THE PINCHAS SAPIR CENTER
FOR DEVELOPMENT
Tel Aviv University

המרכז לפיתוח ע"ש פנחס ספיר
ליד אוניברסיטת תל אביב
"עמותה רשומה"
580010221

THE PINHAS SAPIR CENTER FOR DEVELOPMENT
TEL AVIV UNIVERSITY

Concentrated Risk: Misallocation and Granular Business Cycles

Tomer Ifergane

Discussion Paper No. 3-2026

Concentrated Risk: Misallocation and Granular Business Cycles*

Tomer Ifergane†

July 4, 2026

Abstract

When firm-level shocks affect aggregate outcomes, distortions reshaping the firm-size distribution affect not only output but also aggregate volatility. The sign depends on the correlation between distortions and productivity: compressing high-productivity firms reduces concentration and volatility, while compressing low-productivity firms raises both. In a U.S. calibration, positively correlated distortions reduce aggregate volatility resulting from firm-level shocks by 42 percent relative to a zero-distortion benchmark. Normatively, dispersion in marginal products need not imply inefficiency. Rather, in a granular economy, the risk-averse planner optimally diversifies production away from high-productivity firms, incurring an output loss in exchange for reduced aggregate risk.

JEL: E32, D24, O47
Keywords: Business cycles, Misallocation, Granularity, Stabilization policies, Size-dependent policies.

*I am especially grateful to Ben Moll for his extensive feedback and continued support during my work on this paper. I also thank Daniel Albuquerque, Marco Bellifemine, Nadav Ben Zeev, Vasco Carvalho, Adrien Couturier, Maarten De Ridder, Wouter den Haan, Xavier Gabaix, Lukas Freund, Jonas Gathen, Basile Grassi, Hugo Hopenhayn, David Lagziel, Ethan Ilzetzki, Guido Menzio, Dmitry Mukhin, Hugo Reichardt, Jane Olmstead Rumsey, Anthony Savagar, Aditya Soenarjo, Yannick Schindler, Roy (JJ) Shmueli, Ricardo Reis, and participants in several conferences and seminars for their helpful feedback, comments and suggestions. I am also grateful to Matan Levintov and Jonah Weiniger for outstanding research assistance. A previous version of this paper had been circulated under the title 'Efficiency and Macroeconomic Stability in Granular Economies'. This research was supported by THE ISRAEL SCIENCE FOUNDATION (grant No. 51/22 and No. 922/23) and benefited from financial support from the Pinhas Sapir Center for Development, and the Foerder Institute for Economic Research.

†Tel Aviv University, The Eitan Berglas School of Economics & London School of Economics and Political Science, Centre for Macroeconomics. iftomer@gmail.com

1 Introduction

Misallocation has emerged as a central explanation for cross-country differences in income and productivity and is therefore critical for understanding long-run outcomes. Much less is known about its implications for short-run fluctuations. In a granular economy, where a few large firms account for a significant share of economic activity, firm-level shocks can have a substantial impact on aggregate volatility. I show that distortions associated with misallocation change how firm-level shocks affect aggregate volatility, since they alter concentration and thereby the mapping from firm-level shocks to aggregate risk.

The sign of the volatility effect depends on the correlation between distortions and firm productivity. Distortions that concentrate production among the largest firms amplify volatility, whereas those that diversify production away from them dampen volatility. The empirical literature typically points to the latter case: distortions are positively correlated with productivity, compressing the largest and most productive firms. This mechanism reveals a trade-off between production efficiency and macroeconomic stability since those same positively correlated distortions that reduce output can also dampen volatility. For a risk-averse social planner, the optimal allocation in a granular economy may tilt production away from the most productive firms, generating dispersion in marginal products and sacrificing some expected output to reduce aggregate volatility. Ultimately, what appears as misallocation ex post may partly reflect ex ante stabilization against aggregate risk, dampening volatility in the economy.

I begin by proposing a theory of aggregate volatility arising endogenously from firms' optimal choice of size. The theory combines the canonical firm dynamics model à la Hopenhayn (1992) with misallocation in the form of implicit taxes or wedges distorting the firm's input choices as conceptualized in Restuccia and Rogerson (2008) and Hsieh and Klenow (2009). Throughout this paper's positive analysis, I use the term misallocation to refer to dispersion in marginal products, as is commonly done in the literature. I study a repeated static economy in which firms with heterogeneous productivity are exposed to idiosyncratic shocks and demand labor without observing these shocks. The model yields tractable aggregation results and provides a micro-foundation for the result in Gabaix (2011) in which aggregate fluctuations are governed by the severity of firm-level shocks and the degree of concentration.

Using the model, I characterize how distortions affect volatility from firm-level shocks, demonstrating that the correlation between distortions and firm productivity determines their effect on aggregate volatility. Distortions are positively (negatively) correlated with productivity when the resulting allocation is such that more productive firms exhibit a systematically higher (lower) marginal product of labor. Introducing positively correlated distortions to the economy leads to an equilibrium in which the relative size of the most productive firms is smaller, thereby diversifying production away from them, lowering concentration, and reducing aggregate volatility resulting from firm-level shocks. In contrast, negatively correlated distortions increase concentration and raise aggregate volatility.

Positively correlated distortions have played a central role in many papers in the misallocation literature, e.g., Restuccia and Rogerson (2008); Hsieh and Klenow (2014); Bartelsman et al. (2013); Buera and Fattal Jaef (2018); Bento and Restuccia (2017); Poschke (2018), and David and Venkateswaran (2019), suggesting that the stabilization effect is the empirically relevant benchmark. Note that the notion of stabilization relevant throughout this paper is ex-ante stabilization, whereby positively correlated distortions reduce unconditional aggregate volatility.

At the policy level, this result illuminates how size-dependent policies favoring small firms stabilize volatility arising from firm-level shocks. Such policies include size-dependent employment protection legislation, small and medium enterprise (SME) subsidies, product market regulations, and other institutions that disproportionately favor small firms or penalize large firms. The intuition is simple: these policies reduce concentration, making the firm-size distribution less skewed and thereby stabilizing activity.

Quantitatively, the stabilization effect of correlated distortions on aggregate volatility can be substantial. I calibrate the model to the U.S. economy by combining moments of the firm-size distribution and firm-level volatility from Carvalho and Grassi (2019) and the estimate of the elasticity of distortions with respect to productivity from Hsieh and Klenow (2014), who find that distortions are positively correlated with productivity. My baseline U.S. calibration implies that, consistent with the literature, firm-level shocks can account for an annual total factor productivity (TFP) volatility of 0.38% compared to an empirical TFP volatility of about 1%. Counterfactually removing the existing positively correlated distortions, while holding the underlying productivity distribution and other parameters fixed, increases TFP volatility from the baseline of 0.38% reported above to 0.65%. Equivalently, distortions reduce the component of TFP volatility generated by firm-level shocks by 42% compared to the zero-distortion benchmark.

This stabilization effect is by no means a free lunch: by compressing the firm-size distribution, these positively correlated distortions and size-dependent policies reduce both concentration and expected output. Thus, they are only valuable if aggregate risk is relevant for aggregate welfare, forcing the social planner to confront a trade-off between production efficiency and macroeconomic stability.

In the final part of the paper, I study the normative implications of this trade-off. In a granular economy, a risk-averse social planner has an incentive to create an allocation in which the marginal products of labor are not equalized between firms. In particular, in the granular economy, the allocation of labor does not merely affect aggregate output; it also pins down aggregate consumption risk. Therefore, the risk-averse social planner allocates labor so that high-productivity firms have systematically higher marginal products, thereby reducing their impact on consumption risk.

Furthermore, I show that the optimal allocation can be implemented by a schedule of firm-specific wedges, as used in the misallocation literature, such that the implicit tax imposed on the firm is increasing in its productivity. This positive correlation arises because it is precisely these high-productivity firms whose idiosyncratic outcomes are more strongly correlated with the marginal utility of consumption. This result implies that the empirically observed systematic dispersion in marginal products across firms can, at least in part, arise from welfare-maximizing insurance considerations, rather than being a symptom of inefficiency as it is commonly interpreted in the misallocation literature. Ultimately, this resulting optimal allocation demonstrates a trade-off between production efficiency and macro stability: the planner is willing to sacrifice some of the expected output level in exchange for reducing consumption risk.

If markets efficiently price firm-level risk, this optimal dispersion would appear in the data as positively correlated wedges even though it is not inefficient. If, however, inefficiencies prevent the allocation from fully reflecting firm-level risk, that is, if firms choose their privately optimal size without accounting for its effects on consumption risk, there is room for policy intervention.

Related literature. This paper is situated at the intersection of the granular business cycles literature and the misallocation literature. The literature on granular business cycles studies how the concentration of economic activity among several large firms leads to firm-level idiosyncratic shocks having aggregate consequences, e.g., Gabaix (2011); Carvalho and Gabaix (2013); di Giovanni et al. (2014, 2018); Carvalho and Grassi (2019); Gaubert and Itskhoki (2021). The most closely related papers are Gabaix (2011), which first proposed the granular hypothesis, and Carvalho and Grassi (2019), which demonstrated that this mechanism is quantitatively meaningful within the context of the canonical heterogeneous firms model. Building on their insights, this paper demonstrates how correlated distortions, which generate systematic dispersion in marginal products, affect concentration and thereby change aggregate volatility in a heterogeneous-firms model.

This paper also contributes to the misallocation literature following the seminal works of Restuccia and Rogerson (2008) and Hsieh and Klenow (2009). It is most closely related to works concerned with distortions changing the size of firms or establishments, inducing systematic dispersion in marginal products such as the work of Bartelsman et al. (2013); Hsieh and Klenow (2014); Bento and Restuccia (2017); Buera and Fattal Jaef (2018), and Poschke (2018).[1] I show that the same distortions studied as sources of output losses can also affect aggregate volatility. In the empirically relevant case of positively correlated distortions, they dampen volatility by reducing concentration, thereby generating a previously unexplored trade-off between production efficiency and macroeconomic stability.

Another body of related work connects misallocation to business cycles. However, most of these are concerned either with the cyclical properties of misallocation, as in, for example, Opp et al. (2014), or with how misallocation affects the transmission of a particular impulse, such as Baqaee and Farhi (2020). Along the first line, recent works demonstrate that misallocation is cyclically affected by granular forces, such as market power in Burstein et al. (2025) and investment irreversibility in Senga and Varotto (2024). In contrast, this work examines how non-cyclical drivers of misallocation affect the unconditional aggregate volatility arising from firm-level shocks in a granular environment. The relevance of such non-cyclical factors was found to be quantitatively important in accounting for misallocation as a whole in the work of David and Venkateswaran (2019). Closest in spirit is Eden (2017), which examines how misallocation affects the transmission of aggregate shocks to factor quantities and argues that misallocation can amplify them by affecting expected marginal returns. This paper studies a distinct concentration-based mechanism: distortions alter the transmission of firm-level shocks to the aggregate, implying that distortions can either amplify or dampen aggregate volatility depending on their correlation with productivity.

The normative analysis in the paper relates to works that find that markup dispersion or deviations from equalization of marginal products need not reflect inefficiency. Factors such as adjustment costs (Asker et al., 2014), consumer search frictions (Menzio, 2024), and non-linear pricing (Bornstein and Peter, 2024) can generate such a result. The most relevant work in this strand is David et al. (2022), which demonstrates that exposure to

[1] For a comprehensive review of this literature, see Hopenhayn (2014).

aggregate risk at the firm level generates systematic deviations in the marginal products of capital due to higher risk premiums. My mechanism runs in the reverse direction. Whereas David et al. (2022) study how exposure to aggregate risk generates firm-level risk premia and marginal-product dispersion, I show that, in a granular economy, idiosyncratic firm-level risk itself becomes aggregate consumption risk because large firms move aggregates. The risk-averse planner therefore equalizes expected risk-weighted marginal products rather than expected marginal products. Related works by Boar et al. (2022); Boar and Midrigan (2024); Di Tella et al. (2025) also explore the aggregate and normative implications of firm-level risk, creating systematic distortions but without discussing its business cycle implications or how firm-level risk determines aggregate risk.

This work also contributes to the literature on size-dependent policies, such as employment protection legislation and SME subsidies. Existing studies emphasize their effects on the firm-size distribution, TFP, and output (Guner et al., 2008; Garicano et al., 2016; Gourio and Roys, 2014; García-Santana and Pijoan-Mas, 2014). This paper identifies a new business-cycle implication of such policies: by changing concentration, they change aggregate volatility. Policies that compress large firms dampen granular volatility, whereas policies that favor large firms amplify it.

Finally, this paper is conceptually and methodologically related to Hulten's theorem (Hulten, 1978) and to follow-up work discussing the transmission of shocks through production networks Acemoglu et al. (2012); Grassi (2018); Baqaee and Farhi (2019) and Baqaee and Farhi (2020). While the present paper abstracts from the presence of production networks, the analysis conducted here can be extended by Hulten's original theorem to be the first-order effect in any arbitrary production structure.[2]

This paper is organized as follows. Section 2 presents the model and characterizes the role of distortions in modulating granular business cycles. Section 3 shows the stabilization role of size-dependent policies. Section 4 quantifies the effect of misallocation in the form of correlated distortions on aggregate volatility. Building on these insights, section 5 studies the optimal allocation in a granular economy with a risk-averse social planner. The final section concludes.

[2]My baseline is a model in which Hulten's theorem holds exactly. Doing so allows me to draw sharp predictions from a model involving a discrete number of firms without alluding to the law of large numbers, which would nullify the granular mechanism. However, in Appendix D.2 I employ the insights gleaned by Baqaee and Farhi (2019) to demonstrate how and to what extent potential deviations from Hulten's theorem affect the results.

2 Misallocation and Aggregate Volatility

This section presents the model and the paper's main positive result on the effect of distortions on aggregate volatility. I begin by presenting the main properties of the efficient production economy without distortions and illustrate how, in an economy with a finite number of firms, concentration affects aggregate volatility. Next, I show how distortions affect size choices and concentration in the economy. Last, I compare the aggregate volatility in an economy with and without distortions, demonstrating that the correlation between distortions and productivity determines the sign of the volatility effect.

2.1 Environment: Efficient Production Benchmark

Technology. There are N firms in the economy, each with a decreasing returns to scale production technology used to produce a single homogeneous final consumption numeraire good y_i. Each firm i produces its output by hiring labor[3] l_i using the production function

$$y_i = \tilde{z}_i l_i{}^{\gamma}, \quad \log(\tilde{z}_i) = \log(a_i) + \tilde{x}_i, \tag{1}$$

where $0 < \gamma < 1$ denotes the degree of decreasing returns in the economy, and $\tilde{z}_i$ is the productive ability of firm i. $\tilde{z}_i$ is composed of a firm's deterministic ability a_i and a stochastic idiosyncratic component $e^{\tilde{x}_i}$ with $\mathbb{E}\left[e^{\tilde{x}_i}\right] = 1$, such that its log is a random variable with mean $\mathbb{E}[\tilde{x}_i] = \bar{x}$ and variance σ_x^2, common to all N firms.[4] Realizations of $\tilde{x}_i$ are assumed to be i.i.d. such that for every pair of firm i and j we have $\tilde{x}_i \perp \tilde{x}_j$. The vector $\tilde{\mathbf{x}} \in \mathbb{R}^N$ denotes the aggregate state of the economy listing all realized values of $\tilde{x}_i$. Throughout this paper, I adopt the convention that tildes denote stochastic variables, boldface letters denote column vectors, and capital letters denote aggregates.

Under these assumptions, the model economy experiences no aggregate shocks in the conventional sense; instead, all aggregate fluctuations arise solely from firm-level shocks manifesting in the aggregate. The model is that of a repeated static economy, and thus volatility arises from differences in the cross-sectional distribution of firm-level shocks. Aggregate volatility is defined and discussed with respect to the stochastic aggregates.

The form $\log(\tilde{z}_i) = \log(a_i) + \tilde{x}_i$ is particularly convenient since it allows one to consider aggregate fluctuations around a stable ergodic firm productivity distribution, char-

[3] Labor represents a mobile factor; replacing labor with a bundle of other factors does not affect the result.

[4] Appendix D.1 demonstrates that the qualitative intuition is not sensitive to the common variance assumption, but, quantitatively, the magnitudes in section 4 are affected.

acterized by the time-invariant cross-sectional dispersion of a_i. Under most conventional firm-heterogeneity models, the distribution of a_i would, in equilibrium, generate the ergodic firm-size distribution and the measures of concentration in the economy. Thus, the shocks $\tilde{x}_i$ can be viewed as fluctuations around that ergodic firm-size distribution. Suppose one contemplates the business cycle as fluctuations around a balanced growth path. In that case, one can still interpret the distribution of a_i as the de-trended productivity distribution and the shocks $\tilde{x}_i$ as the deviations from trend growth in each firm. Therefore, the model is well-suited to study short-run fluctuations but not long-run dynamics or industry-specific trends.

Decision problem of the firm. The firm chooses how much labor to hire to maximize *expected* profits. The firm's problem is as follows

$$\max_{l_i} \quad \mathbb{E}\left[\tilde{z}_i {l_i}^{\gamma}\right] - w \times l_i. \tag{2}$$

Expectations are taken with respect to $\tilde{x}_i$ because the firm is assumed to hire labor without knowledge of its idiosyncratic state, making the production choice ex-ante efficient. Note that the assumptions on the process governing $\tilde{x}_i$ imply that labor is chosen based on the deterministic component a_i. Each firm takes the wage rate w as given. Labor is supplied inelastically in quantity L by the household, and the wage rate is such that the labor market clears according to

$$\sum_{i=1}^{N} l_i = L. \tag{3}$$

Thus stated, production is ex-ante efficient with the expected marginal product of labor equalized ex ante across all production units. The resulting equilibrium yields the maximum *expected* output in the economy. I will refer to this case as that of efficient production in the output-maximizing sense.

This problem is standard, and the resulting production economy exhibits the following aggregate properties.

Lemma 1. *Aggregate properties of the efficient production economy:*

1. *The production economy with N firms and a given labor supply L allocates labor according to the allocation rule*

$$l_j^{\star} = \frac{{a_j}^{\frac{1}{1-\gamma}}}{\sum_{i=1}^{N} {a_i}^{\frac{1}{1-\gamma}}} L. \tag{4}$$

2. *This labor allocation lends itself to an aggregate production function representation where aggregate output $Y_{\tilde{\mathbf{x}}}$ in aggregate state $\tilde{\mathbf{x}}$ is given by*

$$Y_{\tilde{\mathbf{x}}} = \sum_{i=1}^{N} y_i \; = \frac{\sum_{i=1}^{N} a_i^{\frac{1}{1-\gamma}} e^{\tilde{x}_i}}{\left[\sum_{i=1}^{N} a_i^{\frac{1}{1-\gamma}}\right]^{\gamma}} L^{\gamma} = \underbrace{Z_{\tilde{\mathbf{x}}}}_{TFP} \times L^{\gamma}, \quad Z_{\tilde{\mathbf{x}}} = \frac{\sum_{i=1}^{N} a_i^{\frac{1}{1-\gamma}} e^{\tilde{x}_i}}{\left[\sum_{i=1}^{N} a_i^{\frac{1}{1-\gamma}}\right]^{\gamma}}. \tag{5}$$

3. *Hulten's theorem: the effect of a one-percent shock to the productivity of the j^{th} firm $\tilde{z}_j$ on $Z_{\tilde{\mathbf{x}}}$, denoted by $\eta_j(\tilde{\mathbf{x}})$, is given by*

$$\eta_j(\tilde{\mathbf{x}}) = \frac{\partial \log Z_{\tilde{\mathbf{x}}}}{\partial \log \tilde{z}_j} = \frac{a_j^{\frac{1}{1-\gamma}} e^{\tilde{x}_j}}{\sum_{i=1}^{N} a_i^{\frac{1}{1-\gamma}} e^{\tilde{x}_i}}, \tag{6}$$

which incidentally is also the sales share of firm j or $y_j/Y_{\tilde{\mathbf{x}}}$.

4. *Aggregate volatility: The standard deviation of log TFP is given by*

$$\sigma_Z \quad \approx \quad \underbrace{\sigma_x}_{\text{micro volatility}} \quad \times \quad \underbrace{\Psi}_{\text{amplification term (HHI)}} \quad , \Psi = \sqrt{\sum_{i=1}^{N} \overline{\eta}_i^2}, \tag{7}$$

where $\overline{\eta}_i = \eta_i(\tilde{\mathbf{x}})|_{\tilde{\mathbf{x}}=\overline{\mathbf{x}}}$. Given that $\overline{\eta}_i$ is the sales share of i, then Ψ is also the square root of the Herfindahl–Hirschman index (HHI) of sales in the economy when all shocks are at their expected level.

Lemma 1 is a special case of the economy with misallocation discussed in Lemma 2, which is presented further in this section. The two proofs are presented jointly along with a step-by-step derivation in Appendix A.1.

The above Lemma establishes several properties of the stochastic heterogeneous firms model. It is worthwhile to emphasize that the model of this section and the following extensions are designed to follow as closely as possible the canonical firm dynamics model à la Hopenhayn (1992), which is commonly employed in the misallocation literature.[5]

Several features of the resulting economy are noteworthy. First, as is standard in the literature, the model generates a non-degenerate firm-size distribution since $l_i^\star > 0, \forall i$. Second, the allocation rule for labor is scale invariant. That is, scaling up or down the productivity of each firm by a constant factor $(1+g)$, where g is a common growth rate, leaves

[5]For further discussion see the review in Hopenhayn (2014).

the allocation of labor identical, and *relative* productive ability is all that matters. Third, *expected aggregate total factor productivity* (TFP)[6] is given by $\mathbb{E}[Z_{\tilde{\mathbf{x}}}] = \left[\sum_{i=1}^{N} a_i^{\frac{1}{1-\gamma}}\right]^{1-\gamma}$, which is not the same as the realized TFP, given by $Z_{\tilde{\mathbf{x}}}$. Finally, the resulting economy aggregates into a single-factor real business cycle (RBC) economy with an inelastic labor supply. Since all the positive results of this paper are related to TFP and not output, focusing on the inelastic case is not a limitation, as the labor allocation rule dictates relative factor quantities proportional to labor supply.

