Nonrivalry and the Economics of Data

Charles I. Jones
Stanford GSB and NBER

Christopher Tonetti*
Stanford GSB and NBER

July 6, 2018 — Version 0.5

Abstract

Data is nonrival: a person's location history, medical records, and driving data can be used by any number of firms simultaneously without being depleted. Nonrivalry leads to increasing returns and implies an important role for market structure and property rights. Who should own data? What restrictions should apply to the use of data? We show that in equilibrium, firms may not adequately respect the privacy of consumers. In addition, fearing creative destruction, firms may choose to hoard data they own. Yet because of nonrivalry, there may be large social gains to sharing data across firms, even in the presence of privacy considerations. In a simple numerical example, firms owning data is substantially worse than consumers owning data, which in turn is close to optimal. A law that outlaws sharing is particularly harmful.

*We are grateful to V.V. Chari, Ben Hebert, Pete Klenow, Hannes Malmberg, and especially Sebastian Di Tella for helpful comments.

1. Introduction

In recent years, the importance of data in the economy has become increasingly apparent. More powerful computers, the growth of networks, and advances such as machine learning have led to an explosion in the usefulness of data. Examples include self-driving cars, real-time language translation, medical diagnoses, product recommendations, and social networks.

This paper develops a simple theoretical framework to study the economics of data. We are particularly interested in how different property rights for data determine its use in the economy, and thus affect output, privacy, and consumer welfare. The starting point for our analysis is the observation that data is nonrival. That is, at a technological level, data is not depleted through use. Most goods in economics are rival: if a person consumes a kilogram of rice or an hour of an accountant's time, some resource with a positive opportunity cost is used up. In contrast, existing data can be used by any number of firms or people simultaneously, without being diminished. Consider a collection of a million labeled images, the human genome, the U.S. Census, or the data generated by 10,000 cars driving 10,000 miles. Any number of firms, people, or machine learning algorithms can use this data simultaneously without reducing the amount of data available to anyone else.

The key finding in our paper is that policies related to data have important economic consequences. When firms own data, they may not adequately respect the privacy of consumers. But nonrivalry leads to other consequences that are less obvious. Because data is nonrival, there are potentially large gains to sharing data. Markets for data provide financial incentives that promote sharing, but if selling data increases the rate of creative destruction, firms may hoard data in ways that are socially inefficent. Data that could be productively used at low social cost by many others is not made available.

Another allocation we consider is one in which a government — perhaps out of concern for privacy — sharply limits the use of consumer data by firms. While this policy succeeds in generating privacy gains, it has an even larger cost because of the inefficiency that arises from a nonrival input not being used at the appropriate scale. Not only is an economy without data sharing poorer in the long run, but this policy actually reduces long-run growth in income per person. This is true even though our

learning-by-doing setting is one of semi-endogenous growth in which most policies leave the long-run growth rate unchanged.

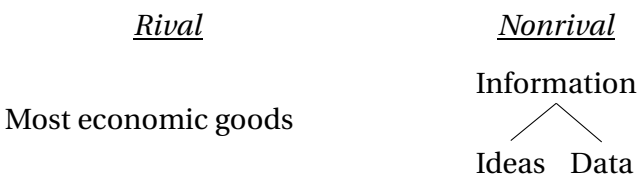
Finally, we consider an institutional arrangement in which consumers own the data associated with their behavior. Consumers then balance their concerns for privacy against the economic gains that come from selling data to all interested parties. This equilibrium results in substantial data sharing across firms, taking advantage of the nonrivalry of data and generating consumption and welfare that is close to optimal.

The remainder of the paper is structured as follows. The introduction continues by discussing the similarities and differences between data and ideas — another nonrival good — and provides a literature review. Section 2 presents the economic environment. Section 3 examines the allocation chosen by the social planner. Section 4 turns to a decentralized equilibrium in which firms own data and shows that it may be privately optimal for firms to limit data sharing while overusing their own data. Section 5 instead considers an allocation in which consumers own data and, weighing privacy considerations, sell some of it to multiple firms. Section 6 shows what happens if the government outlaws data sharing. Section 7 collects and discusses our main theoretical results while Section 8 presents a numerical simulation of our model to illustrate the various forces at work. Section 9 discusses the broader implications of our results in the context of industrial organization and cross-country patterns of growth, with a focus on the boundaries of data sharing across firms and countries. Section 10 concludes.

