1.

For results that refer to more data, we mean a derivative with respect to the number of data points about a firm, n_{di}.

A.2. Solving the model with firm-specific shocks and public information

As mentioned earlier, we show our results for two specifications of the model. In the first specification, there are firm-specific shocks and all data is public. Therefore, instead of facing the average market price $\bar{p}_j^M$ for attribute j, each firm i faces a different price $\tilde{p}_{i,j} = \bar{p}_j^M + b_{i,j}$ where $b_{i,j}$ is mean-zero shock distributed randomly with variance 1. The vector of prices faced by firm i is

$$\tilde{\mathbf{p}}_i = [\tilde{p}_1^M, \tilde{p}_2^M, ..., \tilde{p}_N^M] + \boldsymbol{b_i} \tag{34}$$

When other firms' outputs are known, the only uncertainty is about the b shock. Therefore conditional mean and variance of prices is a linear function of the mean and variance of b. For firm i and attribute j,

$$\begin{aligned}\mathbb{E}\left[\tilde{p}_{i,j}|\mathcal{I}_i\right] &= \bar{p}_j + \mathbb{E}\left[b_{i,j}|\mathcal{I}_i\right] - \frac{1}{\phi}\sum_{i'=1}^{n_F}\tilde{q}_{i',j} = \bar{p}_j + K_{i,j}s_{i,j} - \frac{1}{\phi}\sum_{i'=1}^{n_F}\tilde{q}_{i',j} \\ \mathbb{V}\left[\tilde{p}_{i,j}|\mathcal{I}_i\right] &= \mathbb{V}\left[b_{i,j}|\mathcal{I}_i\right] = \frac{1}{1+n_{di}}\end{aligned} \tag{35}$$

where $K_{i,j}$ is defined to be the weight that firm i puts on its signal about attribute j, when forming its expectation about the price. In this model, $K_{i,j} = \frac{n_{di}}{n_{di}+1}$. When others' actions are not observed, K takes another form.

MAXIMIZING RISK-ADJUSTED PROFIT Taking first-order condition of firm's utility function, we get an expression for optimal attribute choices.

$$\tilde{q}_{i,j} = \left(\rho_i \mathbb{V}\left[\tilde{p}_{i,j}|\mathcal{I}_i\right] - \frac{\partial \mathbb{E}\left[\tilde{p}_{i,j}|\mathcal{I}_i\right]}{\partial \tilde{q}_{i,j}}\right)^{-1}\left(\mathbb{E}\left[\tilde{p}_{i,j}|\mathcal{I}_i\right] - \tilde{c}_{i,j}\right) \tag{36}$$

Differentiating the inverse demand curve $\tilde{p}_{i,j} = \bar{p}_j + b_{i,j} - \frac{1}{\phi}\sum_{i'=1}^{n_F}\tilde{q}_{i',j}$ reveals that market power is constant:

$$\frac{\partial \mathbb{E}\left[\tilde{p}_{i,j}|\mathcal{I}_i\right]}{\partial \tilde{q}_{i,j}} = \frac{\partial \mathbb{E}\left[p_{i,j}|\mathcal{I}_i\right]}{\partial q_{i,j}} = -\frac{1}{\phi} \tag{37}$$

Define the sensitivity of supply to a change in the expected profit as:

$$\hat{H}_{i,j} \coloneqq \left(\frac{1}{\phi} + \rho_i \mathbb{V}\left[\tilde{p}_{i,j}|\mathcal{I}_i\right]\right)^{-1}. \tag{38}$$

Substituting this constant market power into the first order condition for optimal output yields the next expression for optimal attribute production. But this expression has $\tilde{q}_{i,j}$ on both the left and the right sides of the equality. It arises on the right side because firm i's production choice $\tilde{q}_{i,j}$ affects the expected price $\mathbb{E}\left[\tilde{p}_{i,j}|\mathcal{I}_i\right]$. Therefore, we substitute in the price and re-arrange to collect all $\tilde{q}_{i,j}$ terms and reveal the optimal production choice. Use (37) to substitute out $\partial \mathbb{E}\left[\tilde{p}_{i,j}|\mathcal{I}_i\right]/\partial \tilde{q}_{i,j}$. Then use Bayes law to replace the expectation $\mathbb{E}\left[b_{i,j}|\mathcal{I}_i\right]$ with the weighted sum of signals $K_{i,j}s_{i,j}$, with the Bayesian updating weight $K_{i,j} = \frac{n_{di}}{n_{di}+1}$. Using these three substitutions, we can rewrite (36) as:

$$\tilde{q}_{i,j} = \hat{H}_{i,j}\left(\bar{p}_j + K_{i,j}s_{i,j} - \frac{1}{\phi}\sum_{i'=1}^{n_F}\tilde{q}_{i',j} - \tilde{c}_{i,j}\right) \tag{39}$$

The solution above generates the best-response function, given the aggregate output. When all data is public, this aggregate output is known. But firm i output choice still shows up on the left and right sides. Before continuing on to solve for aggregate output and prices, we first stop to correctly express firm i's best response, as a function of all other firms' output choices. Define $H_{i,j} = \left(\rho_i \mathbb{V}\left[\tilde{p}_{i,j}|\mathcal{I}_i\right] + \frac{2}{\phi}\right)^{-1}$. Then, we can collect the terms in $\tilde{q}_{i,j}$ to get the best response:

$$\tilde{q}_{i,j} = H_{i,j}\left(\bar{p}_j + K_{i,j}s_{i,j} - \frac{1}{\phi}\sum_{i'\neq i}\tilde{q}_{i',j} - \tilde{c}_{i,j}\right) \tag{40}$$

SUB-GAME EQUILIBRIUM We solve the sub-game Nash equilibrium by summing both sides of (39) over all firms to express the aggregate output:

$$\begin{aligned}
\sum_i \tilde{q}_{i,j} &= \sum_i \hat{H}_{i,j}\left(\bar{p}_j + K_{i,j}s_{i,j} - \frac{1}{\phi}\sum_{i=1}^{n_F}\tilde{q}_{i,j} - \tilde{c}_{i,j}\right) \\
\left(1+\frac{1}{\phi}\sum_i \hat{H}_{i,j}\right)\sum_i \tilde{q}_{i,j} &= \sum_i \hat{H}_{i,j}\left(\bar{p}_j + K_{i,j}s_{i,j} - \tilde{c}_{i,j}\right) \\
\sum_i \tilde{q}_{i,j} &= \left(1+\frac{1}{\phi}\sum_i \hat{H}_{i,j}\right)^{-1}\left(\bar{p}_j\sum_i \hat{H}_{i,j} + \sum_i \hat{H}_{i,j}K_{i,j}s_{i,j} - \sum_i \hat{H}_{i,j}\tilde{c}_{i,j}\right)
\end{aligned} \tag{41}$$

Denote,

$$\begin{aligned}
C_j &= \sum_i \hat{H}_{i,j}\tilde{c}_{i,j} \\
\kappa_j &= \sum_i \hat{H}_{i,j}K_{i,j}s_{i,j} \\
\bar{H}_j &= \sum_i \hat{H}_{i,j}
\end{aligned} \tag{42}$$

As $\hat{H}_{i,j}$ denotes the sensitivity of supply to a change in the expected profit, C_j represents the sensitivity weighted average cost of attribute j, hereafter referred to as the average cost of attribute j. Similarly, κ_j represents the average demand shock sensitivity for attribute j, and $\bar{H}_j$ represents the average sensitivity.

Then, aggregate output can be expressed as

$$\sum_i \tilde{q}_{i,j} = \frac{\bar{p}_j\bar{H}_j + \kappa_j - C_j}{1+\frac{\bar{H}_j}{\phi}} \tag{43}$$

From aggregate output, we can express the market price of each attribute as $p_j^M = \bar{p}_j - \frac{1}{\phi}\sum_i \tilde{q}_{i,j}$. Using the expressions derived above, this can be written as

$$\begin{aligned}
p_j^M &= \bar{p}_j - \frac{1}{\phi}\left(\frac{\bar{p}_j\bar{H}_j + \kappa_j - C_j}{1+\frac{\bar{H}_j}{\phi}}\right) \\
&= \frac{\bar{p}_j + \frac{C_j}{\phi}}{1+\frac{\bar{H}_j}{\phi}} - \frac{\frac{\kappa_j}{\phi}}{1+\frac{\bar{H}_j}{\phi}}
\end{aligned} \tag{44}$$

Define $\bar{p}_j^M$ to be the expected market price for attribute j while p_j^M denotes the realized market price. Given that the signals $s_{i,j}$ are, on average, equal to zero, κ_j equals zero. Therefore, the expected market price is

$$\bar{p}_j^M = \frac{\bar{p}_j + \frac{C_j}{\phi}}{1+\frac{\bar{H}_j}{\phi}} = \left(1+\frac{1}{\phi}\sum_i \hat{H}_{i,j}\right)^{-1}\left(\bar{p}_j + \frac{1}{\phi}\sum_i \hat{H}_{i,j}\tilde{c_{i,j}}\right) \tag{45}$$

Note that this also shows the how the realized market price p_j^M differs from its expectation $\bar{p}_j^M$ because of a weighted sum of the firms' random data realizations $s_{i,j}$. More specifically,

$$\begin{aligned}
p_j^M &= \bar{p}_j^M - \frac{\kappa_j}{\phi(1+\frac{\bar{H}_j}{\phi})} \\
&= \bar{p}_j^M - \frac{1}{\phi}\left(1+\frac{1}{\phi}\sum_i \hat{H}_{i,j}\right)^{-1}\sum_i \hat{H}_{i,j}K_{i,j}s_{i,j}
\end{aligned} \tag{46}$$

Finally, we can express the equilibrium output as functions of marginal costs, parameters and firms' data $s_{i,j}$. Start-

ing from (39), we substitute the definition of p^M to obtain:

$$\begin{aligned}\tilde{q}_{i,j} &= \hat{H}_{i,j}\left(p_j^M + K_{i,j}s_{i,j} - \tilde{c}_{i,j}\right) \\ &= \hat{H}_{i,j}\left(\bar{p}_j^M - \frac{\kappa_j}{\phi(1+\frac{\bar{H}_j}{\phi})} + K_{i,j}s_{i,j} - \tilde{c}_{i,j}\right) \\ &= \hat{H}_{i,j}(\bar{p}_j^M - \tilde{c}_{i,j}) + \hat{H}_{i,j}K_{i,j}s_{i,j} - \frac{\hat{H}_{i,j}\kappa_j}{\phi(1+\frac{\bar{H}_j}{\phi})} \\ &= \hat{H}_{i,j}(\bar{p}_j^M - \tilde{c}_{i,j}) + \hat{H}_{i,j}K_{i,j}s_{i,j} - \frac{\hat{H}_{i,j}}{\phi}\left(1+\frac{1}{\phi}\sum_{i'=1}^{n_F}\hat{H}_{i',j}\right)^{-1}\sum_{i'=1}^{n_F}\hat{H}_{i',j}K_{i',j}s_{i',j}\end{aligned} \tag{47}$$

where the second equality uses the relationship between market price and aggregate quantity, and the third equality uses (45). Similarly, equilibrium price available to firm i for attribute j is given by:

$$\tilde{p}_{i,j} \equiv p_j^M + b_{i,j} = \bar{p}_j^M + b_{i,j} - \frac{1}{\phi}\left(1+\frac{1}{\phi}\sum_{i'=1}^{n_F}\hat{H}_{i',j}\right)^{-1}\sum_{i'=1}^{n_F}\hat{H}_{i',j}K_{i',j}s_{i',j} \tag{48}$$

A.3. Solving the model with aggregate shocks and private information

So far, we have established the equilibrium solution for the model with private shocks and public data. With aggregate shocks and private data, there are two main changes. First, the vector for market price vector p is given by the following inverse demand function

$$\tilde{\boldsymbol{p}} = \bar{\boldsymbol{p}} + \boldsymbol{b} - \frac{1}{\phi}\sum_{i=1}^{n_F}\tilde{\boldsymbol{q}}_i \tag{49}$$

where $\boldsymbol{b} = [b_1, \ldots, b_N]$ is the vector of aggregate shocks to the market price vector. We assume that these shocks are normally distributed with mean 0 and variance 1. The second change is that instead of the data being public, now each firm sees a private signal $\boldsymbol{s}_i = \boldsymbol{b} + \boldsymbol{\varepsilon}_i$ where the variance of $\boldsymbol{\varepsilon}_i$ depends on the number of data points (n_{di}) available to firm i and equals $1/n_{di}$. With this background, we can now prove Lemma 1 which provides the equilibrium solution for this model.

Proof of Lemma 1: Pricing with Aggregate Shocks Since firm i could only observe $\boldsymbol{s}_i$, its expectation of the price is $\mathbb{E}[\boldsymbol{p}|\boldsymbol{s}_i] = \bar{\boldsymbol{p}}^M + \boldsymbol{K}_i\boldsymbol{s}_i$, where

$$\boldsymbol{K}_i = \mathbf{Cov}(\boldsymbol{p}, \boldsymbol{s}_i)\mathbf{Var}(\boldsymbol{s}_i)^{-1} \tag{50}$$

The variance of price forecast error is

$$\mathbf{Var}[\boldsymbol{p}|\boldsymbol{s}_i] = \mathbf{Var}(\boldsymbol{p}) - \mathbf{Cov}(\boldsymbol{p}, \boldsymbol{s}_i)\mathbf{Var}(\boldsymbol{s}_i)^{-1}\mathbf{Cov}(\boldsymbol{p}, \boldsymbol{s}_i)' \tag{51}$$

SOLUTION: We guess and verify a linear price function and then solve for the coefficients at the end. A linear ansatz takes the following form with coefficients $\bar{\boldsymbol{p}}^M, \boldsymbol{F}$ and $\{\boldsymbol{h}_i\}_{i=1,\ldots,n_F}$

$$\bar{\boldsymbol{p}} = \bar{\boldsymbol{p}}^M + \boldsymbol{F}\boldsymbol{b} + \sum_{i=1}^{n_F}\boldsymbol{h}_i\boldsymbol{\varepsilon}_i \tag{52}$$

We can rearrange the pricing equation (4) as $\frac{1}{\phi}\sum_{i=1}^{n_F}\tilde{\boldsymbol{q}}_i = \bar{\boldsymbol{p}} + \boldsymbol{b} - \tilde{\boldsymbol{p}}$ and then substitute in the first order condition (7) and the linear pricing rule guess to obtain

$$\begin{aligned}&\frac{1}{\phi}\sum_{i=1}^{n_F}\hat{\boldsymbol{H}}_i(\bar{\boldsymbol{p}}^M - \boldsymbol{c}_i + \boldsymbol{K}_i\boldsymbol{s}_i) = \bar{\boldsymbol{p}} + \boldsymbol{b} - \left(\bar{\boldsymbol{p}}^M + \boldsymbol{F}\boldsymbol{b} + \sum_{i=1}^{n_F}\boldsymbol{h}_i\boldsymbol{\varepsilon}_i\right) \\ &\left(\boldsymbol{F} - \boldsymbol{I}_N + \frac{1}{\phi}\sum_{i=1}^{n_F}\hat{\boldsymbol{H}}_i\boldsymbol{K}_i\right)\boldsymbol{b} + \sum_{i=1}^{n_F}\left(\boldsymbol{h}_i + \frac{1}{\phi}\hat{\boldsymbol{H}}_i\boldsymbol{K}_i\right)\boldsymbol{\varepsilon}_i + \left(\boldsymbol{I}_N + \frac{1}{\phi}\sum_{i=1}^{n_F}\hat{\boldsymbol{H}}_i\right)\bar{\boldsymbol{p}}^M - \bar{\boldsymbol{p}} - \frac{1}{\phi}\sum_{i=1}^{n_F}\hat{\boldsymbol{H}}_i\boldsymbol{c}_i = 0\end{aligned} \tag{53}$$

Collecting terms in p delivers the pricing equation $\boldsymbol{p} = \bar{\boldsymbol{p}}^M + \boldsymbol{F}\boldsymbol{b} - \frac{1}{\phi}\sum_{i=1}^{n_F} \hat{\boldsymbol{H}}_i \boldsymbol{K}_i \boldsymbol{\varepsilon}_i$, where $\bar{\boldsymbol{p}}^M$ is given by (45). Matching coefficients proves lemma 1 with the following coefficients:

$$\begin{aligned}
\boldsymbol{F} &= \boldsymbol{I}_N - \frac{1}{\phi}\sum_{i=1}^{n_F} \hat{\boldsymbol{H}}_i \boldsymbol{K}_i \\
\boldsymbol{K}_i &= \left(\boldsymbol{F} n_{di} - \frac{1}{\phi}\hat{\boldsymbol{H}}_i \boldsymbol{K}_i\right)\frac{1}{n_{di}+1} \\
\hat{\boldsymbol{H}}_i &= \left[\rho_i\left(\boldsymbol{F}\boldsymbol{F}' + \frac{1}{n_{di}\phi^2}\sum_{i=1}^{n_F}(\hat{\boldsymbol{H}}_i\boldsymbol{K}_i)^2 - \frac{n_{di}}{n_{di}+1}\left(\boldsymbol{F} - \frac{1}{\phi n_{di}}\hat{\boldsymbol{H}}_i\boldsymbol{K}_i\right)\left(\boldsymbol{F} - \frac{1}{\phi n_{di}}\hat{\boldsymbol{H}}_i\boldsymbol{K}_i\right)'\right) + \frac{\boldsymbol{I}_N}{\phi}\right]^{-1}
\end{aligned} \tag{54}$$

□

As we can see, apart from the differences in the expression for $\hat{\boldsymbol{H}}_i$ and $\hat{\boldsymbol{K}}_i$, the expressions for expected price and expected quantity take similar form as for the model with private shocks and public data. For example, output and expected output are

$$\begin{aligned}
\tilde{\boldsymbol{q}}_i &= \hat{\boldsymbol{H}}_i(\bar{\boldsymbol{p}}^M - \boldsymbol{c}_i + \boldsymbol{K}_i \boldsymbol{s}_i) \\
\mathbb{E}[\tilde{\boldsymbol{q}}_i] &= \hat{\boldsymbol{H}}_i(\bar{\boldsymbol{p}}^M - \boldsymbol{c}_i).
\end{aligned} \tag{55}$$

Of course, these expressions still contain endogenous variables. Appendix C digs deeper to express price in terms of underlying parameters. It also shows conditions under which H, F and K have derivatives with respect to more data that have the same sign as in the public information model. Because H, F and K react similarly to more data, many of the the same results will hold for both the models.

A.4. Markups

In this section, we define the markups based on different levels of aggregation and characterize how various components of the markups are affected by data.

PRODUCT-LEVEL MARKUP The product-level markup of product j for firm i is defined as $M_{i,j}^{\tilde{p}} := E[\tilde{p}_{i,j}]/\tilde{c}_{i,j}$. The expected price of product j by firm i is the expectation of (48). Note that $b_{i,j}$ and $s_{i,j}$ are mean-zero random variables. Thus from (48), the mean of $\tilde{p}_{i,j}$ is $\bar{p}_j^M$. Thus, the product-level markup, averaged over products (attributes) and over firms is

$$\overline{M}^{\tilde{p}} = \frac{1}{N}\frac{1}{n_F}\sum_{i=1}^{n_F}\sum_{j=1}^{N} M_{i,j}^{\tilde{p}} = \frac{1}{n_F N}\sum_{i=1}^{n_F}\sum_{j=1}^{N}\frac{E[\tilde{p}_{i,j}]}{\tilde{c}_{i,j}} = \frac{1}{n_F N}\sum_{i=1}^{n_F}\sum_{j=1}^{N}\frac{\bar{p}_j^M}{\tilde{c}_{i,j}} \tag{56}$$

To see the effect of obtaining more data points on the average product level mark-up, we need to calculate $\partial\overline{M}^{\tilde{p}}/\partial n_{di}$. Using the expression above, this becomes

$$\frac{\partial\overline{M}^{\tilde{p}}}{\partial n_{di}} = \frac{1}{n_F N}\sum_{i=1}^{n_F}\sum_{j=1}^{N}\frac{1}{\tilde{c}_{i,j}}\frac{\partial\bar{p}_j^M}{\partial n_{di}} \tag{57}$$

Using (45), we get

$$\begin{aligned}
\frac{\partial\bar{p}_j^M}{\partial n_{di}} &= \frac{(1+\frac{\bar{H}_j}{\phi})\frac{1}{\phi}\frac{\partial C_j}{\partial n_{di}} - (\bar{p}_j + \frac{C_j}{\phi})\frac{1}{\phi}\frac{\partial\bar{H}_j}{\partial n_{di}}}{(1+\frac{\bar{H}_j}{\phi})^2} \\
&= \frac{(1+\frac{\bar{H}_j}{\phi})\frac{1}{\phi}\frac{\partial C_j}{\partial n_{di}} - (1+\frac{\bar{H}_j}{\phi})\frac{\bar{p}_j^M}{\phi}\frac{\partial\bar{H}_j}{\partial n_{di}}}{(1+\frac{\bar{H}_j}{\phi})^2} \\
&= \frac{\frac{1}{\phi}\frac{\partial C_j}{\partial n_{di}} - \frac{\bar{p}_j^M}{\phi}\frac{\partial\bar{H}_j}{\partial n_{di}}}{(1+\frac{\bar{H}_j}{\phi})}
\end{aligned} \tag{58}$$

Using the expressions for $\bar{H}_j$ and C_j from (42), we get

$$\begin{aligned}\frac{\partial \bar{H}_j}{\partial n_{di}} &= \frac{\partial \hat{H}_{i,j}}{\partial n_{di}} \\ \frac{\partial C_j}{\partial n_{di}} &= \tilde{c}_{i,j}\frac{\partial \hat{H}_{i,j}}{\partial n_{di}}\end{aligned} \tag{59}$$

Now, using (35) to replace the variance term in (38) in terms of n_{di} and taking the partial derivative yields

$$\frac{\partial \hat{H}_{i,j}}{\partial n_{di}} = \rho_i \left(\frac{\hat{H}_{i,j}}{1+n_{di}}\right)^2 > 0 \tag{60}$$

Using the two expressions above, we can now write:

$$\begin{aligned}\frac{\partial \bar{p}_j^M}{\partial n_{di}} &= \frac{1}{\phi}\frac{\partial \hat{H}_{i,j}}{\partial n_{di}}\left(1+\frac{\bar{H}_j}{\phi}\right)^{-1}(\tilde{c}_{i,j}-\bar{p}_j^M) \\ &= \frac{\rho_i}{\phi}\left(\frac{\hat{H}_{i,j}}{1+n_{di}}\right)^2\left(1+\frac{\bar{H}_j}{\phi}\right)^{-1}(\tilde{c}_{i,j}-\bar{p}_j^M) < 0\end{aligned} \tag{61}$$

where the last inequality follows from the fact that for a firm to have non-negative expected profits, the cost $\tilde{c}_{i,j}$ must be less than the average market price $\bar{p}_j^M$. If it were not so, the firm i would not choose to produce positive quantity (see (36)).

Using this result in (57), we obtain that $\partial \overline{M}^{\tilde{p}}/\partial n_{di} < 0$

FIRM-LEVEL MARKUP The firm-level markup for firm i is the quantity-weighted prices divided by quantity-weighted costs:

$$M_i^f = \frac{\mathbb{E}[\tilde{\boldsymbol{q}}_i'\tilde{\boldsymbol{p}}_i]}{\mathbb{E}[\tilde{\boldsymbol{q}}_i'\tilde{\boldsymbol{c}}_i]} = \frac{\mathbb{E}[\tilde{\boldsymbol{q}}_i]'\mathbb{E}[\tilde{\boldsymbol{p}}_i] + \textbf{Trace}[\textbf{Cov}\,(\tilde{\boldsymbol{p}}_i, \tilde{\boldsymbol{q}}_i)]}{\mathbb{E}[\tilde{\boldsymbol{q}}_i'\tilde{\boldsymbol{c}}_i]} \tag{62}$$

As for the denominator, note that $\mathbb{E}\tilde{q}_{i,j} = \hat{H}_{i,j}(\bar{p}_j^M - \tilde{c}_{i,j})$. So, the equilibrium output increases with more data since

$$\begin{aligned}\frac{\partial \mathbb{E}\tilde{q}_{i,j}}{\partial n_{di}} &= (\bar{p}_j^M - \tilde{c}_{i,j})\frac{\partial \hat{H}_{i,j}}{\partial n_{di}} + \hat{H}_{i,j}\frac{\partial \bar{p}_j^M}{\partial n_{di}} \\ &= (\bar{p}_j^M - \tilde{c}_{i,j})\frac{\partial \hat{H}_{i,j}}{\partial n_{di}} + \frac{\hat{H}_{i,j}}{\phi}\left(\frac{\tilde{c}_{i,j}-\bar{p}_j^M}{1+\frac{\bar{H}_j}{\phi}}\right)\frac{\partial \hat{H}_{i,j}}{\partial n_{di}} \\ &= (\bar{p}_j^M - \tilde{c}_{i,j})\frac{\partial \hat{H}_{i,j}}{\partial n_{di}}\left(1-\frac{\hat{H}_{i,j}}{\phi+\bar{H}_j}\right) > 0\end{aligned} \tag{63}$$

where the second equality follows from (58) and (59), and the last inequality uses $\bar{H}_j = \sum_{i'}\hat{H}_{i',j} > \hat{H}_{i,j}$ and $\bar{p}_j^M > \tilde{c}_{i,j}$ as argued earlier.