Aggregate volatility. Lemma 1 (3) verifies Hulten's theorem (Hulten, 1978), which states that the effect of a shock to a sector or a firm on aggregate output is the firm's Domar weight or the ratio of its sales to GDP. While conventionally stated in terms of the effect on output, given equation (5), the effects on TFP and output coincide throughout this paper. Unlike Hulten's original result, the above derivation involves no envelope condition. It is the exact solution to the aggregate representation of the model economy and thus sidesteps the critique of Baqaee and Farhi (2019). Furthermore, results obtained in this environment generalize via a Hulten's theorem logic to more complex settings, as these results are the first-order representation of TFP volatility in an economy with an arbitrary production network.

I interpret the shocks described throughout this paper as the usual churn of economic activity "garden-variety fluctuations", if you will, and emphatically not as those induced by catastrophic events like the 2007-2008 financial crisis or the COVID-19 recession, which are arguably "true" aggregate shocks.

Observe that the derivation of the elasticities $\eta_j(\tilde{\mathbf{x}})$ holds regardless of the distributional assumptions on a_i and $\tilde{x}_i$. However, these assumptions affect the resulting amplification term Ψ, given in equation (7), which is a concentration indicator in the resulting economy, namely the square root of the HHI when $\tilde{\mathbf{x}} = \overline{\mathbf{x}}$.[7] Ψ governs aggregate volatility in the economy and maps the realizations of firm-level shocks into a single measure of aggregate volatility, thus providing a possible micro-foundation for the "granular hypothesis". The distribution of $\overline{\eta}_i$, induced by the distribution of deterministic ability a_i, determines aggregate volatility in the economy by pinning down Ψ. This key insight originates from

[6]Lemma 1 implies that expected TFP is increasing in the number of firms, reflecting scale effects in the economy. I refer to this expression as TFP rather than using $Z_{\tilde{\mathbf{x}}}/N$ as is sometimes done in the literature. My reasons are twofold: first, $Z_{\tilde{\mathbf{x}}}$ is the correct theoretical counterpart in my environment to the classical Solow residual, and second I do not discuss entry and exit so the two are proportional, and no insight would be gained from this distinction.

[7]Equation (7) is similar in flavor to equations (4) and (5) in Gabaix (2011).

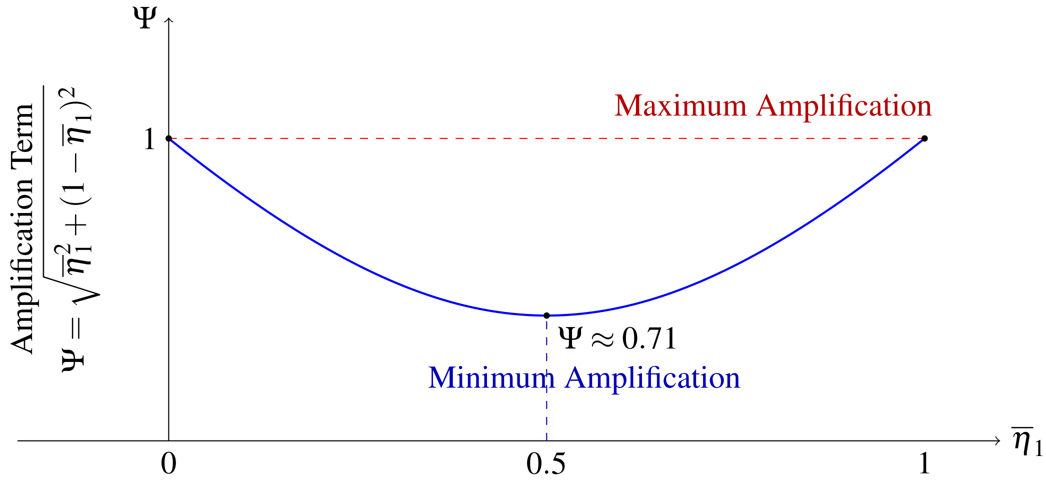

Figure 1: The Amplification Term Ψ as a Function of $\overline{\eta}$ in a Two-Firm Economy

the seminal work of Gabaix (2011), which illustrates that if sales shares are distributed according to a fat-tailed distribution, firm-level shocks of a reasonable magnitude can generate aggregate volatility of a quantitatively plausible scale, even in the presence of a large number of firms.

Note that all of these elasticities sum up to unity, $\sum_{i=1}^{N} \eta_i(\tilde{\mathbf{x}}) = 1$, implying that reducing the productive ability of all firms by 1% is equivalent to a 1% aggregate TFP shock. It is the concentration of these elasticities, however, which governs the excess sensitivity of the resulting economy to shocks to one particular firm. To demonstrate this idea, let us briefly consider a two-firm economy.

2.2 Intuition for the Relationship Between Misallocation and Aggregate Volatility: A Two-Firm Example

Suppose there were only two firms in the economy ($N = 2$). Aggregate volatility in this economy is given by $\Psi = \sqrt{\overline{\eta}_1{}^2 + (1-\overline{\eta}_1)^2}$. Figure 1 reports the value of the amplification term for different levels of the market share of the first firm $\overline{\eta}_1$. Intuitively, aggregate volatility is lowest when both firms are the same size and highest when one firm controls the entire market. This intuition generalizes to the N-firm case. The more concentrated the firm-size distribution is, the more volatile the economy will be, all else being equal.

Observation. The firms' choice of relative size, i.e., the share of labor they can utilize in equilibrium, generates aggregate volatility through the resulting HHI, as demonstrated in Lemma 1. In so doing, *the firm-size distribution maps firm-level shocks to a macro-level shock. Any friction or element of the economic environment that distorts the firms' choice of size also changes the volatility of TFP*. This effect operates in addition to the friction's effect on the level of TFP, which is the conventional object of interest in the misallocation literature. This observation motivates the rest of the analysis.

To build intuition going forward, one can ask: how would aggregate volatility in the economy with inefficient production differ from that in the efficient benchmark? Through the lens of Figure 1, the question amounts to asking how production inefficiency affects the market share of the larger firm. Increasing the share of the larger firm increases volatility. Conversely, if the largest firm's market share decreases due to inefficiencies, volatility would decrease. In what follows, I formalize this intuition.

2.3 Misallocation and Aggregate Volatility

I now introduce distortions into the benchmark model and study its implications for aggregate volatility in the resulting economy. I demonstrate that the correlation between distortions and productive ability determines the direction of the effect: positively correlated distortions dampen aggregate volatility, while negatively correlated distortions amplify it.

Compared to the efficient benchmark, the production technology remains unchanged; there are still N firms producing with the same decreasing returns-to-scale technology. However, the firm now faces an implicit output tax $\tau_i \in [0,1]$, which introduces a wedge between marginal revenue and marginal cost. The firm solves the following decision problem:

$$\max_{l_i} \ \mathbb{E}[y_i(1-\tau_i)] - wl_i. \tag{8}$$

Production is still carried out according to equation (1) such that $y_i = \tilde{z}_i l_i^{\gamma}$ and the wedge $(1-\tau_i)$ only affects input choices. Thus, the wedge alters the firm's input choice and, consequently, its chosen size, compared to the efficient production case, leading to misallocation as in Restuccia and Rogerson (2008), Hsieh and Klenow (2009), and subsequent work. This modeling tool has proven particularly useful as it lends itself to straightforward empirical interpretation in which observing a higher average revenue product of labor for a firm, sector, or establishment implies a higher τ_i. The aggregate properties of the distorted economy can be summarized as follows.

Lemma 2. *Aggregate properties of the distorted production economy*

1. *The production economy with N firms and a given labor supply L allocates labor according to the allocation rule*

$$l_j^{d\star} = L\frac{\left(a_j\left(1-\tau_j\right)\right)^{\frac{1}{1-\gamma}}}{\sum_{i=1}^{N}\left(a_i(1-\tau_i)\right)^{\frac{1}{1-\gamma}}}, \tag{9}$$

2. *This labor allocation lends itself to an aggregate production function representation where aggregate output $Y_{\tilde{\mathbf{x}}}^d$ is produced according to*

$$Y_{\tilde{\mathbf{x}}}^d = \sum_{i=1}^{N} y_i = Z_{\tilde{\mathbf{x}}}^d \times L^{\gamma}, \quad Z_{\tilde{\mathbf{x}}}^d = \frac{\sum_{i=1}^{N} a_i^{\frac{1}{1-\gamma}}(1-\tau_i)^{\frac{\gamma}{1-\gamma}} e^{\tilde{x}_i}}{\left[\sum_{i=1}^{N}\left(a_i(1-\tau_i)\right)^{\frac{1}{1-\gamma}}\right]^{\gamma}}. \tag{10}$$

3. *Hulten's theorem: the effect of a one-percent shock to the productivity of the j^{th} firm $\tilde{z}_j$ on $Z_{\tilde{\mathbf{x}}}^d$, denoted by $\delta_j(\tilde{\mathbf{x}})$, is given by*

$$\delta_j(\tilde{\mathbf{x}}) = \frac{\partial \log\left({Z_{\tilde{\mathbf{x}}}}^d\right)}{\partial \log \tilde{z}_j} = \frac{{a_j}^{\frac{1}{1-\gamma}}\left(1-\tau_j\right)^{\frac{\gamma}{1-\gamma}} e^{\tilde{x}_j}}{\sum_{i=1}^{N} a_i^{\frac{1}{1-\gamma}}(1-\tau_i)^{\frac{\gamma}{1-\gamma}} e^{\tilde{x}_i}}, \tag{11}$$

   which incidentally is also the sales share of firm j or $y_j^d/Y_{\tilde{\mathbf{x}}}^d$.

4. *Aggregate volatility: The standard deviation of log TFP is given by*

$$\sigma_Z^d \approx \sigma_x \times \Psi^d, \quad \Psi^d = \sqrt{\sum_{i=1}^{N} \overline{\delta}_i^2}, \tag{12}$$

   where $\overline{\delta}_i = \delta_i(\tilde{\mathbf{x}})|_{\tilde{\mathbf{x}}=\overline{\mathbf{x}}}$ and Ψ^d is the square root of the sales HHI in the distorted economy.

For proof, see Appendix A.1. To distinguish the efficient production and distorted production economies, I denote all sizes pertaining to the distorted economy with superscript d and let the distorted counterpart of the elasticity $\eta_i(\tilde{\mathbf{x}})$ be denoted by $\delta_i(\tilde{\mathbf{x}})$. Note that Lemma 1 is a special case of Lemma 2 in which $\tau_i = \tau_j$ for every i and j. The aggregate production function is such that wedges affect TFP by inducing a different allocation of labor. In so doing, they affect the firm-size distribution and market concentration, inducing a different sales HHI, and thus result in different levels of aggregate volatility σ_Z^d even while holding constant the productive ability of the N firms. This environment facilitates analytical characterization since factors are allocated ex ante, thus creating an environment with misallocation in which Hulten's theorem still holds, as is demonstrated by equation (11).[8]

2.4 Volatility in a Distorted Economy vs. the Efficient Benchmark

To compare aggregate volatility in the efficient production economy with that in the distorted one really amounts to understanding how the distortions affect concentration, as measured by Ψ or the HHI, in both economies. Changing the HHI implies that the wedges distort the firms' size choices unevenly. If misallocation involves drawing more factors into less productive firms, i.e., firms with low $\overline{\eta}_j$, and away from high-productivity firms, with high $\overline{\eta}_j$, then misallocation dampens aggregate volatility. The converse is true. When misallocation draws factors into more productive firms, thus making them inefficiently large, it also exacerbates aggregate volatility.

To formalize this intuition, I define a *relative distortions vector* $\mathbf{d} \in \mathbb{R}^N$. A relative distortion d_j faced by firm j maps the elasticity $\overline{\eta}_j$, which is also its sales share in the efficient production case, into its counterpart in the distorted case $\overline{\delta}_j$, as follows

$$\overline{\delta}_j = \frac{a_j^{\frac{1}{1-\gamma}}\left(1-\tau_j\right)^{\frac{\gamma}{1-\gamma}}}{\sum_{i=1}^N a_i^{\frac{1}{1-\gamma}}\left(1-\tau_i\right)^{\frac{\gamma}{1-\gamma}}} = \frac{a_j^{\frac{1}{1-\gamma}}}{\sum_{i=1}^N a_i^{\frac{1}{1-\gamma}}} \frac{\left(1-\tau_j\right)^{\frac{\gamma}{1-\gamma}}}{\left(1-\hat{\tau}\right)^{\frac{\gamma}{1-\gamma}}} \equiv \overline{\eta}_j \sqrt{1-d_j}, \tag{13}$$

where the value of $\hat{\tau}$ is defined from the following weighted average wedge $(1-\hat{\tau})^{\frac{\gamma}{1-\gamma}} = \sum_{i=1}^N \overline{\eta}_i (1-\tau_i)^{\frac{\gamma}{1-\gamma}}$. Given the assumptions on a_j and τ_j, which ensure $\overline{\eta}_j$ and $\overline{\delta}_j$ are strictly positive, we can conclude that $\sqrt{1-d_j} > 0$.

In the misallocation literature broadly, the term distortion is usually applied to the wedge $1-\tau_j$. Distortions are said to be positively correlated if misallocation disproportionately harms the high-ability firms, i.e., a firm with a higher value of a_j has a higher τ_j, and as a result, a high measured marginal revenue product of labor. Negatively correlated distortions imply the converse. My newly-defined relative distortion d_j is analogous but expressed in terms of the elasticities or relative firm size instead of absolute firm size in the economy where $\tilde{\mathbf{x}} = \overline{\mathbf{x}}$.

The equivalence between the two notions is demonstrated by isolating d_j from equation

[8] Deviations from this assumption are discussed in appendix D.2.

(13) as follows

$$d_j = 1 - \left(\frac{1-\tau_j}{1-\hat{\tau}}\right)^{\frac{2\gamma}{1-\gamma}}. \tag{14}$$

Through equation (14), $d_j = 0$ if $\tau_j = \hat{\tau}$, leading the firm's relative size to be the same whether production decisions are distorted or efficient. When all τ_j are identical, we collapse to the efficient production case with $d_j = 0, \forall j$. Dispersion in the wedges implies dispersion in the relative distortions. When a firm has $\tau_j > \hat{\tau}$, we have that $d_j > 0$, implying that the firm faces a higher than average implicit tax and in the distorted case it is smaller in relative terms than it would have been in the efficient production case. $d_j < 0$ implies the opposite. Observing a positive correlation between relative size and relative distortion requires that a high d_j firm also has a high $\overline{\eta}_j$. Note that $\overline{\eta}_j$ is a strictly increasing function of firm j's productive ability a_j, and that d_j is strictly increasing in τ_j (as long as there are i and j such that $\tau_i \neq \tau_j$). Therefore, correlated relative distortions in my setting are conceptually equivalent, up to the exact correlation measure used, to the correlated distortions employed by the misallocation literature, which relates to the correlation between τ_j and a_j.[9]

We can now compare the aggregate volatility in the distorted and efficient production economies by substituting equation (13) into the definition of Ψ^d in equation (12)

$$\Psi^d = \sqrt{\sum_{i=1}^{N} \overline{\delta}_i^{\,2}} = \sqrt{\sum_{i=1}^{N} \overline{\eta}_i^{\,2}(1-d_i)}\,. \tag{15}$$

By squaring both sides of the above equation and using the definition of Ψ in equation (7), we obtain the following:

$$\underbrace{\Psi^2 - {\Psi^d}^2}_{\Delta\text{HHI}} = \sum_{i=1}^{N} \overline{\eta}_i^{\,2} d_i \ = \|\mathbf{d}\| \times \|\overline{\boldsymbol{\eta}}^{\mathbf{2}}\| \times \cos(\theta)\,. \tag{16}$$

Where $\|\cdot\|$ denotes the Euclidean norm and vector powers denote elementwise operations. Since $\sum_{i=1}^{N} \overline{\eta}_i^{\,2} d_i$ is the dot product of two vectors, we can express the difference in terms of the cosine similarity where θ is the angle between the two vectors as follows $\cos(\theta) = \frac{\sum_{i=1}^{N} \overline{\eta}_i^{\,2} d_i}{\|\mathbf{d}\| \times \|\overline{\boldsymbol{\eta}}^{\mathbf{2}}\|}$.[10] Cosine similarity measures how aligned two vectors are in direction, irrespective of their magnitude. When both vectors are mean-centered, cosine similarity co-

[9] In section 4.2, I also demonstrate how the two notions coincide using a parametric form used widely in the misallocation literature.

[10] Generally, for any two vectors $\mathbf{v}, \mathbf{u} \in \mathbb{R}^N$ we can state their dot product in terms of their cosine similarity,

incides with the Pearson correlation coefficient. In equation (16), it serves as a measure of how relative distortions correlate with (squared) relative sizes. Thus, we can conclude that distortions affect aggregate volatility as follows:

Proposition 1. *Distortions dampen (amplify) aggregate volatility arising from firm-level shocks if relative distortions are positively (negatively) correlated with firms' productive ability. The strength of this dampening (amplification) effect increases with:*

1. *the magnitude of the correlation, as measured by the absolute value of the cosine similarity* $|\cos(\theta)|$;
2. *the dispersion in (squared) sales shares, given by* $\|\overline{\eta}^2\|$;
3. *the dispersion of relative distortions, measured by* $\|\mathbf{d}\|$.

The proof follows immediately from equation (16).

The change in concentration $\Delta\text{HHI} = \Psi^2 - {\Psi^d}^2$ is proportional to the change in the variance of TFP given by $\sigma_Z^2 - (\sigma_Z^d)^2 = \sigma_x^2 \times \left(\Psi^2 - {\Psi^d}^2\right)$. Additionally, observe that when distortions are all identical, or when distortions are uncorrelated, the wedges do not introduce changes to the economy's volatility, so only correlated distortions, ones inducing a systematic variation in marginal products of labor, affect aggregate volatility. The immediate policy implication of Proposition 1 is that any policy which induces a positively correlated distortion dampens aggregate volatility.[11] The next section demonstrates this with concrete policy examples.

3 Size-Dependent Policies as Ex Ante Automatic Stabilizers

Having established that distortions affect granular business cycles if they are correlated with productivity, and that positively correlated distortions reduce aggregate volatility while negatively correlated distortions amplify it, I now turn to specific policies that can generate this effect. The most directly relevant set of policy instruments is size-dependent policies

or the cosine of the angle α between them, as $\sum_i^N v_i u_i = \|\mathbf{u}\| \times \|\mathbf{v}\| \times \cos(\alpha)$.

[11] Appendix B discusses the correlation structure between productivity and distortions generated by different frictions, such as market power and financial frictions à la Kiyotaki and Moore (1997). The paper throughout is agnostic about the root cause of misallocation.

broadly defined. That is, any regulation, political structure, or friction that disproportionately favors firm i over firm j because of its size can be considered a size-dependent policy. In this section, I focus on tax policies and subsidies as they are the most commonly used policy instruments and are more straightforward to model in the benchmark environment. The analysis can be extended to other types of size-dependent policies as well.

Size-dependent policies. Consider now an economy whereby all firms are uniformly subject to a revenue tax at a rate of t_0 and either (i) the government decides to subsidize small businesses or (ii) large firms get tax credit or subsidies to maintain large local manufacturers and prevent them from moving abroad. The former case will be referred to as 'SME subsidies', and the latter as large-business subsidies. Note that SME subsidies are an instance of positively correlated distortion, whereas large business subsidies are negatively correlated. Unlike the general statement in Proposition 1, these taxes and subsidies are now made explicit. Both of these examples are stylized representations of size-dependent policies more broadly. I now restrict attention to examples with a constant subsidy elasticity as follows

$$s(l_i) = \left(\frac{l_i}{C_0}\right)^{-\nu} - 1, \tag{17}$$

where $s_i(l_i)$ is the size-dependent subsidy rate. Stated in logs: $\log(1+s_i) = \alpha_0 - \nu \log(l_i)$, with C_0 and α_0 denoting normalizing constants such that $1 + s_i(l_i) \geqslant 1$ for every i. Importantly, ν is the elasticity of the subsidy with respect to size. I use size as measured by employment here for tractability. When $\nu > 0$, we consider SME subsidies; when $\nu < 0$, we consider large-business subsidies. The flat tax rate t_0 is chosen so that the entire tax system generates zero equilibrium revenue, making it revenue-neutral and purely redistributive. This structure streamlines the presentation considerably and implies that the firms are now solving the following modified problem

$$\max_{l_i} \quad \mathbb{E}\left[(1-t_0)\left(1+s(l_i)\right) y_i\right] - w \times l_i, \tag{18}$$

where, as previously, $y_i = \tilde{z}_i {l_i}^{\gamma}$. These policies affect aggregate volatility as follows.

Proposition 2. *When all firms face the decision problem* (18) *given the tax structure given by equation* (17)*, and* $0 < \gamma - \nu < 1$*, such that* (18) *is well defined, the resulting equilibrium* *allocation of labor is given by*

$$l_j^\star = \frac{a_j^{\frac{1}{1-(\gamma-\nu)}}}{\sum_{i=1}^{N} a_i^{\frac{1}{1-(\gamma-\nu)}}} L. \tag{19}$$

The above allocation rule reduces concentration in the economy as measured by the HHI compared to the baseline efficient production environment of Lemma 1 when $\nu > 0$ and increases it when $\nu < 0$. Therefore, SME subsidies dampen aggregate volatility, whereas large-business subsidies amplify it.

For proof, and complete derivation, see Appendix A.2.

Note that first-order conditions now require that the firm's behavior follows a modified demand for labor, internalizing the subsidy structure. While the tax rate t_0 affects all firms uniformly and therefore has no effect on their relative equilibrium sizes, the subsidies alter *relative* sizes. The implied allocation rule reflects this such that it is isomorphic to a reduction of the span of control parameter γ.

The result is intuitive; volatility in the granular economy is governed by the degree of concentration as given by the HHI or Ψ^2. SME subsidies diversify production away from large firms, reduce concentration and, therefore, reduce Ψ. Conversely, large-business subsidies or favorable tax treatment for big businesses increase concentration, thereby increasing the volatility arising from firm-level shocks. Viewed thus, SME subsidies are ex-ante automatic stabilizers in the sense of McKay and Reis (2021). They are steady-state policy instruments that reduce the volatility of TFP one can expect in the economy ex ante, and do not require adjustments conditional on the realization of shocks (ex post).

To conclude this discussion, when taxes, subsidies, or regulations are size-dependent, they can affect the volatility of TFP by altering the degree of concentration in the economy. When these policies disproportionately favor small firms, they induce positively correlated distortions and act as ex-ante automatic stabilizers. In the next section, I take a broader view and quantify the extent to which existing distortions affect aggregate volatility in the U.S. economy, without committing to a specific source of misallocation or policy.