1.1 Data versus Ideas

Romer (1990) emphasized that ideas are nonrival. We add data to this taxonomy, as summarized in Figure 1. We find it helpful to define information as the set of all economic goods that are nonrival. That is, *information* consists of economic goods that can be entirely represented as bit strings, i.e., as sequences of ones and zeros. Ideas and data are types of information. An *idea* is a piece of information that is a set of instructions for making an economic good, which may include other ideas. *Data* denotes the remaining forms of information. It includes things like driving data, medical records, location data, and consumption history that are not themselves instructions for making a new good but that may still be useful in the production process itself. *An*

Figure 1: Taxonomy of Economic Goods



idea is a production function whereas data is a factor of production.[1]

Some examples distinguishing data from ideas might be helpful. First, consider a million images of cats, rainbows, kids, buildings, etc., labeled with their main subject. Data like this is extremely useful for training machine learning algorithms, but these labeled images are clearly not themselves ideas, i.e., not blueprints. The same is true of the hourly heart-rate history of a thousand people or the speech samples of a population. It seems obvious at this level that data and ideas are distinct.

Second, consider the efforts to build a self-driving car. The essence is a machine learning algorithm, which can be thought of as a collection of nonlinear regressions attempting to forecast what actions an expert driver will take given the data from various sensors including cameras, lidar, GPS, and so on. Data in this example includes both the collection of sensor readings and the actions taken by expert drivers. The nonlinear regression estimates a large number of parameters to produce the best possible forecasts. A successful self-driving car algorithm — a computer program, and hence an idea — is essentially just the forecasting rules that come from using data to estimate the parameters of the nonlinear model. The data and the idea are distinct: the software algorithm is the idea that is embedded in the self-driving cars of the future; data is an input used to produce this idea. Once the model is estimated, the data can be thrown away.

Another dimension along which ideas and data can differ is the extent to which they are excludable. On the one hand, it seems technologically easier to transmit data than to transmit ideas. Data can be sent at the press of button over the internet, whereas

[1] Perhaps confusingly, ideas can also be an input into the production function of new ideas: researchers use existing ideas (and data and other things) to make new ideas. Another example illustrating the fuzziness of our dichotomy is DNA. Clearly this is a set of instructions and so might be classified as an idea. However, for many medical applications, genome sequences can be thought of as data. Furthermore, data is obviously useful in producing ideas.

we invest many resources in education to learn ideas. On the other hand, data can be encrypted. Engineers change jobs and bring knowledge with them; people move and communicate causing ideas to diffuse, at least eventually. Data, in contrast, especially when it is "big," may be more easily monitored and made to be highly excludable. The "idea" of machine learning is public, whereas the driving data that is fed into the machine learning algorithm is kept private; each firm is gathering its own data.

1.2 Relation to the Literature

The "economics of data" is a new but rapidly-growing field. In this paper we take a more macroeconomic view of the importance of data, remaining silent on many of the interesting related topics in industrial organization. A surely-incomplete list of papers is mentioned here; please send us more references. Varian (2018) provides a general discussion of the economics of data and machine learning. He emphasizes that data is nonrival and refers to a common notion that "data is the new oil." Varian notes that this nonrivalry means that "data access" may be more important than "data ownership" and suggests that while markets for data are relatively limited at this point, some types of data (like maps) are currently licensed by data providers to other firms. Our paper explores these and other insights in a formal model. Our results suggest that data ownership is likely to influence data access.

Acquisti, Taylor and Wagman (2016) discuss the economics of privacy and how consumers value the privacy of their data. Farboodi and Veldkamp (2017) study the implications of expanding access to data for financial markets. Begenau, Farboodi and Veldkamp (2017) suggest that access to big data has lowered the cost of capital for large firms relative to small ones, leading to a rise in firm-size inequality. Chiou and Tucker (2017) study how the length of time that search engines keep their server logs affects the accuracy of their subsequent searches and find little evidence of a large impact. Gentzkow, Kelly and Taddy (2017) provide an overview of how text can be used as data, illustrating both statistical techniques and summarizing various interesting applications including forecasting stock prices, measuring central bank sentiment, and measuring economic policy uncertainty. Arrieta Ibarra, Goff, Jimenez Hernandez, Lanier and Weyl (2018) suggest that "free data" leads to problems and discuss the possibility of markets for data. Bajari, Chernozhukov, Hortasu and Suzuki (2018) examine how the amount of

data impacts weekly retail sales forecasts for product categories at Amazon. They find that forecasts for a given product improve with the square-root of the number of weeks of data on that product. However, forecasts of sales for a given category do not seem to improve much as the number of products within the category grows.