Although price decreases with more data, the revenue rises.

$$\begin{aligned}
\frac{\partial \mathbb{E}\tilde{q}_{i,j}\mathbb{E}\tilde{p}_{i,j}}{\partial n_{di}} &= \mathbb{E}\tilde{q}_{i,j}\frac{\partial \mathbb{E}\tilde{p}_{i,j}}{\partial n_{di}} + \mathbb{E}\tilde{p}_{i,j}\frac{\partial \mathbb{E}\tilde{q}_{i,j}}{\partial n_{di}} \\
&= \hat{H}_{i,j}(\bar{p}_j^M - \tilde{c}_{i,j})\frac{\partial \bar{p}_j^M}{\partial n_{di}} + \bar{p}_j^M(\bar{p}_j^M - \tilde{c}_{i,j})\frac{\partial \hat{H}_{i,}j}{\partial n_{di}}\left(1 - \frac{\hat{H}_{i,j}}{\phi + \bar{H}_j}\right) \\
&= (\bar{p}_j^M - \tilde{c}_{i,j})\left(\frac{\hat{H}_{i,j}}{\phi}\frac{\partial \hat{H}_{i,j}}{\partial n_{di}}\left(1 + \frac{\bar{H}_j}{\phi}\right)^{-1}(\tilde{c}_{i,j} - \bar{p}_j^M) + \bar{p}_j^M\frac{\partial \hat{H}_{i,}j}{\partial n_{di}}\left(1 - \frac{\hat{H}_{i,j}}{\phi + \bar{H}_j}\right)\right) \\
&= \frac{\partial \hat{H}_{i,j}}{\partial n_{di}}\left(\frac{\bar{p}_j^M - \tilde{c}_{i,j}}{\phi + \bar{H}}\right)\left(\hat{H}_{i,j}(\tilde{c}_{i,j} - \bar{p}_j^M) + \bar{p}_j^M\left(\phi + \bar{H}_j - \hat{H}_{i,j}\right)\right) \\
&= \frac{\partial \hat{H}_{i,j}}{\partial n_{di}}\left(\frac{\bar{p}_j^M - \tilde{c}_{i,j}}{\phi + \bar{H}}\right)\left(\hat{H}_{i,j}\tilde{c}_{i,j} + \bar{p}_j^M\left(\phi - \hat{H}_{i,j} + \sum_{k\neq i}\hat{H}_{k,j}\right)\right) > 0
\end{aligned} \tag{64}$$

The second equality above uses (63), the third equality uses (61), and the final inequality follows from the fact that $\phi > \hat{H}_{i,j}$ by definition (see (38)).

COST-WEIGHTED INDUSTRY MARKUP The industry markup weighted by cost is

$$M^c := \frac{\mathbb{E}\left[\sum_{i=1}^{n_F}\tilde{\boldsymbol{q}}_i'\tilde{\boldsymbol{p}}_i\right]}{\mathbb{E}\left[\sum_{i=1}^{n_F}\tilde{\boldsymbol{q}}_i'\tilde{\boldsymbol{c}}_i\right]} = \frac{\sum_{i=1}^{n_F}\mathbb{E}\left[\tilde{\boldsymbol{q}}_i'\tilde{\boldsymbol{p}}_i\right]}{\sum_{i=1}^{n_F}\mathbb{E}\left[\tilde{\boldsymbol{q}}_i'\tilde{\boldsymbol{c}}_i\right]} = \sum_{i=1}^{n_F} w_i^{cost}M_i^f \quad \text{where} \quad w_i^{cost} = \frac{\mathbb{E}\left[\tilde{\boldsymbol{q}}_i'\tilde{\boldsymbol{c}}_i\right]}{\sum_{i=1}^{n_F}\mathbb{E}\left[\tilde{\boldsymbol{q}}_i'\tilde{\boldsymbol{c}}_i\right]}. \tag{65}$$

Denote $w_{i,j}^{cost} = \frac{\mathbb{E}[\tilde{q}_{i,j}\tilde{c}_{i,j}]}{\sum_{i=1}^{n_F}\mathbb{E}\left[\tilde{\boldsymbol{q}}_i'\tilde{\boldsymbol{c}}_i\right]}$. Then, $w_i^{cost} = \sum_{j=1}^{N} w_{i,j}^{cost}$. The weight $w_{i,j}^{cost}$ increases with more data n_{di} as

$$\frac{\partial w_{i,j}^{cost}}{\partial n_{di}} = \frac{\tilde{\boldsymbol{c}}_{i,j}}{\left(\sum_{i=1}^{n_F}\mathbb{E}\left[\tilde{\boldsymbol{q}}_i'\tilde{\boldsymbol{c}}_i\right]\right)^2}\left[\frac{\partial \mathbb{E}\tilde{\boldsymbol{q}}_{i,j}}{\partial n_{di}}\left(\sum_{k=1,k\neq i}^{n_F}\mathbb{E}\left[\tilde{\boldsymbol{q}}_k'\tilde{\boldsymbol{c}}_k\right]\right) - \mathbb{E}\tilde{\boldsymbol{q}}_{i,j}\left(\sum_{k=1,k\neq i}^{n_F}\frac{\partial \mathbb{E}(\tilde{\boldsymbol{q}}_{k,j})}{\partial n_{di}}\tilde{\boldsymbol{c}}_{k,j}\right)\right] > 0 \tag{66}$$

The last inequality is due to the existing results $\frac{\partial \mathbb{E}\tilde{\boldsymbol{q}}_{i,j}}{\partial n_{di}} > 0$ and $\frac{\partial \mathbb{E}(\tilde{\boldsymbol{q}}_{k,j})}{\partial n_{di}} = \hat{H}_{k,j}\frac{\partial \bar{\boldsymbol{p}}_j^M}{\partial n_{di}} < 0$.

SALES-WEIGHTED INDUSTRY MARKUP The industry markup weighted by sales is

$$M^s := \sum_{i=1}^{n_F} w_i^s M_i^f = \frac{\sum_{i=1}^{n_F}\frac{\mathbb{E}^2[\tilde{\boldsymbol{q}}_i'\tilde{\boldsymbol{p}}_i]}{\mathbb{E}\left[\tilde{\boldsymbol{q}}_i'\tilde{\boldsymbol{c}}_i\right]}}{\sum_{i=1}^{n_F}\mathbb{E}\left[\tilde{\boldsymbol{q}}_i'\tilde{\boldsymbol{p}}_i\right]} \quad \text{where} \quad w_i^s = \frac{\mathbb{E}\left[\tilde{\boldsymbol{q}}_i'\tilde{\boldsymbol{p}}_i\right]}{\sum_{i=1}^{n_F}\mathbb{E}\left[\tilde{\boldsymbol{q}}_i'\tilde{\boldsymbol{p}}_i\right]}. \tag{67}$$

EXPECTED RISK-ADJUSTED FIRM PROFIT To solve for the firms' cost choices, we need to solve for expected utility of each firm. Substituting in the definition of firm profit into the objective function (3)

$$\begin{aligned}
\mathbb{E}[U_i] &= \mathbb{E}\left[\tilde{\boldsymbol{q}}_i'\left(\mathbb{E}[\tilde{\boldsymbol{p}}_i|\mathcal{I}_i] - \tilde{\boldsymbol{c}}_i\right)\right] - \frac{\rho_i}{2}\mathbb{E}\left[\tilde{\boldsymbol{q}}_i'\mathbb{V}\left[\tilde{\boldsymbol{p}}_i|\mathcal{I}_i\right]\tilde{\boldsymbol{q}}_i\right] - g(\chi_c, \tilde{\boldsymbol{c}}_i) \\
&= \mathbb{E}\left[\sum_{j=1}^{N}\tilde{q}_{i,j}\left(\mathbb{E}[\tilde{p}_{i,j}|\mathcal{I}_i] - \tilde{c}_{i,j}\right)\right] - \frac{\rho_i}{2}\mathbb{E}\left[\sum_{j=1}^{N}\tilde{q}_{i,j}\mathbb{V}\left[\tilde{p}_{i,j}|\mathcal{I}_i\right]\tilde{q}_{i,j}\right] - \sum_{j=1}^{N} g_j(\chi_c, \tilde{c}_{i,j}) \\
&= \sum_{j=1}^{N}\left(\mathbb{E}\left[\tilde{q}_{i,j}\left(\mathbb{E}[\tilde{p}_{i,j}|\mathcal{I}_i] - \tilde{c}_{i,j}\right)\right] - \frac{\rho_i}{2}\mathbb{E}\left[\tilde{q}_{i,j}^2\mathbb{V}\left[\tilde{p}_{i,j}|\mathcal{I}_i\right]\right] - g_j(\chi_c, \tilde{c}_{i,j})\right) \\
&= \sum_{j=1}^{N}\mathbb{E}[U_{i,j}]
\end{aligned} \tag{68}$$

where $U_{i,j} = \left[\tilde{q}_{i,j}\left(\mathbb{E}[\tilde{p}_{i,j}|\mathcal{I}_i] - \tilde{c}_{i,j}\right)\right] - \frac{\rho_i}{2}\mathbb{E}\left[\tilde{q}_{i,j}^2\mathbb{V}\left[\tilde{p}_{i,j}|\mathcal{I}_i\right]\right] - g_j(\chi_c, \tilde{c}_{i,j})$ represents the utility of firm i from attribute j based on firm's data. In deriving the above expression, we have used the assumption that the signal noises are uncorrelated among attributes which makes $\mathbb{V}\left[\tilde{\boldsymbol{p}}_i|\mathcal{I}_i\right]$ a diagonal matrix. We have also used the assumption that $g(\chi_c, \tilde{\boldsymbol{c}}_i)$

is additively separable w.r.t to the attributes. Because of the independence of attributes, we can now simplify each $\mathbb{E}[U_{i,j}]$ term separately and add them up to get the total expected utility.

Note that the first term in the expected profits expression is $\mathbb{E}\left[\tilde{q}_{i,j}\left(\mathbb{E}\left[\tilde{p}_{i,j}|\mathcal{I}_i\right]-\tilde{c}_{i,j}\right)\right]$. Using the first order condition (36), we can substitute $\left(\mathbb{E}\left[\tilde{p}_{i,j}|\mathcal{I}_i\right]-\tilde{c}_{i,j}\right)$ out with $\hat{H}_{i,j}^{-1}\tilde{q}_{i,j}$. That substitution allows us to write the firm objective for attribute j as

$$\begin{aligned} U_{i,j} &= \tilde{q}_{i,j}^2\left(\hat{H}_{i,j}^{-1}-\frac{\rho_i}{2}\mathbb{V}\left[\tilde{p}_{i,j}|\mathcal{I}_i\right]\right)-g_j(\chi_c,\tilde{c}_{i,j}) \\ &= \tilde{q}_{i,j}^2\left(\frac{1}{\phi}+\frac{\rho_i}{2}\mathbb{V}\left[\tilde{p}_{i,j}|\mathcal{I}_i\right]\right)-g_j(\chi_c,\tilde{c}_{i,j}) \\ &= \frac{1}{2}\tilde{q}_{i,j}^2 H_{i,j}^{-1}-g_j(\chi_c,\tilde{c}_{i,j}) \end{aligned} \tag{69}$$

The expression above is the utility from attribute j conditional on a firm's data. To choose marginal cost, we need to compute expected utility that is not conditional on the firm's signals because firms choose cost before signals are observed. This utility could be expressed as expected profit minus the price of risk. Substituting in the expected profit expression above, we get

$$\begin{aligned} \mathbb{E}[U_{i,j}] &= \frac{1}{2}\mathbb{E}\left[\tilde{q}_{i,j}^2\right]H_{i,j}^{-1}-g_j(\chi_c,\tilde{c}_{i,j}) \\ &= \frac{H_{i,j}^{-1}}{2}\left(\mathbb{E}[\tilde{q}_{i,j}]^2+\mathbb{V}[\tilde{q}_{i,j}]\right)-g_j(\chi_c,\tilde{c}_{i,j}) \end{aligned} \tag{70}$$

We can compute the mean of the firm's quantity choice by taking an expectation of (47), using the fact that the prior means of all data points s_i and s_j are zero:

$$E[\tilde{q}_{i,j}] = \hat{H}_{i,j}(\bar{p}_j^M-\tilde{c}_{i,j}) \tag{71}$$

To work out the variance term, use the first order condition (36) to rewrite $\mathbb{V}[\tilde{q}_{i,j}]=\hat{H}_{i,j}^2\mathbb{V}\left(\tilde{p}_{i,j}\right)$, since the only ex-ante unknown variable in the first order condition is the price. Next, recognize that the price is a sum of two independent terms, the market price and the demand shock: $\tilde{p}_{i,j}=p_j^M+b_{i,j}$, where p_j^M is given by (46). Thus,

$$\mathbb{V}\left(\tilde{p}_{i,j}\right)=1+\frac{1}{\phi^2}\left(1+\frac{1}{\phi}\bar{H}_j\right)^{-2}\sum_{k=1}^{n_F}(\hat{H}_{k,j}K_{k,j})^2\mathbb{V}(s_{k,j}) \tag{72}$$

If there are n_{dk} data points about each attribute, each with precision 1, then the variance of a firm's average data point $s_{k,j}$ is $(1+1/n_{dk})$. In this case, the formula for the variance of output is $\mathbb{V}[\tilde{q}_{i,j}]=\hat{H}_{i,j}^2\mathbb{V}\left(\tilde{p}_{i,j}\right)$, with the variance of the price given by (72). This yields

$$\mathbb{V}[\tilde{q}_{i,j}]=\hat{H}_{i,j}^2\left[1+\frac{1}{\phi^2}\left(1+\frac{1}{\phi}\bar{H}_j\right)^{-2}\sum_{k=1}^{n_F}(\hat{H}_{k,j}K_{k,j})^2\left(\frac{n_{dk}}{n_{dk}+1}\right)\right] \tag{73}$$

Notice that this part of expected utility is independent of the firm's cost choices.

OPTIMAL CHOICES OF MARGINAL COST The first and second order condition for the optimal marginal cost choice $\tilde{c}_{i,j}$ is

$$\begin{aligned} \frac{\partial\mathbb{E}[U_{i,j}]}{\partial\tilde{c}_{i,j}} &= \frac{1}{2}\frac{\partial\mathbb{E}[\tilde{q}_{i,j}]^2H_{i,j}^{-1}}{\partial\tilde{c}_{i,j}}-\frac{\partial g_j(\chi_c,\tilde{c}_{i,j})}{\partial\tilde{c}_{i,j}}=0 \\ \frac{\partial^2\mathbb{E}[U_{i,j}]}{\partial\tilde{c}_{i,j}^2} &= \frac{1}{2}\frac{\partial^2\mathbb{E}[\tilde{q}_{i,j}]^2H_{i,j}^{-1}}{\partial\tilde{c}_{i,j}^2}-\frac{\partial^2 g_j(\chi_c,\tilde{c}_{i,j})}{\partial\tilde{c}_{i,j}^2}\leq 0 \end{aligned} \tag{74}$$

Since signal noise is diagonal, we have $H_{i,j}^{-1} = \frac{2}{\phi} + \rho_i \mathbb{V}\left[b_i|\mathcal{I}_i\right]$ and $\mathbb{V}\left[b_i|\mathcal{I}_i\right] = (1+n_{di})^{-1}$. Using the fact that $\mathbb{E}\tilde{q}_{i,j} = \hat{H}_{i,j}(\bar{p}_j^M - \tilde{c}_{i,j})$, the FOC and SOC could be written as

$$\begin{aligned}
\frac{\partial \mathbb{E}[U_{i,j}]}{\partial c_{i,j}^{\sim}} &= H_{i,j}^{-1}\mathbb{E}[\tilde{q}_{i,j}]\frac{\partial \mathbb{E}[\tilde{q}_{i,j}]}{\partial \tilde{c}_{i,j}} - \frac{\partial g_j(\chi_c, \tilde{c}_{i,j})}{\partial \tilde{c}_{i,j}} \\
&= \left(\bar{p}_j^M - \tilde{c}_{i,j}\right) H_{i,j}^{-1}\hat{H}_{i,j}^2\left(\frac{\frac{1}{\phi}\hat{H}_{i,j}}{1+\frac{1}{\phi}\bar{H}} - 1\right) - \frac{\partial g_j(\chi_c, \tilde{c}_{i,j})}{\partial \tilde{c}_{i,j}} = 0 \\
\frac{\partial^2 \mathbb{E}[U_i]}{\partial \tilde{c}_{i,j}^2} &= \hat{H}_{i,j}^2 H_{i,j}^{-1}\left(\frac{\frac{1}{\phi}\hat{H}_{i,j}}{1+\frac{1}{\phi}\bar{H}} - 1\right)^2 - \frac{\partial^2 g_j(\chi_c, \tilde{c}_{i,j})}{\partial \tilde{c}_{i,j}^2}
\end{aligned} \tag{75}$$

since the average market price p_j^M for attribute j is

$$\bar{p}_j^M = \frac{\bar{p}_j + \frac{1}{\phi}\sum_{s=1}^{n_F}\hat{H}_{s,j}\tilde{c}_{s,j}}{1+\frac{1}{\phi}\bar{H}} \quad \text{and} \quad \frac{\partial \bar{p}_j^M}{\partial \tilde{c}_{i,j}} = \frac{\frac{1}{\phi}\hat{H}_{i,j}}{1+\frac{1}{\phi}\bar{H}} \tag{76}$$

B. Proofs

In Appendix A., we established the equilibrium solution. In this appendix, we prove the main results discussed in the paper.

Proof of Lemma 1: Pricing with Aggregate Shocks Since firm i observes data summarized by $\boldsymbol{s}_i$, its expectation of the price is $\mathbf{E}[\boldsymbol{p}|\boldsymbol{s}_i] = \bar{\boldsymbol{p}}^M + \boldsymbol{K}_i\boldsymbol{s}_i$, where

$$\boldsymbol{K}_i = \mathbf{Cov}(\boldsymbol{p}, \boldsymbol{s}_i)\mathbf{Var}(\boldsymbol{s}_i)^{-1} \tag{77}$$

The variance of price forecast error is

$$\mathbf{Var}[\boldsymbol{p}|\boldsymbol{s}_i] = \mathbf{Var}(\boldsymbol{p}) - \mathbf{Cov}(\boldsymbol{p}, \boldsymbol{s}_i)\mathbf{Var}(\boldsymbol{s}_i)^{-1}\mathbf{Cov}(\boldsymbol{p}, \boldsymbol{s}_i)' \tag{78}$$

SOLUTION We can rearrange the pricing equation (4) as $\frac{1}{\phi}\sum_{i=1}^{n_F}\tilde{\boldsymbol{q}}_i = \bar{\boldsymbol{p}} + \boldsymbol{b} - \tilde{\boldsymbol{p}}$ and then substitute in the first order condition (7) and the linear pricing rule guess from lemma 1:

$$\begin{aligned}
&\frac{1}{\phi}\sum_{i=1}^{n_F}\hat{\boldsymbol{H}}_i(\bar{\boldsymbol{p}}^M - \boldsymbol{c}_i + \boldsymbol{K}_i\boldsymbol{s}_i) = \bar{\boldsymbol{p}} + \boldsymbol{b} - \left(\bar{\boldsymbol{p}}^M + \boldsymbol{F}\boldsymbol{b} + \sum_{i=1}^{n_F}\boldsymbol{h}_i\boldsymbol{\varepsilon}_i\right) \\
&\left(\boldsymbol{F} - \boldsymbol{I}_N + \frac{1}{\phi}\sum_{i=1}^{n_F}\hat{\boldsymbol{H}}_i\boldsymbol{K}_i\right)\boldsymbol{b} + \sum_{i=1}^{n_F}\left(\boldsymbol{h}_i + \frac{1}{\phi}\hat{\boldsymbol{H}}_i\boldsymbol{K}_i\right)\boldsymbol{\varepsilon}_i + \left(\boldsymbol{I}_N + \frac{1}{\phi}\sum_{i=1}^{n_F}\hat{\boldsymbol{H}}_i\right)\bar{\boldsymbol{p}}^M - \bar{\boldsymbol{p}} - \frac{1}{\phi}\sum_{i=1}^{n_F}\hat{\boldsymbol{H}}_i\boldsymbol{c}_i = 0
\end{aligned} \tag{79}$$

Collecting terms in p delivers the pricing equation $\boldsymbol{p} = \bar{\boldsymbol{p}}^M + \boldsymbol{F}\boldsymbol{b} - \frac{1}{\phi}\sum_{i=1}^{n_F}\hat{\boldsymbol{H}}_i\boldsymbol{K}_i\boldsymbol{\varepsilon}_i$, where $\bar{\boldsymbol{p}}^M$ is given by (45). Matching coefficients proves lemma 1 with the following coefficients:

$$\begin{aligned}
\boldsymbol{F} &= \boldsymbol{I}_N - \frac{1}{\phi}\sum_{i=1}^{n_F}\hat{\boldsymbol{H}}_i\boldsymbol{K}_i \\
\boldsymbol{K}_i &= \left(\boldsymbol{F}n_{di} - \frac{1}{\phi}\hat{\boldsymbol{H}}_i\boldsymbol{K}_i\right)\frac{1}{n_{di}+1} \\
\hat{\boldsymbol{H}}_i &= \left[\rho_i\left(\boldsymbol{F}\boldsymbol{F}' + \frac{1}{n_{di}\phi^2}\sum_{i=1}^{n_F}(\hat{\boldsymbol{H}}_i\boldsymbol{K}_i)^2 - \frac{n_{di}}{n_{di}+1}\left(\boldsymbol{F} - \frac{1}{\phi n_{di}}\hat{\boldsymbol{H}}_i\boldsymbol{K}_i\right)\left(\boldsymbol{F} - \frac{1}{\phi n_{di}}\hat{\boldsymbol{H}}_i\boldsymbol{K}_i\right)'\right) + \frac{\boldsymbol{I}_N}{\phi}\right]^{-1}
\end{aligned} \tag{80}$$

Proof of Lemma 2: Data-Investment Complementarity

Proof. To show this complementarity between information and costs, we first differentiate $\mathbb{E}\left[U_{i,j}\right]$ in (70) with respect to marginal cost. Here, $\tilde{c}_{ij}$ denotes firm i's marginal cost of producing attribute j. $\widehat{H}_{ij}$ denotes the jj-th entry of the

diagonal matrix $\widehat{\boldsymbol{H}}_i$, which captures the sensitivity of i's production of attribute j to a marginal change in the expected profit of producing attribute j. We can simplify the expression derived in (75) a bit further as follows:

$$\begin{aligned}
\frac{\partial \mathbb{E}[U_{i,j}]}{\partial \tilde{c}_{i,j}} &= \left(\bar{p}_j^M - \tilde{c}_{i,j}\right) H_{i,j}^{-1} \hat{H}_{i,j}^2 \left(\frac{\frac{1}{\phi}\hat{H}_{i,j}}{1+\frac{1}{\phi}\bar{H}_j} - 1\right) - \frac{\partial g_j(\chi_c, \tilde{c}_{i,j})}{\partial \tilde{c}_{i,j}} \\
\frac{\partial^2 \mathbb{E}[U_{i,j}]}{\partial \tilde{c}_{i,j} \partial \hat{H}_{i,j}} &= \left(\bar{p}_j^M - \tilde{c}_{i,j}\right) \frac{\partial}{\partial \hat{H}_{i,j}} \left(H_{i,j}^{-1} \hat{H}_{i,j}^2 \left(\frac{\frac{1}{\phi}\hat{H}_{i,j}}{1+\frac{1}{\phi}\bar{H}_j} - 1\right)\right) \\
&= \left(\bar{p}_j^M - \tilde{c}_{i,j}\right) \frac{\partial}{\partial \hat{H}_{i,j}} \left(\left(1+\frac{\hat{H}_{i,j}}{\phi}\right) \hat{H}_{i,j} \left(\frac{\frac{1}{\phi}\hat{H}_{i,j}}{1+\frac{1}{\phi}\bar{H}_j} - 1\right)\right) \\
&= \frac{\bar{p}_j^M - \tilde{c}_{i,j}}{\phi} \frac{\partial}{\partial \hat{H}_{i,j}} \left(\left(\phi + \hat{H}_{i,j}\right) \hat{H}_{i,j} \left(\frac{\hat{H}_{i,j}}{\phi + \bar{H}_j} - 1\right)\right)
\end{aligned} \tag{81}$$

where the third equality uses the fact that $H_{i,j}^{-1} = \hat{H}_{i,j}^{-1} + \frac{1}{\phi}$.