4 Quantifying the Stabilizing Effect of Correlated Distortions

In this section, I quantify the degree to which existing positively correlated distortions stabilize aggregate volatility. This quantification is challenging since one needs to compute Ψ without perfectly observing $\mathbf{a}$, and one must hold this vector constant when computing counterfactual economies, to guarantee that the baseline and the counterfactual differ only in the severity of distortions and not in firm-level productivity. I propose a portable approximation method requiring only a statistical or model-generated description of the firm-size distribution.

Applying this method to a U.S. calibration of the model implies that firm-level shocks in the U.S. generate annual TFP volatility of 0.38%, compared with an observed annual TFP volatility of 1%. Combined with estimates from the misallocation literature, one can infer that removing existing positively correlated distortions in the U.S. would raise TFP volatility from firm-level shocks to 0.65%. This effect suggests that positively correlated distortions have a powerful stabilization role against firm-level shocks. Appendix C validates the approximation method by relating it to existing findings in the literature. Appendix D shows that the main qualitative results remain robust when firm-level volatility declines with size and when labor can be reallocated ex post, although the magnitudes vary with the assumptions.

4.1 Approximation Strategy

In the efficient production case of Lemma 1, the amplification term is defined by $\Psi = \sqrt{\sum_{i=1}^{N} \overline{\eta}_i^2}$. Therefore, Ψ is a function of the firm-level deterministic ability $\mathbf{a}$ and not of the stochastic shocks $\tilde{\mathbf{x}}$. Suppose now that $\mathbf{a}$ is an i.i.d. sample of size N drawn from an ability distribution with CDF $G(a) : \mathbb{R}_{++} \rightarrow [0,1]$. Holding G and N constant, Ψ can vary depending on the exact realizations of a_i. Therefore, it is useful to consider the expected value of Ψ across realizations of $\mathbf{a}$. That is, with some abuse of notation, we can consider the function $\tilde{\Psi}(\mathbf{a})$, and define $\mathbb{E}\left[\tilde{\Psi}(\mathbf{a})\right] = \overline{\Psi}$, where expectations are now taken over the draws $\mathbf{a}$, rather than over the shocks $\tilde{\mathbf{x}}$. This $\overline{\Psi}$ is a function of G and N and can be thought of as the average amplification term across all possible realizations of $\mathbf{a} \in \mathbb{R}_{++}^{N}$ drawn from G.

To approximate this expected value $\overline{\Psi}$, define the quantile function Q as the inverse of G such that $Q:[0,1]\to\mathbb{R}_{++}$. For a random sample $\mathbf{a}$ we can define the appropriate vector $\mathbf{q}\in[0,1]^N$ such that $a_i=Q(q_i)$ or $G(a_i)=q_i$. Without loss of generality, we can order the vectors such that $a_1\geqslant a_2,\ldots,\geqslant a_N$. Since q_i are now quantiles of a random sample of N i.i.d. draws, they are uniformly distributed, and their ordering implies that they also correspond to order statistics. The ordered quantiles q_i are distributed as $q_i\sim\text{Beta}\,(N+1-i,i)$ with expectations given by $\overline{q}_i=\mathbb{E}\,[q_i]=\frac{N+1-i}{N+1}$.[12] Since the expected values $\overline{q}_i$ are known regardless of G, using the change of variable into the quantiles vector $\mathbf{q}$, we can approximate $\overline{\Psi}$ tractably as follows.

Lemma 3. *The function $\tilde{\Psi}(\mathbf{a})$ of the random i.i.d. sample $\mathbf{a}\in\mathbb{R}^N_{++}$ drawn from G can be expressed as a function of the quantiles vector $\mathbf{q}$ and the quantile function Q as follows $\tilde{\Psi}(\mathbf{a})=\tilde{\Psi}(Q(\mathbf{q}))$. Therefore, the expected value of $\tilde{\Psi}(\mathbf{a})$ can be approximated using*

$$\overline{\Psi}=\mathbb{E}\left[\tilde{\Psi}(\mathbf{a})\right]\approx\tilde{\Psi}(Q(\overline{\mathbf{q}}))\tag{20}$$

where $\overline{\mathbf{q}}$ is the vector of expected order statistics with $\overline{q}_i=\mathbb{E}\,[q_i]=\frac{N+1-i}{N+1}$.

The proof is straightforward. Taking a first-order Taylor expansion of $\tilde{\Psi}$ around $\overline{\mathbf{q}}$ as $\tilde{\Psi}(Q(\mathbf{q}))\approx\tilde{\Psi}(Q(\overline{\mathbf{q}}))+\sum_{i=1}^N\frac{\partial\tilde{\Psi}(Q(\overline{\mathbf{q}}))}{\partial q_i}(q_i-\overline{q}_i)$ and noting that $\mathbb{E}\,[(q_i-\overline{q}_i)]=0$, we obtain $\mathbb{E}\left[\tilde{\Psi}(\mathbf{a})\right]\approx\tilde{\Psi}(Q(\overline{\mathbf{q}}))$.[13] With this approximation, we can compute the expected amplification term $\overline{\Psi}$ as a function of N and G without perfectly observing $\mathbf{a}$. This approximation can also be applied to the case with distortions, but one must account for the relationship between distortions and productivity.[14]

This approximation has some desirable features. First, it accommodates fat-tailed distributions or distributions without finite moments. Using the quantiles vector $\mathbf{q}$ allows us to take expectations even if G itself does not have finite moments, since $\mathbf{q}$ always represents order statistics of a uniform distribution, thus allowing us to evaluate expectations of functions of potentially pathological distributions by working through the quantile function Q. Second, the approximation does not truncate the support of G. To demonstrate,

[12]This is a well-known result on the distribution of order statistics in finite random samples. For a formal treatment of this with opposite ordering of order statistics, see Gentle (2009).

[13]Higher order approximations are conceptually possible since the second derivatives and covariances of order statistics are also straightforward to obtain, see Gentle (2009) for exact formulas. However, these are computationally infeasible because the variance-covariance matrix required has N^2 entries.

[14]Formally, one can define $\tilde{\Psi}(\mathbf{a},\tau(\mathbf{a}))$, where letting $\tau(\mathbf{a})$ allows one to model the explicit correlation of distortions and productivity in specifying the wedges. Lemma 3 can be applied to this case as well, and we can approximate $\tilde{\Psi}(\mathbf{a},\tau(\mathbf{a}))$ as a function of the quantiles vector $\mathbf{q}$ and the quantile function Q as follows $\tilde{\Psi}(\mathbf{a},\tau(\mathbf{a}))=\tilde{\Psi}(Q(\mathbf{q}),\tau(\mathbf{Q}(\mathbf{q})))$. Therefore, the expected value of $\tilde{\Psi}$ in the case with distortions can be approximated using $\overline{\Psi}^d=\mathbb{E}\left[\tilde{\Psi}(\mathbf{a},\tau(\mathbf{a}))\right]\approx\tilde{\Psi}(Q(\overline{\mathbf{q}}),\tau(\mathbf{Q}(\overline{\mathbf{q}})))$.

$\overline{q} = (\frac{1}{N+1}, \ldots, \frac{N}{N+1})$, where the last term monotonically increases with the sample size. Appendix C further validates this approximation by showing its consistency with various results in the literature; both in terms of large N behavior given a Pareto tail for firm productivity as in Gabaix (2011), and in terms of quantitative results on the severity of granular business cycles as in the work of Carvalho and Grassi (2019). This approximation will be used throughout the paper to evaluate the magnitude of the amplification term in various scenarios.

4.2 How Strong is the Stabilization Effect of Positively Correlated Distortions on Volatility from Granular Shocks?

Now with the approximation strategy in place, I can evaluate the effect of correlated distortions on aggregate volatility. Doing so requires calibrating the distribution of firm-level productivity G, the number of firms in the economy N, the span of control γ, the volatility of firm-level shocks σ_x, and the correlation structure of distortions. These parameters and functional forms are discussed below, and the parameter values are summarized in Table 1.

Correlation of distortions and productivity. The bulk of the misallocation literature, starting from Restuccia and Rogerson (2008), has argued that a positive correlation of distortions and ability is an important feature. Hsieh and Klenow (2014) use plant life-cycle data from India, Mexico, and the United States and find a positive correlation between distortions and ability. They also report model-implied positive elasticities for the United States. Similar positive elasticities are used to match establishment sizes and firm sizes in various countries, e.g., Bartelsman et al. (2013), Buera and Fattal Jaef (2018); Bento and Restuccia (2017); Poschke (2018), and David and Venkateswaran (2019). I therefore follow the literature and assume that distortions are positively correlated with productivity.[15]

To evaluate quantitatively the effects of correlated distortions on aggregate volatility, I follow Bento and Restuccia (2017), Poschke (2018), and Buera and Fattal Jaef (2018) in specifying that distortions are positively correlated with productivity, using the following single parameter specification for the wedge $1-\tau_i$ as a function of the firm's ability a_i:

$$(1-\tau_i) = a_i{}^{-\phi}, \tag{21}$$

[15] Appendix B also shows how endogenous market power and financial frictions can generate positively correlated distortions. However, the direction and magnitude of the correlation induced by each friction depend on the specifics of the modeling approach.

where $\phi \in [0,1)$ corresponds to the elasticity of the distortions with respect to productive ability. An increase in a_i implies a higher value of τ_i.

When shocks are at their expected level, equation (21) and the allocation rule for labor derived in Lemma 2 jointly imply that firm-level employment is proportional to $a_i^{\frac{1}{1-\gamma}}(1-\tau_i)^{\frac{1}{1-\gamma}}$ or using equation (21) that $l_i^{d\star} \propto a_i^{\frac{1-\phi}{1-\gamma}}$. When firm size is measured instead by sales y_i^d or the relative sizes $\overline{\delta}_i$, the relative size of firm i is proportional to $a_i^{\frac{1}{1-\gamma}}(1-\tau_i)^{\frac{\gamma}{1-\gamma}}$, and therefore using equation (21) yields $y_i^d, \overline{\delta}_i \propto a_i^{\frac{1-\gamma\phi}{1-\gamma}}$.

Firm-level productivity distribution. I follow the literature on firm dynamics and granular business cycles in particular and assume the distribution of firm-level productivity follows a Pareto form. Given that the model is a repeated static economy, there is no need to commit to the particular mechanism generating a Pareto tail or power law behavior for the firm size distribution.

Suppose that the ability distribution in the economy takes Pareto form with a tail parameter $\zeta_a > 0$ and a scale parameter normalized to unity[16], i.e., the CDF for a_i is given by $G(a) = \text{Prob}(a_i \leqslant a) = 1 - a^{-\zeta_a}$. Then we have the following as a direct result.

Lemma 4. *When firm-level deterministic ability a_i is Pareto distributed with tail ζ_a, the firm-size distribution is also a Pareto distribution with tail parameter depending on the definition of size as follows*

$$\zeta_{emp} = \zeta_a \frac{1-\gamma}{1-\phi}, \quad \textit{and} \quad \zeta_{sales} = \zeta_a \frac{1-\gamma}{1-\phi\gamma}. \tag{22}$$

ζ_{emp} and ζ_{sales} denote the tail parameters for the firm-size distribution when size is measured by employment and sales, respectively.

For a formal proof, see Appendix A.3.

This parametric example provides additional intuition to Proposition 1. Positively correlated distortions reduce concentration and aggregate volatility in the economy by modifying the firm-size distribution. Reducing ϕ from a positive level to zero or alleviating the misallocation induced by the wedges reduces ζ_{sales}, which increases concentration and volatility due to firm-level shocks.

These formulas for tail parameters are also useful since they correspond to the empirically observed tail of the firm-size distribution. While given the multiplicative structure,

[16]This assumption is without loss of generality because aggregate volatility is governed by the distribution of relative, and not absolute, sizes.

the parameters ζ_a, γ, and ϕ cannot be separately identified from the firm-size distribution alone. These empirical estimates are still informative about the direction of the correlation between distortions and productivity.

To demonstrate, consider the efficient production case ($\phi = 0$) in which both measures of the tail parameter converge to $\zeta_a(1-\gamma)$, a known result in the literature (see, for example, the model of Carvalho and Grassi (2019)). However, when including distortions, the two measures diverge. Specifically, when distortions are positively correlated or $\phi > 0$, we have that $\zeta_{\text{emp}} > \zeta_{\text{sales}}$, which is consistent with the estimates of Axtell (2001) reporting $\zeta_{\text{emp}} = 1.059$ and $\zeta_{\text{sales}} = 0.994$ for the U.S. economy. While Axtell's estimates are noisy, the findings of di Giovanni et al. (2011), who use three different estimation procedures and French data, also consistently show that $\zeta_{\text{emp}} > \zeta_{\text{sales}}$ and yield more precise estimates. Crucially, ζ_{sales} is the value that governs the amplification term Ψ since it is the distribution of sales shares that modulates aggregate volatility due to firm-level shocks. Therefore, the fact that ζ_{sales} is smaller than ζ_{emp} is consistent with the presence of positively correlated distortions and a smaller amplification term in the economy with misallocation than in the efficient production case. To quantify the extent of this reduced volatility, I now proceed to calibrate the model and evaluate the effect of misallocation on aggregate volatility.

Parameter values. I follow Carvalho and Grassi (2019) and calibrate the model to have $N = 4.5 \times 10^6$ firms, with an observed Pareto tail for the firm-size distribution measured by employment of $\zeta_{\text{emp}} = 1.097$. I set the span of control parameter to $\gamma = 0.8$, which is well within the range accepted in the literature. Additionally, using the estimates from Hsieh and Klenow (2014), I calibrate the distortion elasticity in the U.S. to $\phi = 0.09$, which is an estimate obtained for U.S. establishments. This estimate is also used by Bento and Restuccia (2017).

A critical input that governs the degree of aggregate volatility due to firm-level shocks is σ_x, which is the volatility of the firm-level Solow residual. In the model, firm-level output is given by $y_i = e^{\tilde{x}_i} a_i l_i^{\gamma}$, thus the firm-level Solow residual or the residual term from regressing output changes on input changes (in logs) in the model is $\tilde{x}_i$. Carvalho and Grassi (2019) survey the literature estimating σ_x, citing estimates ranging between 0.09 and 0.2, and use a lower bound estimate of 0.08 in their benchmark calibration. I follow their calibration and set σ_x to 0.08 as well. This estimate is conservative, and all the results that follow scale up linearly with σ_x; thus, using a higher estimate would only amplify the effects of correlated distortions on aggregate volatility. But, importantly, changing σ_x does not affect the relative volatility in the distorted economy compared with the efficient

Table 1: Calibration of the Baseline Economy with Positively Correlated Distortions

	Parameter	**Value**	**Source**
N	Number of firms	4.5×10^{6}	Carvalho and Grassi (2019)
ζ_{emp}	Observed Pareto tail	1.097	Carvalho and Grassi (2019)
γ	Span of control	0.80	Carvalho and Grassi (2019)
σ_x	Std. of firm-level shock	8%	Carvalho and Grassi (2019)
ϕ	Distortion elasticity	0.09	Hsieh and Klenow (2014)

production benchmark since $\frac{\sigma_Z^d-\sigma_Z}{\sigma_Z}=\frac{\Psi^d-\Psi}{\Psi}$, which is unaffected by σ_x.

Effect of Misallocation on TFP Volatility. The calibrated model exhibits a TFP volatility of $\sigma_Z = 0.38\%$ due to firm-level shocks alone. Counterfactually removing the distortions increases TFP volatility from 0.38% to 0.65%. Therefore, the presence of positively correlated distortions in the U.S. dampens TFP volatility resulting from firm-level shocks by 42% of its potential level. To put these numbers in perspective, Carvalho and Grassi (2019) estimate the annual volatility of TFP to be 1% in the U.S.; thus, the model implies that firm-level shocks account for 38% of aggregate TFP volatility, and the implied increase in TFP volatility from removing positively correlated distortions corresponds to a 27 percentage point increase in aggregate TFP volatility. This sizable effect suggests that positively correlated distortions have a powerful stabilization role against firm-level shocks, even with the relatively mild distortion present in the U.S. economy.

To assess the robustness of this sizable stabilization effect, Figure 2 reports how aggregate volatility changes in response to removing the positively correlated distortions for various values of ϕ and γ. The stabilization effect, represented in the figure by the difference between the dark red and light blue bars for each scenario, is monotone and increasing in ϕ and γ. The parameter ranges are chosen to be consistent with the existing literature.

Estimates for ϕ in the literature are naturally bounded between zero and unity; Hsieh and Klenow (2014) report values of 0.5 for India and 0.66 for Mexico. Bento and Restuccia (2017) report estimates for various countries with the U.S. value of $\phi = 0.09$ as the lower bar, and most other countries ranging between $\phi = 0.3$ and $\phi = 0.7$. I thus consider values ranging between $\phi = 0$ and $\phi = 0.2$ for my U.S.-based calibration. The stabilization effect is monotonically increasing in ϕ.

The value $\gamma = 0.8$ is conventional in the literature. The admissible range for γ ranges between 0.6 and 0.9. $\gamma = 0.6$ corresponds to the U.S. labor share and serves as a natural lower bound. A natural upper bound for the span of control parameter is one, at the constant

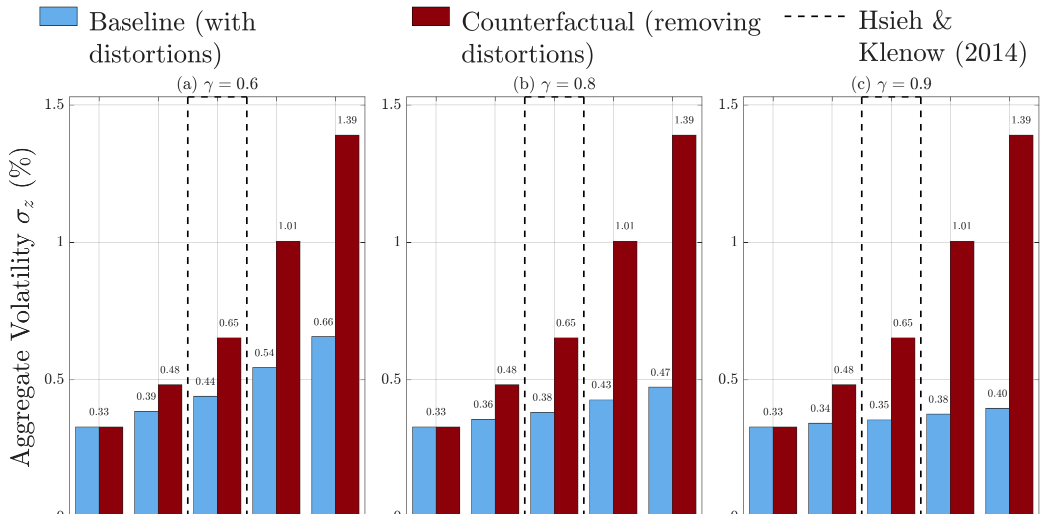

Figure 2: The Effect of Correlated Distortions on Aggregate Volatility Due to Granular Shocks

Note: Aggregate volatility due to firm-level shocks with positively correlated distortions (in light blue) and when these distortions are counterfactually removed (in dark red) using different assumptions on γ and ϕ.

returns to scale limit. However, given that corporate pre-tax profits are in the order of magnitude of 10%, $\gamma = 0.9$ serves as a better upper bound, as with decreasing returns to scale the profit share is $1-\gamma$. Figure 2 illustrates that the higher γ is, the stronger the stabilizing effect of correlated distortions. Still, even under the lowest value of $\gamma = 0.6$, a distortion elasticity of $\phi = 0.09$ implies a sizable effect from removing correlated distortions with the volatility due to firm-level shocks rising from the benchmark of 0.44% to a counterfactual level of 0.65%.

Appendix D shows that the stabilizing role of positively correlated distortions is robust to changing two critical assumptions underlying the above exercise. First, Appendix D.1 demonstrates that the stabilization role is present even when the volatility of firm-level shocks is lower for larger firms; however, the total volatility attributable to firm-level shocks is also lower. Second, Appendix D.2 demonstrates that the stabilization effect does not rely on the assumption that labor is allocated before the shock is realized. Ultimately, these results highlight the importance of accounting for correlated distortions when studying aggregate fluctuations in granular economies. Given the size of this stabilization effect, I now turn to the normative implications of these results.

5 Social Optimum in the Granular Economy

The analysis thus far establishes that when firm-level shocks are a source of aggregate risk, distortions that are correlated with productivity affect this risk. Such distortions are studied in relation to misallocation and are often inferred from the data using observed systematic dispersion in marginal products as an indication of inefficiency. Having demonstrated that positively correlated distortions can dampen this risk and act as stabilizers, one may ask: if aggregate risk is consequential to social welfare, then, in the presence of a finite number of firms (granularity), would dispersion in marginal products be indeed a sign of inefficiency or a desirable feature? In other words, can what we call misallocation in the literature emerge endogenously as an optimal response to the presence of large firms in the economy? In what follows, I show that the answer is: yes.

Before proceeding, I want to note that my focus in this section is to make a qualitative statement. The model presented exhibits a small welfare cost of business cycles, a property known since the work of Lucas (1987), and uses similar tools. The mechanism, however, is more general than the specific model and applies in more sophisticated frameworks in which concentration maps to aggregate risk.