2. Economic Environment

The economic environment that we work with throughout the paper is summarized in Table 1. There is a representative consumer with log utility over per capita consumption, c_t. There are N_t varieties of consumer goods that combine to enter utility with a constant elasticity of substitution (CES) aggregator.

Privacy considerations also enter flow utility in two ways, as seen in equation (2). The first is via x_{it}, which denotes the fraction of an individual's data on consumption of variety i that is used by the firm producing that variety. The second is through $\tilde{x}_{it}$, which denotes the fraction of a person's data on variety i that is shared with *other firms* in the economy. Privacy costs enter via a quadratic loss function, where κ and $\tilde{\kappa}$ capture the weight on privacy versus consumption. Because there are N varieties, we add up the privacy costs across all varieties and then assume the utility cost of privacy depends on the average. There is an additional $1/N$ scaling of the x_{it} privacy cost. Because $\tilde{x}_{it}$ reflects costs associated with sharing data with all other (N) firms in the economy, it is natural that there is a factor of N difference between these costs, and this formulation generates interior solutions along the balanced growth path.

Where does data come from? Each unit of consumption is assumed to generate one unit of data. This is our "learning by doing" formulation and is captured in equation (4): $J_{it} = c_{it}L_t = Y_{it}$, where J_{it} is data created about variety i.

Firm i produces variety i according to equation (6) in the table:

$$Y_{it} = D_{it}^{\eta} L_{it}, \text{ with } \eta \in (0,1)$$

where D_{it} is the amount of data used in producing variety i. This is one of the places where the nonrivalry of data shows up in the environment: there are constant returns to scale in the rivalrous inputs — here just labor — and increasing returns to both labor and data taken together. Imagine that each worker sets up her own factory for making

Table 1: The Economic Environment

Utility	$\int_0^\infty e^{-\rho t} L_t u(c_t, x_{it}, \tilde{x}_{it}) dt$	(1)
Flow Utility	$u(c_t, x_{it}, \tilde{x}_{it}) = \log c_t - \frac{\kappa}{2} \frac{1}{N_t^2} \int_0^{N_t} x_{it}^2 \, di - \frac{\tilde{\kappa}}{2} \frac{1}{N_t} \int_0^{N_t} \tilde{x}_{it}^2 \, di$	(2)
Consumption per person	$c_t = \left(\int_0^{N_t} c_{it}^{\frac{\sigma-1}{\sigma}} di \right)^{\frac{\sigma}{\sigma-1}}$ with $\sigma > 1$	(3)
Data creation	$J_{it} = c_{it} L_t$	(4)
Variety resource constraint	$c_{it} = Y_{it} / L_t$	(5)
Firm production	$Y_{it} = D_{it}^{\eta} L_{it}$ with $\eta \in (0,1)$	(6)
Data used by firm i	$D_{it} \leq \alpha x_{it} J_{it} + (1-\alpha) B_t$	(7)
Data on variety i used by others	$D_{sit} \leq \tilde{x}_{it} J_{it}$	(8)
Data bundle	$B_t = \left(N_t^{-\frac{1}{\epsilon}} \int_0^{N_t} D_{sit}^{\frac{\epsilon-1}{\epsilon}} \, di \right)^{\frac{\epsilon}{\epsilon-1}}$ with $\epsilon > 1$	(9)
Innovation (new varieties)	$\dot{N}_t = \frac{1}{\chi} \cdot L_{et}$	(10)
Labor resource constraint	$L_{et} + L_{pt} = L_t$ where $L_{pt} := \int_0^{N_t} L_{it} \, di$	(11)
Population growth (exogenous)	$L_t = L_0 e^{g_L t}$	(12)
Aggregate output	$Y_t := c_t L_t$	(13)
Creative destruction	$\delta(\tilde{x}_{it}) = \frac{\delta_0}{2} \tilde{x}_{it}^2$	(14)

variety i. Because data is nonrival, each worker gets to use *all* the data: data can be used at any scale of production without being depleted. The parameter η captures the importance of data. We will show some evidence in Section 8 suggesting that η might take a value of 0.03 to 0.15; we think of it as a small positive number.

Data used by firm i is the sum of two terms:

$$D_{it} \leq \alpha x_{it} J_{it} + (1-\alpha) B_t.$$

The first term captures the amount of firm i's own data that is used to help it produce. In some of our allocations, firm i will be able to use all the variety i data — for example if firms own data. However, if consumers own data, they may restrict the amount of data that firms are able to use ($x_{it} < 1$). The second part of the equation incorporates data from other varieties that is used by firm i. If a firm is building a self-driving car, then data from other self-driving car companies may also be useful. Shared data on other varieties is aggregated into a bundle, B_t. The weights α and $1-\alpha$ govern the importance of own versus others' data. Importantly, this expression incorporates a second role for the nonrivalry of data: the bundle B_t can be used by any number of firms simultaneously without being depleted; hence it does not have an i subscript.