Let $T = \hat{H}_{i,j}(\phi + \hat{H}_{i,j}) \left(\frac{\hat{H}_{i,j}}{\phi+\bar{H}_j} - 1\right)$. Taking log and differentiating w.r.t $\hat{H}_{i,j}$, we obtain

$$\begin{aligned}
\log T &= \log \hat{H}_{i,j} + \log(\phi + \hat{H}_{i,j}) + \log\left(\frac{\hat{H}_{i,j}}{\phi + \bar{H}_j} - 1\right) \\
\frac{1}{T}\frac{\partial T}{\hat{H}_{i,j}} &= \frac{1}{\hat{H}_{i,j}} + \frac{1}{\phi + \hat{H}_{i,j}} + \left(\frac{\hat{H}_{i,j}}{\phi + \bar{H}_j} - 1\right)^{-1} \left(\frac{(\phi + \bar{H}_j) - \hat{H}_{i,j} \sum_{k=1}^{n_F} \frac{\partial \hat{H}_{k,j}}{\partial \hat{H}_{i,j}}}{(\phi + \bar{H}_j)^2}\right) \\
&= \frac{1}{\hat{H}_{i,j}} + \frac{1}{\phi + \hat{H}_{i,j}} + \left(\frac{\phi + \bar{H}_j}{\hat{H}_{i,j} - \phi + \bar{H}_j}\right) \left(\frac{(\phi + \bar{H}_j) - \hat{H}_{i,j} \sum_{k=1}^{n_F} \frac{\partial \hat{H}_{k,j}}{\partial \hat{H}_{i,j}}}{(\phi + \bar{H}_j)^2}\right)
\end{aligned} \tag{82}$$

For the model with private shocks and public data, for $i \neq k$, $\frac{\partial \hat{H}_{i,j}}{\partial \hat{H}_{k,j}} = 0$. So, the above equality becomes

$$\begin{aligned}
\frac{1}{T}\frac{\partial T}{\partial \hat{H}_{i,j}} &= \frac{1}{\hat{H}_{i,j}} + \frac{1}{\phi + \hat{H}_{i,j}} + \frac{1}{\hat{H}_{i,j} - \phi - \bar{H}_j} \left(\frac{\phi + \bar{H}_j - \hat{H}_{i,j}}{\phi + \bar{H}_j}\right) \\
&= \frac{1}{\hat{H}_{i,j}} + \frac{1}{\phi + \hat{H}_{i,j}} - \frac{1}{\phi + \bar{H}_j} \\
\implies \frac{\partial T}{\partial \hat{H}_{i,j}} &= T\left(\frac{1}{\hat{H}_{i,j}} + \frac{1}{\phi + \hat{H}_{i,j}} - \frac{1}{\phi + \bar{H}_j}\right)
\end{aligned} \tag{83}$$

Note that the as $\bar{H}_j = \sum_{k=1}^{n_F} \hat{H}_{k,j} > \hat{H}_{i,j}$, we have that $\frac{1}{\phi+\hat{H}_{i,j}} > \frac{1}{\phi+\bar{H}_j}$. So, the second term on the RHS above is positive. However, $T = \hat{H}_{i,j}(\phi + \hat{H}_{i,j}) \left(\frac{\hat{H}_{i,j}}{\phi+\bar{H}_j} - 1\right) < 0$ as $\phi + \sum_{k=1}^{n_F} \hat{H}_{k,j} > \hat{H}_{i,j}$. Hence, we get that $\frac{\partial T}{\partial \hat{H}_{i,j}} < 0$ and therefore, $\frac{\partial^2 \mathbb{E}[U_i]}{\partial \tilde{c}_{ij} \partial \hat{H}_{i,j}} < 0$, which means the marginal benefit from reducing costs is higher (more negative) when firms have better information (higher sensitivity $\hat{H}_{i,j}$).

Another way to look at this result is to note that from FOC (75), and the fact that $H_{i,j}^{-1} = \hat{H}_{i,j}^{-1} + \frac{1}{\phi}$, we have

$$\frac{\partial \mathbb{E}[U_{i,j}]}{\partial \tilde{c}_{i,j}} = \left(\bar{p}_j^M - \tilde{c}_{i,j}\right) \left(1 + \frac{\hat{H}_{i,j}}{\phi}\right) \hat{H}_{i,j} \left(\frac{\frac{1}{\phi}\hat{H}_{i,j}}{1+\frac{1}{\phi}\bar{H}_j} - 1\right) - \frac{\partial g_j(\chi_c, \tilde{c}_{i,j})}{\partial \tilde{c}_{i,j}} = 0 \tag{84}$$

Define $F(\tilde{c}_{i,j}, \hat{H}_{i,j}) = \left(\bar{p}_j^M - \tilde{c}_{i,j}\right) H_{i,j}^{-1} \hat{H}_{i,j}^2 \left(\frac{\frac{1}{\phi}\hat{H}_{i,j}}{1+\frac{1}{\phi}\bar{H}_j} - 1\right) - \frac{\partial g_j(\chi_c, \tilde{c}_{i,j})}{\partial \tilde{c}_{i,j}}$ Close to the optimal choice of $(\tilde{c}_{i,j}, \hat{H}_{i,j})$, we have

that $F(\tilde{c}_{i,j}, \hat{H}_{i,j}) = 0$. Using implicit function theorem, we can obtain

$$\frac{d\tilde{c}_{i,j}}{d\hat{H}_{i,j}} = -\frac{\frac{\partial F}{\partial \hat{H}_{i,j}}}{\frac{\partial F}{\partial \tilde{c}_{i,j}}} \tag{85}$$

Using the notation above, $\frac{\partial F}{\partial \hat{H}_{i,j}} = (\bar{p}_j^M - \tilde{c}_{i,j})\frac{\partial T}{\partial \hat{H}_{i,j}} < 0$. Also, $\frac{\partial F}{\partial \tilde{c}_{i,j}} = \frac{\partial}{\partial \tilde{c}_{i,j}}\left(\frac{\partial \mathbb{E}[U_{i,j}]}{\partial \tilde{c}_{i,j}}\right) < 0$ by the second order condition (75). Combining the above two results, we get the required result

$$\frac{d\tilde{c}_{i,j}}{d\hat{H}_{i,j}} < 0 \tag{86}$$

□

Proof of Lemma 3: Greater investment raises a firm's product markup.

Proof. More investment would lower marginal cost $\tilde{c}_{i,j}$. The markup effect is

$$\frac{\partial M_{i,j}^{\tilde{p}}}{\partial \tilde{c}_{i,j}} = \frac{\frac{\partial \bar{p}_j^M}{\partial \tilde{c}_{i,j}}\tilde{c}_{i,j} - \bar{p}_j^M}{\tilde{c}_{i,j}^2} = \frac{\frac{1}{\phi}\hat{H}_{i,j}\tilde{c}_{i,j} - \bar{p}_j - \frac{1}{\phi}\sum_{s=1}^{n_F}\hat{H}_{s,j}\tilde{c}_{s,j}}{\tilde{c}_{i,j}^2\left(1+\bar{H}_j\right)} = -\frac{\bar{p}_j + \frac{1}{\phi}\sum_{s=1,s\neq i}^{n_F}\hat{H}_{s,j}\tilde{c}_{s,j}}{\tilde{c}_{i,j}^2\left(1+\bar{H}_j\right)} \leq 0 \tag{87}$$

The negative derivative confirms that more investment leads to higher attribute-level markup. Similarly, for the other attributes j' we have $\frac{\partial M_{i,j'}^{\tilde{p}}}{\partial \tilde{c}_{i,j}} = 0$.

Next, differentiate this product markup with respect to the marginal cost of attribute j. Consider the markup on product k that used attribute j ($A_{kj} > 0$). This markup is $\sum_j A_{k,j}\mathbb{E}[\tilde{p}_{i,j}]/(\sum_j A_{kj}\tilde{c}_{i,j})$. Its derivative is

$$\frac{dM_{i,k}}{d\tilde{c}_{i,j}} = \frac{[\sum_j A_{kj}\tilde{c}_{i,j}]A_{kj}\frac{\partial}{\partial \tilde{c}_{i,j}}\mathbb{E}[\tilde{p}_{i,j}] - [\sum_j A_{k,j}\mathbb{E}[\tilde{p}_{i,j}]]A_{kj}}{[\sum_j A_{kj}\tilde{c}_{i,j}]^2} \tag{88}$$

$$= \frac{A_{kj}}{\sum_j A_{kj}\tilde{c}_{i,j}}\left[\frac{\partial}{\partial \tilde{c}_{i,j}}\mathbb{E}[\tilde{p}_{i,j}] - M_{i,k}\right]. \tag{89}$$

We know that $\frac{\partial}{\partial \tilde{c}_{i,j}}\mathbb{E}[p_{i,j}] < M_{i,j}^{\tilde{p}}$ because earlier in the proof, we established that

$$\frac{dM_{i,j}^{\tilde{p}}}{d\tilde{c}_{i,j}} = \frac{1}{\tilde{c}_{i,j}}\left[\frac{\partial}{\partial \tilde{c}_{i,j}}\mathbb{E}[\tilde{p}_{i,j}] - M_{i,j}^{\tilde{p}}\right] \leq 0. \tag{90}$$

Therefore, (89) is negative if the markup on product k is greater than the markup on attribute j: $M_{i,k} \geq M_{i,j}^{\tilde{p}}$. □

Proof of Lemma 4: (Risk premium channel) Product-level markup decreases in data. When investment is sufficiently inflexible (high χ_c), and product i loads positively on all attributes ($a_{ij} \geq 0$), then the product markup $\mathbb{E}(p_i/c_i) = \mathbb{E}(p_i)/c_i$ is decreasing in data.

Proof. Assume each firm is endowed with a fixed investment ($\boldsymbol{c}_i$). By continuity, the result will extend to cases where the investment is close to fixed, which is when χ_c is sufficiently high. The markup on the attribute j, produced by firm i is $M_{i,j}^{\tilde{p}} := \mathbb{E}[\tilde{p}_{i,j}]/\tilde{c}_{i,j}$. The average markup on the attributes is

$$\overline{M}^{\tilde{p}} = \frac{1}{N}\frac{1}{n_F}\sum_{i=1}^{n_F}\sum_{j=1}^{N}M_{i,j}^{\tilde{p}} = \frac{1}{n_F N}\sum_{i=1}^{n_F}\sum_{j=1}^{N}\frac{\mathbb{E}[\tilde{p}_{i,j}]}{\tilde{c}_{i,j}} = \frac{1}{n_F N}\sum_{i=1}^{n_F}\sum_{j=1}^{N}\frac{\bar{p}_j^M}{\tilde{c}_{i,j}} \tag{91}$$

The j^{th} term of equilibrium price $\bar{\boldsymbol{p}}^M$ is

$$\bar{p}_j^M = \frac{\phi\bar{p}_j + \sum_{i=1}^{n_F}\hat{H}_{i,j}\tilde{c}_{i,j}}{\phi + \bar{H}_j} \quad \text{where} \quad \hat{H}_{i,j} = \left(\rho_i \mathbf{Var}[\tilde{p}_{i,j}|\mathcal{I}_i] + \frac{1}{\phi}\right)^{-1} \quad \text{and} \quad \bar{H}_j = \sum_{i=1}^{n_F}\hat{H}_{i,j} \tag{92}$$

The positive output means $\bar{p}_j^M \geq \tilde{c}_{i,j}$, thus

$$\begin{aligned}
\frac{\partial \bar{p}_j^M}{\partial \hat{H}_{i,j}} &= (\phi + \bar{H}_j)^{-2} \left((\phi + \bar{H}_j) \sum_{k=1}^{n_F} \tilde{c}_{k,j} \frac{\partial \hat{H}_{k,j}}{\partial \hat{H}_{i,j}} - \left(\phi \bar{p}_j + \sum_{i=1}^{n_F} \hat{H}_{i,j} \tilde{c}_{i,j} \right) \sum_{k=1}^{n_F} \frac{\partial \hat{H}_{k,j}}{\partial \hat{H}_{i,j}} \right) \\
&= (\phi + \bar{H}_j)^{-2} \left((\phi + \bar{H}_j) \sum_{k=1}^{n_F} \tilde{c}_{k,j} \frac{\partial \hat{H}_{k,j}}{\partial \hat{H}_{i,j}} - \bar{p}_j^M (\phi + \bar{H}_j) \sum_{k=1}^{n_F} \frac{\partial \hat{H}_{k,j}}{\partial \hat{H}_{i,j}} \right) \\
&= (\phi + \bar{H}_j)^{-1} \sum_{k=1}^{n_F} (\tilde{c}_{k,j} - \bar{p}_j^M) \frac{\partial \hat{H}_{k,j}}{\partial \hat{H}_{i,j}} \\
&= (\phi + \bar{H}_j)^{-1} \left(\tilde{c}_{k,j} - \bar{p}_j^M + \sum_{k \neq i} (\tilde{c}_{k,j} - \bar{p}_j^M) \frac{\partial \hat{H}_{k,j}}{\partial \hat{H}_{i,j}} \right)
\end{aligned} \tag{93}$$

We know from earlier results that for the private shocks model, $\frac{\partial \hat{H}_{k,j}}{\partial \hat{H}_{i,j}} = 0$ for $i \neq k$. For the aggregate shocks model also, $\frac{\partial \hat{H}_{k,j}}{\partial \hat{H}_{i,j}}$ is negligible and can be ignored in comparison to the effect of increase in $\hat{H}_{i,j}$. Therefore,

$$\frac{\partial \bar{p}_j^M}{\partial \hat{H}_{i,j}} = (\phi + \bar{H}_j)^{-1} \left(\tilde{c}_{k,j} - \bar{p}_j^M \right) \leq 0 \tag{94}$$

Since the price of a good is a_i times the vector of attribute prices, and all the attribute prices are decreasing in data, the good price and thus the product-level markup is decreasing in data as well.

We prove the negative first order derivative for fixed choices of cost $\tilde{c}_i$, which corresponds to infinitely high marginal cost $\chi_c \to \infty$. This result is strictly negative and continuous in $\tilde{c}_i$. If we assume χ_c is sufficiently high, this is arbitrarily close to fixed c. By continuity, the inequality will still hold. □

□

Proof of Proposition 1: Product markups increase or decrease in data (net change).

Proof. The product-level markup is $M_{i,j}^{\tilde{p}} = \mathbb{E}[\tilde{p}_{i,j}] / \tilde{c}_{i,j} = \bar{p}_j^M / \tilde{c}_{i,j}$. Its partial derivative to the sensitivity $\hat{H}_{i,j}$ is

$$\frac{\partial M_{i,j}^{\tilde{p}}}{\partial n_{di}} = \frac{1}{\tilde{c}_{i,j}^2} \left(\frac{\partial \bar{p}_j^M}{\partial \hat{H}_{i,j}} \tilde{c}_{i,j} - \bar{p}_j^M \frac{\partial \tilde{c}_{i,j}}{\partial \hat{H}_{i,j}} \right) \frac{\partial \hat{H}_{i,j}}{\partial n_{di}} \tag{95}$$

We have already established that under the conditions considered for this proposition, $\frac{\partial \hat{H}_{i,j}}{\partial n_{di}} > 0$. From (45), we have $\bar{p}_j^M = \left(\phi + \bar{H}_j \right)^{-1} \left(\phi \bar{p}_j + \sum_i \hat{H}_{i,j} \tilde{c_{i,j}} \right)$. For the private-shocks-public-data model, we have

$$\begin{aligned}
\frac{\partial \bar{p}_j^M}{\partial \hat{H}_{i,j}} &= \frac{(\phi + \bar{H}_j) \tilde{c}_{i,j} - (\phi \bar{p}_j + \sum_k \hat{H}_{k,j} \tilde{c}_{k,j})}{(\phi + \bar{H}_j)^2} \\
&= \frac{(\phi + \bar{H}_j) \tilde{c}_{i,j} - \bar{p}_j^M (\phi + \bar{H}_j)}{(\phi + \bar{H}_j)^2} \\
&= \frac{\tilde{c}_{i,j} - \bar{p}_j^M}{\phi + \bar{H}_j} < 0
\end{aligned} \tag{96}$$

Now, from Lemma 2, we have

$$\frac{\partial \tilde{c}_{i,j}}{\partial \hat{H}_{i,j}} < 0 \tag{97}$$

If marginal cost $\tilde{c}_{i,j}$ or price of risk ρ_i is sufficiently low, the second term in the numerator $-\bar{p}_j^M \frac{\partial \tilde{c}_{i,j}}{\partial n_{di}} > 0$ dominates the marginal effect, thus increasing product markups.

□

Proof of Proposition 2: The firm-level markup wedge increases in data.

Proof. The firm-level markup wedge is given by

$$M_i^f - \bar{M}_i^p = \frac{\mathbb{E}[\tilde{\boldsymbol{q}}_i'\tilde{\boldsymbol{p}}_i]}{\mathbb{E}[\tilde{\boldsymbol{q}}_i'\tilde{\boldsymbol{c}}_i]} - \frac{1}{N}\sum_{j=1}^{n_F}\mathbb{E}\left[\frac{\tilde{p}_{i,j}}{\tilde{c}_{i,j}}\right] \tag{98}$$

Data has free disposal. So, the expected utility $\mathbb{E}[U_i]$ of firm i must be (weakly) increasing in its data (holding the data of all other firms fixed). By Lemma 2 (data-investment complementarity), the upfront investment $(g(\chi_c, \tilde{\boldsymbol{c}}_i))$ is increasing in data. Therefore, it follows from (3) that when the price of risk is low, the expected profits of firm i must be increasing in data i.e. the the following must hold

$$\frac{\partial}{\partial n_{di}}(\mathbb{E}\left[\tilde{\boldsymbol{q}}_i'\tilde{\boldsymbol{p}}_i\right] - \mathbb{E}\left[\tilde{\boldsymbol{q}}_i'\tilde{\boldsymbol{c}}_i\right]) > 0 \tag{99}$$

(99) indicates that starting from any level of $(\tilde{\boldsymbol{p}}_i, \tilde{\boldsymbol{q}}_i, \tilde{\boldsymbol{c}}_i)$, when a firm i expects to get more data (n_{di}), either $\mathbb{E}\left[\tilde{\boldsymbol{q}}_i'\tilde{\boldsymbol{p}}_i\right]$ goes up or $\mathbb{E}\left[\tilde{\boldsymbol{q}}_i'\tilde{\boldsymbol{c}}_i\right]$ goes down, or both. Any of these three changes would result in an increase in firm markup M_i^f. Therefore, M_i^f goes up unambiguously. By Lemma 4, the average product mark-up $\bar{M}_i^p$ goes down with data (n_{di}). Therefore, the firm-level markup wedge $M_i^f - \bar{M}_i^p$ increases in data.

□

Proof of Proposition 3a: Wedge between cost-weighted firm markup and average firm markup.

This proof shows that high-data firms produce more on average. Therefore, they have larger impacts on cost-weighted industry markup, increasing the industry-level markup wedge.

Proof. The cost weight for firm i is

$$w_i^{cost} = \frac{\mathbb{E}\left[\tilde{\boldsymbol{q}}_i'\tilde{\boldsymbol{c}}_i\right]}{\sum_{k=1}^{n_F}\mathbb{E}\left[\tilde{\boldsymbol{q}}_k'\tilde{\boldsymbol{c}}_k\right]} = \frac{\sum_{l=1}^{N}\mathbb{E}[\tilde{q}_{i,l}]\tilde{c}_{i,l}}{\sum_{k=1}^{n_F}\sum_{l=1}^{N}\mathbb{E}[\tilde{q}_{k,l}]\tilde{c}_{k,l}} \tag{100}$$

We show below that this weight is increasing in data for the firm i. Taking the partial derivative of the weight with respect to the number of data points n_{di}, we get

$$\frac{\partial w_i^{cost}}{\partial n_{di}} = \frac{(\sum_{k=1}^{n_F}\mathbb{E}\left[\tilde{\boldsymbol{q}}_k'\tilde{\boldsymbol{c}}_k\right])\left(\sum_{j=1}^{N}\tilde{c}_{i,j}\frac{\partial\mathbb{E}[\tilde{q}_{i,j}]}{\partial n_{di}}\right) - \mathbb{E}[\tilde{\boldsymbol{q}}_i'\tilde{\boldsymbol{c}}_i]\left(\sum_{j=1}^{N}\tilde{c}_{i,j}\frac{\partial\mathbb{E}[\tilde{q}_{i,j}]}{\partial n_{di}} + \sum_{k=1,k\neq i}^{n_F}\sum_{j=1}^{N}\tilde{c}_{k,j}\frac{\partial\mathbb{E}[\tilde{q}_{k,j}]}{\partial n_{di}}\right)}{(\sum_{k=1}^{n_F}\mathbb{E}\left[\tilde{\boldsymbol{q}}_k'\tilde{\boldsymbol{c}}_k\right])^2} \tag{101}$$

Now, we use that fact that $\mathbb{E}[\tilde{q}_{k,j}] = \hat{H}_{k,j}(\bar{p}_j^M - \tilde{c}_{k,j})$. Differentiating this w.r.t n_{di}, we obtain

$$\frac{\partial\mathbb{E}[\tilde{q}_{k,j}]}{\partial n_{di}} = (\bar{p}_j^M - \tilde{c}_{k,j})\frac{\partial\hat{H}_{k,j}}{\partial n_{di}} + \hat{H}_{k,j}\frac{\partial\bar{p}_j^M}{\partial n_{di}} \tag{102}$$

(61) shows that $\frac{\partial\bar{p}_j^M}{\partial n_{di}} < 0$. For the model with private shocks and public data, $\hat{H}_{k,j}$ does not depend on the amount of data for any other firm. Therefore, for $k \neq i$, $\frac{\partial\hat{H}_{k,j}}{\partial n_{di}} = 0$. For the model with aggregate shocks and private data, $k \neq i$, $\frac{\partial\hat{H}_{k,j}}{\partial n_{di}}$ is negligible in comparison to the change in market price $\frac{\partial\bar{p}_j^M}{\partial n_{di}}$. Therefore, in either case, for $k \neq i$, $\frac{\partial\mathbb{E}[\tilde{q}_{k,j}]}{\partial n_{di}} < 0$. Using this fact in the (101), we obtain

$$\begin{aligned}\frac{\partial w_i^{cost}}{\partial n_{di}} &\geq \frac{(\sum_{k=1}^{n_F}\mathbb{E}\left[\tilde{\boldsymbol{q}}_k'\tilde{\boldsymbol{c}}_k\right])\left(\sum_{j=1}^{N}\tilde{c}_{i,j}\frac{\partial\mathbb{E}[\tilde{q}_{i,j}]}{\partial n_{di}}\right) - \mathbb{E}[\tilde{\boldsymbol{q}}_i'\tilde{\boldsymbol{c}}_i]\left(\sum_{j=1}^{N}\tilde{c}_{i,j}\frac{\partial\mathbb{E}[\tilde{q}_{i,j}]}{\partial n_{di}}\right)}{(\sum_{k=1}^{n_F}\mathbb{E}\left[\tilde{\boldsymbol{q}}_k'\tilde{\boldsymbol{c}}_k\right])^2} \\ &\geq \frac{(\sum_{k=1,k\neq i}^{n_F}\mathbb{E}\left[\tilde{\boldsymbol{q}}_k'\tilde{\boldsymbol{c}}_k\right])\left(\sum_{j=1}^{N}\tilde{c}_{i,j}\frac{\partial\mathbb{E}[\tilde{q}_{i,j}]}{\partial n_{di}}\right)}{(\sum_{k=1}^{n_F}\mathbb{E}\left[\tilde{\boldsymbol{q}}_k'\tilde{\boldsymbol{c}}_k\right])^2} > 0\end{aligned} \tag{103}$$

where the last inequality follows because we know from (63) that $\forall j \in [1,\ldots,N]$, $\frac{\partial\mathbb{E}[\tilde{q}_{i,j}]}{\partial n_{di}} > 0$. We know from (63) that high-data firms produce more on average and the result above indicates that these firms have larger impacts on cost-weighted industry markup than their low-data counterparts. We know from Proposition 2 that the firm-level markup increases in data if the price of risk ρ is sufficiently small. Since more data increases both w_i^{cost} and M_i^f for a firm,

it makes the expected product $E[w_i^{cost} M_i^f]$ greater than the unweighted sum $\bar{M}^f$. This logic holds for fixed costs $\tilde{c}_i$. However, Lemma 3 shows that data reduces the costs firms choose, which increases markups. Thus the cost channel increases markups even more, for the highly-weighted firms.

□

Proof of Proposition 3b: Sales weighted vs cost-weighted markup

Proof. The result assumes that firms are ex ante identical. Under this assumption, when one firm gets more data, the wedge between the sales weighted markup and cost weighted markup goes up. Using the expressions for sales weighted markup (19) and cost weighted markup (18), we can express the wedge as

$$\begin{aligned} M^s - M^c &= \sum_{k=1}^{n_F} \left(\underbrace{\frac{\mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{p}}_k\right]}{\sum_{k=1}^{n_F} \mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{p}}_k\right]}}_{w_k^s} - \underbrace{\frac{\mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{c}}_k\right]}{\sum_{k=1}^{n_F} \mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{c}}_k\right]}}_{w_k^c} \right) \underbrace{\frac{\mathbb{E}\left[\tilde{\boldsymbol{q}}_i' \tilde{\boldsymbol{p}}_k\right]}{\mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{c}}_k\right]}}_{M_k^f} \\ &= \sum_{k=1}^{n_F} \left(w_k^s - w_k^c\right) M_k^f = \sum_{k=1}^{n_F} \hat{w}_k M_k^f \end{aligned} \tag{104}$$

where $\hat{w}_k \equiv w_k^s - w_k^c$ denotes the difference in weights for firm k. It is easy to see that $\sum_{k=1}^{n_F} \hat{w}_k = 0$ as $\sum_{k=1}^{n_F} w_k^s = \sum_{k=1}^{n_F} w_k^c = 1$. Next, we define the average firm markup as $\bar{M}^f = \frac{1}{n_F} \sum_{k=1}^{n_F} M_k^f$. We can now rewrite the wedge as

$$\begin{aligned} M^s - M^c &= \sum_{k=1}^{n_F} \hat{w}_k (M_k^f - \bar{M}^f + \bar{M}^f) \\ &= \sum_{k=1}^{n_F} \hat{w}_k (M_k^f - \bar{M}^f) + \bar{M}^f \underbrace{\sum_{k=1}^{n_F} \hat{w}_k}_{=0} \\ &= \sum_{k=1}^{n_F} \hat{w}_k (M_k^f - \bar{M}^f) = \sum_{k=1}^{n_F} \hat{w}_k \hat{M}_k^f \end{aligned} \tag{105}$$

where $\hat{M}_k^f \equiv M_k^f - \bar{M}^f$. When firms are identical, this wedge is zero as by symmetry, $M_k^f = \bar{M}^f \implies \hat{M}_k^f = 0$. Therefore, it is sufficient to show that if some firm i gets more data, the wedge becomes positive. We prove this by showing that an increase in n_{di} makes each term of the summation in (105) positive.