Consider the welfare of a representative household with flow utility from consumption $u(C_{\tilde{\mathbf{x}}})$ that is monotonically increasing and concave, with output evolving as in the previous sections. Due to goods market clearing $Y_{\tilde{\mathbf{x}}} = C_{\tilde{\mathbf{x}}}$ thus output volatility also implies consumption volatility. Let household preferences exhibit CRRA with parameter $\chi = -C_{\tilde{\mathbf{x}}} u''(C_{\tilde{\mathbf{x}}}) \big/ u'(C_{\tilde{\mathbf{x}}})$, and suppose we normalize $L = 1$. Therefore, $Y_{\tilde{\mathbf{x}}} = Z_{\tilde{\mathbf{x}}}$ and thus welfare is solely a function of TFP with $u(C_{\tilde{\mathbf{x}}}) = u(Z_{\tilde{\mathbf{x}}})$. Using standard second-order approximations, one can obtain that welfare in the economy with distortions can be approximated by

$$\mathbb{E}\left[u\left(Z^d_{\tilde{\mathbf{x}}}\right)\right] \approx u\left(\overline{Z^d}\right) + \frac{{\sigma_x}^2}{2}\overline{Z^d}\,u'\left(\overline{Z^d}\right)\left[1 - \chi {\Psi^d}^2\right] \tag{23}$$

where $\overline{Z^d} = Z^d_{\tilde{\mathbf{x}}}\big|_{\tilde{\mathbf{x}}=\overline{\mathbf{x}}}$. Note that the efficient production case is analogous but with $Z_{\tilde{\mathbf{x}}}$ and Ψ substituted instead of $Z^d_{\tilde{\mathbf{x}}}$ and Ψ^d.[17] The log utility case, in which $\chi = 1$ and $\overline{Z^d}\,u'\left(\overline{Z^d}\right) = 1$, is particularly insightful since when we subtract the case with distortions from the one with efficient production we have that the welfare gain or loss from alleviating distortions is

[17] For a detailed derivation see appendix A.4.

given by

$$\underbrace{\mathbb{E}\left[u\left(Z_{\tilde{\mathbf{x}}}\right)\right]-\mathbb{E}\left[u\left(Z_{\tilde{\mathbf{x}}}^{d}\right)\right]}_{\text{Counterfactual welfare gain}} \approx \underbrace{u\left(\overline{Z}\right)-u\left(\overline{Z^{d}}\right)}_{\substack{\text{Production efficiency}\\ \text{Gain}}} \underbrace{-\frac{{\sigma_x}^2}{2}\left[\Psi^2-{\Psi^{d}}^2\right]}_{\substack{\text{Stability effect}\\ \Delta\text{HHI}}}. \tag{24}$$

The two equations above demonstrate that stabilizing aggregate fluctuations resulting from firm-level shocks is valuable to a risk-averse social planner.

Works in the misallocation literature often ignore the second term on the right-hand side of equation (24) when evaluating counterfactuals, which is correct only if the planner is risk neutral or if firm-level shocks have no aggregate effects. When ignoring this term, Proposition 1 tells us that we ignore the stability effect, which is signed by the correlation between distortions and productivity. If distortions are positively correlated, we obtain overstated welfare gains. Likewise, if the distortions are negatively correlated, the welfare gains are understated.

The above description of the economy is naïve in the sense that the household does not allocate labor to maximize expected utility but rather expected output, as in the production economies described in section 2. One should consider the above derivation only as an intermediate step that provides intuition for the following result.

Let us now consider a social planner solving the optimal allocation of labor without knowledge of the shocks $\tilde{\mathbf{x}}$. Still, the planner understands how the firm-size distribution governs consumption risk via exposure to these shocks. Formally, let the planner solve

$$\begin{aligned} \max_{\mathbf{l}} \quad & \mathbb{E}\left[u\left(Y_{\tilde{\mathbf{x}}}\right)\right] \\ \text{s.t.} \quad & Y_{\tilde{\mathbf{x}}}=\sum_{i=1}^{N}\tilde{z}_i l_i^{\gamma}, \quad \sum_{i=1}^{N} l_i = L, \quad \mathbf{l}\in\mathbb{R}_{+}^{N}, \quad \tilde{z}_i = a_i e^{\tilde{x}_i} \end{aligned} \tag{25}$$

where u is assumed to be a monotonically increasing and concave utility function. Note that the first order condition for this problem does not involve equalizing marginal products across firms, but rather equalizing risk-weighted expected marginal products in marginal utility terms as

$$\mathbb{E}[u'\left(Y_{\tilde{\mathbf{x}}}\right)\tilde{z}_i\gamma l_i^{\gamma-1}] = \mathbb{E}[u'\left(Y_{\tilde{\mathbf{x}}}\right)\tilde{z}_j\gamma l_j^{\gamma-1}] = \lambda, \quad \forall i,j\in\{1,\ldots,N\}, \tag{26}$$

where λ is the Lagrange multiplier associated with labor market clearing.

Proposition 3. *The optimal allocation which solves the planner's problem in (25), can be decentralized and implemented using the production economy with distortions given in Lemma 2, if the implicit taxes $\tau \in \mathbb{R}_{+}^{N}$ are set such that for every pair of firms i and j the following condition holds*

$$\frac{1-\tau_i}{1-\tau_j} = \frac{\mathbb{E}[u'(Y_{\tilde{\mathbf{x}}})\,\tilde{z}_i]/\mathbb{E}[\tilde{z}_i]}{\mathbb{E}\left[u'(Y_{\tilde{\mathbf{x}}})\,\tilde{z}_j\right]/\mathbb{E}\left[\tilde{z}_j\right]}. \tag{27}$$

Multiple wedge vectors can be constructed to follow this condition, all proportional to each other. A natural choice is to normalize the scale of the distortions such that $\tau_i = 0$ when $u'(Y_{\tilde{\mathbf{x}}}) \perp \tilde{z}_i$, or when firm-specific shocks are uncorrelated with marginal utility of consumption. This can be achieved by setting $(1-\tau_i) = \frac{\mathbb{E}[u'(Y_{\tilde{\mathbf{x}}})\tilde{z}_i]}{\mathbb{E}[u'(Y_{\tilde{\mathbf{x}}})]\mathbb{E}[\tilde{z}_i]}$, which obeys equation (27), and yields the convenient interpretation of the optimal wedge for firm i as follows

$$1-\tau_i = 1 + \frac{\mathbb{COV}(u'(Y_{\tilde{\mathbf{x}}}),\tilde{z}_i)}{\mathbb{E}[u'(Y_{\tilde{\mathbf{x}}})]\mathbb{E}[\tilde{z}_i]}. \tag{28}$$

For proof, see Appendix A.5.

The covariance term in equation (28) is necessarily negative since a positive firm-level shock increases output and decreases the marginal utility of consumption. Consequently, a larger covariance in absolute value leads to a higher τ_i. It is therefore optimal for the planner to set a higher wedge for firms that are more correlated with the marginal utility of consumption, or, equivalently, with higher aggregate consumption risk. Thus, when more productive firms have a larger effect on aggregate consumption risk, reallocating labor towards low-productivity firms can provide partial insurance. Ultimately, the covariance term leads the planner to optimally allocate labor to induce a dispersion in marginal products, which empirically manifests as if distortions are positively correlated with firm-level productivity. A cleaner way to see this is to assume a CRRA utility to obtain the following approximation for the optimal wedge.

Corollary 1. *If the utility function is CRRA with parameter χ, the optimal wedge for firm i is given by*

$$1-\tau_i = 1 + \frac{\mathbb{COV}(u'(Y_{\tilde{\mathbf{x}}}),\tilde{z}_i)}{\mathbb{E}[u'(Y_{\tilde{\mathbf{x}}})]\mathbb{E}[\tilde{z}_i]} \approx 1 - \chi\overline{\eta}_i{\sigma_x}^2. \tag{29}$$

See proof in Appendix A.5.

The above corollary demonstrates that the optimal, welfare-maximizing, allocation of labor in the granular economy internalizes a trade-off between production efficiency and business-cycle risk. A high-productivity firm contributes more to aggregate output, but has a higher value of $\overline{\eta}_i$, indicating that it also exerts a larger influence on aggregate volatility. Diversifying labor away from high-productivity firms yields a loss in output that the planner is willing to incur in exchange for a reduction in business-cycle risk. In fact, the implicit tax τ_i is higher for firms with higher $\overline{\eta}_i$ or with higher productive ability. Therefore, the dispersion in marginal products of labor in the economy is optimal in the granular economy.

To conclude, a risk-averse social planner or an efficient market mechanism will optimally reduce the size of large productive businesses relative to the efficient production benchmark in Lemma 1. Note that this is so without assuming anything about these large firms' market power or political influence. Instead, the planner is concerned about the presence of large firms in the economy solely because of their size. The optimal allocation exhibits dispersion of the marginal products of labor as if there are positively correlated distortions in the economy. However, these distortions are not evidence of misallocation in the welfare-suboptimal sense, but rather of the planner fully internalizing the effects of large firms on the economy and optimizing factor allocations accordingly.

One could conceive of market failures that would prevent the planner or market mechanism from internalizing the effect of large firms on aggregate risk and thus from implementing the optimal allocation. In such cases, large firms would create an externality that merits intervention. The extent to which such potential inefficiency is a concern depends on the degree to which aggregate fluctuations resulting from firm-level shocks are detrimental to welfare.

6 Concluding Remarks

The key idea of this paper is that distortions to firm size choices affect not only the level of aggregate output but also aggregate fluctuations by reshaping concentration in the economy. I formalize this idea within the canonical firm-dynamics model à la Hopenhayn (1992), in which correlated distortions change the firm-size distribution and therefore the economy's exposure to firm-level shocks. Positively correlated distortions, which compress large firms, dampen aggregate volatility; negatively correlated distortions do the opposite. Since the empirical literature on misallocation typically finds distortions to be positively correlated with productivity, these distortions primarily stabilize the economy against firm-level shocks. In the baseline U.S. calibration, removing positively correlated distortions raises TFP volatility from firm-level shocks from 0.38 percent to 0.65 percent, implying that these distortions reduce this component of volatility by about 42 percent.

The analysis challenges the standard normative interpretation of marginal product dispersion as evidence of inefficiency. In a granular economy, such dispersion can arise in an optimal, welfare-maximizing, allocation that internalizes firms' contributions to aggregate risk. A risk-averse social planner does not equalize marginal products, but risk-weighted expected marginal products in marginal-utility terms, because the allocation of labor determines both expected output and aggregate volatility. What appears as misallocation ex post may therefore partly reflect ex ante stabilization: a deliberate sacrifice of expected output in order to reduce aggregate risk.

If decentralized markets fail to price firms' contributions to aggregate risk, private size choices need not implement the efficient allocation. In such cases, size-dependent policies that favor smaller firms or compress large ones can act as ex ante stabilizers. I show that policies such as SME subsidies can reduce aggregate volatility, although their desirability depends on the efficiency of the market mechanism, the welfare cost of aggregate fluctuations, and the implementation costs and dynamic effects not considered here. More broadly, the results show that the firm-size distribution is both a determinant of aggregate productivity and a determinant of aggregate risk. Policies and counterfactuals that reshape it should therefore be evaluated in terms of both the level and volatility of aggregate outcomes.

References

Acemoglu, D., V. M. Carvalho, A. Ozdaglar, and A. Tahbaz-Salehi (2012). The network origins of aggregate fluctuations. *Econometrica 80*(5), 1977–2016.

Amaral, L. A. N., S. V. Buldyrev, S. Havlin, P. Maass, M. A. Salinger, H. E. Stanley, and M. H. Stanley (1997). Scaling behavior in economics: the problem of quantifying company growth. *Physica A: Statistical Mechanics and its Applications 244*(1-4), 1–24.

Asker, J., A. Collard-Wexler, and J. D. Loecker (2014). Dynamic inputs and resource (mis)allocation. *Journal of Political Economy 122*(5), 1013–1063.

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

Axtell, R. L. (2001). Zipf distribution of us firm sizes. *science 293*(5536), 1818–1820.

Baqaee, D. and E. Farhi (2019). The macroeconomic impact of microeconomic shocks: Beyond hulten's theorem. *Econometrica 87*(4), 1155–1203.

Baqaee, D. and E. Farhi (2020). Productivity and misallocation in general equilibrium. *Quarterly Journal of Economics*.

Bartelsman, E., J. Haltiwanger, and S. Scarpetta (2013, February). Cross-country differences in productivity: The role of allocation and selection. *American Economic Review 103*(1), 305–34.

Bento, P. and D. Restuccia (2017). Misallocation, establishment size, and productivity. *American Economic Journal: Macroeconomics 9*(3), 267–303.

Boar, C., D. Gorea, and V. Midrigan (2022). Why are returns to private business wealth so dispersed? Technical report, National Bureau of Economic Research.

Boar, C. and V. Midrigan (2024). Markups and inequality. *Review of Economic Studies*, rdae103.

Bornstein, G. and A. Peter (2024). Nonlinear pricing and misallocation. Technical report, National Bureau of Economic Research.

Buera, F. J. and R. N. Fattal Jaef (2018, June 29). The dynamics of development: Innovation and reallocation. *World Bank Policy Research Working Paper* (8505).

Buera, F. J. and Y. Shin (2013). Financial frictions and the persistence of history: A quantitative exploration. *Journal of Political Economy 121*(2), 221–272.

Burstein, A., V. M. Carvalho, and B. Grassi (2025). Bottom-up markup fluctuations. *The Quarterly Journal of Economics*, qjaf029.

Carvalho, V. and X. Gabaix (2013, August). The great diversification and its undoing. *American Economic Review 103*(5), 1697–1727.

Carvalho, V. M. and B. Grassi (2019, April). Large firm dynamics and the business cycle. *American Economic Review 109*(4), 1375–1425.

Clementi, G. L. and H. A. Hopenhayn (2006). A theory of financing constraints and firm dynamics. *The Quarterly Journal of Economics 121*(1), 229–265.

Cooley, T. F. and V. Quadrini (2001). Financial markets and firm dynamics. *American Economic Review 91*(5), 1286–1310.

David, J. M., L. Schmid, and D. Zeke (2022). Risk-adjusted capital allocation and misallocation. *Journal of Financial Economics 145*(3), 684–705.

David, J. M. and V. Venkateswaran (2019, July). The sources of capital misallocation. *American Economic Review 109*(7), 2531–67.

di Giovanni, J., A. A. Levchenko, and I. Mejean (2014). Firms, destinations, and aggregate fluctuations. *Econometrica 82*(4), 1303–1340.

di Giovanni, J., A. A. Levchenko, and I. Mejean (2018, January). The micro origins of international business-cycle comovement. *American Economic Review 108*(1), 82–108.

di Giovanni, J., A. A. Levchenko, and R. Rancière (2011). Power laws in firm size and openness to trade: Measurement and implications. *Journal of International Economics 85*(1), 42–52.

Di Tella, S., C. Malgieri, and C. Tonetti (2025). Risk markups. Technical report, National Bureau of Economic Research.

Drechsel, T. (2023, April). Earnings-based borrowing constraints and macroeconomic fluctuations. *American Economic Journal: Macroeconomics 15*(2), 1–34.

Eden, M. (2017, February). Misallocation and the distribution of global volatility. *American Economic Review 107*(2), 592–622.

Edmond, C., V. Midrigan, and D. Y. Xu (2015, October). Competition, markups, and the gains from international trade. *American Economic Review 105*(10), 3183–3221.

Evans, D. S. (1987). Tests of alternative theories of firm growth. *The Journal of Political Economy 95*(4), 657–674.

Gabaix, X. (2011). The granular origins of aggregate fluctuations. *Econometrica 79*(3), 733–772.

García-Santana, M. and J. Pijoan-Mas (2014). The reservation laws in india and the misallocation of production factors. *Journal of Monetary Economics 66*, 193–209.

Garicano, L., C. Lelarge, and J. Van Reenen (2016, November). Firm size distortions and the productivity distribution: Evidence from france. *American Economic Review 106*(11), 3439–79.

Gaubert, C. and O. Itskhoki (2021). Granular comparative advantage. *Journal of Political Economy 129*(3), 871–939.

Gentle, J. (2009). *Computational Statistics*. Statistics and Computing. Springer New York.

Gourio, F. and N. Roys (2014). Size-dependent regulations, firm size distribution, and reallocation. *Quantitative Economics 5*(2), 377–416.

Grassi, B. (2018). IO in I-O: Size, Industrial Organization, and the Input-Output Network Make a Firm Structurally Important. Working Papers 619, IGIER (Innocenzo Gasparini Institute for Economic Research), Bocconi University.

Guner, N., G. Ventura, and Y. Xu (2008). Macroeconomic implications of size-dependent policies. *Review of Economic Dynamics 11*(4), 721–744.

Haltiwanger, J., R. S. Jarmin, and J. Miranda (2013). Who creates jobs? small versus large versus young. *Review of Economics and Statistics 95*(2), 347–361.

Hopenhayn, H. A. (1992, September). Entry, Exit, and Firm Dynamics in Long Run Equilibrium. *Econometrica 60*(5), 1127–1150.

Hopenhayn, H. A. (2014). Firms, misallocation, and aggregate productivity: A review. *Annual Review of Economics 6*(1), 735–770.

Hsieh, C.-T. and P. Klenow (2009). Misallocation and manufacturing tfp in china and india. *The Quarterly Journal of Economics 124*(4), 1403–1448.

Hsieh, C.-T. and P. Klenow (2014). The life cycle of plants in india and mexico. *The Quarterly Journal of Economics 129*(3), 1035–1084.

Hulten, C. R. (1978, 10). Growth Accounting with Intermediate Inputs. *The Review of Economic Studies 45*(3), 511–518.

Kiyotaki, N. and J. Moore (1997). Credit cycles. *Journal of Political Economy 105*(2), 211–248.

Koren, M. and S. Tenreyro (2013, February). Technological diversification. *American Economic Review 103*(1), 378–414.

Lian, C. and Y. Ma (2021). Anatomy of corporate borrowing constraints. *The Quarterly Journal of Economics 136*(1), 229–291.

Lucas, R. E. J. (1987). *Models of Business Cycles*. Yrjö Jahnsson Lectures. London: Blackwell Publishing.

McKay, A. and R. Reis (2021). Optimal automatic stabilizers. *The Review of Economic Studies 88*(5), 2375–2406.

Menzio, G. (2024). Markups: A search-theoretic perspective. Technical report, National Bureau of Economic Research.

Moll, B. (2014, October). Productivity losses from financial frictions: Can self-financing undo capital misallocation? *American Economic Review 104*(10), 3186–3221.

Olley, G. S. and A. Pakes (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica 64*(6), 1263–1297.

Opp, M. M., C. A. Parlour, and J. Walden (2014). Markup cycles, dynamic misallocation, and amplification. *Journal of Economic Theory 154*, 126–161.

Poschke, M. (2018). The firm size distribution across countries and skill-biased change in entrepreneurial technology. *American Economic Journal: Macroeconomics 10*(3), 1–41.

Restuccia, D. and R. Rogerson (2008, October). Policy Distortions and Aggregate Productivity with Heterogeneous Plants. *Review of Economic Dynamics 11*(4), 707–720.

Senga, T. and I. Varotto (2024). Idiosyncratic shocks and investment irreversibility: Capital misallocation over the business cycle.

Stanley, M., L. Amaral, S. Havlin, H. Leschhorn, P. Maass, M. Salinger, and H. Stanley (1996, 02). Scaling behavior in the growth of companies. *Nature 379*.

Sutton, J. (2002). The variance of firm growth rates: the 'scaling'puzzle. *Physica a: statistical mechanics and its applications 312*(3-4), 577–590.

Yeh, C. (2023). Revisiting the origins of business cycles with the size-variance relationship. *Review of Economics and Statistics*, 1–28.

Appendix A Proofs

A.1 Proof of Lemmas 1 and 2 and a Step by Step Derivation

Proof. Allocation rule for labor. The first order condition for the production problem (8), is given by

$$\gamma a_j\big(1-\tau_j\big)l_j^{\gamma-1} = w. \tag{30}$$

We can rearrange this expression and use the market clearing condition for labor to obtain

$$\sum_{i=1}^{N} l_i \;=\left(\frac{\gamma}{w}\right)^{\frac{1}{1-\gamma}} \sum_{i=1}^{N} \left(a_i(1-\tau_i)\right)^{\frac{1}{1-\gamma}} \;= L. \tag{31}$$

Thus, in equilibrium, we have that

$$l_j^{d\star} = \frac{\big(a_j\big(1-\tau_j\big)\big)^{\frac{1}{1-\gamma}}}{\sum_{i=1}^{N}\left(a_i(1-\tau_i)\right)^{\frac{1}{1-\gamma}}}L, \quad y_j^d = \frac{a_j^{\frac{1}{1-\gamma}}\big(1-\tau_j\big)^{\frac{\gamma}{1-\gamma}} e^{\tilde{x}_j}}{\left(\sum_{i=1}^{N}\left(a_i(1-\tau_i)\right)^{\frac{1}{1-\gamma}}\right)^{\gamma}} L^{\gamma}, \tag{32}$$

which is the allocation rule given in equation (9), and firm-level output.

Aggregate production function representation. Using the above firm-level output y_j and summing across all N firms allows one to obtain the aggregate production function in equation (10) as

$$Y^d_{\tilde{\mathbf{x}}} = \sum_{i=1}^{N} y_i \ = Z^d_{\tilde{\mathbf{x}}} \times L^{\gamma}, \quad Z^d_{\tilde{\mathbf{x}}} = \frac{\sum_{i=1}^{N} a_i^{\frac{1}{1-\gamma}} (1-\tau_i)^{\frac{\gamma}{1-\gamma}} e^{\tilde{x}_i}}{\left[\sum_{i=1}^{N} \left(a_i(1-\tau_i)\right)^{\frac{1}{1-\gamma}}\right]^{\gamma}}. \tag{33}$$

Hulten's theorem. The elasticity of the resulting economy's log TFP, to a one-percent shock to the productive ability of the j^{th} firm, is derived as:

$$\delta_j(\tilde{\mathbf{x}}) = \frac{\partial \log\left({Z_{\tilde{\mathbf{x}}}}^d\right)}{\partial \log \tilde{z}_j} = \frac{\partial \log\left({Z_{\tilde{\mathbf{x}}}}^d\right)}{\partial \tilde{x}_j} = \frac{1}{{Z_{\tilde{\mathbf{x}}}}^d} \frac{a_j^{\frac{1}{1-\gamma}} (1-\tau_j)^{\frac{\gamma}{1-\gamma}} e^{\tilde{x}_j}}{\left[\sum_{i=1}^{N} \left(a_i(1-\tau_i)\right)^{\frac{1}{1-\gamma}}\right]^{\gamma}} = \frac{a_j^{\frac{1}{1-\gamma}} (1-\tau_j)^{\frac{\gamma}{1-\gamma}} e^{\tilde{x}_j}}{\sum_{i=1}^{N} a_i^{\frac{1}{1-\gamma}} (1-\tau_i)^{\frac{\gamma}{1-\gamma}} e^{\tilde{x}_i}}. \tag{34}$$

Observe that $s^d_{Y,j} = y^d_j / Y^d_{\tilde{\mathbf{x}}}$ is given by

$$s^d_{Y,j} = \frac{y_j}{Y^d_{\tilde{\mathbf{x}}}} = \frac{a_j^{\frac{1}{1-\gamma}} \left(1-\tau_j\right)^{\frac{\gamma}{1-\gamma}} e^{\tilde{x}_j}}{\left(\sum_{i=1}^{N} \left(a_i(1-\tau_i)\right)^{\frac{1}{1-\gamma}}\right)^{\gamma}} L^{\gamma} \times \left[\frac{\sum_{i=1}^{N} a_i^{\frac{1}{1-\gamma}} (1-\tau_i)^{\frac{\gamma}{1-\gamma}} e^{\tilde{x}_i}}{\left[\sum_{i=1}^{N} \left(a_i(1-\tau_i)\right)^{\frac{1}{1-\gamma}}\right]^{\gamma}} \times L^{\gamma}\right]^{-1} = \delta_j,$$

thus yielding Lemma 2 (3), and demonstrating that Hulten's theorem holds exactly in the environment.