How is the bundle of data created? Let D_{sit} denote the data about variety i that is "shared" (hence the "s" subscript) and used by other firms to produce their varieties. Then, $D_{sit} \leq \tilde{x}_{it} J_{it}$. Shared data is bundled together via a CES production function with elasticity of substitution ϵ:

$$B_t = \left(N_t^{-\frac{1}{\epsilon}} \int_0^{N_t} D_{sit}^{\frac{\epsilon-1}{\epsilon}} \, di \right)^{\frac{\epsilon}{\epsilon-1}}.$$

We divorce the returns to variety from the elasticity of substitution in this CES function using the method suggested by Benassy (1996). In particular, this formulation implies that B will scale in direct proportion to N, which simplifies the analysis.

For tractability, we set up the model so that data produced today is used to produce output today, i.e., roundabout production. We think of this as a within-period timing assumption. We also assume that data depreciates fully every period. These two assumptions imply that data is not a state variable, simplifying the analysis.

The creation of new varieties is straightforward: χ units of labor are needed to create

a new variety. Total labor used for entry, L_{et}, plus total labor used in production, L_{pt}, equals total labor available in the economy, L_t. Population grows exogenously at rate g_L.

The last piece of the economic environment in Table 1 is simply a definition. Aggregate output in the economy, Y_t, equals aggregate consumption; there is no capital or investment.

Notice that in our environment, ideas and data are well-defined and distinct. An idea is a blueprint for producing a new variety, and each new blueprint is created by χ units of labor. Data is a byproduct of consumption, and each time a good is consumed, one unit of data is created that is in turn useful for improving productivity. A new idea is a new production function for producing a variety while data is a factor of production.

For allocations to be well-defined, we require that $\sigma\eta < 1$. This restriction can be rewritten as $\eta < 1/\sigma$ and has its bite in the monopolistic competition of the decentralized equilibrium. As the firm expands in size, its monopoly price falls at rate $1/\sigma$. This rate of decline must be sufficiently large relative to the increasing returns associated with data (via D_{it}^{η}) so that individual firms remain finite in size.

Finally, equation (14) is not actually part of the economic environment, but it is an important feature of the economy. We've already mentioned one downside to data sharing — the privacy cost to individuals. Data sharing also increases the rate of creative destruction: ownership of variety i changes according to a Poisson process with an arrival rate $\delta(\tilde{x}_{it})$. The more that competitors know about an incumbent firm, the greater the chance that the incumbent firm is displaced by an entrant.

A question that comes up immediately in this paper is why the Coase theorem does not apply: why does it matter whether firms or consumers own data initially? With trade and monetary transfers, why isn't the allocation the same in either case? One could certainly set up the model so that this would be true. However, to explore the subtleties of production with data, we make an additional important modification to the environment: neither firms nor consumers can commit to not selling or using the data they own. When firms own data, for example, they cannot charge consumers a higher price in exchange for the firm limiting its use of data. Similarly, if consumers own data, they cannot commit to sell the data to only a single firm. This lack of commitment serves to illustrate various properties of an economy with data. How it plays out in

the real world is a distinct and interesting question, but we simply note that there are many recent episodes in the news in which firms display a remarkable inability to avoid selling or using data that they have access to, often at odds with public statements on data-use policy, so this assumption seems reasonable.

3. The Optimal Allocation

The optimal allocation in our environment is easy to define and characterize. Using symmetry, the production structure of the economy can be simplified considerably. Consumption per person is

$$c_t = N_t^{\frac{\sigma}{\sigma-1}} c_{it} = N_t^{\frac{\sigma}{\sigma-1}} \frac{Y_{it}}{L_t}. \tag{15}$$

Moreover, the production of a variety is

$$Y_{it} = D_{it}^{\eta} L_{it} = D_{it}^{\eta} \cdot \frac{L_{pt}}{N_t}. \tag{16}$$

Combining these two expressions, aggregate GDP in the symmetric economy is

$$Y_t = N_t^{\frac{1}{\sigma-1}} D_{it}^{\eta} L_{pt}. \tag{17}$$

Next, symmetry allows us to further simplify the data component:

$$\begin{aligned} D_{it} &= \alpha x_{it} Y_{it} + (1-\alpha) N_t \tilde{x}_{it} Y_{it} \\ &= [\alpha x_{it} + (1-\alpha)\tilde{x}_{it} N_t] Y_{it} \end{aligned} \tag{18}$$