First note that by definition $\sum_{k=1}^{n_F} \hat{M}_k^f = 0$. Differentiating this expression with respect to n_{di}, we obtain

$$\frac{\partial \hat{M}_i^f}{\partial n_{di}} + \sum_{k \neq i} \frac{\partial \hat{M}_k^f}{\partial n_{di}} = 0 \tag{106}$$

From Proposition 2, we know that the firm-markup increases in firm i's data n_{di}. Therefore, $\frac{\partial \hat{M}_i^f}{\partial n_{di}} > 0$. Now, note that starting from all firms being identical, the $n_F - 1$ firms which do not get additional data are still identical to each other. Using this in the expression above, we obtain for any firm $k \neq i$,

$$\begin{aligned} \frac{\partial \hat{M}_i^f}{\partial n_{di}} + (n_F - 1) \frac{\partial \hat{M}_k^f}{\partial n_{di}} &= 0 \\ \frac{\partial \hat{M}_k^f}{\partial n_{di}} &= -\frac{1}{n_F - 1} \frac{\partial \hat{M}_i^f}{\partial n_{di}} < 0 \end{aligned} \tag{107}$$

By similar calculations for $\sum_{k=1}^{n_F} \hat{w}_k$, we obtain that

$$\frac{\partial \hat{w}_k}{\partial n_{di}} = -\frac{1}{n_F - 1} \frac{\partial \hat{w}_i}{\partial n_{di}} \tag{108}$$

As we start the analysis from identical firms where $\hat{M}_k^f = \hat{w}_k = 0, \forall k \in \{1, \ldots, n_F\}$, the above inequalities imply that if a firm i gets more data, $\hat{M}_i^f$ becomes positive and $\hat{M}_k^f$ becomes negative for $k \neq i$. For the weight differences, we have that the change in $\hat{w}_k$ for all the other firms $k \neq i$ goes in the opposite direction to the change for firm i. To obtain the

required result, it is now sufficient to show that an increase in n_{di} makes $\hat{w}_i$ positive. To this end, note that

$$\frac{\partial \hat{w}_i}{\partial n_{di}} = \frac{\partial w_i^s}{\partial n_{di}} - \frac{\partial w_i^c}{\partial n_{di}} \tag{109}$$

We can calculate the expressions for each of these terms separately and then combine them. For the first term, we obtain from the definition of sales weight

$$\begin{aligned}
\frac{\partial w_i^s}{\partial n_{di}} &= \frac{\partial}{\partial n_{di}} \frac{\mathbb{E}\left[\tilde{\boldsymbol{q}}_i' \tilde{\boldsymbol{p}}_i\right]}{\sum_{k'=1}^{n_F} \mathbb{E}\left[\tilde{\boldsymbol{q}}_{k'}' \tilde{\boldsymbol{p}}_{k'}\right]} \\
&= w_i^s \left(\frac{1}{\mathbb{E}\left[\tilde{\boldsymbol{q}}_i' \tilde{\boldsymbol{p}}_i\right]} \frac{\partial \mathbb{E}\left[\tilde{\boldsymbol{q}}_i' \tilde{\boldsymbol{p}}_i\right]}{\partial n_{di}} - \frac{1}{\sum_{k'} \mathbb{E}\left[\tilde{\boldsymbol{q}}_{k'}' \tilde{\boldsymbol{p}}_{k'}\right]} \sum_{k'} \frac{\partial \mathbb{E}\left[\tilde{\boldsymbol{q}}_{k'}' \tilde{\boldsymbol{p}}_{k'}\right]}{\partial n_{di}} \right) \\
&= w_i^s \left(\frac{1}{\mathbb{E}\left[\tilde{\boldsymbol{q}}_i' \tilde{\boldsymbol{p}}_i\right]} \frac{\partial \mathbb{E}\left[\tilde{\boldsymbol{q}}_i' \tilde{\boldsymbol{p}}_i\right]}{\partial n_{di}} - \frac{1}{\sum_{k'} \mathbb{E}\left[\tilde{\boldsymbol{q}}_{k'}' \tilde{\boldsymbol{p}}_{k'}\right]} \frac{\partial \mathbb{E}\left[\tilde{\boldsymbol{q}}_i' \tilde{\boldsymbol{p}}_i\right]}{\partial n_{di}} - \frac{1}{\sum_{k'} \mathbb{E}\left[\tilde{\boldsymbol{q}}_{k'}' \tilde{\boldsymbol{p}}_{k'}\right]} \sum_{k' \neq i} \frac{\partial \mathbb{E}\left[\tilde{\boldsymbol{q}}_{k'}' \tilde{\boldsymbol{p}}_{k'}\right]}{\partial n_{di}} \right)
\end{aligned}$$

Using the assumption that all firms are ex-ante identical and all the $n_F - 1$ firms which do not get additional data are ex-post identical to each other, we obtain

$$\frac{\partial w_i^s}{\partial n_{di}} = \frac{n_F - 1}{n_F^2 \mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{p}}_k\right]} \left(\frac{\partial \mathbb{E}\left[\tilde{\boldsymbol{q}}_i' \tilde{\boldsymbol{p}}_i\right]}{\partial n_{di}} - \frac{\partial \mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{p}}_k\right]}{\partial n_{di}} \right) \tag{110}$$

Similarly for the cost weight, we get

$$\frac{\partial w_i^c}{\partial n_{di}} = \frac{n_F - 1}{n_F^2 \mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{c}}_k\right]} \left(\frac{\partial \mathbb{E}\left[\tilde{\boldsymbol{q}}_i' \tilde{\boldsymbol{c}}_i\right]}{\partial n_{di}} - \frac{\partial \mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{c}}_k\right]}{\partial n_{di}} \right) \tag{111}$$

Combining the results for firm i, we get

$$\begin{aligned}
\frac{\partial}{\partial n_{di}}(w_i^s - w_i^c) &= \frac{n_F - 1}{n_F^2 \mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{p}}_k\right]} \left(\frac{\partial \mathbb{E}\left[\tilde{\boldsymbol{q}}_i' \tilde{\boldsymbol{p}}_i\right]}{\partial n_{di}} - \frac{\partial \mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{p}}_k\right]}{\partial n_{di}} \right) \\
&\quad - \frac{n_F - 1}{n_F^2 \mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{c}}_k\right]} \left(\frac{\partial \mathbb{E}\left[\tilde{\boldsymbol{q}}_i' \tilde{\boldsymbol{c}}_i\right]}{\partial n_{di}} - \frac{\partial \mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{c}}_k\right]}{\partial n_{di}} \right) \\
&= \frac{n_F - 1}{n_F^2} \left(\frac{\frac{\partial \mathbb{E}\left[\tilde{\boldsymbol{q}}_i' \tilde{\boldsymbol{p}}_i\right]}{\partial n_{di}} - \frac{\partial \mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{p}}_k\right]}{\partial n_{di}}}{\mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{p}}_k\right]} - \frac{\frac{\partial \mathbb{E}\left[\tilde{\boldsymbol{q}}_i' \tilde{\boldsymbol{c}}_i\right]}{\partial n_{di}} - \frac{\partial \mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{c}}_k\right]}{\partial n_{di}}}{\mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{c}}_k\right]} \right)
\end{aligned} \tag{112}$$

where k is any firm different from i. Next, we use the results derived for markups earlier in the proof to show that the RHS of the above expression (112) is positive. From Proposition 2 and (107), we know that $\frac{\partial \hat{M}_i^f}{\partial n_{di}} > 0$ and $\frac{\partial \hat{M}_k^f}{\partial n_{di}} < 0$. Therefore,

$$\begin{aligned}
& \frac{\partial \hat{M}_i^f}{\partial n_{di}} > \frac{\partial \hat{M}_k^f}{\partial n_{di}} \\
\iff \quad & \frac{\partial M_i^f}{\partial n_{di}} - \frac{\partial \bar{M}^f}{\partial n_{di}} > \frac{\partial M_k^f}{\partial n_{di}} - \frac{\partial \bar{M}^f}{\partial n_{di}} \\
\iff \quad & \frac{\partial}{\partial n_{di}} \frac{\mathbb{E}\left[\tilde{\boldsymbol{q}}_i' \tilde{\boldsymbol{p}}_i\right]}{\mathbb{E}\left[\tilde{\boldsymbol{q}}_i' \tilde{\boldsymbol{c}}_i\right]} > \frac{\partial}{\partial n_{di}} \frac{\mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{p}}_k\right]}{\mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{c}}_k\right]} \\
\iff \quad & \frac{1}{\mathbb{E}\left[\tilde{\boldsymbol{q}}_i' \tilde{\boldsymbol{c}}_i\right]} \frac{\partial \mathbb{E}\left[\tilde{\boldsymbol{q}}_i' \tilde{\boldsymbol{p}}_i\right]}{\partial n_{di}} - \frac{\mathbb{E}\left[\tilde{\boldsymbol{q}}_i' \tilde{\boldsymbol{p}}_i\right]}{(\mathbb{E}\left[\tilde{\boldsymbol{q}}_i' \tilde{\boldsymbol{c}}_i\right])^2} \frac{\partial \mathbb{E}\left[\tilde{\boldsymbol{q}}_i' \tilde{\boldsymbol{c}}_i\right]}{\partial n_{di}} > \frac{1}{\mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{c}}_k\right]} \frac{\partial \mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{p}}_k\right]}{\partial n_{di}} - \frac{\mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{p}}_k\right]}{(\mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{c}}_k\right])^2} \frac{\partial \mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{c}}_k\right]}{\partial n_{di}} \\
\iff \quad & \frac{1}{\mathbb{E}\left[\tilde{\boldsymbol{q}}_i' \tilde{\boldsymbol{c}}_i\right]} \frac{\partial \mathbb{E}\left[\tilde{\boldsymbol{q}}_i' \tilde{\boldsymbol{p}}_i\right]}{\partial n_{di}} - \frac{1}{\mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{c}}_k\right]} \frac{\partial \mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{p}}_k\right]}{\partial n_{di}} > \frac{\mathbb{E}\left[\tilde{\boldsymbol{q}}_i' \tilde{\boldsymbol{p}}_i\right]}{(\mathbb{E}\left[\tilde{\boldsymbol{q}}_i' \tilde{\boldsymbol{c}}_i\right])^2} \frac{\partial \mathbb{E}\left[\tilde{\boldsymbol{q}}_i' \tilde{\boldsymbol{c}}_i\right]}{\partial n_{di}} - \frac{\mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{p}}_k\right]}{(\mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{c}}_k\right])^2} \frac{\partial \mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{c}}_k\right]}{\partial n_{di}} \\
\iff \quad & \frac{1}{\mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{p}}_k\right]} \left(\frac{\partial \mathbb{E}\left[\tilde{\boldsymbol{q}}_i' \tilde{\boldsymbol{p}}_i\right]}{\partial n_{di}} - \frac{\partial \mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{p}}_k\right]}{\partial n_{di}} \right) > \frac{1}{\mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{c}}_k\right]} \left(\frac{\partial \mathbb{E}\left[\tilde{\boldsymbol{q}}_i' \tilde{\boldsymbol{c}}_i\right]}{\partial n_{di}} - \frac{\partial \mathbb{E}\left[\tilde{\boldsymbol{q}}_k' \tilde{\boldsymbol{c}}_k\right]}{\partial n_{di}} \right)
\end{aligned} \tag{113}$$

where the last inequality follows because we have assumed that the firms are ex-ante identical. In such a setting, the

expected costs and revenues are the same for all the firms a priori (i.e. $\mathbb{E}\left[\tilde{\boldsymbol{q}}_i'\tilde{\boldsymbol{c}}_i\right] = \mathbb{E}\left[\tilde{\boldsymbol{q}}_k'\tilde{\boldsymbol{c}}_k\right]$ and $\mathbb{E}\left[\tilde{\boldsymbol{q}}_i'\tilde{\boldsymbol{p}}_i\right] = \mathbb{E}\left[\tilde{\boldsymbol{q}}_k'\tilde{\boldsymbol{p}}_k\right]$)

Using the above result (113) in (112) calculated earlier, we conclude that $\hat{w}_i$, the difference in sales-weight and cost-weight for firm i, increases with data n_{di} and for $k \neq i$, $\hat{w}_k$ decreases with n_{di} (using 108). Therefore, we have shown that when firm i gets more data, $\hat{M}_i^f$ and $\hat{w}_i$ become positive and for all the remaining firms $k \neq i$, $\hat{M}_k^f$ and $\hat{w}_k$ become negative. Rewriting (105) below

$$M^s - M^c = \sum_{k=1}^{n_F} \hat{w}_k \hat{M}_k^f \tag{114}$$

we note that each term in the summation is positive which means that $M^s - M^c$ is positive. When all firms are identical, the wedge $M^s - M^c = 0$ and when one firm gets more data, the wedge becomes positive. Therefore, an increase in one firm's data increases the wedge. □

Proof of Proposition 3c: Sales-weighted vs. industry aggregates markup The reason this corollary follows directly from Proposition 3b, that the cost-weighted industry markup and the aggregate markup are the same, in our setting. This is a version of the aggregation results of Edmond, Midrigan, and Xu (2019), extended to our linear demand system. The proof is just algebraic manipulation:

$$M^{ag} := \frac{\mathbb{E}\left[\sum_{i=1}^N \boldsymbol{q}_i'\boldsymbol{p}_i\right]}{\mathbb{E}\left[\sum_{i=1}^N \boldsymbol{q}_i'\boldsymbol{c}_i\right]} = \frac{\sum_{i=1}^N \mathbb{E}\left[\boldsymbol{q}_i'\boldsymbol{p}_i\right]}{\sum_{i=1}^N \mathbb{E}\left[\boldsymbol{q}_i'\boldsymbol{c}_i\right]} = \sum_{i=1}^N w_i^m M_i^f = M^m \quad \text{where} \quad w_i^c = \frac{\mathbb{E}\left[\boldsymbol{q}_i'\boldsymbol{c}_i\right]}{\sum_{i=1}^N \mathbb{E}\left[\boldsymbol{q}_i'\boldsymbol{c}_i\right]}. \tag{115}$$

Proof of proposition 4: Cyclical Markups Part a: product markups are increasing in demand variance and converge to a constant.

Proof. Let $\sigma_b I_N$ denote the variance of demand shocks b. According to the definition of $\hat{\boldsymbol{H}}_i$, we have

$$\begin{aligned} \hat{\boldsymbol{H}}_i &= \left(\frac{\boldsymbol{I}_N}{\phi} + \rho_i \mathbf{Var}(\tilde{\boldsymbol{p}}_i|\mathcal{I}_i)\right)^{-1} \quad \text{and} \quad \mathbf{Var}(\tilde{\boldsymbol{p}}_i|\mathcal{I}_i) = \left(\sigma_b^{-1} + n_{di}\right)^{-1} \\ \Rightarrow \lim_{\sigma_b \to \infty} \mathbf{Var}(\tilde{\boldsymbol{p}}_i|\mathcal{I}_i) &= 1/n_{di}, \quad \tilde{\hat{\boldsymbol{H}}}_i := \lim_{\sigma_b \to \infty} \hat{\boldsymbol{H}}_i = \left(\frac{\boldsymbol{I}_N}{\phi} + \rho_i / n_{di}\right)^{-1} \end{aligned} \tag{116}$$

The equilibrium price is given by

$$\mathbb{E}\left[\tilde{\boldsymbol{p}}_i\right] = \tilde{\boldsymbol{p}}^M = \left(\boldsymbol{I}_N + \frac{1}{\phi}\sum_{i=1}^{n_F} \hat{\boldsymbol{H}}_i\right)^{-1} \left(\bar{\boldsymbol{p}} + \frac{1}{\phi}\sum_{i=1}^{n_F} \hat{\boldsymbol{H}}_i \boldsymbol{c}_i\right) \tag{117}$$

It clearly converges due to convergent $\hat{\boldsymbol{H}}_i$, so we have

$$\begin{aligned} \tilde{\boldsymbol{p}} := \lim_{\sigma_b \to \infty} \mathbb{E}\left[\tilde{\boldsymbol{p}}_i\right] &= \left(\boldsymbol{I}_N + \frac{1}{\phi}\sum_{i=1}^{n_F} \lim_{\sigma_b \to \infty} \hat{\boldsymbol{H}}_i\right)^{-1} \left(\bar{\boldsymbol{p}} + \frac{1}{\phi}\sum_{i=1}^{n_F} \lim_{\sigma_b \to \infty} \hat{\boldsymbol{H}}_i \boldsymbol{c}_i\right) \\ &= \left[\boldsymbol{I}_N + \sum_{i=1}^{n_F} \left(\boldsymbol{I}_N + \phi\rho_i/n_{di}\right)^{-1}\right]^{-1} \left[\bar{\boldsymbol{p}} + \sum_{i=1}^{n_F} \boldsymbol{c}_i \left(\boldsymbol{I}_N + \phi\rho_i/n_{di}\right)^{-1}\right] \end{aligned} \tag{118}$$

This result implies convergent product-level markup on the attributes as $\lim_{\sigma_b \to \infty} \overline{M}^p$ exists. Since equilibrium price on the goods is a linear combination of weight matrix $\boldsymbol{A}$ and $\tilde{\boldsymbol{p}}_i$, the product-level markup on the goods converges.

$$\boldsymbol{q}_i = \boldsymbol{A}\tilde{\boldsymbol{q}}_i \quad \text{and} \quad \boldsymbol{p}_i = \boldsymbol{A}\tilde{\boldsymbol{p}}_i \Rightarrow \overline{M}^p = \frac{1}{N}\frac{1}{n_F}\sum_{i=1}^{n_F}\sum_{j=1}^{N} \frac{\left(\boldsymbol{A}\mathbb{E}\left[\tilde{\boldsymbol{p}}_i\right]\right)_j}{\left(\boldsymbol{A}\boldsymbol{c}_i\right)_j} \quad \text{converges.} \tag{119}$$

If all the firms have identical sizes ($\boldsymbol{c}_i = \bar{\boldsymbol{c}}$), the derivative of equilibrium price for specific attribute j is

$$\frac{\partial \mathbb{E}[\tilde{\boldsymbol{p}}_{i,j}]}{\partial \Sigma_{b,j}} = \frac{(\bar{\boldsymbol{c}}_j - \bar{\boldsymbol{p}}_j)\frac{1}{\phi}\sum_{i=1}^{n_F} \frac{\partial \hat{\boldsymbol{H}}_{i,j}}{\partial \Sigma_{b,j}}}{\left(1 + \frac{1}{\phi}\sum_{i=1}^{n_F} \hat{\boldsymbol{H}}_{i,j}\right)^2} \quad \text{and} \quad \frac{\partial \hat{\boldsymbol{H}}_{i,j}}{\partial \Sigma_{b,j}} = -\frac{\hat{\boldsymbol{H}}_{i,j}^2 \rho_i / n_{di}^2}{\left(\Sigma_{b.j} + 1/n_{di}\right)^2} \leq 0 \tag{120}$$

Since positive production implies lower marginal cost ($\bar{\boldsymbol{c}}_j \leq \bar{\boldsymbol{p}}_j$), the numerator of the derivative is positive. □

Part b: Firm and industry level markups are increasing in demand variance. They asymptote to a linearly increasing function of demand variance.

Proof. First, We will show that the trace of the covariance $\mathbf{tr}[\mathbf{Cov}\left(\tilde{\boldsymbol{p}}_i, \tilde{\boldsymbol{q}}_i\right)]$ is always positive.

$$\begin{aligned}\mathbf{Cov}\left(\tilde{\boldsymbol{p}}_i, \tilde{\boldsymbol{q}}_i\right) = &\left(\boldsymbol{I}_N + \sum_{j=1}^{n_F} \frac{\hat{\boldsymbol{H}}_j}{\phi}\right)^{-1} \sum_{j=1}^{n_F} \hat{\boldsymbol{H}}_j \mathbf{Var}(\boldsymbol{K}_j \boldsymbol{s}_j) \hat{\boldsymbol{H}}_j \left(\boldsymbol{I}_N + \sum_{j=1}^{n_F} \frac{\hat{\boldsymbol{H}}_j}{\phi}\right)^{-1} \frac{\hat{\boldsymbol{H}}_i}{\phi^2} + \mathbf{Var}(\boldsymbol{K}_i \boldsymbol{s}_i) \hat{\boldsymbol{H}}_i \\ &- \left(\boldsymbol{I}_N + \sum_{j=1}^{n_F} \frac{\hat{\boldsymbol{H}}_j}{\phi}\right)^{-1} \hat{\boldsymbol{H}}_i \mathbf{Var}(\boldsymbol{K}_i \boldsymbol{s}_i) \frac{\hat{\boldsymbol{H}}_i}{\phi} - \mathbf{Var}(\boldsymbol{K}_i \boldsymbol{s}_i) \hat{\boldsymbol{H}}_i \left(\boldsymbol{I}_N + \sum_{j=1}^{n_F} \frac{\hat{\boldsymbol{H}}_j}{\phi}\right)^{-1} \frac{\hat{\boldsymbol{H}}_i}{\phi}\end{aligned} \tag{121}$$

Denote $\boldsymbol{Y}$ the sum of price impacts $\boldsymbol{Y} = \boldsymbol{I}_N + \sum_{j=1}^{n_F} \frac{\hat{\boldsymbol{H}}_j}{\phi}$. The trace covariance becomes

$$\begin{aligned}&\mathbf{tr}[\mathbf{Cov}\left(\tilde{\boldsymbol{p}}_i, \tilde{\boldsymbol{q}}_i\right)] = \\ &\mathbf{tr}\left[\boldsymbol{Y}^{-1} \sum_{j=1}^{n_F} \hat{\boldsymbol{H}}_j \mathbf{Var}(\boldsymbol{K}_j \boldsymbol{s}_j) \hat{\boldsymbol{H}}_j \boldsymbol{Y}^{-1} \frac{\hat{\boldsymbol{H}}_i}{\phi^2}\right] + \mathbf{tr}\left[\mathbf{Var}(\boldsymbol{K}_i \boldsymbol{s}_i) \hat{\boldsymbol{H}}_i\right] - \mathbf{tr}\left[\boldsymbol{Y}^{-1} \hat{\boldsymbol{H}}_i \mathbf{Var}(\boldsymbol{K}_i \boldsymbol{s}_i) \frac{\hat{\boldsymbol{H}}_i}{\phi}\right] - \mathbf{tr}\left[\mathbf{Var}(\boldsymbol{K}_i \boldsymbol{s}_i) \hat{\boldsymbol{H}}_i \boldsymbol{Y}^{-1} \frac{\hat{\boldsymbol{H}}_i}{\phi}\right] \\ \geq &\mathbf{tr}\left[\boldsymbol{Y}^{-1} \hat{\boldsymbol{H}}_i \mathbf{Var}(\boldsymbol{K}_i \boldsymbol{s}_i) \hat{\boldsymbol{H}}_i \boldsymbol{Y}^{-1} \frac{\hat{\boldsymbol{H}}_i}{\phi^2}\right] + \mathbf{tr}\left[\mathbf{Var}(\boldsymbol{K}_i \boldsymbol{s}_i) \hat{\boldsymbol{H}}_i\right] - \mathbf{tr}\left[\boldsymbol{Y}^{-1} \hat{\boldsymbol{H}}_i \mathbf{Var}(\boldsymbol{K}_i \boldsymbol{s}_i) \frac{\hat{\boldsymbol{H}}_i}{\phi}\right] - \mathbf{tr}\left[\mathbf{Var}(\boldsymbol{K}_i \boldsymbol{s}_i) \hat{\boldsymbol{H}}_i \boldsymbol{Y}^{-1} \frac{\hat{\boldsymbol{H}}_i}{\phi}\right] \\ = &\mathbf{tr}\left[\mathbf{Var}(\boldsymbol{K}_i \boldsymbol{s}_i) \hat{\boldsymbol{H}}_i \boldsymbol{Y}^{-1} \frac{\hat{\boldsymbol{H}}_i}{\phi^2} \boldsymbol{Y}^{-1} \hat{\boldsymbol{H}}_i\right] + \mathbf{tr}\left[\mathbf{Var}(\boldsymbol{K}_i \boldsymbol{s}_i) \hat{\boldsymbol{H}}_i\right] - \mathbf{tr}\left[\mathbf{Var}(\boldsymbol{K}_i \boldsymbol{s}_i) \frac{\hat{\boldsymbol{H}}_i}{\phi} \boldsymbol{Y}^{-1} \hat{\boldsymbol{H}}_i\right] - \mathbf{tr}\left[\mathbf{Var}(\boldsymbol{K}_i \boldsymbol{s}_i) \hat{\boldsymbol{H}}_i \boldsymbol{Y}^{-1} \frac{\hat{\boldsymbol{H}}_i}{\phi}\right] \\ = &\phi \mathbf{tr}\left[\mathbf{Var}(\boldsymbol{K}_i \boldsymbol{s}_i) \left(\frac{\hat{\boldsymbol{H}}_i}{\phi} \boldsymbol{Y}^{-1} \frac{\hat{\boldsymbol{H}}_i}{\phi} \boldsymbol{Y}^{-1} \frac{\hat{\boldsymbol{H}}_i}{\phi} + \frac{\hat{\boldsymbol{H}}_i}{\phi} - \frac{\hat{\boldsymbol{H}}_i}{\phi} \boldsymbol{Y}^{-1} \frac{\hat{\boldsymbol{H}}_i}{\phi} - \frac{\hat{\boldsymbol{H}}_i}{\phi} \boldsymbol{Y}^{-1} \frac{\hat{\boldsymbol{H}}_i}{\phi}\right)\right] \\ = &\phi \mathbf{tr}\left[\mathbf{Var}(\boldsymbol{K}_i \boldsymbol{s}_i) \frac{\hat{\boldsymbol{H}}_i}{\phi} \left(\boldsymbol{Y}^{-1} \frac{\hat{\boldsymbol{H}}_i}{\phi} - \boldsymbol{I}_N\right)^2\right] \geq 0\end{aligned} \tag{122}$$