Aggregate volatility. With some abuse of notation, we can express

$$Z^d_{\tilde{\mathbf{x}}} = \sum_{i=1}^{N} a_i^{\frac{1}{1-\gamma}} (1-\tau_i)^{\frac{\gamma}{1-\gamma}} e^{\tilde{x}_i} \times \left[\sum_{i=1}^{N} \left(a_i(1-\tau_i)\right)^{\frac{1}{1-\gamma}}\right]^{-\gamma} = Z^d(\tilde{\mathbf{x}}), \tag{35}$$

where $\tilde{\mathbf{x}}$ is the vector containing all the N i.i.d. realizations of $\tilde{x}_j$. Using a Taylor series expansion around the mean point $\overline{\mathbf{x}} \in \mathbb{R}^N$, we can derive the approximation given in Lemma 2 (4) as follows

$$\mathbb{VAR}\left[\log\left(Z^d(\tilde{\mathbf{x}})\right)\right] \approx \mathbb{VAR}\left[\log\left(Z^d(\overline{\mathbf{x}})\right) + \sum_{i=1}^{N} \frac{\partial \log\left(Z^d(\overline{\mathbf{x}})\right)}{\partial \tilde{x}_i} (\tilde{x}_i - \overline{x})\right] = {\sigma_x}^2 \sum_{i=1}^{N} \overline{\delta}_i^2, \tag{36}$$

where $\overline{\delta}_i = \delta_i(\tilde{\mathbf{x}})|_{\tilde{\mathbf{x}}=\overline{\mathbf{x}}}$, and σ_x^2 is the variance of $\tilde{x}_j$. Taking the square root of equation (36) yields the volatility σ^d_Z in equation (12) of Lemma 2 (4), thus concluding the proof.

Note that setting $\tau_i = \tau_j, \forall i, j \in \{1, \ldots, N\}$, in the above expressions yields that $l_i^{d\star} = l_i^{\star}$, $y^d_i = y_i$, $Y^d_{\tilde{\mathbf{x}}} = Y_{\tilde{\mathbf{x}}}$, $Z^d_{\tilde{\mathbf{x}}} = Z_{\tilde{\mathbf{x}}}$, $s^d_{Y,i} = s_{Y,i}$, $\delta_i(\tilde{\mathbf{x}}) = \eta_i(\tilde{\mathbf{x}})$, $\overline{\delta}_i = \overline{\eta}_i$, $\sigma^d_Z = \sigma_Z$, and $\Psi^d = \Psi$, as stated in Lemma 1. $\square$

A.2 Proof of Proposition 2

Proof. To prove that SME subsidies dampen aggregate volatility and that large-business subsidies amplify it, we must first derive the aggregate properties of the economy in question. The firm problem in (18) yields the following first order condition

$$l_j = a_j^{\frac{1}{1-(\gamma-\nu)}} \left(C_0{}^{\nu} \frac{(1-t_0)(\gamma-\nu)}{w} \right)^{\frac{1}{1-(\gamma-\nu)}}. \tag{37}$$

Using labor market clearing as

$$L = \sum_{i=1}^{N} l_i = \left(C_0{}^{\nu} \frac{(1-t_0)(\gamma-\nu)}{w} \right)^{\frac{1}{1-(\gamma-\nu)}} \sum_{i=1}^{N} a_i^{\frac{1}{1-(\gamma-\nu)}}, \tag{38}$$

we obtain the labor allocation rule in equation (19). It is trivial that the resulting allocation of labor is identical to the one in an economy with a different span of control parameter γ. Note that output at the firm level is given by

$$y_j = e^{\tilde{x}_j} a_j^{\frac{1+\nu}{1-(\gamma-\nu)}} \left[\sum_{i=1}^{N} a_i^{\frac{1}{1-(\gamma-\nu)}} \right]^{-\gamma} L^{\gamma}, \tag{39}$$

and therefore, aggregate output and productivity are given by

$$Y_{\tilde{\mathbf{x}}}^{\mathrm{SME}} = \sum_{i=1}^{N} y_i = Z_{\tilde{\mathbf{x}}}^{\mathrm{SME}} L^{\gamma}, \quad Z_{\tilde{\mathbf{x}}}^{\mathrm{SME}} = \sum_{i=1}^{N} \frac{e^{\tilde{x}_i} a_i^{\frac{1+\nu}{1-(\gamma-\nu)}}}{\left[\sum_{i=1}^{N} a_i^{\frac{1}{1-(\gamma-\nu)}} \right]^{\gamma}}. \tag{40}$$

Similarly to the derivations in Appendix A.1, we can compute the elasticity of TFP with respect to shocks to firm j as

$$\eta_j^{\mathrm{SME}}(\tilde{\mathbf{x}}) = \frac{\partial \log Z_{\tilde{\mathbf{x}}}^{\mathrm{SME}}}{\partial \tilde{x}_j} = \frac{1}{Z_{\tilde{\mathbf{x}}}^{\mathrm{SME}}} \frac{e^{\tilde{x}_j} a_j^{\frac{1+\nu}{1-(\gamma-\nu)}}}{\left[\sum_{i=1}^{N} a_i^{\frac{1}{1-(\gamma-\nu)}} \right]^{\gamma}} = \frac{e^{\tilde{x}_j} a_j^{\frac{1+\nu}{1-(\gamma-\nu)}}}{\sum_{i=1}^{N} e^{\tilde{x}_i} a_i^{\frac{1+\nu}{1-(\gamma-\nu)}}}. \tag{41}$$

Repeating the approximation in equation (36) allows us to state that aggregate volatility, in this case, is given by

$$\sigma_Z^{\mathrm{SME}} = \sigma_x \times \Psi^{\mathrm{SME}}, \quad \Psi^{\mathrm{SME}} = \sqrt{\sum_{i=1}^{N} \left(\overline{\eta}_i^{\mathrm{SME}} \right)^2} \tag{42}$$

where $\overline{\eta}_j^{\text{SME}}$ again denote the values of η_j^{SME} when all the shocks $\tilde{\mathbf{x}}$ are set to their expected value.

With some abuse of notation, let $p = \frac{1+\nu}{1-(\gamma-\nu)}$ and define the auxiliary function

$$h(p) = \Psi^{\text{SME}}(p) = \sqrt{\sum_{i=1}^{N} {a_i}^{2p}} \times \left[\sum_{i=1}^{N} {a_i}^{p}\right]^{-1}. \tag{43}$$

where the efficient production benchmark in equation (7) is given by $h(p = \frac{1}{1-\gamma}) = \Psi$. To prove that stronger SME subsidies dampen aggregate volatility and large-business subsidies amplify it, it is sufficient to show that Ψ^{SME} is a decreasing function of ν or that h is an increasing function of p since we have that $\frac{dp}{d\nu} = \frac{-\gamma}{(1-(\gamma-\nu))^2} < 0$ for all values of ν obeying the regularity condition $\gamma - 1 < \nu < \gamma$. Given the definitions of γ and ν, p can take values between one at the limit where $\gamma \to 0$ and infinity when $\gamma \to 1+\nu$.[18] We now state technical claim 1.

Claim 1. *Let $\mathbf{a} \in \mathbb{R}^N_{++}$. The function: $h(p) = \frac{\|\mathbf{a}^p\|_2}{\|\mathbf{a}^p\|_1}$ where, $\mathbf{a}^p$ is the elementwise power of $\mathbf{a}$ by p and $\|\cdot\|_1$ and $\|\cdot\|_2$ denote the L^1 and L^2 norms correspondingly. Then, $h'(p) \geqslant 0, \forall p > 0$ with strict inequality if $\exists j,k \in \{1,\ldots,N\}$ for which $a_j \neq a_k$.*

The proof follows below. This concludes the proof of Proposition 2. $\square$

A.2.1 Proof of Claim 1

Proof. Since $h(p) > 0$, the chain rule gives $\frac{d}{dp}h(p)^2 = 2h(p)h'(p)$, so $h'(p)$ and $\frac{d}{dp}h(p)^2$ have the same sign. It is, therefore, sufficient to show $\frac{d}{dp}h(p)^2 \geqslant 0$. Define firm weights and log-productivities as follows:

$$w_i(p) = \frac{a_i^p}{\sum_{k=1}^{N} a_k^p}, \qquad b_i = \log a_i.$$

These satisfy $w_i(p) > 0$ and $\sum_{i=1}^{N} w_i(p) = 1$. The squared ratio $h(p)^2$ is the HHI of these weights given by:

$$h(p)^2 = \frac{\sum_{i=1}^{N} a_i^{2p}}{\left(\sum_{i=1}^{N} a_i^p\right)^2} = \sum_{i=1}^{N} w_i(p)^2.$$

[18] In the latter case, the bounds also impose that $\nu \to 0$ as $\gamma - 1 < \nu$.

It is useful to define $\mu_p = \sum_{j=1}^{N} w_j(p)b_j$ as the w-weighted mean of log-productivities b_i. Differentiating $w_i(p)$ yields $w_i'(p) = w_i(p)(b_i - \mu_p)$, so

$$\frac{d}{dp}h(p)^2 = 2\sum_{i=1}^{N} w_i(p)w_i'(p) = 2\sum_{i=1}^{N} w_i(p)^2(b_i - \mu_p).$$

Substituting the definition of μ_p and using $\sum_{j=1}^{N} w_j(p) = 1$ to write $\sum_{i=1}^{N} w_i(p)^2 b_i = \sum_{i=1}^{N}\sum_{j=1}^{N} w_i(p)^2 w_j(p)b_i$, this becomes

$$\frac{d}{dp}h(p)^2 = 2\sum_{i=1}^{N}\sum_{j=1}^{N} w_i(p)^2 w_j(p)\left(b_i - b_j\right).$$

Swapping i and j gives an equal expression $2\sum_{i=1}^{N}\sum_{j=1}^{N} w_j(p)^2 w_i(p)(b_j - b_i)$. Adding the two and dividing by 2, each (i,j) pair contributes

$$w_i(p)^2 w_j(p)(b_i - b_j) + w_j(p)^2 w_i(p)(b_j - b_i) = w_i(p)w_j(p)\big(w_i(p) - w_j(p)\big)\left(b_i - b_j\right),$$

so

$$\frac{d}{dp}h(p)^2 = \sum_{i=1}^{N}\sum_{j=1}^{N} w_i(p)w_j(p)\big(w_i(p) - w_j(p)\big)\left(b_i - b_j\right).$$

Since the denominator of $w_i(p)$ is common across firms, and $w_i(p), b_i$ are strictly increasing in a_i for any fixed $p > 0$, we have that

$$\big(w_i(p) - w_j(p)\big)\left(b_i - b_j\right) \geqslant 0 \quad \forall i, j,$$

with strict inequality whenever $a_i \neq a_j$. Since all weights are strictly positive, every term in the double sum is nonnegative, and at least one is strictly positive if there is at least one pair such that $a_i \neq a_j$. Hence $\frac{d}{dp}h(p)^2 \geqslant 0$, which gives $h'(p) \geqslant 0$, both with strict inequality unless all a_i are equal. □

A.3 Proof of Lemma 4

Proof. If b is proportional to a random variable a raised to the power α such that a is Pareto distributed with tail parameter ζ_a and $\alpha > 0$, i.e., $b \propto a^{\alpha}$, then it follows that

$$\text{Prob}(b > x) \propto \text{Prob}(a^{\alpha} > x) = \text{Prob}\left(a > x^{\frac{1}{\alpha}}\right) \propto x^{-\frac{\zeta_a}{\alpha}}. \tag{44}$$

Therefore, b is also Pareto distributed with tail parameter $\zeta_b = \frac{\zeta_a}{\alpha}$. Thus, it follows that since firm-level employment in the model when shocks are at their expected level $l_i^{d\star} \propto a_i^{\frac{1-\phi}{1-\gamma}}$ then the tail of the firm-size distribution when size is measured by employment is given by $\zeta_{\text{emp}} = \zeta_a \frac{1-\gamma}{1-\phi}$. Alternatively, when measured using sales y_i^d or the relative sizes $\overline{\delta}_i$, the relative size of firm i is such that $y_i^d, \overline{\delta}_i \propto a_i^{\frac{1-\gamma\phi}{1-\gamma}}$ and thus $\zeta_{\text{sales}} = \zeta_a \frac{1-\gamma}{1-\gamma\phi}$. This concludes the proof of Lemma 4. $\square$

A.4 Step by step derivation of the approximation in Equation (23)

Consider the following second-order Taylor series approximation for welfare, where the approximation is taken around $\overline{Z^d} = Z^d_{\tilde{\mathbf{x}}}\big|_{\tilde{\mathbf{x}}=\overline{\mathbf{x}}}$

$$\begin{aligned} u\left(Z^d_{\tilde{\mathbf{x}}}\right) &\approx u\left(\overline{Z^d}\right) + u'\left(\overline{Z^d}\right) \sum_{i=1}^{N} \frac{\partial Z^d_{\tilde{\mathbf{x}}}}{\partial \tilde{x}_i}\bigg|_{\tilde{\mathbf{x}}=\overline{\mathbf{x}}} (\tilde{x}_i - \overline{x}) \\ &+ \frac{1}{2}\sum_{j=1}^{N}\sum_{i=1}^{N}\left(u''\left(\overline{Z^d}\right) \frac{\partial Z^d_{\tilde{\mathbf{x}}}}{\partial \tilde{x}_i}\bigg|_{\tilde{\mathbf{x}}=\overline{\mathbf{x}}} \frac{\partial Z^d_{\tilde{\mathbf{x}}}}{\partial \tilde{x}_j}\bigg|_{\tilde{\mathbf{x}}=\overline{\mathbf{x}}} + u'\left(\overline{Z^d}\right) \frac{\partial^2 Z^d_{\tilde{\mathbf{x}}}}{\partial \tilde{x}_i \partial \tilde{x}_j}\bigg|_{\tilde{\mathbf{x}}=\overline{\mathbf{x}}} \right) (\tilde{x}_i - \overline{x})\left(\tilde{x}_j - \overline{x}\right) . \end{aligned} \tag{45}$$

Taking expectations around the above and exploiting the facts that $\mathbb{E}[(\tilde{x}_i - \overline{x})] = 0$, $\mathbb{E}\left[(\tilde{x}_i - \overline{x})^2\right] = {\sigma_x}^2$ and all shocks are i.i.d. thus all the covariance terms are zero yields

$$\mathbb{E}\left[u\left(Z^d_{\tilde{\mathbf{x}}}\right)\right] \approx u\left(\overline{Z^d}\right) + \frac{{\sigma_x}^2}{2}\sum_{i=1}^{N}\left(u''\left(\overline{Z^d}\right)\left[\frac{\partial Z^d_{\tilde{\mathbf{x}}}}{\partial \tilde{x}_i}\bigg|_{\tilde{\mathbf{x}}=\overline{\mathbf{x}}}\right]^2 + u'\left(\overline{Z^d}\right) \frac{\partial^2 Z^d_{\tilde{\mathbf{x}}}}{(\partial \tilde{x}_i)^2}\bigg|_{\tilde{\mathbf{x}}=\overline{\mathbf{x}}} \right) . \tag{46}$$

We can use the elasticities $\delta_i(\tilde{\mathbf{x}})$ to express the derivatives as $\frac{\partial Z^d_{\tilde{\mathbf{x}}}}{\partial \tilde{x}_i} = \delta_i(\tilde{\mathbf{x}}) Z^d_{\tilde{\mathbf{x}}}$. This expression allows one to derive that $\frac{\partial^2 Z^d_{\tilde{\mathbf{x}}}}{(\partial \tilde{x}_i)^2} = \frac{\partial \delta_i(\tilde{\mathbf{x}})}{\partial \tilde{x}_i} Z^d_{\tilde{\mathbf{x}}} + \delta_i(\tilde{\mathbf{x}}) \frac{\partial Z^d_{\tilde{\mathbf{x}}}}{\partial \tilde{x}_i}$. Additionally, one can compute $\frac{\partial \delta_i(\tilde{\mathbf{x}})}{\partial \tilde{x}_i}$ and obtain that $\frac{\partial^2 Z^d_{\tilde{\mathbf{x}}}}{(\partial \tilde{x}_i)^2} = \delta_i(\tilde{\mathbf{x}}) Z^d_{\tilde{\mathbf{x}}}$.[19] We can combine those derivatives and the defi-

[19] This can be shown directly from

$$\frac{\partial \delta_j(\tilde{\mathbf{x}})}{\partial \tilde{x}_j} = \frac{a_j^{\frac{1}{1-\gamma}}(1-\tau_j)^{\frac{\gamma}{1-\gamma}} e^{\tilde{x}_j}\left[\sum_{i=1}^N a_i^{\frac{1}{1-\gamma}}(1-\tau_i)^{\frac{\gamma}{1-\gamma}} e^{\tilde{x}_i}\right] - \left[a_j^{\frac{1}{1-\gamma}}(1-\tau_j)^{\frac{\gamma}{1-\gamma}} e^{\tilde{x}_j}\right]^2}{\left[\sum_{i=1}^N a_i^{\frac{1}{1-\gamma}}(1-\tau_i)^{\frac{\gamma}{1-\gamma}} e^{\tilde{x}_i}\right]^2} =$$
$$\frac{\left(a_j^{\frac{1}{1-\gamma}}(1-\tau_j)^{\frac{\gamma}{1-\gamma}} e^{\tilde{x}_j}\right)}{\left[\sum_{i=1}^N a_i^{\frac{1}{1-\gamma}}(1-\tau_i)^{\frac{\gamma}{1-\gamma}} e^{\tilde{x}_i}\right]} \frac{\left[\sum_{i=1}^N a_i^{\frac{1}{1-\gamma}}(1-\tau_i)^{\frac{\gamma}{1-\gamma}} e^{\tilde{x}_i}\right] - \left[a_j^{\frac{1}{1-\gamma}}(1-\tau_j)^{\frac{\gamma}{1-\gamma}} e^{\tilde{x}_j}\right]}{\left[\sum_{i=1}^N a_i^{\frac{1}{1-\gamma}}(1-\tau_i)^{\frac{\gamma}{1-\gamma}} e^{\tilde{x}_i}\right]} = \delta_j(\tilde{\mathbf{x}})(1-\delta_j(\tilde{\mathbf{x}})),$$

nition of Ψ^d to obtain that

$$\mathbb{E}\left[u\left(Z^d_{\tilde{\mathbf{x}}}\right)\right] = u\left(\overline{Z^d}\right) + \frac{1}{2}\left(\Psi^d \sigma_x\right)^2 \overline{Z^d}^2 u''\left(\overline{Z^d}\right) + \frac{{\sigma_x}^2}{2}\overline{Z^d}\, u'\left(\overline{Z^d}\right). \tag{47}$$

Finally, by exploiting the CRRA utility specification, i.e., $\chi = -\frac{\overline{Z^d}\, u''\left(\overline{Z^d}\right)}{u'\left(\overline{Z^d}\right)}$, we have that

$$\mathbb{E}\left[u\left(Z^d_{\tilde{\mathbf{x}}}\right)\right] = u\left(\overline{Z^d}\right) + \frac{{\sigma_x}^2}{2}\overline{Z^d}\, u'\left(\overline{Z^d}\right) - \chi\frac{1}{2}\left(\Psi^d \sigma_x\right)^2 \overline{Z^d}\, u'\left(\overline{Z^d}\right). \tag{48}$$

Log utility case. When utility takes log form we have that $\chi = 1$ and $\overline{Z^d}\, u'\left(\overline{Z^d}\right) = 1$ thus,

$$\mathbb{E}\left[u\left(Z^d_{\tilde{\mathbf{x}}}\right)\right] = u\left(\overline{Z^d}\right) + \frac{{\sigma_x}^2}{2} - \frac{1}{2}\left(\Psi^d \sigma_x\right)^2, \tag{49}$$

whereas in the case without misallocation, one can analogously derive that

$$\mathbb{E}\left[u\left(Z_{\tilde{\mathbf{x}}}\right)\right] = u\left(\overline{Z}\right) + \frac{{\sigma_x}^2}{2} - \frac{1}{2}\left(\Psi \sigma_x\right)^2. \tag{50}$$

Equation (24) is the result of subtracting the two.

A.5 Proof of Proposition 3 and Corollary 1

Proof. Solving the planner's problem in equation (25) implies that the allocation rule for labor must satisfy the first order condition in equation (26) such that

$$\gamma {l_i}^{\gamma-1}\mathbb{E}\left[u'\left(Y_{\tilde{\mathbf{x}}}\right)\tilde{z}_i\right] = \lambda. \tag{51}$$

For the production economy described in Lemma 2, the first order condition of the firm's profit maximization problem is

$$\mathbb{E}[\tilde{z}_i]\gamma {l_i}^{\gamma-1}(1-\tau_i) = w. \tag{52}$$

Thus, to implement the optimal allocation solving the planner's problem, the distortion τ_i must be set such that $(1-\tau_i) = \kappa\frac{\mathbb{E}(u'(Y_{\tilde{\mathbf{x}}})\tilde{z}_i)}{\mathbb{E}[\tilde{z}_i]}$, with κ being a positive constant. Dividing the

which implies $\frac{\partial^2 Z^d_{\tilde{\mathbf{x}}}}{(\partial \tilde{x}_i)^2} = \frac{\partial \delta_i(\tilde{\mathbf{x}})}{\partial \tilde{x}_i} \times Z^d_{\tilde{\mathbf{x}}} + \delta_i(\tilde{\mathbf{x}})\frac{\partial Z^d_{\tilde{\mathbf{x}}}}{\partial \tilde{x}_i} = \frac{\partial \delta_i(\tilde{\mathbf{x}})}{\partial \tilde{x}_i} \times Z^d_{\tilde{\mathbf{x}}} + \delta_i(\tilde{\mathbf{x}})^2 \times Z^d_{\tilde{\mathbf{x}}} = Z^d_{\tilde{\mathbf{x}}} \times \left[\delta_i(\tilde{\mathbf{x}})(1-\delta_i(\tilde{\mathbf{x}})) + \delta_i(\tilde{\mathbf{x}})^2\right] = \delta_i(\tilde{\mathbf{x}}) Z^d_{\tilde{\mathbf{x}}}$.

distortions of firm i and j yields the expression in equation (27), thus concluding the proof of the first part of the proposition.