This expression can be substituted into the production function for variety i in (16) to yield

$$Y_{it} = [(\alpha x_{it} + (1-\alpha)\tilde{x}_{it} N_t)^{\eta} L_{it}]^{\frac{1}{1-\eta}}. \tag{19}$$

The increasing returns associated with data shows up in the $1/(1-\eta)$ exponent. Also, the term $\alpha x_{it} + (1-\alpha)\tilde{x}_{it} N_t$ will appear frequently whenever data is shared. This derivation shows that the αx_{it} piece reflects firms using their own data while the $(1-$

$\alpha)\tilde{x}_{it}N_t$ piece reflects firms using data from other varieties. Moreover, when data is shared, this data term scales with the measure of varieties, N_t. This ultimately provides an extra scale effect associated with data nonrivalry.

Finally, substituting the expression for D_{it} into the aggregate production function in (17) and using $L_{it} = L_{pt}/N_t$ yields

$$Y_t = N_t^{\frac{1}{\sigma-1}} \left(\frac{\alpha x_{it}}{N_t} + (1-\alpha)\tilde{x}_{it} \right)^{\frac{\eta}{1-\eta}} L_{pt}^{\frac{1}{1-\eta}}. \tag{20}$$

This equation captures the two sources of increasing returns in our model. The $N_t^{\frac{1}{\sigma-1}}$ is the standard increasing returns from love-of-variety associated with the nonrivalry of ideas. The $L_{pt}^{\frac{1}{1-\eta}}$ captures the increasing returns associated with data. In the optimal allocation, both play important roles.

We can now state the social planner problem concisely. The key allocations that need to be determined are how to allocate labor between production and entry and how much data to share. The optimal allocation solves

$$\max_{\{L_{pt}, x_{it}, \tilde{x}_{it}\}} \int_0^\infty e^{-\tilde{\rho}t} L_0 u(c_t, x_{it}, \tilde{x}_{it})\, dt, \quad \tilde{\rho} := \rho - g_L \tag{21}$$

s.t.

$$\begin{aligned}
c_t &= Y_t/L_t \\
Y_t &= N_t^{\frac{1}{\sigma-1}} \left(\frac{\alpha x_{it}}{N_t} + (1-\alpha)\tilde{x}_{it} \right)^{\frac{\eta}{1-\eta}} L_{pt}^{\frac{1}{1-\eta}} \\
\dot{N}_t &= \frac{1}{\chi}(L_t - L_{pt}) \\
L_t &= L_0 e^{g_L t}
\end{aligned}$$

The planner wants to share variety i data with firm i because that increases productivity and output. Similarly, the planner wants to share variety i data with all other firms to take advantage of the nonrivalry of data, increasing all firms' productivity and output. Tempering the planner's desire for sharing are consumers' privacy concerns. Finally, the planner weighs the gains from new varieties against the gains from producing more of the existing varieties when allocating labor to production and entry. The optimal allocation is given in Proposition 1.

Proposition 1 (The Optimal Allocation): *Along a balanced growth path, as N_t grows large, the optimal allocation converges to*

$$\tilde{x}_{it} = \tilde{x}_{sp} := \left(\frac{1}{\tilde{\kappa}} \cdot \frac{\eta}{1-\eta}\right)^{1/2} \tag{22}$$

$$x_{it} = x_{sp} := \frac{\alpha}{1-\alpha} \cdot \frac{\tilde{\kappa}}{\kappa} \left(\frac{1}{\tilde{\kappa}} \cdot \frac{\eta}{1-\eta}\right)^{1/2} \tag{23}$$

$$L_i^{sp} = \chi\rho \cdot \frac{\sigma-1}{1-\eta} := \nu_{sp} \tag{24}$$

$$N_t^{sp} = \frac{L_t}{\chi g_L + \nu_{sp}} := \psi_{sp} L_t \tag{25}$$

$$L_{pt}^{sp} = \nu_{sp}\psi_{sp} L_t \tag{26}$$

$$Y_t^{sp} = \left[\nu_{sp}(1-\alpha)^{\eta}\tilde{x}_{sp}^{\eta}\right]^{\frac{1}{1-\eta}} (\psi_{sp} L_t)^{\frac{1}{\sigma-1}+\frac{1}{1-\eta}} \tag{27}$$