We denote $\boldsymbol{x}_i = \frac{\hat{\boldsymbol{H}}_i}{\phi}$ and $Z_i = \mathbf{Var}(\boldsymbol{K}_i \boldsymbol{s}_i) = \frac{\Sigma_b^2}{\sigma_b + \Sigma_i}$ and consider diagonal shock and signal variance. $\boldsymbol{x}_i$, $\boldsymbol{Y}$ and Z_i are diagonal under our assumption. The covariance matrix is simplified as

$$\mathbf{Cov}\left(\tilde{\boldsymbol{p}}_i, \tilde{\boldsymbol{q}}_i\right) = \phi \left[\boldsymbol{Y}^{-1} \sum_{j=1}^{n_F} \boldsymbol{x}_i Z_j \boldsymbol{x}_i \boldsymbol{Y}^{-1} \boldsymbol{x}_i + Z_i \boldsymbol{x}_i - \boldsymbol{Y}^{-1} \boldsymbol{x}_i Z_i \boldsymbol{x}_i - Z_i \boldsymbol{x}_i \boldsymbol{Y}^{-1} \boldsymbol{x}_i\right] \tag{123}$$

The covariance matrix is also diagonal and denote the k^{th} diagonal $\mathbf{Cov}_{i,k}$. Subscript k refers to the k^{th} diagonal value.

$$\begin{aligned}\mathbf{Cov}_{i,k} &:= \mathbf{Cov}\left(\boldsymbol{p}_{i,k}, \tilde{\boldsymbol{q}}_{i,k}\right) \\ &= \phi \left[\boldsymbol{Y}_k^{-1} \sum_{j=1}^{n_F} \boldsymbol{x}_{j,k} Z_{j,k} \boldsymbol{x}_{j,k} \boldsymbol{Y}_k^{-1} \boldsymbol{x}_{i,k} + Z_i \boldsymbol{x}_{i,k} - \boldsymbol{Y}_k^{-1} \boldsymbol{x}_{i,k} Z_{i,k} \boldsymbol{x}_{i,k} - Z_{i,k} \boldsymbol{x}_{i,k} \boldsymbol{Y}_k^{-1} \boldsymbol{x}_{i,k}\right] \\ &= \phi \frac{\boldsymbol{x}_{i,k}}{\boldsymbol{Y}_k^2} \left[\sum_{j \neq i, j=1}^{n_F} \boldsymbol{x}_{j,k}^2 Z_{j,k} + Z_{i,k} \left(\boldsymbol{x}_{i,k} - \boldsymbol{Y}_k\right)^2\right]\end{aligned} \tag{124}$$

The limiting behavioral for all variables are

$$\begin{aligned}\lim_{\sigma_b \to \infty} \boldsymbol{x}_{i,k} &= \left(1 + \phi \rho_i / n_{di}\right)^{-1} \\ \lim_{\sigma_b \to \infty} \boldsymbol{Y}_k &= 1 + \sum_{j=1}^{n_F} \lim_{\sigma_b \to \infty} \boldsymbol{x}_{j,k} = 1 + \sum_{j=1}^{n_F} \left(1 + \phi \rho_j / n_{di}\right)^{-1} \\ \lim_{\sigma_b \to \infty} \frac{Z_{i,k}}{\sigma_b} &= \lim_{\sigma_b \to \infty} \frac{\frac{\Sigma_{b,k}^2}{\sigma_b + \Sigma_{i,k}}}{\sigma_b} = 1\end{aligned} \tag{125}$$

The ratio of covariance to shock variance converges as

$$\lim_{\sigma_b\to\infty}\frac{\mathbf{Cov}_{i,k}}{\sigma_b}=\frac{\phi\left(1+\phi\rho_i/n_{di}\right)^{-1}\left[\sum_{j\neq i,j=1}^{n_F}\left(1+\phi\rho_j/n_{di}\right)^{-2}+\left(1+\sum_{j=1,j\neq i}^{n_F}\left(1+\phi\rho_j/n_{di}\right)^{-1}\right)^2\right]}{\left(1+\sum_{j=1}^{n_F}\left(1+\phi\rho_j/n_{di}\right)^{-1}\right)^2} \tag{126}$$

we have

$$\begin{aligned}M_i^f&=\frac{\mathbb{E}[\tilde{\boldsymbol{q}}_i'\tilde{\boldsymbol{p}}_i]}{\mathbb{E}[\tilde{\boldsymbol{q}}_i'\boldsymbol{c}_i]}=\frac{\mathbb{E}[\tilde{\boldsymbol{q}}_i]'\mathbb{E}[\boldsymbol{p}]+\mathbf{tr}\left[\mathbf{Cov}(\tilde{\boldsymbol{p}}_i,\tilde{\boldsymbol{q}}_i)\right]}{\mathbb{E}[\tilde{\boldsymbol{q}}_i'\boldsymbol{c}_i]}\\&=\frac{\sum_{j=1}^N(\mathbb{E}(\tilde{\boldsymbol{p}}_{i,j})-\boldsymbol{c}_{i,j})\mathbb{E}(\tilde{\boldsymbol{p}}_{i,j})\hat{\boldsymbol{H}}_{i,j}+\sum_{j=1}^N\mathbf{Cov}_{i,j}}{\sum_{j=1}^N(\mathbb{E}(\tilde{\boldsymbol{p}}_{i,j})-\boldsymbol{c}_{i,j})\boldsymbol{c}_{i,j}\hat{\boldsymbol{H}}_{i,j}}\end{aligned} \tag{127}$$

We assume the diagonal values of shock variance are the same, so the asymptote of M_i^f is

$$\begin{aligned}\alpha_i&\coloneqq\lim_{\sigma_b\to\infty}\frac{M_i^f}{\sigma_b}=\frac{\sum_{j=1}^N\lim_{\Sigma_b\to\infty}\frac{\mathbf{Cov}_{i,j}}{\sigma_b}}{\sum_{j=1}^N(\tilde{\boldsymbol{p}}_j-\boldsymbol{c}_{i,j})\boldsymbol{c}_{i,j}\tilde{\hat{\boldsymbol{H}}}_{i,j}}>0\\\gamma_i&\coloneqq\lim_{\sigma_b\to\infty}\left(M_i^f-\alpha_i\sigma_b\right)=\frac{\sum_{j=1}^N(\tilde{\boldsymbol{p}}_j-\boldsymbol{c}_{i,j})\tilde{\boldsymbol{p}}_j\tilde{\hat{\boldsymbol{H}}}_{i,j}+\widetilde{\mathbf{Cov}}_{i,j}}{\sum_{j=1}^N(\tilde{\boldsymbol{p}}_j-\boldsymbol{c}_{i,j})\boldsymbol{c}_{i,j}\tilde{\hat{\boldsymbol{H}}}_{i,j}}\end{aligned} \tag{128}$$

where the difference $\widetilde{\mathbf{Cov}}_i$ is defined as

$$\begin{aligned}\widetilde{\mathbf{Cov}}_i&\coloneqq\lim_{\sigma_b\to\infty}\left(\mathbf{Cov}_{i,k}-\left(\lim_{\sigma_b\to\infty}\frac{\mathbf{Cov}_{i,k}}{\sigma_b}\right)\sigma_b\right)\\&=-\frac{\phi\left(1+\phi\rho_i/n_{di}\right)^{-1}\left[\sum_{j\neq i,j=1}^{n_F}\left(1+\phi\rho_j/n_{di}\right)^{-2}n_{di}^{-1}+\left(1+\sum_{j=1,j\neq i}^{n_F}\left(1+\phi\rho_j/n_{di}\right)^{-1}\right)^2/n_{di}\right]}{\left(1+\sum_{j=1}^{n_F}\left(1+\phi\rho_j/n_{di}\right)^{-1}\right)^2}\end{aligned} \tag{129}$$

The average firm-level markup $\overline{M}^f=(1/n_F)\sum_{i=1}^{n_F}M_i^f$ approaches $\sum_{i=1}^{n_F}\frac{\alpha_i}{n_F}\sigma_b+\sum_{i=1}^{n_F}\frac{\gamma_i}{n_F}$ in the long run. The economy-level markup is $M^m=\sum_{i=1}^{n_F}w^{\boldsymbol{H}_i}M_i^f$ with $w^{\boldsymbol{H}_i}=\frac{\mathbb{E}[\tilde{\boldsymbol{q}}_i'\boldsymbol{c}_i]}{\sum_{i=1}^{n_F}\mathbb{E}[\tilde{\boldsymbol{q}}_i'\boldsymbol{c}_i]}$. The weight $w^{\boldsymbol{H}_i}$ converges to w_i as shock variance goes to infinity, implying an asymptote of economy-level markup.

$$w_i\coloneqq\lim_{\sigma_b\to\infty}w^{\boldsymbol{H}_i}=\frac{\sum_{j=1}^N(\tilde{\boldsymbol{p}}_j-\boldsymbol{c}_{i,j})\boldsymbol{c}_{i,j}\tilde{\hat{\boldsymbol{H}}}_{i,j}}{\sum_{i=1}^{n_F}\sum_{j=1}^N(\tilde{\boldsymbol{p}}_j-\boldsymbol{c}_{i,j})\boldsymbol{c}_{i,j}\tilde{\hat{\boldsymbol{H}}}_{i,j}}\Rightarrow M^m\text{ approaches }\sum_{i=1}^{n_F}w_i\alpha_i\sigma_b+\sum_{i=1}^{n_F}w_i\gamma_i \tag{130}$$

Finally, the derivative of each component of covariance is

$$\begin{aligned}\frac{\partial\boldsymbol{x}_{i,k}}{\partial\sigma_b}&=-\phi\rho_i\boldsymbol{x}_{i,k}^2\left(n_{di}\sigma_b+I_N\right)^{-2}=-\frac{\boldsymbol{x}_{i,k}(1-\boldsymbol{x}_{i,k})}{\sigma_b(\sigma_b n_{di}+I_N)}\\\frac{\partial\boldsymbol{Y}_k}{\partial\sigma_b}&=\sum_{j=1}^{n_F}\frac{\partial\boldsymbol{x}_{j,k}}{\partial\sigma_b}=-\sum_{j=1}^{n_F}\frac{\boldsymbol{x}_{j,k}(1-\boldsymbol{x}_{j,k})}{\sigma_b(n_{di}\sigma_b+I_N)}\\\frac{\partial Z_{i,k}}{\partial\sigma_b}&=\frac{\sigma_b(n_{di}\sigma_b+2I_N)}{(n_{di}\sigma_b+I_N)^2}=\frac{Z_{i,k}(n_{di}\sigma_b+2I_N)}{\Sigma_{b,k}\left(n_{di}\sigma_b+I_N\right)}\\\frac{\partial\frac{\boldsymbol{x}_{i,k}}{\boldsymbol{Y}_k^2}}{\partial\sigma_b}&=\frac{\boldsymbol{x}_{i,k}}{\sigma_b\boldsymbol{Y}_k^2}\left[\frac{2}{\boldsymbol{Y}_k}\sum_{j=1}^{n_F}\frac{\boldsymbol{x}_{j,k}(1-\boldsymbol{x}_{j,k})}{n_{di}\sigma_b+I_N}-\frac{(1-\boldsymbol{x}_{i,k})}{n_{di}\sigma_b+I_N}\right]\\\frac{\partial\sum_{j\neq i,j=1}^{n_F}\boldsymbol{x}_{j,k}^2Z_{j,k}}{\partial\sigma_b}&=\sum_{j\neq i,j=1}^{n_F}\frac{\boldsymbol{x}_{j,k}^2Z_{j,k}}{(n_{di}\sigma_b+I_N)\sigma_b}\left[n_{di}\sigma_b+2I_N-2(1-\boldsymbol{x}_{j,k})\right]\\\frac{\partial Z_{i,k}\left(\boldsymbol{x}_{i,k}-\boldsymbol{Y}_k\right)^2}{\partial\sigma_b}&=\frac{Z_{i,k}\left(\boldsymbol{x}_{i,k}-\boldsymbol{Y}_k\right)^2(n_{di}\sigma_b+2I_N)}{(n_{di}\sigma_b+2I_N)\sigma_b}-2\left(\boldsymbol{Y}_k-\boldsymbol{x}_{i,k}\right)\frac{Z_{i,k}}{\sigma_b}\sum_{j\neq i,j=1}^{n_F}\boldsymbol{x}_{j,k}(1-\boldsymbol{x}_{j,k})(n_{di}\sigma_b+I_N)^{-1}\end{aligned} \tag{131}$$

So the derivative of covariance $\mathbf{Cov}_{i,k}$ could be decomposed into two parts

$$\frac{\partial \mathbf{Cov}_{i,k}}{\partial \sigma_b} = \phi \frac{\boldsymbol{x}_{i,k}}{\boldsymbol{Y}_k^2} \left[\boldsymbol{G}_1 + \boldsymbol{G}_2\right] \tag{132}$$

where

$$\begin{aligned}
\boldsymbol{G}_1 &\coloneqq Z_{i,k} \left(\boldsymbol{x}_{i,k} - \boldsymbol{Y}_k\right)^2 \frac{n_{di}\sigma_b + (1+\boldsymbol{x}_{i,k})}{n_{di}\sigma_b + 1} - 2Z_{i,k} \left(\boldsymbol{Y}_k - \boldsymbol{x}_{i,k}\right) \frac{\boldsymbol{x}_{i,k}}{\boldsymbol{Y}_k} \sum_{j\neq i,j=1}^{n_F} \frac{\boldsymbol{x}_{j,k}(1-\boldsymbol{x}_{j,k})}{n_{di}\sigma_b + 1} \\
\boldsymbol{G}_2 &\coloneqq \frac{n_{di}\sigma_b + \boldsymbol{x}_{i,k}}{n_{di}\sigma_b + 1} \sum_{j\neq i,j=1}^{n_F} \boldsymbol{x}_{j,k}^2 Z_{j,k} + \frac{2}{\boldsymbol{Y}_k} \sum_{j=1}^{n_F} \frac{\boldsymbol{x}_{j,k}(1-\boldsymbol{x}_{j,k})}{\sigma_b + 1} \sum_{j\neq i,j=1}^{n_F} \boldsymbol{x}_{j,k}^2 Z_{j,k} \\
&\qquad + \sum_{j\neq i,j=1}^{n_F} \boldsymbol{x}_{j,k}^2 Z_{j,k} \frac{1}{n_{di}\sigma_b + 1} \left[1 - 2(1-\boldsymbol{x}_{j,k})\right]
\end{aligned} \tag{133}$$

We can prove that $\boldsymbol{G}_1$ is always positive

$$\boldsymbol{G}_1 \geq 0 \Leftrightarrow Z_{i,k} \left(\boldsymbol{x}_{i,k} - \boldsymbol{Y}_k\right)^2 \frac{\sigma_b + (1+\boldsymbol{x}_{i,k})}{n_{di}\sigma_b + 1} \geq 2Z_{i,k} \left(\boldsymbol{Y}_k - \boldsymbol{x}_{i,k}\right) \frac{\boldsymbol{x}_{i,k}}{\boldsymbol{Y}_k} \sum_{j\neq i,j=1}^{n_F} \frac{\boldsymbol{x}_{j,k}(1-\boldsymbol{x}_{j,k})}{n_{di}\sigma_b + 1} \tag{134}$$

Since $\boldsymbol{Y}_k \geq 1 + \boldsymbol{x}_{i,k}$ and $0 \leq \boldsymbol{x}_{i,k} \leq 1$, we have

$$\begin{aligned}
Z_{i,k} \left(\boldsymbol{x}_{i,k} - \boldsymbol{Y}_k\right)^2 \frac{n_{di}\sigma_b + (1+\boldsymbol{x}_{i,k})}{n_{di}\sigma_b + 1} &\geq Z_{i,k} \left(\boldsymbol{Y}_k - \boldsymbol{x}_{i,k}\right)^2 \\
&\geq Z_{i,k} \left(\boldsymbol{Y}_k - \boldsymbol{x}_{i,k}\right) \frac{2\boldsymbol{x}_{i,k}}{\boldsymbol{Y}_k} \left(1 + \sum_{j\neq i,j=1}^{n_F} \boldsymbol{x}_{j,k}\right) \\
&\geq Z_{i,k} \left(\boldsymbol{Y}_k - \boldsymbol{x}_{i,k}\right) \frac{2\boldsymbol{x}_{i,k}}{\boldsymbol{Y}_k} \sum_{j\neq i,j=1}^{n_F} \frac{\boldsymbol{x}_{j,k}(1-\boldsymbol{x}_{j,k})}{n_{di}\sigma_b + 1}
\end{aligned} \tag{135}$$

As for the $\boldsymbol{G}_2$, large shock variance ($\sigma_b \geq 1/n_{di}, \forall j$) guarantees its positivity since

$$\begin{aligned}
&\sigma_b \geq 1/n_{di} \Rightarrow \frac{n_{di}\sigma_b}{n_{di}\sigma_b + 1} \geq \frac{1}{n_{di}\sigma_b + n_{di}} \\
\Rightarrow &\frac{n_{di}\sigma_b + \boldsymbol{x}_{i,k}}{n_{di}\sigma_b + 1} \sum_{j\neq i,j=1}^{n_F} \boldsymbol{x}_{j,k}^2 Z_{j,k} \geq \sum_{j\neq i,j=1}^{n_F} \boldsymbol{x}_{j,k}^2 Z_{j,k} \frac{1}{n_{di}\sigma_b + 1} \\
\Rightarrow &\boldsymbol{G}_2 \geq \sum_{j\neq i,j=1}^{n_F} \boldsymbol{x}_{j,k}^2 Z_{j,k} \frac{1}{n_{di}\sigma_b + 1} \left[2 - 2(1-\boldsymbol{x}_{j,k})\right] \geq 0
\end{aligned} \tag{136}$$

So the derivative of covariance $\mathbf{Cov}_{i,k}$ is positive when shock variance is large enough.

This proof held marginal costs $\tilde{c}$ fixed. If we assume the marginal cost of adjusting c is sufficiently high, by continuity, the inequality will still hold. □

Cyclical Markups with Efficient Investment : The trade-off between the risk premium and motivation effect still exists here. When the variances of shocks increase, the firm has a tendency to charge a higher price in order to compensate for the increasing risk. On the other hand, they will become less willing to invest, which leads to higher production costs and thus drive markups down. Depending on the parameters, aggregated markups can increase or decrease in booms.

Proof of Proposition 5: Welfare Firm profits are given by (69), where the production decisions are given by (7).

Consumer surplus is the area under the demand curve that lies above the equilibrium price. Since there are N attributes, we sum up the surplus from each attribute. In (4), we showed that the demand curve is a linear function, $\tilde{p}_j^M = \bar{p} - \frac{1}{\phi}\sum_i \widetilde{q}_i$. Consumer surplus is the triangle under this demand curve, that lies above the equilibrium price. In the private shocks model, firm i gets price $\tilde{p}_{i,j} = \tilde{p}_j^M + b_i =$. For the firm i, product j, the height of this triangle is $\bar{p} - (\bar{p} - 1/\phi \sum_k \widetilde{q}_{k,j} + b_i) = \frac{1}{\phi}\sum_k \widetilde{q}_k + b_{i,j}$. The width of this triangle is $\widetilde{q}_{i,j}$. So, the area of this triangle is $1/2$ its base times the height: $\frac{1}{2}\widetilde{q}_{i,j}(\frac{1}{\phi}\sum_k \widetilde{q}_{k,j} + b_{i,j})$. The total consumer surplus is the sum of this expression over all firms i.e.:

$\frac{1}{2}\sum_i \tilde{q}_{i,j}(\frac{1}{\phi}\sum_{k,j}\tilde{q}_{k,j} + b_{i,j})$. We take the expectation of this product as the measure of expected consumer surplus.

$$\mathbb{E}CS = \mathbb{E}\left[\frac{1}{2}\sum_i \tilde{q}_{i,j}(\frac{1}{\phi}\sum_k \tilde{q}_{k,j} + b_{i,j})\right] \tag{137}$$

$$= \frac{1}{2}\left[\frac{1}{\phi}\mathbb{E}[(\sum_i \tilde{q}_{i,j})^2] + \sum_i \mathbb{E}[\tilde{q}_{i,j}b_{i,j}]\right] \tag{138}$$

We separately calculate both the terms on RHS and then put them together. From, (41),

$$\sum_i \tilde{q}_{i,j} = \sum_i \hat{H}_{i,j}\left(\bar{p}_j + K_{i,j}s_{i,j} - \frac{1}{\phi}\sum_{i=1}^{n_F}\tilde{q}_{i,j} - \tilde{c}_{i,j}\right) \tag{139}$$

For symmetric firms, this becomes

$$\sum_i \tilde{q}_{i,j} = n_F\zeta_j(\tilde{p}_j - \tilde{c}_j) - n_F\zeta_j\frac{1}{\phi}\sum_i \tilde{q}_{k,j} + \zeta_j K_j\sum_i s_{i,j} \tag{140}$$

where $\zeta_j = \hat{H}_{i,j}, \tilde{c}_j = \tilde{c}_{i,j}, K_j = K_{i,j}$ for all firms i by symmetry. Rearranging to isolate total output yields

$$\sum_i \tilde{q}_{i,j} = \left(1 + \frac{n_F\zeta_j}{\phi}\right)^{-1}\left(n_F\zeta_j(\tilde{p}_j - \tilde{c}_j) + \zeta_j K_j\sum_i s_{i,j}\right) \tag{141}$$

Next, we square both sides and take expectations. Note that $\mathbb{E}[s_{i,j}] = 0$, $\mathbb{E}[(s_{i,j})^2] = \mathbb{V}[s_{i,j}]$, and independence of shocks and noise across firms implies that $\forall i \neq k, \mathbb{E}[s_{i,j}s_{k,j}] = 0$. Therefore,

$$\begin{aligned}\mathbb{E}[(\sum_i \tilde{q}_{i,j})^2] &= \left(1 + \frac{n_F\zeta_j}{\phi}\right)^{-2}\left(n_F^2\zeta_j^2(\tilde{p}_j - \tilde{c}_j)^2 + \zeta_j^2K_j^2\sum_i \mathbb{V}[s_{i,j}]\right) \\ &= \left(1 + \frac{n_F\zeta_j}{\phi}\right)^{-2}\left(n_F^2\zeta_j^2(\tilde{p}_j - \tilde{c}_j)^2 + n_F\zeta_j^2K_j^2V_j\right)\end{aligned} \tag{142}$$

where V_j denotes $\mathbb{V}[s_{i,j}]$ for all i.