The exact wedge in equation (28) is set by a normalization. Such that $(1-\tau_i)=1$ when $u'(Y_{\tilde{\mathbf{x}}}) \perp \tilde{z}_i$. This normalization is obtained by setting $(1-\tau_i)=\frac{\mathbb{E}[u'(Y_{\tilde{\mathbf{x}}})\tilde{z}_i]}{\mathbb{E}[u'(Y_{\tilde{\mathbf{x}}})]\mathbb{E}[\tilde{z}_i]}$. Note that the above expression can be restated as

$$1-\tau_i=\frac{\mathbb{E}[u'(Y_{\tilde{\mathbf{x}}})]\mathbb{E}[\tilde{z}_i]+\mathbb{COV}(u'(Y_{\tilde{\mathbf{x}}}),\tilde{z}_i)}{\mathbb{E}[u'(Y_{\tilde{\mathbf{x}}})]\mathbb{E}[\tilde{z}_i]}, \tag{53}$$

such that $(1-\tau_i)=1$ when $\mathbb{COV}(u'(Y_{\tilde{\mathbf{x}}}),\tilde{z}_i)=0$ or equivalently when $u'(Y_{\tilde{\mathbf{x}}}) \perp \tilde{z}_i$. This concludes the proof of the second part of the proposition.

To prove Corollary 1, we can leverage the following approximation for CRRA utility around the point $\overline{\mathbf{x}}$. To better connect with Lemma 2, I now specify that the resulting output is $Y^d_{\tilde{\mathbf{x}}}$.

$$\begin{aligned}u'\left(Y^d_{\tilde{\mathbf{x}}}\right) &\approx u'\left(Y^d_{\overline{\mathbf{x}}}\right)+u''\left(Y^d_{\overline{\mathbf{x}}}\right)\sum_{i=1}^{N}\left.\frac{\partial Y^d_{\tilde{\mathbf{x}}}}{\partial \tilde{x}_i}\right|_{\tilde{\mathbf{x}}=\overline{\mathbf{x}}}(\tilde{x}_i-\overline{x}) = u'\left(Y^d_{\overline{\mathbf{x}}}\right)\left[1-\chi\frac{1}{Y^d_{\overline{\mathbf{x}}}}\sum_{i=1}^{N}\left.\frac{\partial Y^d_{\tilde{\mathbf{x}}}}{\partial \tilde{x}_i}\right|_{\tilde{\mathbf{x}}=\overline{\mathbf{x}}}(\tilde{x}_i-\overline{x})\right]\\ &= u'\left(Y^d_{\overline{\mathbf{x}}}\right)\left[1-\chi\sum_{i=1}^{N}\left.\frac{\partial \log Y^d_{\tilde{\mathbf{x}}}}{\partial \tilde{x}_i}\right|_{\tilde{\mathbf{x}}=\overline{\mathbf{x}}}(\tilde{x}_i-\overline{x})\right] = u'\left(Y^d_{\overline{\mathbf{x}}}\right)\left[1-\chi\sum_{i=1}^{N}\overline{\delta}_i(\tilde{x}_i-\overline{x})\right].\end{aligned} \tag{54}$$

This approximation comes from a first-order Taylor expansion around $\overline{\mathbf{x}}$. The first equality sign leverages the definition of CRRA utility as $\chi=-\frac{u''\left(Y^d_{\overline{\mathbf{x}}}\right)}{u'\left(Y^d_{\overline{\mathbf{x}}}\right)}Y^d_{\overline{\mathbf{x}}}$, the second follows from applying the chain rule such that $\frac{1}{Y^d_{\overline{\mathbf{x}}}}\frac{\partial Y^d_{\tilde{\mathbf{x}}}}{\partial \tilde{x}_i}=\frac{\partial \log Y^d_{\tilde{\mathbf{x}}}}{\partial \tilde{x}_i}$, and the third uses the definition of $\overline{\delta}_i=\left.\frac{\partial \log Y^d_{\tilde{\mathbf{x}}}}{\partial \tilde{x}_i}\right|_{\tilde{\mathbf{x}}=\overline{\mathbf{x}}}$. Note also that for small shocks we can approximate $\tilde{z}_i \approx a_i(1+\tilde{x}_i)$ and thus the covariance term in equation (28) can be approximated by

$$\mathbb{COV}\left(u'\left(Y^d_{\tilde{\mathbf{x}}}\right),\tilde{z}_i\right)\approx\mathbb{COV}\left({Y^d_{\overline{\mathbf{x}}}}^{-\chi}\left(1-\chi\sum_{j=1}^{N}\overline{\delta}_j(\tilde{x}_j-\overline{x})\right),a_i(1+\tilde{x}_i)\right)=-\chi{Y^d_{\overline{\mathbf{x}}}}^{-\chi}a_i\overline{\delta}_i{\sigma_x}^2.$$

Exploiting, $\mathbb{E}[\tilde{z}_i]=a_i$ and $\mathbb{E}\left[u'\left(Y^d_{\overline{\mathbf{x}}}\right)\right]={Y^d_{\overline{\mathbf{x}}}}^{-\chi}$, allows us to derive

$$1-\tau_i=1+\frac{\mathbb{COV}\left(u'\left(Y^d_{\tilde{\mathbf{x}}}\right),\tilde{z}_i\right)}{\mathbb{E}\left[u'\left(Y^d_{\tilde{\mathbf{x}}}\right)\right]\mathbb{E}[\tilde{z}_i]}\approx 1-\chi\overline{\delta}_i{\sigma_x}^2,$$

or that $\tau_i \approx \chi\overline{\delta}_i{\sigma_x}^2$. Finally, to obtain a more meaningful expression, we can, with some abuse of notation, write $\overline{\delta}_i(\tau)$ and take a first-order Taylor series approximation of $\overline{\delta}_i$ at $\tau = \mathbf{0}$ as

$$\overline{\delta}_i(\tau) \approx \overline{\delta}_i(\mathbf{0}) + \sum_{j=1}^{N} \left. \frac{\partial \overline{\delta}_i(\tau)}{\partial \tau_j} \right|_{\tau=\mathbf{0}} \tau_j = \overline{\eta}_i + \sum_{j=1}^{N} \left. \frac{\partial \overline{\delta}_i(\tau)}{\partial \tau_j} \right|_{\tau=\mathbf{0}} \tau_j = \overline{\eta}_i + \mathscr{O}({\sigma_x}^2), \tag{55}$$

where the first equality sign comes from noting that when $\tau = \mathbf{0}$ we have $\overline{\delta}_i = \overline{\eta}_i$, and the second from noting that $\frac{\partial \overline{\delta}_i(\tau)}{\partial \tau_j}$ does not depend on the volatility of firm-level shocks, or that it is $\mathscr{O}(1)$ with respect to σ_x and that $\tau_i \approx \chi \overline{\delta}_i {\sigma_x}^2$ which is $\mathscr{O}({\sigma_x}^2)$. Therefore, we have that

$$\tau_i \approx \chi \overline{\delta}_i {\sigma_x}^2 \approx \chi \overline{\eta}_i {\sigma_x}^2 + \mathscr{O}({\sigma_x}^4), \tag{56}$$

or that to a first order, $1 - \tau_i \approx 1 - \chi \overline{\eta}_i {\sigma_x}^2$, which is the expression in Corollary 1, thus concluding the proof. $\square$

Supplemental Appendices

Appendix B The Stabilizing or Destabilizing Role of Specific Frictions

Two compelling frictions often described in the misallocation literature are market power and financial frictions. These two can potentially affect aggregate volatility via the mechanism described in Proposition 1, so the question is: do these frictions imply that distortions are positively or negatively correlated with productivity?

Market Power. Consider, for example, the endogenous-markup model of Atkeson and Burstein (2008). In this type of model, more productive firms within a sector can charge higher markups and are under-producing compared to an efficient production benchmark. This would manifest as a higher τ_i since the markup $\mu_i = \frac{1}{1-\tau_i}$.[20] In particular Burstein et al. (2025), show that μ_i is monotonically increasing in the sales share of firm i in its sector. Therefore, if all markup dispersion is within sector, the markup is increasing in productivity, as more productive firms could capture a higher market share. If sectoral components also drive the distribution of markups, it would still be true that more productive firms within a sector have higher markups, but the correlation between markups and productivity would be attenuated. Thus, it is possible that market power generates positively correlated distortions, but that the strength of the correlation depends on the exact structure of the economy.

Financial Frictions. For the case of financial frictions, the issue is even more nuanced and requires taking a stance on the exact form of the frictions. In the conventional collateral constraint approach à la Kiyotaki and Moore (1997), as applied in Buera and Shin (2013), and Moll (2014), financial frictions limit the amount of inputs a producer can use as a function of their pledgeable assets. The higher their production ability, the more factors they wish to employ, and thus, the higher the implicit tax imposed on them by the financial friction. To formally argue this point, let us consider a simple version of the model with the following firm problem:

$$\max_{l_i} \quad \mathbb{E}[y_i] - w \times l_i, \quad \text{s.t.} \quad A_i \lambda_c \geqslant w l_i, \tag{57}$$

where A_i is the amount of pledgeable assets of firm i, and λ_c denotes the strength of the

[20] For more on this logic connecting misallocation and markups, see Edmond et al. (2015).

collateral constraint. The above problem is well known in the literature, and the solution is given by $l_i^\star = \min\left\{\left(\frac{\gamma a_i}{w}\right)^{\frac{1}{1-\gamma}}, \frac{A_i\lambda_c}{w}\right\}$. Therefore, if the collateral constraint binds for firm i, we have that $l_i^\star = \frac{A_i\lambda_c}{w}$, which is independent of the firm's productivity. We can construct an implied wedge $1-\tau_i(a_i,A_i)$ such that the wedge leads a firm solving the unconstrained problem to set the same $l_i^\star = \left(a_i(1-\tau(a_i,A_i))\frac{\gamma}{w}\right)^{\frac{1}{1-\gamma}}$. The wedge is given by

$$(1-\tau(a_i,A_i)) = \begin{cases} \frac{1}{a_i}\frac{w}{\gamma}\left(\frac{\lambda_c A_i}{w}\right)^{1-\gamma} & \lambda_c A_i < w\left(a_i\frac{\gamma}{w}\right)^{\frac{1}{1-\gamma}}, \\ 1 & \text{otherwise.} \end{cases} \tag{58}$$

Note that when the constraint binds, the wedge is such that $(1-\tau(a_i,A_i)) \propto \frac{A_i^{1-\gamma}}{a_i}$, which is increasing in A_i and decreasing in a_i. Thus, a financial friction of this form generates a distortion that is positively correlated with productivity since high a_i implies a higher τ_i for constrained firms as long as the constraint binds. This correlation might be confounded by potential correlation between pledgeable assets A_i and productivity a_i. If A_i and a_i are positively correlated, the correlation between the distortion and productivity is attenuated. If they are negatively correlated, the correlation between the distortion and productivity is amplified.

It is important to stress that the above result is specific to the collateral constraint approach to modeling financial frictions. Alternative approaches might generate different correlations between distortions and productivity. For example, consider earnings-based constraints as in the recent works of Drechsel (2023) and Lian and Ma (2021); in those models, the firm problem is

$$\max_{l_i} \quad \mathbb{E}[y_i] - w \times l_i, \quad \text{s.t.} \quad wl_i \leqslant \lambda_e \mathbb{E}[y_i], \tag{59}$$

where λ_e denotes the strength of the earnings-based constraint. It is straightforward to show that when $\gamma > \lambda_e$ the constraint binds for all firms, labor demand is given by $l_i^\star = \left(\frac{\lambda_e}{w}a_i\right)^{\frac{1}{1-\gamma}}$ which implies a uniform wedge across firms. Therefore, in this case, financial frictions do not generate any correlation between distortions and productivity.

Throughout the discussion of financial frictions, I have assumed that the constraint is uniform across all firms, but one can also consider scenarios in which the strength of the friction is heterogeneous across firms and correlated with other primitives. Such a scenario would further complicate the analysis of the correlation between distortions and productivity, and thus the question of whether financial frictions generate positively or negatively correlated distortions.

To conclude, this discussion suggests that financial frictions and market power can generate positively correlated distortions. Still, the strength and the direction of the correlation depend on the exact structure of the economy and the modeling choices. One can model markups differently from Atkeson and Burstein (2008), and one can model financial frictions differently from the two above approaches. Ultimately, the question of whether these frictions generate positively or negatively correlated distortions, and thus reduce or amplify firm-level shocks, is an empirical one. It is sensitive to the exact structure of the economy and the modeling choices. This is an important question for future research.

Appendix C Validation of Approximation Strategy

This appendix provides a conceptual and quantitative validation of the approximation strategy for the amplification term Ψ^d given in Lemma 3. The approximation is based on the idea that the expected value of a function of a random sample can be approximated by the function evaluated at the expected value of the order statistics. In what follows, I first derive a closed-form expression for the expected amplification term $\overline{\Psi}^d$ when the distribution of firm productivities is Pareto, demonstrating that the approximation maintains the asymptotic properties derived in the seminal work of Gabaix (2011). I then proceed to validate the approximation strategy quantitatively by comparing the magnitude of aggregate volatility from firm-level shocks this approximation implies to estimates in the existing literature on granular business cycles.

C.1 Approximating the Amplification Term When Firm-Level Productivity is Pareto

Combining the result from Lemma 3 with the fact that the quantile function of a Pareto distribution with tail parameter ζ_a is given by $G^{-1}(q)$ or $Q(q) = (1-q)^{-\frac{1}{\zeta_a}}$, we can obtain the following approximation for the expected amplification term in an economy populated with N firms:

Lemma 5. *If a_i is Pareto distributed, and distortions are modeled as in equation (21), the* *expected amplification term $\overline{\Psi}^d$ can be approximated as*

$$\overline{\Psi}^d \approx \sqrt{\sum_{i=1}^{N} i^{-\frac{2}{\zeta_{sales}}}} \times \left[\sum_{i=1}^{N} i^{-\frac{1}{\zeta_{sales}}}\right]^{-1}. \tag{60}$$

For proof, see Appendix C.3. Appendix C.3 also proves the following corollary relating the asymptotic properties of this approximation to Proposition 2 in Gabaix (2011), which serves as an important conceptual validation

Corollary 2. *If a_i is Pareto distributed, then as N grows, aggregate volatility decays as follows*

(a) $\sigma_Z \sim 1/constant \quad for \quad \zeta_{sales} < 1.$

(b) $\sigma_Z \sim 1/\log(N) \quad for \quad \zeta_{sales} = 1.$

(c) $\sigma_Z \sim 1/N^{1-\frac{1}{\zeta_{sales}}} \quad for \quad 1 < \zeta_{sales} < 2.$

(d) $\sigma_Z \sim 1/\sqrt{N} \quad for \quad \zeta_{sales} \geqslant 2.$

This asymptotic behavior of TFP volatility in cases (2) through (4) replicates the asymptotic decay properties for TFP volatility given in Proposition 2 of Gabaix (2011), whereas case (1) is new to the best of my knowledge. That is, if the tail of the firm-size distribution is $\zeta_{\text{sales}} < 1$, volatility due to firm-level shocks does not vanish even if the number of firms goes to infinity. Now with this approximation and insights at hand, I turn to validating the approximation against estimates in the existing literature on granular business cycles.

C.2 Quantitative Validation: Aggregate Volatility from Granular Shocks

The above approximation strategy allows one to ask, how much TFP volatility can we attribute to firm-level shocks? For this section only, I focus now on the case without distortions since this is the one considered by existing works on granular business cycles. Note that, abstracting from distortions, the firm-size distribution has an observed tail parameter $\zeta_{\text{emp}} = \zeta_{\text{sales}} = \zeta$. Equation (7) states that aggregate volatility is a function of: the micro-level volatility σ_x, the number of firms N, and the distribution of relative firm-sizes that has a Pareto tail ζ governed by the tail of the productivity distribution. The relationships between σ_x,N, and ζ are illustrated by Figure 3 which demonstrates how aggregate volatility σ_Z changes for given values of σ_x, N, and ζ. Panel (a) shows that aggregate volatility

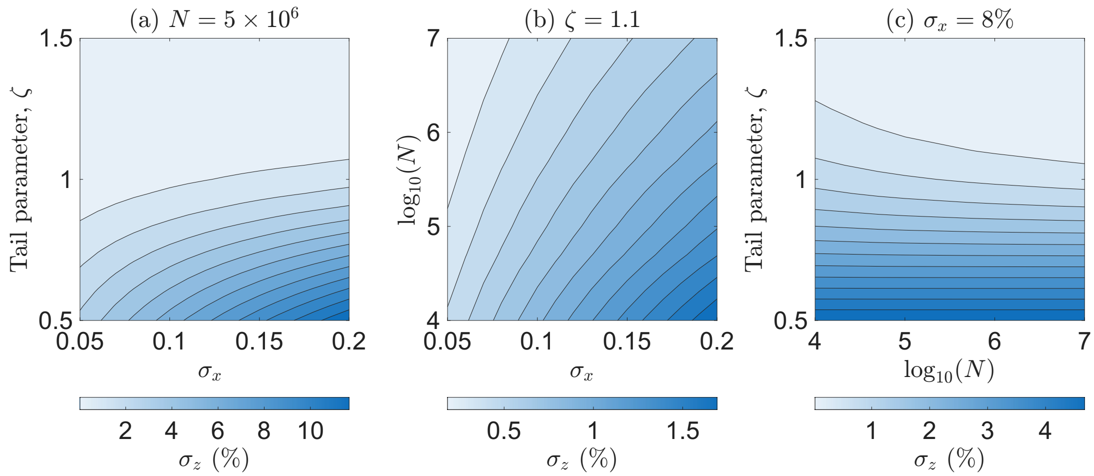

Figure 3: Aggregate Volatility Due to Granular Shocks

Note: Each panel reports the implied values of aggregate volatility $\sigma_Z = \overline{\Psi} \times \sigma_x$ as a function of the parameters on the axes, holding the parameter in the title fixed.

increases when the firm-size distribution is more fat-tailed, i.e., when ζ is lower, and when the firm-level shocks are stronger. Panel (b) reports that holding the tail parameter constant at $\zeta = 1.1$, volatility due to firm-level shocks increases in σ_x and decays as the number of firms grows. Panel (c) fixes σ_x and demonstrates that, as indicated by Corollary 2, volatility does not decay when the tail parameter is small enough.

Validation. To validate the approximation strategy, I utilize estimates from two existing works that employ different approaches to quantify the magnitude of granular business cycles: Gabaix (2011) and Carvalho and Grassi (2019). Results from this exercise are summarized in Table A1.

In his seminal paper, Gabaix cites estimates of the square root of the economy-level HHI (Ψ) in the U.S. as 5.3% and cites firm-level volatility of $\sigma_x = 12\%$. His estimate implies that volatility due to such shocks can account for 0.63%. Using Gabaix's HHI as an input, I use equation (60) to infer that $\zeta = 1.06$, which is precisely the estimate of Axtell (2001). Thus, given an HHI, the approximation is consistent with the observed tail of the firm-size distribution.

Next, the work of Carvalho and Grassi (2019) discusses the business cycle properties of a granular economy, somewhat similar to the one described in the present paper but abstracting from deviations from production efficiency. Carvalho and Grassi (2019) leverage properties of multinomial distributions and simulate an economy with $N = 4.5 \times 10^6$ firms along a productivity grid consisting of 36 distinct bins, thereby approximating a Pareto

Table A1: Validation of Approximation Strategy Against Existing Results

	Paper	Cited estimate σ_Z	Computed σ_Z	σ_x	ζ	Ψ
(1)	Gabaix (2011)	0.630%	0.636%	12% input	1.061 inferred	0.053 input
(2)	Carvalho and Grassi (2019)	0.250%	0.329%	8% input	1.097 input	0.041 inferred

Note: This table summarizes the validation checks reported in this section, comparing the estimated aggregate volatility from Gabaix (2011) and Carvalho and Grassi (2019) using the approximation strategy in equation (60) with $\gamma = 0.8$ and $N = 4.5 \times 10^6$.

distribution with a discrete analog. The authors provide a survey of estimates for σ_x, the firm-level Solow residual in the literature, citing values between 0.09 and 0.2, using a lower-bound estimate of 0.08 as their benchmark. Combined with a tail parameter estimate of $\zeta = 1.097$, they estimate a TFP volatility of 0.25% due to firm-level shocks alone, compared with the total U.S. TFP volatility of 1%.

Using the parameterization of Carvalho and Grassi (2019) and my method, I compute an implied TFP volatility of 0.33%.[21] The two estimates are quite close, and the discrepancy between them is likely due to the bin-based approximation strategy of Carvalho and Grassi (2019), which necessitates truncation of the support, thereby making the large firms slightly too small and eliminating the variance in firm-size within the very top of the firm-size distribution. My strategy avoids this truncation and yields a slightly higher estimate. Furthermore, my approach involves a significantly lower computational burden, requiring only a fraction of a second to compute on a desktop machine.