$$c_t^{sp} = \frac{Y_t}{L_t} = \left[\nu_{sp}(1-\alpha)^{\eta}\tilde{x}_{sp}^{\eta}\right]^{\frac{1}{1-\eta}} \psi_{sp}^{\frac{1}{\sigma-1}+\frac{1}{1-\eta}} L_t^{\frac{1}{\sigma-1}+\frac{\eta}{1-\eta}} \tag{28}$$

$$g_c^{sp} = \left(\frac{1}{\sigma-1} + \frac{\eta}{1-\eta}\right) g_L \tag{29}$$

$$D_i^{sp} = \left[(1-\alpha)\tilde{x}_{sp}\nu_{sp}\psi_{sp} L_t\right]^{\frac{1}{1-\eta}} \tag{30}$$

$$D^{sp} = N D_i = \left[(1-\alpha)\tilde{x}_{sp}\nu_{sp}\right]^{\frac{1}{1-\eta}} (\psi_{sp} L_t)^{1+\frac{1}{1-\eta}} \tag{31}$$

$$Y_i^{sp} = \left[\nu_{sp}(1-\alpha)^{\eta}\tilde{x}_{sp}^{\eta}\right]^{\frac{1}{1-\eta}} (\psi_{sp} L_t)^{\frac{\eta}{1-\eta}} \tag{32}$$

$$U_0 = \frac{1}{\tilde{\rho}}\left(\log c_0 - \frac{\tilde{\kappa}}{2}\tilde{x}_{sp}^2 + \frac{g_c}{\tilde{\rho}}\right) \tag{33}$$

Proof See Appendix **??**.

The most important result in the proposition is the solution for GDP per person in equation (28). In particular, that solution shows that GDP per person is proportional to the size of the economy raised to some power. The exponent, $\frac{1}{\sigma-1} + \frac{\eta}{1-\eta}$, captures the degree of increasing returns to scale in the economy and is the sum of two terms. First is the standard "love of variety" effect that becomes smaller whenever varieties are more substitutable. The second term is new and reflects the increasing returns associated with the nonrivalry of data. It is increasing in η, the importance of data to the economy. A larger economy is richer because it produces more data which then feeds back and

makes all firms more productive. This equation also makes clear why we require $\eta < 1$; if $\eta \geq 1$, then the degree of increasing returns to scale is so large that the economy becomes infinitely rich: more output leads to more data, which leads to more output, and the virtuous circle explodes.

The next equation, (29), expresses the implications for growth: the growth rate of consumption per person, in the long run, is proportional to the growth rate of population, where the factor of proportionality is this degree of increasing returns to scale.

The remaining results in the optimal allocation (and in fact all the allocations we consider) break down in an elegant way. First, optimal data sharing $\tilde{x}_{sp}$ and x_{sp} are decreasing in the privacy costs ($\tilde{\kappa}$ and κ) and increasing in the importance of data in the economy (η), as shown in equations (22) and (23).

Next, equation (25) shows that optimal variety N_t^{sp} is proportional to the population in the economy, and the factor of proportionality is defined to be the parameter ψ_{sp}. Higher entry costs, a higher rate of time preference, and faster population growth all reduce variety along the balanced growth path. A higher elasticity of substitution between varieties makes new varieties less valuable and reduces N_t^{sp}. Finally, if data is more important ($\uparrow \eta$) the economy devotes less resources to entry (which does not create data) and more resources to production (which does).

This is even more apparent in equation (24), which shows employment per firm, L_{it}^{sp}, which equals a combination of parameters that we define to be ν_{sp}. The comparative statics for firm size are essentially the opposite of those for variety. Optimal firm size is constant along a balanced growth path and invariant to the overall population of the economy. This reflects the assumption that the entry cost is a fixed amount of labor that does not change as the economy grows. The fact that the size distribution of firms seems stationary in the U.S. suggests this may be a reasonable assumption; Bollard, Klenow and Li (2016) provide further support. We show later that the key findings of our paper are robust to variations of this assumption.

We will return to these results after discussing other ways to allocate resources in this environment. The ν and ψ parameters for the different allocations will be an important part of that comparison.

4. Firms Own Data

We now explore one possible way to use markets to allocate resources. In this equilibrium, we assume that firms own data and decide whether or not to sell it. Data is bought and sold via a data intermediary that bundles together data from all varieties and resells it to each individual firm. When firms buy bundles of data, they take the price as given, but when they sell, the market structure is monopolistic competition because they are the unique providers of their particular type of data.