For the second term in (138), note that multiplying the equation (39) with the firm specific price shock b_i, we get

$$b_{i,j}\tilde{q}_{i,j} = \hat{H}_{i,j}\left((\bar{p}_j - \tilde{c}_{i,j})b_{i,j} + K_{i,j}s_{i,j}b_{i,j} - \frac{b_{i,j}}{\phi}\sum_{k=1}^{n_F}\tilde{q}_{k,j}\right) \tag{143}$$

Note that $s_{i,j} = b_{i,j} + \epsilon_{i,j}$ and $\mathbb{E}[\epsilon_{i,j}] = \mathbb{E}[b_{i,j}] = \mathbb{E}[\epsilon_{i,j}b_{i,j}] = 0$ and they are uncorrelated and have mean 0. Therefore,

$$\mathbb{E}[b_{i,j}\tilde{q}_{i,j}] = \hat{H}_{i,j}\left(K_{i,j}\mathbb{E}[b_{i,j}^2] - \frac{1}{\phi}\mathbb{E}\left[b_{i,j}\sum_{k=1}^{n_F}\tilde{q}_{k,j}\right]\right) \tag{144}$$

The variance of $b_{i,j}$ is assumed to be 1. Therefore, $\mathbb{E}[b_{i,j}^2] = 1$. For the second, we multiply both sides of (141) by $b_{i,j}$ to obtain

$$\begin{aligned}b_{i,j}\sum_k \tilde{q}_{k,j} &= \left(1 + \frac{n_F\zeta_j}{\phi}\right)^{-1}\left(n_F\zeta_j(\tilde{p}_j - \tilde{c}_j)b_{i,j} + \zeta_j K_j b_{i,j}\sum_k s_{k,j}\right) \\ \implies \quad \mathbb{E}[b_{i,j}\sum_k \tilde{q}_{k,j}] &= \left(1 + \frac{n_F\zeta_j}{\phi}\right)^{-1}\left(\zeta_j K_j\mathbb{E}[b_{i,j}^2]\right) = \left(1 + \frac{n_F\zeta_j}{\phi}\right)^{-1}\zeta_j K_j\end{aligned} \tag{145}$$

Therefore,

$$\begin{aligned}\mathbb{E}[b_{i,j}\tilde{q}_{i,j}] &= \zeta_j\left(K_j - \frac{1}{\phi}\left(1 + \frac{n_F\zeta_j}{\phi}\right)^{-1}\zeta_j K_j\right) \\ \implies \quad \sum_i \mathbb{E}[b_{i,j}\tilde{q}_{i,j}] &= n_F\zeta_j K_j\left(1 - \frac{1}{\phi}\left(1 + \frac{n_F\zeta_j}{\phi}\right)^{-1}\zeta_j\right)\end{aligned} \tag{146}$$

Plugging back these expressions in (138), we obtain

$$\mathbb{E}CS = \frac{1}{2\phi}\left(1+\frac{n_F\zeta_j}{\phi}\right)^{-2}\left(n_F^2\zeta_j^2(\tilde{p}_j-\tilde{c}_j)^2+n_F\zeta_j^2K_j^2V_j\right)+\frac{1}{2}n_F\zeta_jK_j\left(1-\frac{1}{\phi}\left(1+\frac{n_F\zeta_j}{\phi}\right)^{-1}\zeta_j\right) \tag{147}$$

Welfare is consumer surplus (138) plus the sum of all firms' profits (69). From (70), we have

$$\mathbb{E}[U_{i,j}] = \frac{H_{i,j}^{-1}}{2}\mathbb{E}[(\tilde{q}_{i,j})^2] - g_j(\chi_c,\tilde{c}_{i,j}) \tag{148}$$

Since the result considers a case with symmetric firms, we can use the expressions for the symmetric model. Using, (47) with symmetry of firms, we get

$$\tilde{q}_{i,j} = \zeta_j(\bar{p}_j^M-\tilde{c}_j)+\zeta_jK_js_{i,j}-\frac{\zeta_j^2K_j}{\phi}\left(1+\frac{n_F\zeta_j}{\phi}\right)^{-1}\sum_k s_{k,j} \tag{149}$$

Therefore,

$$\begin{aligned}
\mathbb{E}[\tilde{q}_{i,j}] &= \zeta_j(\bar{p}_j^M-\tilde{c}_j)\\
\mathbb{E}[\tilde{q}_{i,j}^2] &= \zeta_j^2(\bar{p}_j^M-\tilde{c}_j)^2+\zeta_j^2K_j^2\mathbb{E}[s_{i,j}^2]+\frac{\zeta_j^4K_j^2}{\phi^2}\left(1+\frac{n_F\zeta_j}{\phi}\right)^{-2}\sum_k\mathbb{E}[s_{k,j}^2]+\\
&\quad 2\frac{\zeta^3K_j^2}{\phi}\left(1+\frac{n_F\zeta_j}{\phi}\right)^{-1}\mathbb{E}[s_{i,j}^2]\\
&= \zeta_j^2(\bar{p}_j^M-\tilde{c}_j)^2+\zeta_j^2K_j^2V_j+\frac{\zeta_j^4K_j^2}{\phi^2}\left(1+\frac{n_F\zeta_j}{\phi}\right)^{-2}n_FV_j+\\
&\quad 2\frac{\zeta_j^3K_j^2}{\phi}\left(1+\frac{n_F\zeta_j}{\phi}\right)^{-1}V_j
\end{aligned} \tag{150}$$

Also, $H_{i,j}^{-1} = \hat{H}_{i,j}^{-1}+\frac{1}{\phi}$. So, for symmetrical firms, we have $H_{i,j}^{-1} = \frac{1}{\zeta_j}+\frac{1}{\phi}$ for all i. Also, in the private shocks model, $K_j = \frac{n_d}{1+n_d}$ and $V_j = 1+\frac{1}{n_d}$. Therefore, $K_jV_j = 1$. Combining these expressions, we have the total profits of all firms for attribute j as

$$\begin{aligned}
\sum_i E[U_{i,j}] &= \frac{n_F\zeta_j^2}{2}(\zeta_j^{-1}+\phi^{-1})(\bar{p}_j^M-\tilde{c}_j)^2+\\
&\quad \frac{n_F\zeta_j^2K_j}{2}(\zeta_j^{-1}+\phi^{-1})\left(1+\frac{\zeta_j^2}{\phi^2}\left(1+\frac{n_F\zeta_j}{\phi}\right)^{-2}n_F+2\frac{\zeta_j}{\phi}\left(1+\frac{n_F\zeta_j}{\phi}\right)^{-1}\right)\\
&\quad -n_Fg_j(\chi_c,\tilde{c}_j)
\end{aligned} \tag{151}$$

The next step is to differentiate each of these terms with respect to n_d, to show that welfare increases in the number of data points. We differentiate each term in the firms' profit function, holding the marginal cost of the firm $\tilde{c}_j$ fixed. Then, we come back at the end to include the marginal effect of data on firms' marginal cost choices as well. First, define

$$\partial\zeta_j/\partial n_d = \partial\hat{H}_{i,j}/\partial n_{di} > 0$$

Then, consider the first term of consumer surplus and the first term of firms' profits jointly. These terms include the surplus that is shifted from firms to consumers when equilibrium prices change.

$$\begin{aligned}
ECS1+E\Pi1 &= \left[\frac{1}{2\phi}\left(1+\frac{n_F\zeta_j}{\phi}\right)^{-2}n_F^2\zeta_j^2+n_F\zeta_j^2\frac{\zeta_j^{-1}+\phi^{-1}}{2}\right](\tilde{p}_j-\tilde{c}_j)^2\\
&= \left[\frac{1}{2\phi}\left(\frac{1}{\zeta_j}+\frac{n_F}{\phi}\right)^{-2}n_F^2+\frac{n_F\zeta_j}{2}\left(1+\frac{\zeta_j}{\phi}\right)\right](\tilde{p}_j-\tilde{c}_j)^2
\end{aligned} \tag{152}$$

The squared difference between $\bar{\boldsymbol{p}}$ and $\bar{\boldsymbol{c}}$ is always positive, as is ϕ. It is easy to see that both the terms inside the brackets are strictly increasing in ζ_j. Therefore, $ECS1+E\Pi1$ is strictly increasing in data. Thus, this part of welfare is increasing in ζ_j. Since $\partial\zeta_j/\partial n_d > 0$, they are increasing in n_d as well, holding the cost, $\bar{c}_j$, fixed, for the moment.

However, this increase masks the fact that it is the consumers who are gaining from lower prices and firms that lose

profits. The firms lose because more data makes firms less uncertain. Less uncertain firms produce more. More output lowers firms' markups. However, firms will still gain from the reduction in risk, if they factor risk pricing into their objective, i.e. if $\rho > 0$.

Next, I show that the second term in (151) is increasing in data. We can rewrite the second term as

$$E\Pi 2 = \frac{n_F \zeta_j K_j}{2}\left(1+\frac{\zeta_j}{\phi}\right)\left(1+\frac{n_F}{(\frac{\phi}{\zeta_j}+n_F)^2}+\frac{2}{(\frac{\phi}{\zeta_j}+n_F)}\right) \tag{153}$$

First note that both ζ_j and K_j are increasing in data. Therefore, the terms $n_F\zeta_j K_j$ and $1+\zeta_j/\phi$ are increasing in data. Also, out of the remaining three terms, the first one is constant and both the second and the third term have ζ_j in the denominator of their denominator. Therefore, these terms are also increasing in data which makes the whole expression strictly increasing in data. The last term in the firm's payoff function is the fixed cost of investment. Holding marginal cost $\tilde{c}_j$ constant, this term is not affected by change in data.

Finally, consider the remaining terms of consumer surplus. The 2nd term is given by

$$\begin{aligned}\mathbb{E}CS2 &= \frac{1}{2\phi}\left(1+\frac{n_F\zeta_j}{\phi}\right)^{-2} n_F\zeta_j^2 K_j^2 V_j \\ &= \frac{n_F K_j}{2\phi\left(\frac{1}{\zeta_j}+\frac{n_F}{\phi}\right)^2}\end{aligned}$$

where we have used the fact that for the private shocks model $K_j V_j = 1$. Now, it is easy to see that the above expression is increasing in ζ_j and hence in data. The third term is given by

$$\begin{aligned}\mathbb{E}CS3 &= \frac{1}{2}n_F\zeta_j K_j\left(1-\frac{1}{\phi}\left(1+\frac{n_F\zeta_j}{\phi}\right)^{-1}\zeta_j\right) \\ &= \frac{1}{2}n_F K_j\left(\frac{\phi+(n_F-1)\zeta_j}{\frac{\phi}{\zeta_j}+n_F}\right)\end{aligned}$$

The numerator is clearly increasing in data as K_j and ζ_j are increasing in data. By similr reasoning, the denominator is decreasing in data. Therefore, $\mathbb{E}CS3$ is increasing in data.

The last step of the proof relaxes the assumption that marginal costs $\tilde{c}_j$ stay fixed. $\tilde{c}_j$ enters through the first term in both firm profits and consumer surplus that we combined as $ECS1 + E\Pi 1$ in (152). Note that firms will not produce anything on average if $\bar{\boldsymbol{p}} - \overline{\boldsymbol{c}} < 0$. So we can restrict attention to cases where that difference is non-negative. Thus, the squared difference is decreasing in $\tilde{c}_j$. Furthermore, all the coefficients multiplying the term are positive. Thus, this term in decreasing in the marginal cost of production $\tilde{c}_j$. Lemma 1 proves data - investment complementarity: $\partial\bar{c}/\partial n_d < 0$.

The final term in the firm's payoff function $g_j(\chi, \tilde{c}_j)$ which represents the fixed cost of investment is decreasing in marginal cost $\tilde{c}_j$. This term enters the payoff function with a negative sign.

Thus, more symmetric data for firms increases welfare, both by raising welfare for a given amount of investment, and by inducing firms to choose lower marginal costs. □

C. Auxiliary Results

C.1. Aggregate Shock Model: Auxiliary Lemmas

This section contains proofs of lemmas that are used in the proofs of propositions and additional welfare analysis. First, we derive the price in terms of parameters. The solution is still an implicit solution to a system of equations. Then, we derive conditions under which the H terms, which represent the sensitivity of quantity to a chance in expected price, increase/decrease in data. Finally, we consumer surplus welfare when firms are asymmetric.

Because the production problem is additively separable in attributes, without loss of generality, we focus on the single-attribute problem. The results the generalize to the the multi-attribute case because they hold for each one of the attributes.

Lemma 5. *In the equilibrium, each firm i's quantity is linear in their signal, that is to say,* $\boldsymbol{q}_i = k_i \mathbf{s}_i + p_i^M$. *And* k_i *and* p_i^M *are*

solutions of the following set of equations.

$$k_i = \frac{\frac{n_{di}}{n_{di}+1} - \frac{1}{\phi}(\sum_{j=1}^{n_F} k_i)}{\rho_i[(1 - \frac{\sum_{j\neq i}^{n_F} k_j}{\phi})^2 \frac{1}{n_{di}+1} + \frac{1}{\phi^2}\sum_{j\neq i}^{n_F} k_j^2 \frac{1}{n_{dj}+1}] + 1/\phi}$$

and

$$p_i^M = \frac{\bar{p} - \frac{1}{\phi}\sum_{j=1}^{n_F} p_i^M - \boldsymbol{c}_i}{\rho_i[(1 - \frac{\sum_{j\neq i}^{n_F} k_j}{\phi})^2 \frac{1}{n_{di}+1} + \frac{1}{\phi^2}\sum_{l\neq i}^{n_F} k_l^2 \frac{1}{n_{dl}+1}] + 1/\phi}.$$

Consider the denominator. The term $\frac{1}{\phi^2}\sum_{j\neq i}^{n_F} k_j^2 \frac{1}{n_{dj}+1}$ represents strategic uncertainty. This is the uncertainty about the price that comes from not knowing what competitors will do. The first term in brackets represents the way in which other firms' choices reduce uncertainty: $(1 - \frac{\sum_{j\neq i}^{n_F} k_j}{\phi})^2 \frac{1}{n_{di}+1}$. Other firms' responses to their data (if $k_l \geq 0 \ \forall l$) reduce the squared term, which is a component of price variance. The idea is that if every firm's quantity choice is positively correlated with their signal, when you receive a positive shock signal, you know the prices will increase, but you also know that other firms are likely to receive positive shock signals so that they will produce more. Others producing more reduces the price and offsets some of the price increase, which reduces price variance.

Proof. Taking the variance of the pricing rule $\boldsymbol{p} = \bar{p} - \frac{1}{\phi}\sum_{i=1}^{n_F} \boldsymbol{q}_i + \mathbf{b}$ yields

$$\begin{aligned}
\mathbf{Var}\left[\boldsymbol{p}|\mathcal{I}_i\right] &= \mathbf{Var}\left[\frac{1}{\phi}\sum_{j\neq i}^{n_F} \boldsymbol{q}_j - \mathbf{b}|\mathbf{s}_i\right] \\
&= \mathbf{Var}\left[\mathbf{E}\left[\frac{1}{\phi}\sum_{j\neq i}^{n_F} \boldsymbol{q}_j|\mathbf{b}\right] - \mathbf{b}|\mathbf{s}_i\right] + \mathbf{E}\left[\mathbf{Var}\left[\frac{1}{\phi}\sum_{j\neq i}^{n_F} \boldsymbol{q}_j|\mathbf{b}\right]|\mathbf{s}_i\right].
\end{aligned}$$

Recall the first order condition $\boldsymbol{q}_i = (\rho_i \mathbf{Var}\left[\boldsymbol{p}|\mathcal{I}_i\right] + 1/\phi)^{-1}(\mathbf{E}\left[\boldsymbol{p}|\mathcal{I}_i\right] - \boldsymbol{c}_i)$. Because our random variables are normal and functions are linear in those random variables, $\mathbf{E}\left[\boldsymbol{p}|\mathbf{s}_i\right]$ should be linear in $\mathbf{s}_i$, and $\mathbf{Var}\left[\boldsymbol{p}|\mathbf{s}_i\right]$ should not change with respect to the realization of $\mathbf{s}_i$. Thus, we can represent equilibrium production as $\boldsymbol{q}_i = k_i\mathbf{s}_i + p_i^M$.

Note conditional on $\mathbf{s}_i$, $\mathbf{b}$ is distributed as $N(\frac{n_{di}}{n_{di}+1}\mathbf{s}_i, \frac{1}{n_{di}+1})$. Thus, we have

$$\begin{aligned}
\mathbf{E}\left[\boldsymbol{p}|\mathbf{s}_i\right] &= \mathbf{E}\left[\bar{p} - \frac{1}{\phi}\sum_{i=1}^{n_F} \boldsymbol{q}_i + \mathbf{b}|\mathbf{s}_i\right] \\
&= \mathbf{E}\left[\bar{p} - \frac{1}{\phi}\sum_{i=1}^{n_F}(k_i\mathbf{s}_i + p_i^M) + \mathbf{b}|\mathbf{s}_i\right] \\
&= \bar{p} + \frac{n_{di}}{n_{di}+1}\mathbf{s}_i - \frac{1}{\phi}(\sum_{j=1}^{n_F} k_i)\mathbf{s}_i - \frac{1}{\phi}\sum_{j=1}^{n_F} p_i^M
\end{aligned}$$

$$\begin{aligned}
\mathbf{Var}\left[\mathbf{E}\left[\frac{1}{\phi}\sum_{j\neq i}^{n_F} \boldsymbol{q}_j|\mathbf{b}\right] - \mathbf{b}|\mathbf{s}_i\right] &= \mathbf{Var}\left[\mathbf{E}\left[\frac{1}{\phi}\sum_{j\neq i}^{n_F}(k_j\mathbf{s}_j + a_j)|\mathbf{b}\right] - \mathbf{b}|\mathbf{s}_i\right] \\
&= \mathbf{Var}\left[\frac{1}{\phi}\sum_{j\neq i}^{n_F}(k_j\mathbf{b} + a_j) - \mathbf{b}|\mathbf{s}_i\right] \\
&= (1 - \frac{\sum_{j\neq i}^{n_F} k_j}{\phi})^2 \frac{1}{n_{di}+1}
\end{aligned}$$

$$\begin{aligned}\mathbf{E}\left[\mathbf{Var}\left[\frac{1}{\phi}\sum_{j\neq i}^{n_F}\boldsymbol{q}_j|\mathbf{b}\right]|\mathbf{s}_i\right] &= \mathbf{E}\left[\mathbf{Var}\left[\frac{1}{\phi}\sum_{j\neq i}^{n_F}\boldsymbol{q}_j|\mathbf{b}\right]|\mathbf{s}_i\right]\\ &= \mathbf{E}\left[\frac{1}{\phi^2}\sum_{j\neq i}^{n_F}\mathbf{Var}\left[\boldsymbol{q}_j|\mathbf{b}\right]|\mathbf{s}_i\right]\\ &= \mathbf{E}\left[\frac{1}{\phi^2}\sum_{j\neq i}^{n_F}k_j^2\frac{1}{n_{dj}+1}|\mathbf{s}_i\right]\\ &= \frac{1}{\phi^2}\sum_{j\neq i}^{n_F}k_j^2\frac{1}{n_{dj}+1}\end{aligned}$$

Thus, we have for any i

$$k_i\mathbf{s}_i + p_i^M = \frac{\bar{p} + \frac{n_{di}}{n_{di}+1}\mathbf{s}_i - \frac{1}{\phi}(\sum_{j=1}^{n_F}k_i)\mathbf{s}_i - \frac{1}{\phi}\sum_{j=1}^{n_F}p_i^M - \boldsymbol{c}_i}{\rho_i[(1-\frac{\sum_{j\neq i}^{n_F}k_j}{\phi})^2\frac{1}{n_{di}+1} + \frac{1}{\phi^2}\sum_{j\neq i}^{n_F}k_j^2\frac{1}{n_{dj}+1}] + 1/\phi}.$$

If we match coefficients, we have

$$k_i = \frac{\frac{n_{di}}{n_{di}+1} - \frac{1}{\phi}(\sum_{j=1}^{n_F}k_i)}{\rho_i[(1-\frac{\sum_{j\neq i}^{n_F}k_j}{\phi})^2\frac{1}{n_{di}+1} + \frac{1}{\phi^2}\sum_{j\neq i}^{n_F}k_j^2\frac{1}{n_{dj}+1}] + 1/\phi}$$

and

$$p_i^M = \frac{\bar{p} - \frac{1}{\phi}\sum_{j=1}^{n_F}p_i^M - \boldsymbol{c}_i}{\rho_i[(1-\frac{\sum_{j\neq i}^{n_F}k_j}{\phi})^2\frac{1}{n_{di}+1} + \frac{1}{\phi^2}\sum_{j\neq i}^{n_F}k_j^2\frac{1}{n_{dj}+1}] + 1/\phi}.$$

□

Lemma 6. ***Data reduces price uncertainty:*** *Suppose ρ_1, ρ_2 are sufficiently small, and n_{d1} is sufficiently large, then $\mathbf{Var}[\boldsymbol{p}|\mathcal{I}_1]$ decreases as n_{d1} increases. Furthermore, if $\rho_1 \gg \rho_2$, then $\frac{\partial\mathbf{Var}[\boldsymbol{p}|\mathcal{I}_1]}{n_{d1}} \gg \frac{\partial\mathbf{Var}[\boldsymbol{p}|\mathcal{I}_2]}{\partial n_{d1}}$.*

Proof. Let

$$f_1(k_1,k_2,n_{d1}) = \rho_1[(1-\frac{1}{\phi}(n_F-1)k_2)^2\frac{1}{n_{d1}+1} + \frac{1}{\phi^2}(n_F-1)k_2^2\frac{1}{n_{d2}+1}] + 1/\phi,$$

then we have

$$\begin{aligned}\frac{\partial f_1}{\partial k_1} &= 0\\ \frac{\partial f_1}{\partial k_2} &= \rho_1[\frac{1}{n_{d1}+1}(\frac{2(n_F-1)^2k_2}{\phi^2} - \frac{2(n_F-1)}{\phi}) + \frac{1}{n_{d2}+1}\frac{1}{\phi^2}2(n_F-1)k_2]\\ \frac{\partial f_1}{\partial n_{d1}} &= -\rho_1(1-\frac{1}{\phi}(n_F-1)k_2)^2\frac{1}{(n_{d1}+1)^2}\end{aligned} \tag{154}$$

Let

$$f_2(k_1,k_2,n_{d1}) = \rho_2[(1-\frac{1}{\phi}(n_F-2)k_2 - \frac{1}{\phi}k_1)^2\frac{1}{n_{d2}+1} + \frac{1}{\phi^2}[(n_F-2)k_2^2\frac{1}{n_{d2}+1} + k_1^2\frac{1}{n_{d1}+1}]] + 1/\phi,$$

then we have

$$\begin{aligned}\frac{\partial f_2}{\partial k_1} &= \rho_2[\frac{1}{n_{d2}+1}\frac{2}{\phi}[\frac{1}{\phi}k_1 + \frac{1}{\phi}(n_F-1)k_2 - 1] + \frac{1}{n_{d1}+1}\frac{1}{\phi^2}2k_1]\\ \frac{\partial f_2}{\partial k_2} &= \rho_2[\frac{1}{n_{d2}+1}\frac{2(n_F-2)}{\phi}[\frac{1}{\phi}(n_F-2)k_2 - 1 + \frac{1}{\phi}k_1)] + \frac{1}{n_{d2}+1}[\frac{1}{\phi^2}(n_F-2)2k_2]]\\ \frac{\partial f_2}{\partial n_{d1}} &= -\rho_2\frac{1}{\phi^2}k_1^2\frac{1}{(n_{d1}+1)^2}\end{aligned}$$

Define

$$g_1(k_1,k_2,d_1)=\frac{\frac{n_{d1}}{n_{d1}+1}-\frac{1}{\phi}k_1-\frac{1}{\phi}(n_F-1)k_2}{\rho_1[(1-\frac{1}{\phi}(n_F-1)k_2)^2\frac{1}{n_{d1}+1}+\frac{1}{\phi^2}(n_F-1)k_2^2\frac{1}{n_{d2}+1}]+1/\phi}-k_1$$

$$g_2(k_1,k_2,d_1)=\frac{\frac{n_{d2}}{n_{d2}+1}-\frac{1}{\phi}k_1-\frac{1}{\phi}(n_F-1)k_2}{\rho_2[(1-\frac{1}{\phi}(n_F-2)k_2-\frac{1}{\phi}k_1)^2\frac{1}{n_{d2}+1}+\frac{1}{\phi^2}[(n_F-2)k_2^2\frac{1}{n_{d2}+1}+k_1^2\frac{1}{n_{d1}+1}]]+1/\phi}-k_2$$

Then

$$\frac{\partial g_1}{\partial k_1}=\frac{-\frac{1}{\phi}f_1-(\frac{n_{d1}}{n_{d1}+1}-\frac{1}{\phi}k_1-\frac{1}{\phi}(n_F-1)k_2)\frac{\partial f_1}{\partial k_1}}{f_1^2}-1=-\frac{\frac{1}{\phi}+k_1\frac{\partial f_1}{\partial k_1}}{f_1}-1$$

$$\frac{\partial g_1}{\partial k_2}=\frac{-\frac{1}{\phi}(n_F-1)f_1-(\frac{n_{d1}}{n_{d1}+1}-\frac{1}{\phi}k_1-\frac{1}{\phi}(n_F-1)k_2)\frac{\partial f_1}{\partial k_2}}{f_1^2}=-\frac{\frac{1}{\phi}(n_F-1)+k_1\frac{\partial f_1}{\partial k_2}}{f_1}$$

$$\frac{\partial g_2}{\partial k_1}=\frac{-\frac{1}{\phi}f_2-(\frac{n_{d2}}{n_{d2}+1}-\frac{1}{\phi}k_1-\frac{1}{\phi}(n_F-1)k_2)\frac{\partial f_2}{\partial k_1}}{f_2^2}=-\frac{\frac{1}{\phi}+k_2\frac{\partial f_2}{\partial k_1}}{f_2}$$

$$\frac{\partial g_2}{\partial k_2}=\frac{-\frac{1}{\phi}(n_F-1)f_2-(\frac{n_{d2}}{n_{d2}+1}-\frac{1}{\phi}k_1-\frac{1}{\phi}(n_F-1)k_2)\frac{\partial f_2}{\partial k_2}}{f_2^2}-1=-\frac{\frac{1}{\phi}(n_F-1)+k_2\frac{\partial f_2}{\partial k_2}}{f_2}-1$$

Thus,

$$-\begin{bmatrix}\frac{\partial g_1}{\partial k_1} & \frac{\partial g_1}{\partial k_2}\\ \frac{\partial g_2}{\partial k_1} & \frac{\partial g_2}{\partial k_2}\end{bmatrix}^{-1}=-[(\frac{\frac{1}{\phi}+k_1\frac{\partial f_1}{\partial k_1}}{f_1}+1)(\frac{\frac{1}{\phi}(n_F-1)+k_2\frac{\partial f_2}{\partial k_2}}{f_2}+1)-\frac{\frac{1}{\phi}(n_F-1)+k_1\frac{\partial f_1}{\partial k_2}}{f_1}\frac{\frac{1}{\phi}+k_2\frac{\partial f_2}{\partial k_1}}{f_2}]$$

$$\begin{bmatrix}-\frac{\frac{1}{\phi}(n_F-1)+k_2\frac{\partial f_2}{\partial k_2}}{f_2}-1 & \frac{\frac{1}{\phi}(n_F-1)+k_1\frac{\partial f_1}{\partial k_2}}{f_1}\\ \frac{\frac{1}{\phi}+k_2\frac{\partial f_2}{\partial k_1}}{f_2} & -\frac{\frac{1}{\phi}+k_1\frac{\partial f_1}{\partial k_1}}{f_1}-1\end{bmatrix}$$

$$\frac{\partial g_1}{\partial n_{d1}}=\frac{\frac{1}{(n_{di}+1)^2}+k_1\rho_1(1-\frac{1}{\phi}(n_F-1)k_2)^2\frac{1}{(n_{d1}+1)^2}}{f_1}$$

$$\frac{\partial g_2}{\partial n_{d1}}=\frac{k_2\rho_2\frac{1}{\phi^2}k_1^2\frac{1}{(n_{d1}+1)^2}}{f_2}$$

First when $\rho_1\approx 0$, $\rho_2\approx 0$, we have

$$\frac{\partial g_1}{\partial n_{d1}}\approx\frac{\frac{1}{(n_{d1}+1)^2}}{1/\phi}$$

$$\frac{\partial g_2}{\partial n_{d1}}\approx 0$$

$$-\begin{bmatrix}\frac{\partial g_1}{\partial k_1} & \frac{\partial g_1}{\partial k_2}\\ \frac{\partial g_2}{\partial k_1} & \frac{\partial g_2}{\partial k_2}\end{bmatrix}^{-1}\approx-[2n_F-(n_F-1)]\begin{bmatrix}-n_F & n_F-1\\ 1 & -2\end{bmatrix}$$

Thus

$$\frac{\partial k_1}{\partial n_{d1}}\approx\frac{[2n_F-(n_F-1)]n_F}{(n_{d1}+1)^2/\phi}$$

$$\frac{\partial k_2}{\partial n_{d1}}\approx-\frac{[2n_F-(n_F-1)]}{(n_{d1}+1)^2/\phi}$$

$$\begin{aligned}\frac{df_1}{dn_{d1}}&=\frac{\partial f_1}{\partial k_1}\frac{\partial k_1}{\partial n_{d1}}+\frac{\partial f_1}{\partial k_2}\frac{\partial k_2}{\partial n_{d1}}+\frac{\partial f_1}{\partial n_{d1}}\\&=\rho_1\frac{1}{(n_{d1}+1)^2}[-[\frac{1}{n_{d1}+1}(\frac{2(n_F-1)^2k_2}{\phi}-2(n_F-1))+\frac{1}{n_{d2}+1}\frac{1}{\phi}2(n_F-1)k_2][2n_F-(n_F-1)]\\&\quad-(1-\frac{1}{\phi}(n_F-1)k_2)^2]\end{aligned}$$

Note as n_{d1} increases, because we let $\rho_1\approx 0$, $\rho_2\approx 0$, k_1/ϕ and k_2/ϕ should converge to some constant. Thus, as long as

n_{d1} is sufficiently large, $\frac{df_1}{dn_{d1}}$ is negative, which implies $\mathbf{Var}[\boldsymbol{p}|\mathcal{I}_1]$ decreases as n_{d1} increases.

$$\begin{aligned}\frac{df_2}{dn_{d1}} &= \frac{\partial f_2}{\partial k_1}\frac{\partial k_1}{\partial n_{d1}} + \frac{\partial f_2}{\partial k_2}\frac{\partial k_2}{\partial n_{d1}} + \frac{\partial f_2}{\partial n_{d1}} \\ &= \rho_2[\frac{2}{n_{d2}+1}[\frac{1}{\phi}k_1 + \frac{1}{\phi}(n_F-1)k_2 - 1] + \frac{1}{n_{d1}+1}\frac{1}{\phi}2k_1]\frac{[2n_F-(n_F-1)]n_F}{(n_{d1}+1)^2} \\ &- \rho_2[\frac{1}{n_{d2}+1}2(n_F-2)[\frac{1}{\phi}(n_F-2)k_2 - 1 + \frac{1}{\phi}k_1)] + \frac{1}{n_{d2}+1}[\frac{1}{\phi}(n_F-2)2k_2]]\frac{[2n_F-(n_F-1)]}{(n_{d1}+1)^2} \\ &- \rho_2\frac{1}{\phi^2}k_1^2\frac{1}{(n_{d1}+1)^2}\end{aligned}$$

If ρ_2 is small, all terms go to zero, except for $\frac{1}{n_{d2}+1}[\frac{1}{\phi}(n_F-2)2k_2]]\frac{[2n_F-(n_F-1)]}{(n_{d1}+1)^2}$. in (154), a similar positive expression is multiplied by ρ_1 to get $\frac{df_1}{dn_{d1}}$. As long as ρ_1 is sufficiently high, then $\frac{df_1}{dn_{d1}} \gg \frac{df_2}{dn_{d1}}$. □

C.2. Asymmetric Welfare Results

The paper explored what happens when all firms have more data. But a key concern for market competition is the possibility that firms have highly unequal stocks of data. In this case, an increase in data can reduce welfare.