C.3 Proof of Lemma 5

Proof. The amplification term as given by the approximation in equation (20), along with the definition of $\overline{q}_i = \mathbb{E}[q_i] = \frac{N+1-i}{N+1}$, and the relationship $a_i = Q(q_i)$ allows one to derive

[21] This implied volatility of 0.33% computed here differs from the headline estimate of 0.38% obtained in the main text due to the fact that having distortions in the economy leads to that if two economies are calibrated using the same ζ_{emp}, the economy with distortions will have a lower value of ζ_{sales} and thus a higher aggregate volatility for given values of σ_x and N.

that given a Pareto ability distribution we have that

$$\overline{\Psi}^d = \sqrt{\sum_{i=1}^{N} \overline{\delta}_i^2} = \frac{\sqrt{\sum_{i=1}^{N} \left(a_i^{\frac{1}{1-\gamma}}(1-\tau_i)^{\frac{\gamma}{1-\gamma}}\right)^2}}{\sum_{i=1}^{N} a_i^{\frac{1}{1-\gamma}}(1-\tau_i)^{\frac{\gamma}{1-\gamma}}} = \frac{\sqrt{\sum_{i=1}^{N} \left(a_i^{\frac{1}{1-\gamma}}(a_i^{-\phi})^{\frac{\gamma}{1-\gamma}}\right)^2}}{\sum_{i=1}^{N} a_i^{\frac{1}{1-\gamma}}(a_i^{-\phi})^{\frac{\gamma}{1-\gamma}}} \tag{61}$$

$$= \frac{\sqrt{\sum_{i=1}^{N} (Q(q_i))^{2\frac{1-\phi\gamma}{1-\gamma}}}}{\sum_{i=1}^{N} (Q(q_i))^{\frac{1-\phi\gamma}{1-\gamma}}} = \frac{\sqrt{\sum_{i=1}^{N} \left(1-\frac{N+1-i}{N+1}\right)^{-\frac{2}{\zeta_a}\frac{1-\phi\gamma}{1-\gamma}}}}{\sum_{i=1}^{N} \left(1-\frac{N+1-i}{N+1}\right)^{-\frac{1}{\zeta_a}\frac{1-\phi\gamma}{1-\gamma}}} = \frac{\sqrt{\sum_{i=1}^{N} i^{-\frac{2}{\zeta_{\text{sales}}}}}}{\sum_{i=1}^{N} i^{-\frac{1}{\zeta_{\text{sales}}}}}.$$

Where the first equality sign is due to equation (12), the second from the definition of $\overline{\delta}_j = \frac{a_j^{\frac{1}{1-\gamma}}(1-\tau_j)^{\frac{\gamma}{1-\gamma}}}{\sum_{i=1}^{N} a_i^{\frac{1}{1-\gamma}}(1-\tau_i)^{\frac{\gamma}{1-\gamma}}}$, the third from substituting in equation (21), the fourth from plugging in the relationship $a_i = Q(q_i)$, the fifth from exploiting that $\overline{q}_i = \mathbb{E}[q_i] = \frac{N+1-i}{N+1}$ and the last from noting that $\zeta_{\text{sales}} = \zeta_a \frac{1-\gamma}{1-\gamma\phi}$. This concludes the proof of Lemma 5.

To obtain Corollary 2, we need to understand the limit behavior of $\overline{\Psi}^d$ as $N \to \infty$. For that purpose, it is worthwhile introducing the following notations. Let $\zeta_{\mathbf{R}}(p) = \sum_{i=1}^{\infty} i^{-p}$ denote the Riemann zeta function. I restrict attention to real values of $p > 0$. When $p > 1$ this function is bounded, and represents a convergent series. For $p = 1$, this function is divergent and corresponds to the harmonic sum which diverges at the same rate as $\log(N)$ since its limit behavior is given by $\gamma_{EM} + \log(N)$, with $\gamma_{EM} \approx 0.577$ being the Euler Mascheroni constant.[22] For values of $p < 1$ we can observe that the series increases polynomially in N since $\sum_{i=1}^{N} \frac{1}{i^p} \approx \int_1^N x^{-p} dx = \frac{N^{1-p}-1}{1-p}$. Now, observe that at the limit we have that

$$\lim_{N\to\infty} \overline{\Psi}^d = \sqrt{\zeta_{\mathbf{R}}\left(\frac{2}{\zeta_{\text{sales}}}\right)} \times \left[\zeta_{\mathbf{R}}\left(\frac{1}{\zeta_{\text{sales}}}\right)\right]^{-1}, \tag{62}$$

thus, the approximate amplification term $\overline{\Psi}^d$ can be described as the ratio of two Riemann zeta functions. The above expression can be analyzed in cases as $N \to \infty$ depending on the value of ζ_{sales} as follows:

1. When $\zeta_{\text{sales}} < 1$ both expressions converge to constants since $\frac{1}{\zeta_{\text{sales}}} > 1$.
2. When $\zeta_{\text{sales}} = 1$ the numerator converges to a constant since $\frac{2}{\zeta_{\text{sales}}} = 2$, but the denominator is given by the harmonic sum such that asymptotically $\overline{\Psi}^d \sim \frac{\text{Constant}}{\log(N)}$.

[22] The constant is in fact defined by the limit of the difference between the harmonic sum up to N and $\log(N)$ as $N \to \infty$.

3. When $1 < \zeta_{\text{sales}} < 2$, the numerator still converges to a constant since $\frac{2}{\zeta_{\text{sales}}} > 1$, but the denominator diverges and its limit behavior is given by $\overline{\Psi}^d \sim \text{Constant} \times N^{-\left(1-\frac{1}{\zeta_{\text{sales}}}\right)}$.

4. When $\zeta_{\text{sales}} \geqslant 2$, the numerator and denominator diverge, with the former behaving as $\sqrt{N^{1-\frac{2}{\zeta_{\text{sales}}}}}$ and the latter as $N^{1-\frac{1}{\zeta_{\text{sales}}}}$, thus the ratio exhibits decay at a rate of $\frac{1}{\sqrt{N}}$.

□

Appendix D Extensions

The analysis presented in the main text is centered around a model of a vertical economy, in which shocks to all firms have identical variance, and factors are allocated before the shock is realized. The fact that Hulten's theorem holds in both the efficient production case in Lemma 1 and the distorted case in Lemma 2 guarantees that, to a first order, positively correlated distortions dampen aggregate volatility in economies with arbitrary production networks.[23] However, the other two limitations, namely, equal variance and no reallocation ex-post, merit a more systematic analysis. The tools developed in the main text facilitate an analysis even in cases where clear analytical results are not possible to derive.

D.1 Size-Volatility Relationship

The baseline model assumes that firm-level volatility, σ_x, is independent of firm size (Gibrat's law). However, the literature has documented systematic deviations from Gibrat's law, especially for young and small firms (e.g., Evans (1987), Haltiwanger et al. (2013)), suggesting that young and small firms are more volatile. Theoretically, one can think of such deviations as arising from a frictional view of firm dynamics, in which firms find it difficult to attain their optimal size quickly. These can result from adjustment costs to capital and labor or from financial frictions as proposed by Cooley and Quadrini (2001) and Clementi and Hopenhayn (2006). Alternatively, such an outcome might emerge from within-firm diversification, as in the model of Koren and Tenreyro (2013).

For the present work, it is more important to describe this statistical regularity and determine whether it alters the model's prediction of the link between correlated distortions and aggregate volatility. Works such as Stanley et al. (1996), Amaral et al. (1997), Sutton

[23] For an analogous argument, see Gabaix (2011), and for a discussion of the limitations of first-order approximations, see Baqaee and Farhi (2019).

(2002), Koren and Tenreyro (2013), and most recently Yeh (2023) measure the elasticity between volatility and size, or estimate α in the following functional form $\sigma_x(\text{Size}) = \sigma_0 \times \text{Size}^{-\alpha}$. All of these works find positive values of α, i.e., that firm size is negatively correlated with firm volatility, with headline estimates ranging from $\alpha = 0.1$ to $\alpha = 0.25$. Yeh (2023) goes even further and compares this constant-elasticity specification to a non-parametric smoother, and finds that the constant-elasticity assumption performs remarkably well. The author also conducts a structural break analysis for the size volatility relationship and finds no evidence of breaks.

To connect this literature to my analysis, I now develop a modified formula for Ψ, adjusted to allow for $\alpha > 0$, and use estimates from the literature to repeat the analysis in Figure 2. I allow for a size-volatility relationship expressed in terms of sales such that $\sigma_x(\bar{y}_j) = \sigma_0 \times \bar{y}_j^{-\alpha}$.[24] I use the parameterization which includes misallocation for generality and to map empirical estimates, based on observed firm size, more easily.

Lemma 6. *Aggregate volatility in an economy with size-volatility relationship such that* $\sigma_x(\bar{y}_j) = \sigma_0 \times \bar{y}_j^{-\alpha}$ *is approximately given by*

$$\sigma_Z^d \approx \frac{\sigma_0}{\overline{Y^d}^{\alpha}} \times \sqrt{\sum_{i=1}^{N} \bar{\delta}_i^{2(1-\alpha)}} = \frac{\sigma_0}{\overline{Y^d}^{\alpha}} \times \Psi^d(\alpha), \tag{63}$$

where $\overline{Y^d} = Y^d_{\tilde{\mathbf{x}}}\big|_{\tilde{\mathbf{x}}=\bar{\mathbf{x}}}$.

Proof in Appendix D.3.1. When $\alpha = 0$ this lemma collapses into the structure given by Lemma 2, equation (12), with $\Psi^d = \Psi^d(\alpha = 0)$. Note that $\Psi^d(\alpha)$ is no longer the HHI but an adjusted measure in which larger firms also have lower volatility and are thus weighted down. Additionally, the volatility is decaying in total output in the case where all shocks are at their expected level $\overline{Y^d}$, which affects the asymptotic properties for large N in this case, when compared with the baseline.

I follow the calibration strategy of Yeh (2023), and target an *average volatility* of firm-level Solow residual of 12% for different values of α, within the empirically likely range. The technical procedure to conduct this calibration exercise is specified in detail in Appendix D.3.1. Results are reported in Figure 4 for values of α ranging between 0.1 and 0.25. All else equal, the level of volatility attributable to firm-level shocks is lower the

[24] Given the repeated static nature of my model, I simplify the definition of size to the expected size when all shocks are switched off, to make the model tractable.

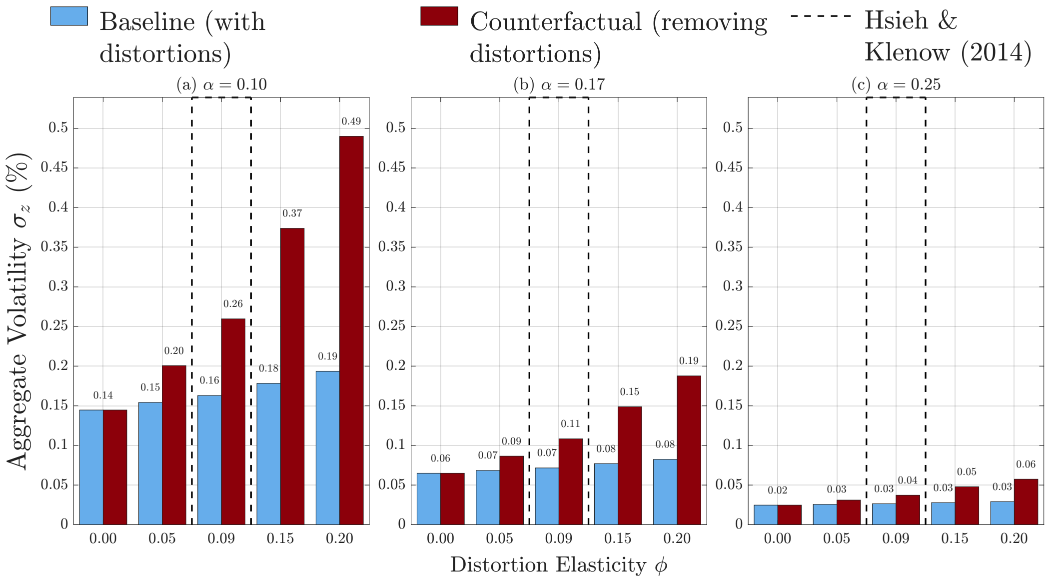

Figure 4: The Effect of Correlated Distortions on Aggregate Volatility Due to Granular Shocks Allowing for a Size-Volatility Relationship

Note: This figure compares aggregate volatility due to firm-level shocks (while allowing for deviations from Gibrat's law) with positively correlated distortions (the benchmark) with the counterfactual volatility measure computed without distortions, under different assumptions on α and ϕ. All computations are done using $\gamma = 0.8$ and $N = 4.5 \times 10^6$ while targeting $\zeta_{\text{emp}} = 1.097$.

higher the value of α. However, across all cases considered, the presence of positively correlated distortions in the benchmark stabilizes volatility relative to the efficient production counterfactual. The increase in volatility is monotonically increasing in the strength of the distortions. The basic intuition behind this result is that as long as the elasticity α is strictly below unity, the empirically likely case, as a large firm grows, in relative terms, by one percent, the economy is more exposed to its idiosyncratic shock by $1-\alpha$ percent, and is thus less diversified. It is possible that removing the distortions raises output so much that volatility decreases, since output, given by $\overline{Y^d}$ in equation (63), increases. However, for this effect to be significant, one must have that both α is large and the distortions themselves are severe. This effect is present but not dominant in all examined scenarios reported in Figure 4.

D.2 Ex-post Reallocation

An assumption underlying the analysis thus far is that labor allocation is done without knowledge of the realized values of $\tilde{x}_i$. In the main text, I assume that factors are allocated ex ante and cannot be reallocated after learning the shock realizations. From a short-run perspective, I believe this is a likely scenario in practice.[25] However, the stylized nature of my model also allows me to gauge the importance of relaxing this assumption.

To illustrate the key differences that arise compared to my analysis in the main text, suppose now that we were to solve a problem similar to (8) but with full knowledge of the values of $\tilde{\mathbf{x}}$. For ease of comparison between the two cases, I use the same notations when possible; for completeness, the full derivation is included in Appendix D.3.2. In both the efficient production case characterized in Lemma 1 and the inefficient case described in Lemma 2, Hulten's theorem holds exactly. Thus, denoting the output share of firm j in the efficient production case by $s_{Y,j}$, for the inefficient production case by $s^d_{Y,j}$, and their corresponding values when $\tilde{\mathbf{x}} = \bar{\mathbf{x}}$ by $\bar{s}^d_{Y,j}$ and $\bar{s}_{Y,j}$. I compared the volatility in both cases by exploiting the following relationship

$$\overline{\delta}_j = \bar{s}^d_{Y,j} = \bar{s}_{Y,j} \times \sqrt{1-d_j} = \overline{\eta}_j \times \sqrt{1-d_j}.$$

However, allowing the choice of labor to be made with full knowledge of the firm-level shocks, we obtain that

$$\overline{\delta}_j = \bar{s}^d_{Y,j} + \frac{\gamma}{1-\gamma}\left(\bar{s}^d_{Y,j} - \bar{s}^d_{L,j}\right), \tag{64}$$

where $\bar{s}^d_{L,j}$ denotes the input share of firm j in the inefficient production case where $\tilde{\mathbf{x}} = \bar{\mathbf{x}}$. For the explicit derivation of the above, see Appendix D.3.2. In the efficient production case, input and output shares are identical, so it would always be the case that $\bar{s}_{Y,j} - \bar{s}_{L,j} = 0$ and Hulten's theorem would hold exactly. However, in the presence of distortions and ex-post reallocation, input shares and output shares are not necessarily aligned and we have the extra higher-order term $\frac{\gamma}{1-\gamma}\left(\bar{s}^d_{Y,j} - \bar{s}^d_{L,j}\right)$.[26]

[25] For example, Google's CEO in announcing major layoffs in January 2023 stated that: "Over the past two years we've seen periods of dramatic growth. To match and fuel that growth, we hired for a different economic reality than the one we face today." The full statement is available at `https://blog.google/inside-google/message-ceo/january-update/`. Thus, hinting that hiring decisions had been made in the absence of information about the present state of the economy.

[26] For formal discussion of Hulten's theorem and higher-order effects in that context see Baqaee and Farhi (2019). The particular expression derived here and the effects it generates concerning the propagation of shocks described in the next paragraphs are similar to the effects detailed in Baqaee and Farhi (2020) con-

The expression in equation (64) is not necessarily positive and depends on the exact values for the implicit taxes or the severity of distortions. When the difference between input and output shares is sufficiently pronounced, the elasticity $\overline{\delta}_j$ might be negative. When a firm has a negative elasticity, it implies that we are so far removed from the efficient production case that the input share of this firm is sufficiently high compared to its output share, or that $\bar{s}^d_{Y,j} < \gamma \bar{s}^d_{L,j}$ to be exact. Suppose now that this firm experiences a positive shock. Such a shock draws more resources into this firm at the expense of other, more productive firms, thus reducing TFP as a result. The converse also holds since a negative shock to such a firm frees up inputs to be more productively employed elsewhere.

These negative elasticities render the previously introduced transformation impractical. This is because we cannot map between the ability of the firm and the size of the squared elasticities. These might be high for large negative or positive values. Thus, breaking the previously established link in this case. This rationale implies that understanding the effects of firm-level shocks on aggregate volatility in an environment with ex-post reallocation hinges on the covariance between input and output shares. Specifically, recall that earlier we had

$$\Psi^d = \sqrt{\sum_{i=1}^{N} \overline{\delta}_i^{\,2}} = \sqrt{\sum_{i=1}^{N} \left(\bar{s}^d_{Y,i}\right)^2}, \tag{65}$$

which is also the HHI of the economy computed using output or sales shares. However, with ex-post reallocation, it follows from equation (64) that

$$\left(\Psi^d\right)^2 = \left(\frac{1}{1-\gamma}\right)^2 \underbrace{\sum_{i=1}^{N} \left(\bar{s}_{Y,i}{}^d\right)^2}_{\text{Sales HHI}} + \left(\frac{\gamma}{1-\gamma}\right)^2 \underbrace{\sum_{i=1}^{N} \left(\bar{s}_{L,i}{}^d\right)^2}_{\text{Input HHI}} - \frac{2\gamma}{(1-\gamma)^2} \underbrace{\sum_{i=1}^{N} \bar{s}_{Y,i}{}^d \bar{s}_{L,i}{}^d}_{\text{Olley-Pakes term}} . \tag{66}$$

Thus, the total volatility in this economy is a function of input and output concentration and the relationship between the two, which is related to the covariance term in the Olley-Pakes decomposition (Olley and Pakes, 1996).[27] The total effect of distortions on aggregate volatility in this case depends crucially on the third term.

Figure 5 uses $\Psi^d = \sqrt{\sum_{i=1}^{N} \overline{\delta}_i^{\,2}}$ and the formula in equation (64) to quantify how positively correlated distortions affect aggregate volatility when allowing firms to reallocate

cerning the propagation of shocks in a horizontal economy in relation to the inverse harmonic mark-up. Examining the formulas derived in Appendix D.3.2 will show similar ratios between weighted averages of the implicit taxes.

[27]The above term is stated with respect to the input and output share and using a non-centered measure, whereas the Olley-Pakes formula uses the covariance between firm-level productivity and output shares.

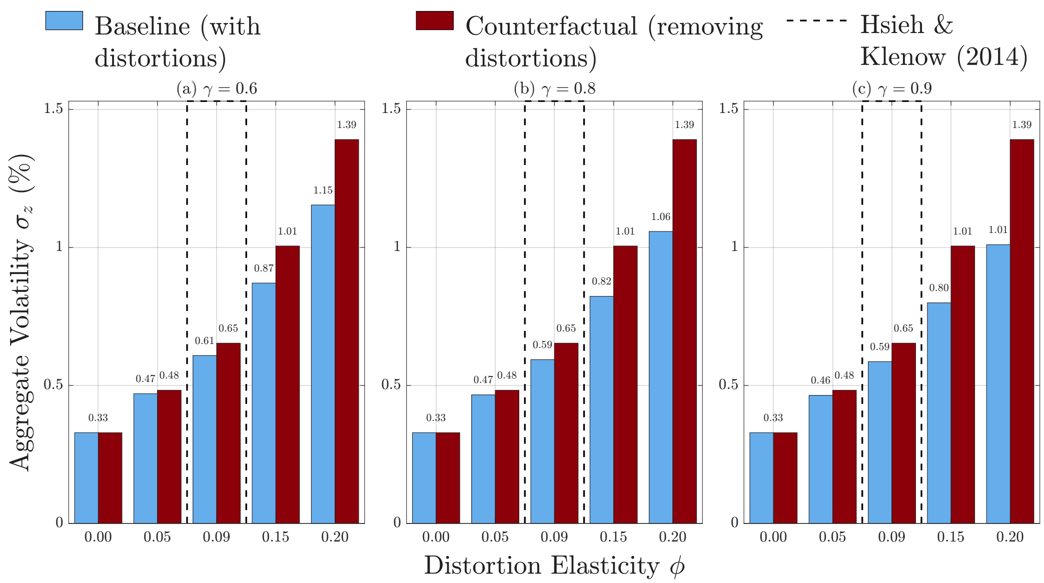

Figure 5: The Effect of Correlated Distortions on Aggregate Volatility Due to Granular Shocks Allowing for Ex-Post Reallocation

Note: This figure presents the aggregate volatility due to firm-level shocks, while allowing for ex-post reallocation. All computations are done targeting $\zeta_{\text{emp}} = 1.097$ and $N = 4.5 \times 10^6$.

factors flexibly in response to the shock. Compared to the results reported in Figure 2, allowing for ex-post reallocation implies a higher degree of volatility at the baseline level. Using the same calibration as in Table 1 implies that volatility due to firm-level shocks in this case is 0.59% compared to 0.38% without ex-post reallocation. Still, removing correlated distortions further increases volatility to 0.65%, implying a strong dampening effect. The result holds in the same direction for all parameter values explored in this section.