4.1 Decision Problems

Household Problem. Households have one unit of labor that they supply inelastically in exchange for the wage w_t. They hold assets that pay a return r_t (these assets in equilibrium are claims on the value of the monopolistically competitive firms). The representative household solves

$$U_0 = \max_{\{c_{it}\}} \int_0^\infty e^{-\tilde{\rho}t} L_0 u(c_t, x_{it}, \tilde{x}_{it}) dt \tag{34}$$

$$\text{s.t.} \quad c_t = \left(\int_0^{N_t} c_{it}^{\frac{\sigma-1}{\sigma}} di \right)^{\frac{\sigma}{\sigma-1}} \tag{35}$$

$$\dot{a}_t = (r_t - g_L) a_t + w_t - \int_0^{N_t} p_{it} c_{it} \, di \tag{36}$$

Notice that households do not choose x_{it} and $\tilde{x}_{it}$ since the firms are the ones who own data in this allocation. Also, the price of c_t is normalized to one so that all prices are expressed in units of c_t.

Firm Problem. Each incumbent firm chooses how much data to buy and sell and how much labor to hire. Each sale generates data: $J_{it} = Y_{it}$. The firm uses the fraction x_{it} of this data itself and sells a fraction $\tilde{x}_{it}$ to the data intermediary at a price p_{sit} that it sets via monopolistic competition. Because of nonrivalry, the firm can both use and sell the same data simultaneously. In addition, the firm buys bundles of data D_{bit} at price p_{bt}, which it takes as given. Finally, each firm takes demand for its variety (aggregating the

FOC from the Household Problem) as given:

$$p_{it} = \left(\frac{c_t}{c_{it}}\right)^{\frac{1}{\sigma}} = \left(\frac{Y_t}{Y_{it}}\right)^{\frac{1}{\sigma}} \tag{37}$$

Letting V_{it} denote the market value of firm i, the incumbent firm problem is:

$$r_t V_{it} = \max_{\{L_{it}, D_{bit}, x_{it}, \tilde{x}_{it}\}} \left(\frac{Y_t}{Y_{it}}\right)^{\frac{1}{\sigma}} Y_{it} - w_t L_{it} - p_{bt} D_{bit} + p_{sit}\tilde{x}_{it} Y_{it} + \dot{V}_{it} - \delta(\tilde{x}_{it}) V_{it} \tag{38}$$

$$\text{s.t.} \quad Y_{it} = D_{it}^{\eta} L_{it} \tag{39}$$

$$D_{it} = \alpha x_{it} Y_{it} + (1-\alpha) D_{bit} \tag{40}$$

$$x_{it} \in [0,1], \tilde{x}_{it} \in [0,1] \tag{41}$$

$$p_{sit} = \lambda_{DI} N_t^{-\frac{1}{\epsilon}} \left(\frac{B_t}{\tilde{x}_{it} Y_{it}}\right)^{\frac{1}{\epsilon}} \tag{42}$$

where the last equation is the downward-sloping demand curve for firm i's data from the data intermediary, which is described next. Firm i takes the aggregates λ_{DI}, B_t, N_t, and Y_t as given in solving this problem.

Each firm wants to use all the data on its own variety: it owns the data already and does not consider consumers' privacy concerns. The firm may also want to sell some of the data on its variety to other firms. Creative destruction limits its desire to sell data. When more information about the firm's variety is available to competitors, the firm is more likely to be replaced by a competitor. The firm may want to buy some of the bundle of other firms' data, weighing the cost of purchase against the gains from increased productivity and sales. Finally, the firm hires labor to reach its desired scale, recognizing the downward sloping demand curve for its variety as governed by the elasticity of substitution across varieties, σ, and that more sales generates more data.

Data Intermediary Problem. The "b" and "s" notation for buying and selling becomes tricky with the data intermediary: D_{bit} is the amount that firm i buys from the data intermediary, so it is the amount the data intermediary sells to firm i. Similarly, p_{sit} is the price at which firm i sells data to the data intermediary, so it is the price at which the data intermediary purchases data.

We originally hoped to model the data intermediary sector as perfectly competitive.

However, the nonrival nature of data makes this impossible: if agents could buy nonrival data at a given price and then sell data at a given price, they would want to buy one unit and sell it an infinite number of times! Nonrivalry poses problems for perfect competition, as in Romer (1990).

Our alternative seeks to minimize the frictions posed by data intermediation. We assume that the data intermediary is a monopolist subject to free entry at a vanishingly small cost, so that the data intermediary earns zero profits. Moreover, we assume the actual and potential data intermediaries take the price at which they buy data from firms, p_{sit}, as given. This setup delivers a limit pricing condition with zero profits even though data is nonrival.