To demonstrate this possibility, we consider an example with two firms. We fix the total number of data points and add data to one firm as we subtract it from second firm. Figure 6 highlights how the economy is affected by data dispersion.

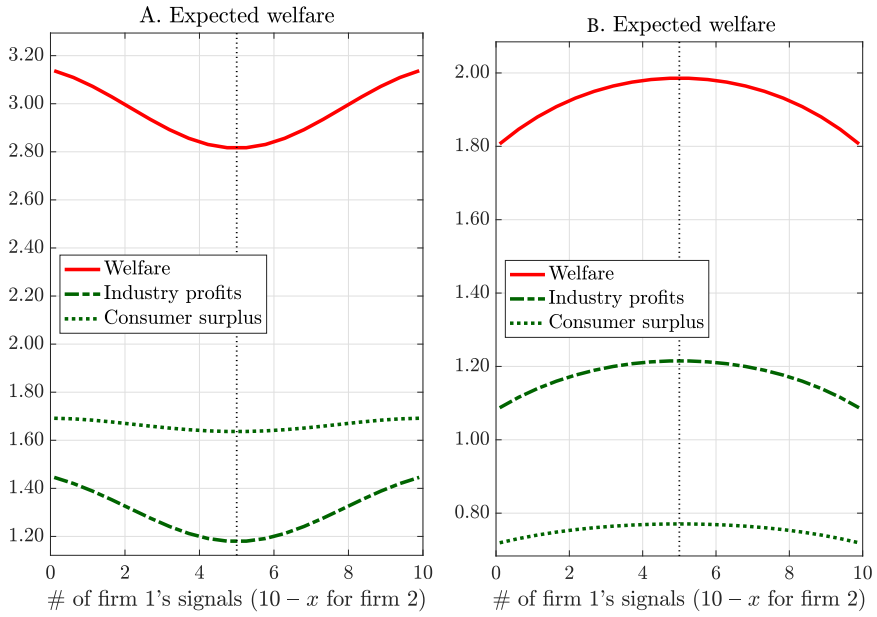


Figure 6: Data asymmetry and welfare with dominant risk channel (left) or investment channel (right).

Notes: This comparative static exercise is constructed over a single-good duopoly example. The investment cost function is assumed as $g(\chi_c, c_i) = \chi_c(\bar{c} - c_i)^2/2$. On the left, $\chi_c = 10$. On the right, $\chi_c = 1$. Other parameters are common to both plots: $\bar{c} = 3$, $\bar{p} = 5$, $\phi = 1$, $\sigma_b = 1$, $\mu_b = 0$, $\sigma_e = 2$, and $\rho_1 = \rho_2 = 1$. See Appendix B. for the computation of welfare.

Welfare with Asymmetric Firms Consider the model with firm-specific shocks and public data, where firms have different numbers of data points available about them. $n_F - 1$ firms in total, have n_{d2} data points, while one firm has asymmetric data, with n_{d1} data points.

In this asymmetric case, the supply sensitivity terms H_i and $\hat{H}_i$ take the same form as in (**??**) and (**??**). But they differ across firms. Define the firm-specific and the aggregate sensitivity of supply to changes in expected profit as:

$$\zeta_i = \frac{1 + n_{di}}{1 + n_{di} + \rho_i \phi} \qquad \zeta_a = \frac{1}{n_F} \left((n_F - 1)\zeta_1 + \zeta_2 \right) \tag{155}$$

Then, $\sum_i \hat{\boldsymbol{H}}_i = n_F \zeta_a \phi$. That allows us to express the expected market price as

$$\tilde{p}^M = \frac{1}{1 + n_F \zeta_a} \left(\overline{\boldsymbol{p}} + (n_F - 1)\zeta_1 \overline{\boldsymbol{c}}_1 + \zeta_2 \overline{\boldsymbol{c}}_2 \right) \tag{156}$$

Then using (41), we can expected express aggregate output as

$$\mathbf{E}[\sum_i \tilde{q}_i] = \frac{1}{1 + n_F \zeta_a} \phi \zeta_a \left(\overline{\boldsymbol{p}} - \overline{\boldsymbol{c}} \right) \tag{157}$$

Using the same substitutions, we can use (41) to express the variance of output and the covariance of output with the aggregate demand shocks $\sum_i b_i$ as

$$\mathbf{V}[\sum_i \tilde{q}_i] = \left(\frac{\phi}{1 + n_F \zeta_a} \right)^2 \sum_i \frac{n_{di}}{n_{di} + 1} \zeta_i^2. \tag{158}$$

$$Cov[\sum_i \tilde{q}_i, \sum_i b_i] = \left(\frac{\phi}{1 + n_F \zeta_a} \right) \sum_i \frac{n_{di} \zeta_i}{n_{di} + 1}. \tag{159}$$

Then, consumer surplus is

$$\mathbb{E}CS = \left(\frac{\phi}{1 + n_F \zeta_a} \right)^2 \left(\phi^2 \zeta_a^2 \left(\overline{\boldsymbol{p}} - \overline{\boldsymbol{c}} \right)' \left(\overline{\boldsymbol{p}} - \overline{\boldsymbol{c}} \right) + \sum_i \frac{n_{di}}{n_{di} + 1} \zeta_i^2 \right) + \left(\frac{2\phi}{1 + n_F \zeta_a} \right) \sum_i \frac{n_{di} \zeta_i}{n_{di} + 1} \tag{160}$$

D. Related Models: Product Innovation and Price Competition

D.1. Choosing A Location in Product Space

In the previous problem, we introduced the idea of product attributes so that a piece of data might be informative about the demand of multiple products. But we held the attributes of each product fixed. In reality, firms can choose the type of product to produce. They choose attributes. We show that the insights of the previous analysis carry over, with one small change. Data will allow a firm to choose a product that has higher-markup attributes. This makes product markups more like firm markups in the original model.

Each firm produces a single product, or bundle of products, with attributes chosen by the firm. Then the firm chooses how many units of the product or product bundle to produce. Formally, firm $i \in \{1, 2, \ldots, n_F\}$ chooses an $n \times 1$ vector a_i that describes their location in the product space, such that $\sum_j a_{ij} = 1$. As before, The jth entry of vector a_i describes how much of attribte i firm i's good contains.

The rest of the model assumptions, including consumer demand and the nature of data are the same as before. Thus, the firm's production problem is

$$max_{a_i, q_i} \mathbf{E}\left[q_i \boldsymbol{a}_i' \left(\tilde{\mathbf{p}} - \boldsymbol{c}_i \right) | \mathcal{I}_i \right] - \frac{\rho_i}{2} \mathbf{Var}\left[q_i \boldsymbol{a}_i' \left(\tilde{\mathbf{p}} - \boldsymbol{c}_i \right) | \mathcal{I}_i \right] - g(\chi_c, \boldsymbol{c}_i), \tag{161}$$

s.t. $\sum_j a_{ij} = 1$.

Just like the previous problem, prior to observing any of their data, each firm also chooses their cost vector $\boldsymbol{c}_i$. Since the data realizations are unknown in this ex-ante invetment stage, the objective is the unconditional expectation of the utility in 3

$$max_{c_i} \mathbf{E}\left[\mathbf{E}\left[q_i \boldsymbol{a}_i' \left(\tilde{\mathbf{p}} - \boldsymbol{c}_i \right) | \mathcal{I}_i \right] - \frac{\rho_i}{2} \mathbf{Var}\left[q_i \boldsymbol{a}_i' \left(\tilde{\mathbf{p}} - \boldsymbol{c}_i \right) | \mathcal{I}_i \right] \right] - g(\chi_c, \boldsymbol{c}_i). \tag{162}$$

SOLUTION Firm i's optimal production from the first order condition looks identical to the one before, except that now it is the the product of quantity and attributes that achieves this solution.

$$q_i \boldsymbol{a}_i = \left(\rho_i \mathbf{Var} \left[\boldsymbol{p}_i | \mathcal{I}_i \right] + \frac{\partial \mathbf{E} \left[\boldsymbol{p}_j | \mathcal{I}_j \right]}{\partial \boldsymbol{q}_i} \right)^{-1} \left(\mathbf{E} \left[\boldsymbol{p}_i | \mathcal{I}_i \right] - \boldsymbol{c}_i \right) \tag{163}$$

This tells us that the solution to the problem is exactly the same. In the previous problem, a firm choice produce any quantity of attributes it wanted with the right mix of products. In this problem, the firm can also choose any quantity of attributes it likes with the right quantity and product location.

The only thing that changes in this formulation of the problem is the interpretation of what constitutes a product. In the previous problem, a product had a fixed set of attributes. In this problem, a product is a fraction of the total output of the firm. Therefore the product markup here is more like what the firm markup was before. In other words, data affects the composition of a product now. Firms with data choose to produce products with higher-value attributes. This is a force that can make markups flat or increasing in data.

Proposition 6. *When firms choose attributes, product markups will increase in data, for a low enough risk aversion ρ_i.*

Proof. Comparing first-order condition (163) with original optimal choice (1), we could solve this extension model by substituting $\tilde{\boldsymbol{q}}_i$ in (1) with $q_i \boldsymbol{a}_i$ and further extend existing propositions for q_i and $\boldsymbol{a}_i$ by one-to-one mapping

$$q_i = \sum_{j=1}^{N} \tilde{\boldsymbol{q}}_{i,j} \quad \text{and} \quad \boldsymbol{a}_i = \frac{\tilde{\boldsymbol{q}}_i}{\sum_{j=1}^{N} \tilde{\boldsymbol{q}}_{i,j}} \tag{164}$$

Since firms optimize their choices in product space, the product markup is then the weighted average of attributes markups

$$M_i^p := \frac{\mathbb{E}[\boldsymbol{a}_i' \tilde{\boldsymbol{p}}_i]}{\mathbb{E}[\boldsymbol{a}_i' \tilde{\boldsymbol{c}}_i]} = \frac{\mathbb{E}[q_i \boldsymbol{a}_i' \tilde{\boldsymbol{p}}_i]}{\mathbb{E}[q_i \boldsymbol{a}_i' \tilde{\boldsymbol{c}}_i]} = \frac{\mathbb{E}[\tilde{\boldsymbol{q}}_i' \tilde{\boldsymbol{p}}_i]}{\mathbb{E}[\tilde{\boldsymbol{q}}_i' \tilde{\boldsymbol{c}}_i]} = M_i^f \tag{165}$$

This tells us that the product markups is equivalent to the firm-level markup of the original model. We already know that data boost firm-level markup with small risk aversion ρ_i (Proposition 2), thus the product markup will increase in data for a low enough risk aversion ρ_i.

This proof held marginal costs $\tilde{c}$ fixed, which corresponds to infinitely high marginal cost of adjusting c: $\chi_c \to \infty$. If we assume χ_c is sufficiently high, by continuity, the inequality will still hold.

□

This result shows why this extension is helpful for the model to match data showing flat or increasing product markups. The fact that markups had to be declining in the previous model was an artifact of the assumption that product characteristics are fixed. While that simplified the model and allowed us to focus on explaining the many other forces at play, the richer model paints a more realistic and data-consistent picture of how data, competition and markups interact.

D.2. Bertrand Competition

Many studies of markup competition use Bertrand, instead of Cournot competition. Therefore, we examine the three main effects in a Bertrand market. We find that the main results about covariance that generate the aggregation effects are similar. Cost choice is similar. But the risk effect can switch signs. Since the risk channel works against our main aggregation results (most of those results assume the price of risk sufficiently low), this strengthens the main effects.

In our model, all final goods are perfect substitutes. We know from the textbook imperfect competition models, that Bertrand with prefect substitutes results in a corner solution where the lowest cost firm captures the entire market at a price equal to the marginal cost of the second lowest cost firm, and the markup is equal to the ratio of the second lowest cost over the lowest cost. In case of a tie in the cost the marginal cost, profits are zero and there is indeterminacy in the exact allocation across firms. Showing that Bertrand competition results in lower prices than Cournot is therefore trivial.

But we know that when goods are imperfectly substitutable, Bertrand pricing is not at a corner, just as under Cournot. In this extension therefore we adjust our baseline model to include the imperfect substitutes. In the setup of our benchmark model all firms produce all goods (possibly with different weights on attributes) which are perfect substitutes. Market power therefore does not originate from the uniqueness of the goods a firm produces. Rather, market power stems from firms' heterogeneous demand shocks and production costs.

So we extend our setting, borrowing from Pellegrino (2020) where each firm produces one good only instead of all goods. First, denote the demand intercept as $\bar{\boldsymbol{p}}$ and the demand shock as $\boldsymbol{b}$. Second, we denote the similarity matrix by $\boldsymbol{\Psi}$. Therefore, we can immediately map $\psi_{ij} = 1/\phi_{ij}$ to our benchmark model where $\phi_{ij} = \phi$ which implies perfect subsitutes.[14]

We first derive a Bertrand solution and then the equivalent equilibrium conditions under Cournot. We then simulate the model to replicate the results.

Bertrand Competition. From the inverse demand function

$$\boldsymbol{p} = \bar{\boldsymbol{p}} - \boldsymbol{\Psi}\boldsymbol{q} + \boldsymbol{b} \tag{166}$$

$$\Rightarrow \boldsymbol{q} = \boldsymbol{\Psi}^{-1}(\bar{\boldsymbol{p}} - \boldsymbol{p} + \boldsymbol{b}) \tag{167}$$

Denote $\boldsymbol{\Gamma} = \boldsymbol{\Psi}^{-1}$, so the demand function for individual firm i can be written as

$$q_i = \sum_{j=1}^{n} \gamma_{ij}(\bar{p}_j - p_j + b_j) \tag{168}$$

The expectation and variance of the quantity is given by

$$\mathbf{E}\left[q_i \mid \mathcal{I}_i\right] = \sum_{j=1}^{n} \gamma_{ij}(\bar{p}_j - p_j + K_j s_j) \tag{169}$$

$$\mathbf{Var}\left[q_i \mid \mathcal{I}_i\right] = \sum_{j=1}^{n} \gamma_{ij}^2 \mathbf{Var}\left[b_j \mid \mathcal{I}_j\right] = \sum_{j=1}^{n} \gamma_{ij}^2 (1 + \Sigma_{\epsilon_j})^{-1} \Sigma_{\epsilon_j}, \tag{170}$$

and the risk-adjusted profit function is

$$U_I = \mathbf{E}\left[q_i \mid \mathcal{I}_i\right](p_i - c_i) - \frac{\rho_i}{2}\mathbf{Var}\left[q_i \mid \mathcal{I}_i\right](p_i - c_i)^2 - g\left(\chi_c, c_i\right) \tag{171}$$

Maximization with respect to the price yields the first-order condition,

$$\frac{\partial U_i}{\partial p_i} = \mathbf{E}\left[q_i \mid \mathcal{I}_i\right] + \frac{\partial \mathbf{E}\left[q_i \mid \mathcal{I}_i\right]}{\partial p_i}(p_i - c_i) - \rho_i \mathbf{Var}\left[q_i \mid \mathcal{I}_i\right](p_i - c_i) = 0 \tag{172}$$

$$p_i = c_i + \left(\rho_i \mathbf{Var}\left[q_i \mid \mathcal{I}_i\right] - \frac{\partial \mathbf{E}\left[q_i \mid \mathcal{I}_i\right]}{\partial p_i}\right)^{-1} \mathbf{E}\left[q_i \mid \mathcal{I}_i\right] \tag{173}$$

$$p_i = c_i + \left(\rho_i \sum_{j=1}^{n} \gamma_{ij}^2 \mathbf{Var}\left[b_j \mid \mathcal{I}_j\right] + \gamma_{ii}\right)^{-1} \sum_{j=1}^{n} \gamma_{ij}(\bar{p}_j - p_j + K_j s_j) \tag{174}$$

Denote $\hat{t}_i = \left(\rho_i \sum_{j=1}^{n} \gamma_{ij}^2 \mathbf{Var}\left[b_j \mid \mathcal{I}_j\right] + \gamma_{ii}\right)^{-1}$, then

$$p_i = c_i + \hat{t}_i \sum_{j=1}^{n} \gamma_{ij}(\bar{p}_j - p_j + K_j s_j), \tag{175}$$

or equivalently in matrix notation:

$$\boldsymbol{p} = \boldsymbol{c} + \hat{\boldsymbol{T}}\boldsymbol{\Gamma}(\bar{\boldsymbol{p}} - \boldsymbol{p} + \boldsymbol{K}\boldsymbol{s}), \tag{176}$$

where $\boldsymbol{p} = [p_1, p_2, ..., p_n]'$, $\bar{\boldsymbol{p}} = [\bar{p}_1, \bar{p}_2, ..., \bar{p}_n]'$, $\boldsymbol{K}\boldsymbol{s} = [K_1 s_1, K_2 s_2, ..., K_n s_n]'$, $\boldsymbol{c} = [c_1, c_2, ..., c_n]'$ and $\hat{\boldsymbol{T}}$ is a matrix where

[14] Pellegrino (2020) denotes the demand intercept by $\boldsymbol{b}$ and the demand similarity matrix by $\boldsymbol{I} + \boldsymbol{\Sigma}$ (note that here $\psi_{ii} = 1$ while $\sigma_{ii} = 0$ in Pellegrino's setup).

the diagonal elements are $\hat{t}_i$, and the other elements are 0, that is

$$\hat{\boldsymbol{T}} \equiv \begin{bmatrix} \hat{t}_1 & 0 & \cdots & 0 \\ 0 & \hat{t}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{t}_n \end{bmatrix}. \tag{177}$$

Then the equilibrium prices and quantities are

$$\boldsymbol{p} = (\boldsymbol{I} + \hat{\boldsymbol{T}}\boldsymbol{\Gamma})^{-1}\boldsymbol{c} + (\boldsymbol{I} + \hat{\boldsymbol{T}}\boldsymbol{\Gamma})^{-1}\hat{\boldsymbol{T}}\boldsymbol{\Gamma}(\bar{\boldsymbol{p}} + \boldsymbol{K}\boldsymbol{s}) \tag{178}$$

$$\boldsymbol{q} = \boldsymbol{\Gamma}(\bar{\boldsymbol{p}} + \boldsymbol{b}) - \boldsymbol{\Gamma}(\boldsymbol{I} + \hat{\boldsymbol{T}}\boldsymbol{\Gamma})^{-1}\boldsymbol{c} - \boldsymbol{\Gamma}(\boldsymbol{I} + \hat{\boldsymbol{T}}\boldsymbol{\Gamma})^{-1}\hat{\boldsymbol{T}}\boldsymbol{\Gamma}(\bar{\boldsymbol{p}} + \boldsymbol{K}\boldsymbol{s}) \tag{179}$$

Lemma 7. ***Data increases price-quantity covariance*** $\partial cov(p_{ij}, q_{ij})/\partial n_{di} > 0$.

Proof: Data n_{di} enters $\boldsymbol{p}$ and $\boldsymbol{q}$ only through $\hat{\boldsymbol{T}}$ and $\boldsymbol{K}$. The terms $\bar{p}$, I and Γ are exogenous.

Step 1: Show that the diagonal elements of $\hat{\boldsymbol{T}}$ and $\boldsymbol{K}$ are increasing in n_{di}. Since Bayesian updating is the same, regardless of market structure, the Bayesian weight on the signal K, is the same as before: $K = n_{di}/(1 + n_{di})I$, which is increasing in data n_{di}. $\hat{\boldsymbol{T}}$ is also a diagonal matrix, with diagonals $\hat{t}_i$ that are decreasing in $\mathbf{Var}\left[b_j \mid \mathcal{I}_j\right]$. Since market structure does not change Bayes' law, we know from Appendix A that data reduces conditional variance (prediction errors) $\partial \mathbf{Var}\left[b_j \mid \mathcal{I}_j\right]/\partial n_{di} < 0$. Thus, the diagonals of $\hat{\boldsymbol{T}}$ are increasing as well: $\partial \hat{t}_i/\partial n_{di} > 0$.

Step 2: Larger diagonal elements of $\hat{\boldsymbol{T}}$ and $\boldsymbol{K}$ raise $cov(p_{ij}, q_{ij})$. Covariance arises from stochastic terms. There are 3 stochastic terms: s in p, and s and b in q. The b term in q is multiplied by Γ, which is exogenous. Data has no effect on that term. The s terms in both p and q are multiplied by $\hat{\boldsymbol{T}}$ and $\boldsymbol{K}$ and other positive terms: $cov(p_i, q_i) = ((\boldsymbol{I} + \hat{\boldsymbol{T}}\boldsymbol{\Gamma})^{-1}\hat{\boldsymbol{T}}\boldsymbol{\Gamma})'\boldsymbol{\Gamma}(\boldsymbol{I} + \hat{\boldsymbol{T}}\boldsymbol{\Gamma})^{-1}\boldsymbol{c} - \boldsymbol{\Gamma}(\boldsymbol{I} + \hat{\boldsymbol{T}}\boldsymbol{\Gamma})^{-1}\hat{\boldsymbol{T}}\boldsymbol{\Gamma}$. Since $(\boldsymbol{I} + \hat{\boldsymbol{T}}\boldsymbol{\Gamma})^{-1}\hat{\boldsymbol{T}}$ is a positive definite matrix, whose eigenvalues are increasing in $\hat{t}_i$, larger $\hat{t}_i$ scales up each diagonal entry of $cov(p_i, q_i)$. Similarly, larger entries of the diagonal matrix K raise each diagonal entry of $cov(p_i, q_i)$. Thus, $cov(p_{ij}, q_{ij})$ are increasing in the diagonal elements of $\hat{\boldsymbol{T}}$ and $\boldsymbol{K}$. □

Markups and profits Next, we show that changing the model of competition can change the nature of the risk effect on markups. When costs are high, risk-averse firms may price low and produce more. However, since the aggregation effects only arise when the risk channel is not too strong, this reversal does not overturn the main aggregation results of the paper.