D.3 Proofs for Appendix D

D.3.1 Proof of Lemma 6 and Related Required Approximations

Proof. The economy with misallocation and size-dependent volatility behaves identically to the one described in Lemma 2, up to the derivation of aggregate volatility. Starting from equation (36), but this time with $\sigma_x(\bar{y}_j) = \sigma_0 \times \bar{y}_j^{-\alpha}$

$$\begin{aligned}
&\mathbb{VAR}\left[\log\left(Z^d(\tilde{\mathbf{x}})\right)\right] \qquad (67)\\
&\approx \mathbb{VAR}\left[\log\left(Z^d(\tilde{\mathbf{x}}=\bar{\mathbf{x}})\right) + \sum_{i=1}^{N} \frac{\partial \log\left(Z^d(\tilde{\mathbf{x}}=\bar{\mathbf{x}})\right)}{\partial \tilde{x}_i}(\tilde{x}_i - \bar{x})\right] = \sum_{i=1}^{N} \bar{\delta}_i^2 \times \sigma_0^2 \bar{y}_i^{-2\alpha}\\
&= \left(\frac{\sigma_0}{\bar{Y}^{d\alpha}}\right)^2 \sum_{i=1}^{N} \bar{\delta}_i^{\,2} \times \left(\frac{\bar{y}_i}{\bar{Y}^d}\right)^{-2\alpha} = \left(\frac{\sigma_0}{\bar{Y}^{d\alpha}}\right)^2 \sum_{i=1}^{N} \bar{\delta}_i^{\,2(1-\alpha)} = \left(\sigma_0 \bar{Y}^{d^{-\alpha}} \times \Psi^d(\alpha)\right)^2.
\end{aligned}$$

The approximation above follows from a first-order Taylor series expansion around the mean point; the first equality sign is the result of substituting in the size-adjusted volatility of each firm, with $\bar{y}_i = y_i|_{\tilde{\mathbf{x}}=\bar{\mathbf{x}}}$ and $\bar{Y}^d = Y^d_{\tilde{\mathbf{x}}}\big|_{\tilde{\mathbf{x}}=\bar{\mathbf{x}}}$; the second equality sign from multiplying and dividing by $\bar{Y}^{d^{-2\alpha}}$; the third from noting that $\bar{\delta}_i = \frac{\bar{y}_i}{\bar{Y}^d}$; and the last from letting $\sqrt{\sum_{i=1}^N \bar{\delta}_i^{\,2(1-\alpha)}} = \Psi^d(\alpha)$. Taking the square root of the above concludes the proof. □

Approximating $\Psi^d(\alpha)$ *Using the Quantile Function.* To approximate the value of $\Psi^d(\alpha)$ in equation (63), with positively correlated distortions as in equation (21) and Pareto distributed ability, we begin by substituting in the expression δ_j from equation (11) such that

$$\begin{aligned}
\bar{\Psi}^d(\alpha) &= \sqrt{\sum_{i=1}^{N} \bar{\delta}_i^{2(1-\alpha)}} = \frac{\sqrt{\sum_{i=1}^N \left(a_i^{\frac{1}{1-\gamma}}(1-\tau_i)^{\frac{\gamma}{1-\gamma}}\right)^{2(1-\alpha)}}}{\left[\sum_{i=1}^N a_i^{\frac{1}{1-\gamma}}(1-\tau_i)^{\frac{\gamma}{1-\gamma}}\right]^{1-\alpha}}\\
&= \frac{\sqrt{\sum_{i=1}^N \left(a_i^{\frac{1}{1-\gamma}}(a_i^{-\phi})^{\frac{\gamma}{1-\gamma}}\right)^{2(1-\alpha)}}}{\left[\sum_{i=1}^N a_i^{\frac{1}{1-\gamma}}(a_i^{-\phi})^{\frac{\gamma}{1-\gamma}}\right]^{1-\alpha}} = \frac{\sqrt{\sum_{i=1}^N \left(a_i^{\frac{1-\phi\gamma}{1-\gamma}}\right)^{2(1-\alpha)}}}{\left[\sum_{i=1}^N a_i^{\frac{1-\phi\gamma}{1-\gamma}}\right]^{1-\alpha}} = \frac{\sqrt{\sum_{i=1}^N (Q(q_i))^{2(1-\alpha)\frac{1-\phi\gamma}{1-\gamma}}}}{\left[\sum_{i=1}^N (Q(q_i))^{\frac{1-\phi\gamma}{1-\gamma}}\right]^{1-\alpha}}\\
&= \frac{\sqrt{\sum_{i=1}^N \left(1-\frac{N+1-i}{N+1}\right)^{-\frac{2(1-\alpha)}{\zeta_a}\frac{1-\phi\gamma}{1-\gamma}}}}{\left[\sum_{i=1}^N \left(1-\frac{N+1-i}{N+1}\right)^{-\frac{1}{\zeta_a}\frac{1-\phi\gamma}{1-\gamma}}\right]^{1-\alpha}} = \frac{\sqrt{\sum_{i=1}^N i^{-\frac{2(1-\alpha)}{\zeta_{\text{sales}}}}}}{\left[\sum_{i=1}^N i^{-\frac{1}{\zeta_{\text{sales}}}}\right]^{1-\alpha}}.
\end{aligned}$$

This equation follows closely the transformations in the proof of Lemma 5 in appendix C.3, with $\zeta_{\text{sales}} = \zeta_a \frac{1-\gamma}{1-\phi\gamma}$.

Calibrating σ_0. To obtain the correct counterpart of σ_0, one needs to use the first-order quantile-based approximation introduced in equation (20), to calibrate to the correct average volatility $\overline{\sigma_x}$ in a sample of N firms. To do so amounts to computing:

$$\overline{\sigma_x} = \frac{\sum_{i=1}^{N} \sigma_x(\bar{y}_i)}{N} = \frac{\sum_{i=1}^{N} \sigma_0 \bar{y}_i^{-\alpha}}{N}. \tag{68}$$

Observe that

$$\bar{y}_j = a_j^{\frac{1}{1-\gamma}} \left(1-\tau_j\right)^{\frac{\gamma}{1-\gamma}} \times \underbrace{\left[\frac{L}{\sum_{i=1}^{N} \left(a_i(1-\tau_i)\right)^{\frac{1}{1-\gamma}}}\right]^{\gamma}}_{C_y} = a_j^{\frac{1-\phi\gamma}{1-\gamma}} C_y, \tag{69}$$

where the last equality sign follows from substituting in $(1-\tau_j) = a_j^{-\phi}$. Next, using the quantile-based approximation method, and $\zeta_{\text{emp}} = \zeta_a \frac{1-\gamma}{1-\phi}$, we have that the expected value of C_y in an economy with N firms with productivity drawn from a Pareto distribution with parameter ζ_a is given by

$$\mathbb{E}[C_y] \approx \left[\frac{L}{(N+1)^{\frac{1}{\zeta_{\text{emp}}}} \sum_{i=1}^{N} i^{-\frac{1}{\zeta_{\text{emp}}}}}\right]^{\gamma}. \tag{70}$$

where the denominator is the quantile-based approximation of $\sum_{i=1}^{N} \left(a_i(1-\tau_i)\right)^{\frac{1}{1-\gamma}}$. To see this, observe that

$$\begin{aligned}
\sum_{i=1}^{N} \left(a_i(1-\tau_i)\right)^{\frac{1}{1-\gamma}} &= \sum_{i=1}^{N} \left(a_i\right)^{\frac{1-\phi}{1-\gamma}} = \sum_{i=1}^{N} \left(Q\left(q_i\right)\right)^{\frac{1-\phi}{1-\gamma}} = \sum_{i=1}^{N} \left(1-q_i\right)^{-\frac{1}{\zeta_a}\frac{1-\phi}{1-\gamma}} \\
\approx \sum_{i=1}^{N} \left(1 - \frac{N+1-i}{N+1}\right)^{-\frac{1}{\zeta_a}\frac{1-\phi}{1-\gamma}} &= \sum_{i=1}^{N} \left(\frac{i}{N+1}\right)^{-\frac{1}{\zeta_a}\frac{1-\phi}{1-\gamma}} = (1+N)^{\frac{1}{\zeta_{\text{emp}}}} \sum_{i=1}^{N} i^{-\frac{1}{\zeta_{\text{emp}}}}.
\end{aligned}$$

Finally, we can target a particular $\overline{\sigma_x}$ and obtain σ_0 via

$$\begin{aligned}
\overline{\sigma_x} &= \frac{\sum \sigma_0 \overline{y_j}^{-\alpha}}{N} = C_y{}^{-\alpha} \sigma_0 \left[\sum_{i=1}^{N} \frac{1}{N} a_i{}^{-\alpha \frac{1-\phi\gamma}{1-\gamma}}\right] = C_y{}^{-\alpha} \sigma_0 \left[\sum_{i=1}^{N} \frac{1}{N} Q^{-\alpha \frac{1-\phi\gamma}{1-\gamma}}\left(q_i\right)\right] \\
&= C_y{}^{-\alpha} \sigma_0 \sum_{i=1}^{N} \frac{1}{N} \left(\frac{i}{N+1}\right)^{\alpha \frac{1}{\zeta_a} \frac{1-\phi\gamma}{1-\gamma}} = \frac{C_y{}^{-\alpha} \sigma_0}{N(N+1)^{\frac{\alpha}{\zeta_{\text{sales}}}}} \sum_{i=1}^{N} i^{\frac{\alpha}{\zeta_{\text{sales}}}},
\end{aligned}$$

implying that

$$\sigma_0 = \overline{\sigma_x} \frac{N(N+1)^{\frac{\alpha}{\zeta_{\text{sales}}}} C_y{}^{\alpha}}{\sum_{i=1}^{N} i^{\frac{\alpha}{\zeta_{\text{sales}}}}}. \tag{71}$$

D.3.2 The Economy With Misallocation When Labor is Allocated Ex post

Allocation rule for labor. The first order condition for the production problem (8) when the value of $\tilde{x}_i$ is known, is given by

$$\gamma a_i e^{\tilde{x}_i}(1-\tau_i) l_i{}^{\gamma-1} = w. \tag{72}$$

we can again rearrange this expression and use the labor market clearing condition to obtain

$$\sum_{i=1}^{N} l_i = \left(\frac{\gamma}{w}\right)^{\frac{1}{1-\gamma}} \sum_{i=1}^{N} \left(a_i e^{\tilde{x}_i}(1-\tau_i)\right)^{\frac{1}{1-\gamma}} = L. \tag{73}$$

Thus, the allocation rule for labor is as follows

$$l_j^d = \frac{\left(a_j e^{\tilde{x}_j}\left(1-\tau_j\right)\right)^{\frac{1}{1-\gamma}}}{\sum_{i=1}^{N} \left(a_i e^{\tilde{x}_i}(1-\tau_i)\right)^{\frac{1}{1-\gamma}}} L. \tag{74}$$

Aggregate production function representation. Firm level output is given by

$$y_j^d = \frac{\left(a_j e^{\tilde{x}_j}\right)^{\frac{1}{1-\gamma}} \left(1-\tau_j\right)^{\frac{\gamma}{1-\gamma}}}{\left(\sum_{i=1}^{N} \left(a_i e^{\tilde{x}_i}(1-\tau_i)\right)^{\frac{1}{1-\gamma}}\right)^{\gamma}} L^{\gamma}. \tag{75}$$

Aggregating this by summing across all production units allows us to obtain

$$Y_{\tilde{\mathbf{x}}}^d = \sum_{i=1}^{N} y_i^d = \frac{\sum_{i=1}^{N} \left(a_i e^{\tilde{x}_i}\right)^{\frac{1}{1-\gamma}} (1-\tau_i)^{\frac{\gamma}{1-\gamma}}}{\left(\sum_{i=1}^{N} \left(a_i e^{\tilde{x}_i}(1-\tau_i)\right)^{\frac{1}{1-\gamma}}\right)^{\gamma}} L^{\gamma} = Z_{\tilde{\mathbf{x}}}^d L^{\gamma}. \tag{76}$$

Transmission of firm-level shocks. To derive the value of the TFP elasticities, observe that TFP is now given by

$$Z_{\tilde{\mathbf{x}}}^d = \frac{\sum_{i=1}^{N} \left(a_i e^{\tilde{x}_i}\right)^{\frac{1}{1-\gamma}} (1-\tau_i)^{\frac{\gamma}{1-\gamma}}}{\left[\sum_{i=1}^{N} \left(a_i e^{\tilde{x}_i}(1-\tau_i)\right)^{\frac{1}{1-\gamma}}\right]^{\gamma}}. \tag{77}$$

Therefore we can derive the elasticity by

$$\frac{\partial \log Z^d_{\tilde{\mathbf{x}}}}{\partial \tilde{x}_j} =$$
$$\frac{1}{Z^d_{\tilde{\mathbf{x}}}} \frac{\frac{1}{1-\gamma}\left(a_j e^{\tilde{x}_j}\right)^{\frac{1}{1-\gamma}}\left(1-\tau_j\right)^{\frac{\gamma}{1-\gamma}} - \gamma\left[\sum_{i=1}^N \left((1-\tau_i)a_i e^{\tilde{x}_i}\right)^{\frac{1}{1-\gamma}}\right]^{-1}\left((1-\tau_j)a_j e^{\tilde{x}_j}\right)^{\frac{1}{1-\gamma}}\frac{1}{1-\gamma}\sum_{i=1}^N (1-\tau_i)^{\frac{\gamma}{1-\gamma}}\left(a_i e^{\tilde{x}_i}\right)^{\frac{1}{1-\gamma}}}{\left[\sum_{i=1}^N \left((1-\tau_i)a_i e^{\tilde{x}_i}\right)^{\frac{1}{1-\gamma}}\right]^{\gamma}},$$

which can be condensed into

$$\frac{1}{Z^d_{\tilde{\mathbf{x}}}} \frac{\left(a_j e^{\tilde{x}_j}\right)^{\frac{1}{1-\gamma}}}{1-\gamma} \frac{\left(1-\tau_j\right)^{\frac{\gamma}{1-\gamma}} - \gamma\left[\sum_{i=1}^N \left((1-\tau_i)a_i e^{\tilde{x}_i}\right)^{\frac{1}{1-\gamma}}\right]^{-1}\left(1-\tau_j\right)^{\frac{1}{1-\gamma}}\sum_{i=1}^N (1-\tau_i)^{\frac{\gamma}{1-\gamma}}\left(a_i e^{\tilde{x}_i}\right)^{\frac{1}{1-\gamma}}}{\left[\sum_{i=1}^N \left((1-\tau_i)a_i e^{\tilde{x}_i}\right)^{\frac{1}{1-\gamma}}\right]^{\gamma}} =$$
$$\frac{\left(a_j e^{\tilde{x}_j}\right)^{\frac{1}{1-\gamma}}}{1-\gamma} \frac{\left(1-\tau_j\right)^{\frac{\gamma}{1-\gamma}} - \gamma\left(\sum_{i=1}^N \left((1-\tau_i)a_i e^{\tilde{x}_i}\right)^{\frac{1}{1-\gamma}}\right)^{-1}\left(1-\tau_j\right)^{\frac{1}{1-\gamma}}\sum_{i=1}^N (1-\tau_i)^{\frac{\gamma}{1-\gamma}}\left(a_i e^{\tilde{x}_i}\right)^{\frac{1}{1-\gamma}}}{\sum_{i=1}^N (1-\tau_i)^{\frac{\gamma}{1-\gamma}}\left(a_i e^{\tilde{x}_i}\right)^{\frac{1}{1-\gamma}}} =$$
$$\frac{\left(a_j e^{\tilde{x}_j}\right)^{\frac{1}{1-\gamma}}}{1-\gamma}\left[\frac{\left(1-\tau_j\right)^{\frac{\gamma}{1-\gamma}}}{\sum_{i=1}^N (1-\tau_i)^{\frac{\gamma}{1-\gamma}}\left(a_i e^{\tilde{x}_i}\right)^{\frac{1}{1-\gamma}}} - \gamma\frac{\left(1-\tau_j\right)^{\frac{1}{1-\gamma}}}{\sum_{i=1}^N \left((1-\tau_i)a_i e^{\tilde{x}_i}\right)^{\frac{1}{1-\gamma}}}\right] =$$
$$\frac{1}{1-\gamma}\left[\frac{\left(1-\tau_j\right)^{\frac{\gamma}{1-\gamma}}\left(a_j e^{\tilde{x}_j}\right)^{\frac{1}{1-\gamma}}}{\sum_{i=1}^N (1-\tau_i)^{\frac{\gamma}{1-\gamma}}\left(a_i e^{\tilde{x}_i}\right)^{\frac{1}{1-\gamma}}} - \gamma\frac{\left(1-\tau_j\right)^{\frac{1}{1-\gamma}}\left(a_j e^{\tilde{x}_j}\right)^{\frac{1}{1-\gamma}}}{\sum_{i=1}^N \left((1-\tau_i)a_i e^{\tilde{x}_i}\right)^{\frac{1}{1-\gamma}}}\right] = \delta_j\left(\tilde{\mathbf{x}}\right).$$

Now, observe that the output share or sales share of firm j is given by

$$s_{Y,j}{}^d = \frac{y_j^d}{Y^d_{\tilde{\mathbf{x}}}} = \frac{\left(1-\tau_j\right)^{\frac{\gamma}{1-\gamma}}\left(a_j e^{\tilde{x}_j}\right)^{\frac{1}{1-\gamma}}}{\sum_{i=1}^N (1-\tau_i)^{\frac{\gamma}{1-\gamma}}\left(a_i e^{\tilde{x}_i}\right)^{\frac{1}{1-\gamma}}}, \tag{78}$$

and its input share is given by

$$s_{L,j}{}^d = \frac{l_j^d}{L} = \frac{\left(1-\tau_j\right)^{\frac{1}{1-\gamma}}\left(a_j e^{\tilde{x}_j}\right)^{\frac{1}{1-\gamma}}}{\sum_{i=1}^N \left((1-\tau_i)a_i e^{\tilde{x}_i}\right)^{\frac{1}{1-\gamma}}}, \tag{79}$$

where the dependence of $s_{L,j}{}^d$ and $s_{Y,j}{}^d$ upon the aggregate state $\tilde{\mathbf{x}}$ is suppressed to economize on notation. Thus, we obtain equation (64)

$$\delta_j\left(\tilde{\mathbf{x}}\right) = s_{Y,j}{}^d + \frac{\gamma}{1-\gamma}\left(s_{Y,j}{}^d - s_{L,j}{}^d\right). \tag{80}$$

Aggregate volatility. To derive the aggregate effect of firm-level shocks, we can proceed as in the proofs of Lemmas 1 and 2, and use the first-order Taylor series approximation of the variance of log TFP. Letting TFP as a function of $\tilde{\mathbf{x}}$ by

$$Z^d_{\tilde{\mathbf{x}}} = \frac{\sum_{i=1}^N \left(a_i e^{\tilde{x}_i}\right)^{\frac{1}{1-\gamma}} (1-\tau_i)^{\frac{\gamma}{1-\gamma}}}{\left[\sum_{i=1}^N \left(a_i e^{\tilde{x}_i}(1-\tau_i)\right)^{\frac{1}{1-\gamma}}\right]^{\gamma}} = Z^d(\tilde{\mathbf{x}}), \tag{81}$$

one can use a Taylor series expansion around the point $\tilde{\mathbf{x}} = \overline{\mathbf{x}}$ to derive that

$$\begin{aligned} &\mathbb{VAR}\left[\log\left(Z^d(\tilde{\mathbf{x}})\right)\right] \approx \\ &\mathbb{VAR}\left[\log\left(Z^d(\tilde{\mathbf{x}}=\overline{\mathbf{x}})\right) + \sum_{i=1}^N \frac{\partial \log\left(Z^d(\tilde{\mathbf{x}}=\overline{\mathbf{x}})\right)}{\partial \tilde{x}_i}(\tilde{x}_i - \overline{x})\right] = {\sigma_x}^2 \sum_{i=1}^N \overline{\delta}_i^2, \end{aligned} \tag{82}$$

where $\overline{\delta}_i = \delta_i(\tilde{\mathbf{x}})|_{\tilde{\mathbf{x}}=\overline{\mathbf{x}}}$, and σ_x^2 is the variance of $\tilde{x}_j$.

Approximating Ψ^d with Ex-Post Reallocation. To apply the approximation method of equation (20), we can proceed as follows. The object of interest is

$$\Psi^d = \sqrt{\sum_{i=1}^N \overline{\delta}_i^2}, \tag{83}$$

it is sufficient to apply the approximation for each value of $\overline{\delta}_j$ and sum them. The approximation is given by

$$\begin{aligned} \overline{\delta}_j &= \frac{1}{1-\gamma}\left[\frac{(1-\tau_j)^{\frac{\gamma}{1-\gamma}} a_j^{\frac{1}{1-\gamma}}}{\sum_{i=1}^N (1-\tau_i)^{\frac{\gamma}{1-\gamma}} a_i^{\frac{1}{1-\gamma}}} - \gamma \frac{(1-\tau_j)^{\frac{1}{1-\gamma}} a_j^{\frac{1}{1-\gamma}}}{\sum_{i=1}^N (a_i(1-\tau_i))^{\frac{1}{1-\gamma}}}\right] \\ &= \frac{1}{1-\gamma}\left[\frac{a_j^{\frac{1-\phi\gamma}{1-\gamma}}}{\sum_{i=1}^N a_i^{\frac{1-\phi\gamma}{1-\gamma}}} - \gamma \frac{a_j^{\frac{1-\phi}{1-\gamma}}}{\sum_{i=1}^N a_i^{\frac{1-\phi}{1-\gamma}}}\right] \approx \frac{1}{1-\gamma}\left[\frac{\left(Q(\overline{q}_j)\right)^{\frac{1-\phi\gamma}{1-\gamma}}}{\sum_{i=1}^N \left(Q(\overline{q}_i)\right)^{\frac{1-\phi\gamma}{1-\gamma}}} - \gamma \frac{\left(Q(\overline{q}_j)\right)^{\frac{1-\phi}{1-\gamma}}}{\sum_{i=1}^N \left(Q(\overline{q}_i)\right)^{\frac{1-\phi}{1-\gamma}}}\right] \\ &= \frac{1}{1-\gamma}\left[\frac{\left(\frac{j}{1+N}\right)^{-\frac{1}{\zeta_a}\frac{1-\phi\gamma}{1-\gamma}}}{\sum_{i=1}^N \left(\frac{i}{1+N}\right)^{-\frac{1}{\zeta_a}\frac{1-\phi\gamma}{1-\gamma}}} - \gamma \frac{\left(\frac{j}{1+N}\right)^{-\frac{1}{\zeta_a}\frac{1-\phi}{1-\gamma}}}{\sum_{i=1}^N \left(\frac{i}{1+N}\right)^{-\frac{1}{\zeta_a}\frac{1-\phi}{1-\gamma}}}\right] \\ &= \frac{1}{1-\gamma}\left[\frac{j^{-\frac{1}{\zeta_a}\frac{1-\phi\gamma}{1-\gamma}}}{\sum_{i=1}^N i^{-\frac{1}{\zeta_a}\frac{1-\phi\gamma}{1-\gamma}}} - \gamma \frac{j^{-\frac{1}{\zeta_a}\frac{1-\phi}{1-\gamma}}}{\sum_{i=1}^N i^{-\frac{1}{\zeta_a}\frac{1-\phi}{1-\gamma}}}\right]. \end{aligned}$$