The risk-adjusted profit in the first stage is

$$\begin{aligned}
\mathbf{E}[U_i] =& \mathbf{E}\left[\mathbf{E}\left[q_i \mid \mathcal{I}_i\right](p_i - c_i) - \frac{\rho_i}{2}\mathbf{Var}\left[q_i \mid \mathcal{I}_i\right](p_i - c_i)^2\right] - g\left(\chi_c, c_i\right) && (180)\\
=& \mathbf{E}\left[\hat{t}_i^{-1}(p_i - c_i)^2 - \frac{\rho_i}{2}\mathbf{Var}\left[q_i \mid \mathcal{I}_i\right](p_i - c_i)^2\right] - g\left(\chi_c, c_i\right) && (181)\\
=& \mathbf{E}\left[(\gamma_{ii} + \frac{\rho_i}{2}\mathbf{Var}\left[q_i \mid \mathcal{I}_i\right])(p_i - c_i)^2\right] - g\left(\chi_c, c_i\right) && (182)\\
=& \mathbf{E}\left[\left(\gamma_{ii} + \frac{\rho_i}{2}\sum_{j=1}^{n}\gamma_{ij}^2\mathbf{Var}\left[b_j \mid \mathcal{I}_j\right]\right)(p_i - c_i)^2\right] - g\left(\chi_c, c_i\right) && (183)\\
=& \left(\gamma_{ii} + \frac{\rho_i}{2}\sum_{j=1}^{n}\gamma_{ij}^2\mathbf{Var}\left[b_j \mid \mathcal{I}_j\right]\right)\mathbf{E}\left[(p_i - c_i)^2\right] - g\left(\chi_c, c_i\right) && (184)\\
=& \left(\gamma_{ii} + \frac{\rho_i}{2}\sum_{j=1}^{n}\gamma_{ij}^2\mathbf{Var}\left[b_j \mid \mathcal{I}_j\right]\right)\left(\mathbf{E}\left[p_i - c_i\right]^2 + \mathbf{Var}\left[p_i - c_i\right]\right) - g\left(\chi_c, c_i\right) && (185)
\end{aligned}$$

where $\mathbf{Var}[p_i - c_i] = \mathbf{Var}[p_i]$ is independent of cost choices.

Then the optimal choices of marginal cost will be

$$\frac{\partial \mathbf{E}[U_i]}{\partial c_i} = \frac{\partial\left(\gamma_{ii} + \frac{\rho_i}{2}\sum_{j=1}^{n}\gamma_{ij}^2\mathbf{Var}\left[b_j \mid \mathcal{I}_j\right]\right)\mathbf{E}\left[p_i - c_i\right]^2}{\partial c_i} - \frac{\partial g\left(\chi_c, c_i\right)}{\partial c_i} = 0 \tag{186}$$

Firm i's markup is defined as $M_i = \mathbf{E}[p_i]/c_i$.

$$\boldsymbol{p} = (\boldsymbol{I} + \hat{\boldsymbol{T}}\boldsymbol{\Gamma})^{-1}\boldsymbol{c} + (\boldsymbol{I} + \hat{\boldsymbol{T}}\boldsymbol{\Gamma})^{-1}\hat{\boldsymbol{T}}\boldsymbol{\Gamma}(\bar{\boldsymbol{p}} + \boldsymbol{K}\boldsymbol{s}) \tag{187}$$

Firm i's markup is defined as $M_i = \mathbf{E}[p_i]/c_i$. Using the Cournot price (203) and $\mathbf{E}[b_i] = 0$,

$$\mathbf{E}[\boldsymbol{p}] = \bar{\boldsymbol{p}} - (\boldsymbol{I} + \boldsymbol{\Psi})(\boldsymbol{I} + \hat{\boldsymbol{H}}\boldsymbol{\Psi})^{-1}\hat{\boldsymbol{H}}(\bar{\boldsymbol{p}} + \boldsymbol{K}\boldsymbol{s} - \boldsymbol{c}) \equiv \bar{\boldsymbol{p}} - \boldsymbol{\Omega}\hat{\boldsymbol{H}}(\bar{\boldsymbol{p}} + \boldsymbol{K}\boldsymbol{s} - \boldsymbol{c}). \tag{188}$$

And we can derive the equivalent condition as Proposition (1) for Cournot:

$$\frac{\partial M_i}{\partial n_{di}} = \frac{\partial (p_i/c_i)}{\partial n_{di}} = \underbrace{\frac{c_i}{c_i^2}\frac{\partial p_i}{\partial n_{di}}}_{\text{Risk premium effect}} - \underbrace{\frac{p_i}{c_i^2}\frac{\partial c_i}{\partial n_{di}}}_{\text{Investment effect}} \tag{189}$$

where we take the price p_i in Bertrand compared to the expected price $\mathbf{E}[p_i]$ in Cournot. That is the only difference between the two expressions.

An Equivalent Cournot Model for Comparison The inverse demand function is

$$\boldsymbol{p} = \bar{\boldsymbol{p}} - \boldsymbol{\Psi}\boldsymbol{q} + \boldsymbol{b} \tag{190}$$

For firm i,

$$p_i = \bar{p}_i - \psi_{ii}q_i - \sum_{j \neq i}\psi_{ij}q_j + b_i \quad (\psi_{ii} = 1) \tag{191}$$

$$= \bar{p}_i - \frac{1}{\phi_{ii}}q_i - \sum_{j \neq i}\frac{1}{\phi_{ij}}q_j + b_i \tag{192}$$

The the expectation and variance of the price are

$$\mathbf{E}\left[p_i \mid \mathcal{I}_i\right] = \bar{p}_i - \psi_{ii}q_i - \sum_{j \neq i}\psi_{ij}q_j + K_i s_i \tag{193}$$

$$\mathbf{Var}\left[p_i \mid \mathcal{I}_i\right] = \mathbf{Var}\left[b_i \mid \mathcal{I}_i\right] = (1 + \Sigma_{\epsilon_i})^{-1}\Sigma_{\epsilon_i}, \tag{194}$$

and the risk-adjusted profit function

$$U_i = q_i(\mathbf{E}\left[p_i \mid \mathcal{I}_i\right] - c_i) - \frac{\rho_i}{2}\mathbf{Var}\left[p_i \mid \mathcal{I}_i\right]q_i^2 - g\left(\chi_c, c_i\right). \tag{195}$$

The first-order condition solves

$$\frac{\partial U_i}{\partial q_i} = \mathbf{E}\left[p_i \mid \mathcal{I}_i\right] - c_i + \frac{\partial \mathbf{E}\left[p_i \mid \mathcal{I}_i\right]}{\partial q_i}q_i - \rho_i\mathbf{Var}\left[b_i \mid \mathcal{I}_i\right]q_i = 0 \tag{196}$$

$$q_i = \left(\rho_i\mathbf{Var}\left[b_i \mid \mathcal{I}_i\right] - \frac{\partial \mathbf{E}\left[p_i \mid \mathcal{I}_i\right]}{\partial q_i}\right)^{-1}\left(\mathbf{E}\left[p_i \mid \mathcal{I}_i\right] - c_i\right) \tag{197}$$

$$q_i = \left(\rho_i\mathbf{Var}\left[b_i \mid \mathcal{I}_i\right] + \psi_{ii}\right)^{-1}\left[\bar{p}_i - \psi_{ii}q_i - \sum_{j \neq i}\psi_{ij}q_j + K_i s_i - c_i\right]. \tag{198}$$

If we denote $\hat{h}_i = \left(\rho_i\mathbf{Var}\left[b_i \mid \mathcal{I}_i\right] + \psi_{ii}\right)^{-1}$, then

$$q_i = \hat{h}_i\left(\bar{p}_i + K_i s_i - c_i - \psi_{ii}q_i - \sum_{j \neq i}\psi_{ij}q_j\right), \tag{199}$$

or equivalently

$$\boldsymbol{q} = \hat{\boldsymbol{H}}(\bar{\boldsymbol{p}} + \boldsymbol{K}\boldsymbol{s} - \boldsymbol{c} - \boldsymbol{\Psi}\boldsymbol{q}) \tag{200}$$

where $\boldsymbol{q} = [q_1, q_2, ..., q_n]'$, $\bar{\boldsymbol{p}} = [\bar{p}_1, \bar{p}_2, ..., \bar{p}_n]'$, $\boldsymbol{K}\boldsymbol{s} = [K_1 s_1, K_2 s_2, ..., K_n s_n]'$, $\boldsymbol{c} = [c_1, c_2, ..., c_n]'$ and $\hat{\boldsymbol{H}}$ is a matrix where

the diagonal elements are $\hat{h}_i$, and the other elements are 0, that is

$$\hat{\boldsymbol{H}} \equiv \begin{bmatrix} \hat{h}_1 & 0 & \cdots & 0 \\ 0 & \hat{h}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{h}_n \end{bmatrix} \tag{201}$$

Then the equilibrium price and quantity are given by

$$\boldsymbol{q} = (\boldsymbol{I} + \hat{\boldsymbol{H}}\boldsymbol{\Psi})^{-1}\hat{\boldsymbol{H}}(\bar{\boldsymbol{p}} + \boldsymbol{K}\boldsymbol{s} - \boldsymbol{c}) \tag{202}$$

$$\boldsymbol{p} = \bar{\boldsymbol{p}} - \boldsymbol{\Psi}(\boldsymbol{I} + \hat{\boldsymbol{H}}\boldsymbol{\Psi})^{-1}\hat{\boldsymbol{H}}(\bar{\boldsymbol{p}} + \boldsymbol{K}\boldsymbol{s} - \boldsymbol{c}) + \boldsymbol{b} \tag{203}$$

and the risk-adjusted profit in the first stage is

$$\mathbf{E}[U_i] = \mathbf{E}\left[q_i(\mathbf{E}\left[p_i \mid \mathcal{I}_i\right] - c_i) - \frac{\rho_i}{2}\mathbf{Var}\left[p_i \mid \mathcal{I}_i\right] q_i^2\right] - g\left(\chi_c, c_i\right) \tag{204}$$

$$= \mathbf{E}\left[\hat{h}_i^{-1} q_i^2 - \frac{\rho_i}{2}\mathbf{Var}\left[b_i \mid \mathcal{I}_i\right] q_i^2\right] - g\left(\chi_c, c_i\right) \tag{205}$$

$$= \mathbf{E}\left[\left(\psi_{ii} + \frac{\rho_i}{2}\mathbf{Var}\left[b_i \mid \mathcal{I}_i\right]\right) q_i^2\right] - g\left(\chi_c, c_i\right) \tag{206}$$

$$= \left(\psi_{ii} + \frac{\rho_i}{2}\mathbf{Var}\left[b_i \mid \mathcal{I}_i\right]\right)\mathbf{E}\left[q_i^2\right] - g\left(\chi_c, c_i\right) \tag{207}$$

$$= \left(\psi_{ii} + \frac{\rho_i}{2}\mathbf{Var}\left[b_i \mid \mathcal{I}_i\right]\right)\left(\mathbf{E}\left[q_i\right]^2 + \mathbf{Var}\left[q_i\right]\right) - g\left(\chi_c, c_i\right) \tag{208}$$

where $\mathbf{Var}[q_i] = h_i^{-1}\mathbf{Cov}(p_i, q_i)$ is independent of cost choices.

Then the optimal investment choice satisfies

$$\frac{\partial \mathbf{E}[U_i]}{\partial c_i} = \frac{\partial\left(\psi_{ii} + \frac{\rho_i}{2}\mathbf{Var}\left[b_i \mid \mathcal{I}_i\right]\right)\mathbf{E}\left[q_i\right]^2}{\partial c_i} - \frac{\partial g\left(\chi_c, c_i\right)}{\partial c_i} = 0. \tag{209}$$

Firm i's markup is defined as $M_i = \mathbf{E}[p_i]/c_i$. Using (203) and $\mathbf{E}[b_i] = 0$,

$$\mathbf{E}[\boldsymbol{p}] = \bar{\boldsymbol{p}} - \boldsymbol{\Psi}(\boldsymbol{I} + \hat{\boldsymbol{H}}\boldsymbol{\Psi})^{-1}\hat{\boldsymbol{H}}(\bar{\boldsymbol{p}} + \boldsymbol{K}\boldsymbol{s} - \boldsymbol{c}) \equiv \bar{\boldsymbol{p}} - \boldsymbol{\Omega}\hat{\boldsymbol{H}}(\bar{\boldsymbol{p}} + \boldsymbol{K}\boldsymbol{s} - \boldsymbol{c}) \tag{210}$$

Then for individual firm i,

$$\mathbf{E}[p_i] = \bar{p}_i - \sum_{j=1}^{n} \omega_{ij}\hat{h}_j(\bar{p}_j + K_j s_j - c_j) \tag{211}$$

$$M_i = \mathbf{E}[p_i]/c_i \tag{212}$$

And we can write the equivalent of the equation in Proposition 1 as

$$\frac{\partial M_i}{\partial n_{di}} = \frac{\partial(\mathbf{E}[p_i]/c_i)}{\partial n_{di}} = \underbrace{\frac{c_i}{c_i^2}\frac{\partial \mathbf{E}[p_i]}{\partial n_{di}}}_{\text{Risk premium effect}} - \underbrace{\frac{\mathbf{E}[p_i]}{c_i^2}\frac{\partial c_i}{\partial n_{di}}}_{\text{Investment effect}} \tag{213}$$

Bertrand Cost Choices The optimal choices of marginal cost will be

$$\frac{\partial \mathbf{E}[U_i]}{\partial c_i} = \frac{\partial\left(\gamma_{ii} + \frac{\rho_i}{2}\sum_{j=1}^{n}\gamma_{ij}^2\mathbf{Var}\left[b_j \mid \mathcal{I}_j\right]\right)\mathbf{E}\left[p_i - c_i\right]^2}{\partial c_i} - \frac{\partial g\left(\chi_c, c_i\right)}{\partial c_i} = 0 \tag{214}$$

Assume that $g\left(\chi_c, c_i\right) = \frac{\chi_c}{2}(c_i - \bar{c})^2$.

$$\left(\gamma_{ii} + \frac{\rho_i}{2}\sum_{j=1}^{n}\gamma_{ij}^2\mathbf{Var}\left[b_j \mid \mathcal{I}_j\right]\right)\frac{\partial \mathbf{E}\left[p_i - c_i\right]^2}{\partial c_i} = \chi_c(c_i - \bar{c}) \tag{215}$$

Since $\boldsymbol{p} = (\boldsymbol{I} + \hat{\boldsymbol{T}}\boldsymbol{\Gamma})^{-1}\boldsymbol{c} + (\boldsymbol{I} + \hat{\boldsymbol{T}}\boldsymbol{\Gamma})^{-1}\hat{\boldsymbol{T}}\boldsymbol{\Gamma}(\bar{\boldsymbol{p}} + \boldsymbol{K}\boldsymbol{s}) \equiv \boldsymbol{\xi}\boldsymbol{c} + \boldsymbol{\kappa}(\bar{\boldsymbol{p}} + \boldsymbol{K}\boldsymbol{s})$,

$$p_i = \xi_{ii}c_i + \xi_{ij}c_j + \kappa_{ii}(\bar{p}_i + K_i s_i) + \kappa_{ij}(\bar{p}_j + K_j s_j) \tag{216}$$

$$p_i - c_i = (\xi_{ii} - 1)c_i + \xi_{ij}c_j + \kappa_{ii}(\bar{p}_i + K_i s_i) + \kappa_{ij}(\bar{p}_j + K_j s_j) \tag{217}$$

$$(p_i - c_i)^2 = (\xi_{ii} - 1)^2 c_i^2 + 2[\xi_{ij}c_j + \kappa_{ii}(\bar{p}_i + K_i s_i) + \kappa_{ij}(\bar{p}_j + K_j s_j)](\xi_{ii} - 1)c_i + \text{terms not related to } c_i \tag{218}$$

$$\mathbf{E}(p_i - c_i)^2 = (\xi_{ii} - 1)^2 c_i^2 + 2(\xi_{ij}c_j + \kappa_{ii}\bar{p}_i + \kappa_{ij}\bar{p}_j)(\xi_{ii} - 1)c_i + \text{terms not related to } c_i \tag{219}$$

$$\frac{\partial \mathbf{E}\left[p_i - c_i\right]^2}{\partial c_i} = 2(\xi_{ii} - 1)^2 c_i + 2(\xi_{ij}c_j + \kappa_{ii}\bar{p}_i + \kappa_{ij}\bar{p}_j)(\xi_{ii} - 1) \tag{220}$$

Then the first order condition will be

$$\left(\gamma_{ii} + \frac{\rho_i}{2}\sum_{j=1}^{n}\gamma_{ij}^2 \mathbf{Var}\left[b_j \mid \mathcal{I}_j\right]\right)\left[2(\xi_{ii} - 1)^2 c_i + 2(\xi_{ij}c_j + \kappa_{ii}\bar{p}_i + \kappa_{ij}\bar{p}_j)(\xi_{ii} - 1)\right] = \chi_c(c_i - \bar{c}) \tag{221}$$

Cournot Cost Choices The optimal choices of marginal cost will be

$$\frac{\partial \mathbf{E}[U_i]}{\partial c_i} = \frac{\partial\left(\psi_{ii} + \frac{\rho_i}{2}\mathbf{Var}\left[b_i \mid \mathcal{I}_i\right]\right)\mathbf{E}\left[q_i\right]^2}{\partial c_i} - \frac{\partial g\left(\chi_c, c_i\right)}{\partial c_i} = 0 \tag{222}$$

Assume that $g\left(\chi_c, c_i\right) = \frac{\chi_c}{2}(c_i - \bar{c})^2$.

$$\left(\psi_{ii} + \frac{\rho_i}{2}\mathbf{Var}\left[b_i \mid \mathcal{I}_i\right]\right)\frac{\partial \mathbf{E}\left[q_i\right]^2}{\partial c_i} = \chi_c(c_i - \bar{c}) \tag{223}$$

Since $\boldsymbol{q} = (\boldsymbol{I} + \hat{\boldsymbol{H}}\boldsymbol{\Psi})^{-1}\hat{\boldsymbol{H}}(\bar{\boldsymbol{p}} + \boldsymbol{K}\boldsymbol{s} - \boldsymbol{c}) \equiv \boldsymbol{\zeta}(\bar{\boldsymbol{p}} + \boldsymbol{K}\boldsymbol{s} - \boldsymbol{c})$,

$$q_i = \zeta_{ij}(\bar{p}_j + K_j s_j - c_j) + \zeta_{ii}(\bar{p}_i + K_i s_i - c_i) \tag{224}$$

$$q_i^2 = \zeta_{ii}^2(\bar{p}_i + K_i s_i - c_i)^2 + \zeta_{ij}^2(\bar{p}_j + K_i s_j - c_j)^2 + 2\zeta_{ii}\zeta_{ij}(\bar{p}_i + K_i s_i - c_i)(\bar{p}_j + K_i s_j - c_j) \tag{225}$$

$$\text{since } s_i \sim N(b_i, \sigma_s^2), \textit{i.i.d.} \text{ and } \mathbf{E}(b_i) \text{ in the first period is 0.} \tag{226}$$

$$\mathbf{E}q_i^2 = \zeta_{ii}^2[(\bar{p}_i - c_i)^2 + K_i^2\sigma_s^2] + \zeta_{ij}^2[(\bar{p}_j - c_j)^2 + K_j^2\sigma_s^2] + 2\zeta_{ii}\zeta_{ij}(\bar{p}_i - c_i)(\bar{p}_j - c_j) \tag{227}$$

$$\frac{\partial \mathbf{E}q_i^2}{\partial c_i} = -2\zeta_{ii}^2(\bar{p}_i - c_i) - 2\zeta_{ii}\zeta_{ij}(\bar{p}_j - c_j) \tag{228}$$

Then the first order condition will be

$$-2(\psi_{ii} + \rho_i \mathbf{Var}\left[b_i \mid \mathcal{I}_i\right])(\zeta_{ii}^2(\bar{p}_i - c_i) + \zeta_{ii}\zeta_{ij}(\bar{p}_j - c_j)) = \chi_c(c_i - \bar{c}) \tag{229}$$

Simulations. Simulations show the combined effect of the cost and risk channels. Figure 7 shows that markups and prices under Bertrand are always below those under Cournot. Because investment determines the marginal cost endogenously, marginal costs under Bertrand are higher than under Cournot. Anticipating lower profits under Bertrand, firms invest less than under Cournot. Note that more information (more draws of data) leads to lower prices but higher markups as uncertainty is reduced. The finding that Cournot pries and markups are higher than under Bertrand are robust to changes in the cost of investment. However, as Figure 8 illustrates, for high investment costs ($\chi_c = 10$) expected markups and expected prices change in opposite directions under Cournot compared to Bertrand when the number of data points increases.

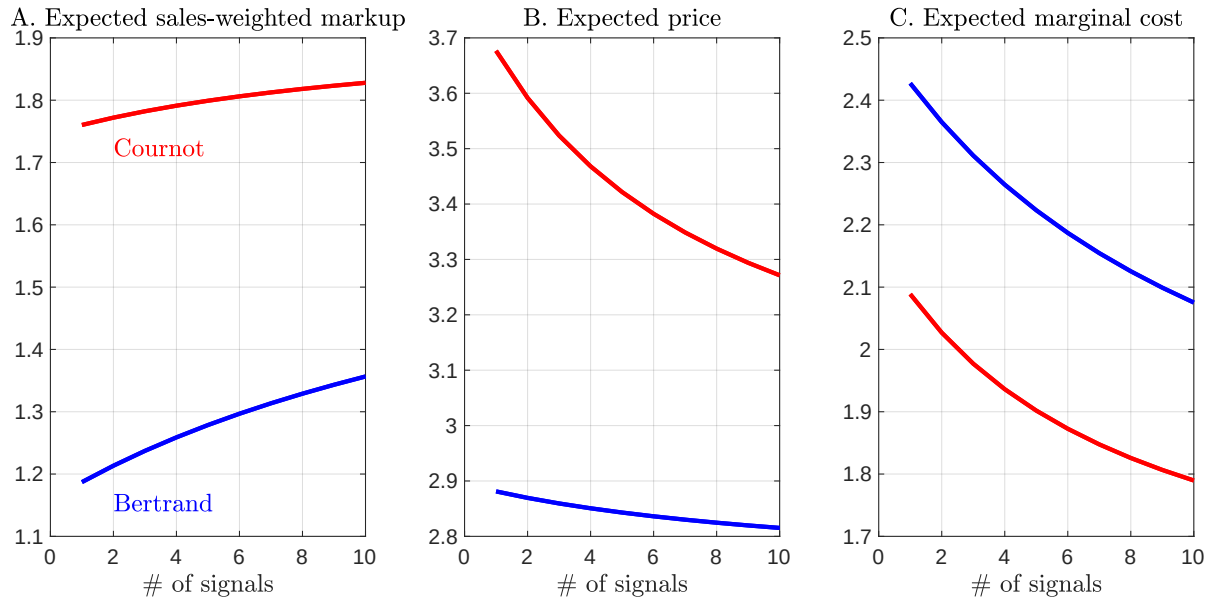


Figure 7: Cournot and Bertrand competition with investment channel dominating the risk channel.
Notes: This comparative static exercise is constructed over a duopoly example. The x-axis is the number of data points that both firms have. The investment cost function is assumed as $g(\chi_c, c_i) = \chi_c (\bar{c} - c_i)^2 / 2$ with $\chi_c = 1$ and $\bar{c} = 3$. Other parameters are: $\bar{p} = 5, \sigma_b = 1, \mu_b = 0, \sigma_e = 2, \rho_1 = \rho_2 = 1$ and $\Psi = [1, 0.5; 0.5, 1]$.

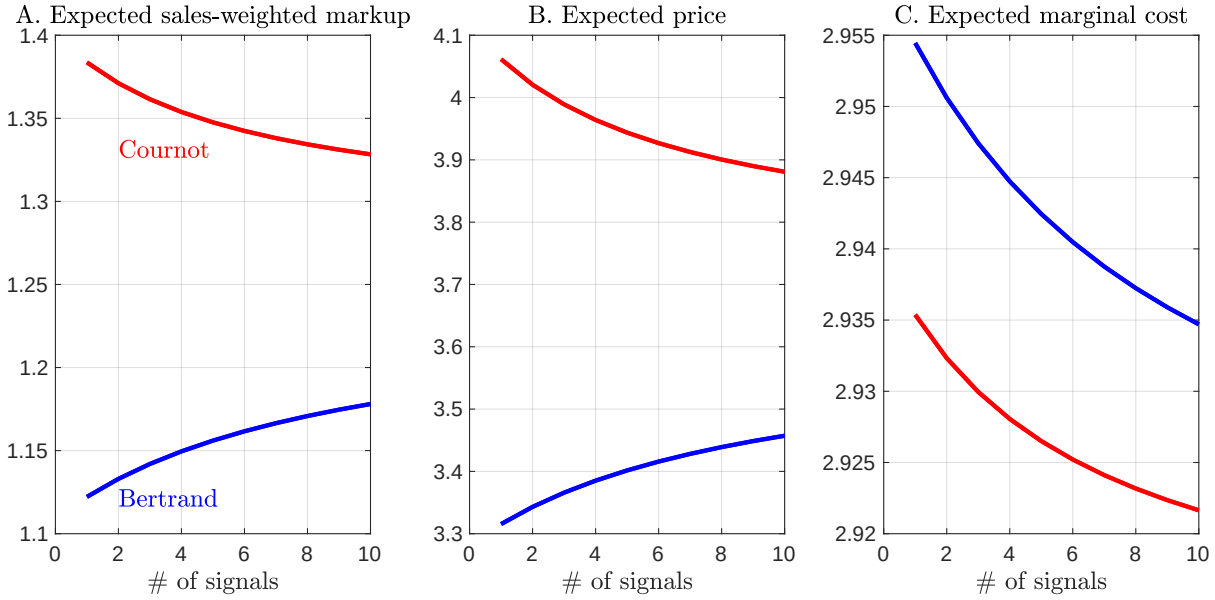


Figure 8: Cournot and Bertrand competition when investment channel matters less than risk.
Notes: This comparative static exercise is constructed over a duopoly example. The x-axis is the number of data points that both firms have. The investment cost function is assumed as $g(\chi_c, c_i) = \chi_c (\bar{c} - c_i)^2 / 2$ with $\chi_c = 10$ and $\bar{c} = 3$. Other parameters are: $\bar{p} = 5, \sigma_b = 1, \mu_b = 0, \sigma_e = 2, \rho_1 = \rho_2 = 1$ and $\Psi = [1, 0.5; 0.5, 1]$.