The data intermediary takes its purchase price of data p_{sit} as given and maximizes profits by choosing the quantity of data to purchase from each firm and the price at which it sells bundles of data to firms:

$$\max_{\{p_{bt}, D_{sit}\}} p_{bt} \int_0^{N_t} D_{bit}\, di - \int_0^{N_t} p_{sit} D_{sit}\, di \tag{43}$$

s.t.

$$D_{bit} \leq B_t = \left(N_t^{-\frac{1}{\epsilon}} \int_0^{N_t} (D_{sit})^{\frac{\epsilon-1}{\epsilon}}\, di \right)^{\frac{\epsilon}{\epsilon-1}} \quad \forall i \tag{44}$$

$$p_{bt} \leq p_{bt}^* \tag{45}$$

subject to the demand curve $p_{bt}(D_{bit})$ from the Firm Problem above, where p_{bt}^* is the limit price associated with the zero profit condition that comes from free entry.

This expression for profits combined with the resource constraint on data in (44) incorporates the fact that the data intermediary can "buy data once and sell it multiple times," i.e., the nonrivalry of data. This is shown in the first term of profits, where revenue essentially equals $N_t p_{bt} B_t$ — the firm is able to sell the same bundle B_t multiple times. For example, location data from consumers can, technologically, be sold to every firm in the economy, not just to the store in which consumers happen to be shopping at the moment.

Firm Entry and the Creation of New Varieties. A new variety can be designed and created at a fixed cost of χ units of labor. In addition, new entrants are the beneficiaries

of business stealing: they obtain the property rights to the varieties that suffer from creative destruction. The free entry condition is then

$$\chi w_t = V_{it} + \frac{\int_0^{N_t} \delta(\tilde{x}_{it}) V_{it}\, di}{\dot{N}_t}. \tag{46}$$

The left side χw_t is the cost of the χ units of labor needed to create a new variety. The right side has two terms. The first is the value of the new variety that is created. The second, is the per-entrant portion of the rents from creative destruction.

4.2 The Equilibrium when Firms Own Data

The equilibrium in which firms own data consists of quantities $\{c_t, Y_t, c_{it}, x_{it}, \tilde{x}_{it}, a_t, Y_{it}, L_{it}, D_{it}, D_{bit}, B_t, D_{sit}, N_t, L_{pt}, L_{et}\}$ and prices $\{p_{it}, p_{bt}, p_{sit}, w_t, r_t, V_{it}\}$ such that

1. $\{c_t, c_{it}, a_t\}$ solve the Household Problem
2. $\{L_{it}, Y_{it}, p_{it}, p_{sit}, D_{bit}, D_{it}, x_{it}, \tilde{x}_{it}, V_{it}\}$ solve the Firm Problem
3. (D_{sit}, B_t) Data markets clear: $D_{bit} = B_t$ and $D_{sit} = \tilde{x}_{it} Y_{it}$
4. (p_{bt}) Free entry into data intermediation gives zero profits there (constrains p_b as a function of p_s)
5. (L_{et}) Free entry into producing a new variety leads to zero profits, as in equation (46)
6. Definition of L_{pt}: $L_{pt} = \int_0^{N_t} L_{it} di$
7. w_t clears the labor market: $L_{pt} + L_{et} = L_t$
8. r_t clears the asset market: $a_t = \int_0^{N_t} V_{it} di / L_t$
9. N_t follows its law of motion: $\dot{N}_t = \frac{1}{\chi}(L_t - L_{pt})$
10. $Y_t := c_t L_t$ denotes aggregate output.

In Section 7, we compare the allocation that results from this equilibrium with the optimal allocation as well as with alternative allocations. Before that, we define the alternative allocations, allowing us to efficiently make the comparisons all at once. For this reason, we turn next to an equilibrium in which consumers own data.

5. Consumers Own Data

We now consider an allocation in which consumers own data associated with their purchases. They can sell data to a data intermediary in a competitive market and choose how much data to sell to balance the gains in income versus the costs to privacy. Firms own zero data as it is created but can purchase data from the data intermediary. As we discussed earlier, consumers cannot commit to sell their data to only a single firm. So it is not possible for firm i to charge consumers a lower price in exchange for the consumers agreeing not to sell their data to others.

5.1 Decision Problems

We now lay out the decision problems of households and firms when consumers own data.

Household Problem. The household problem is similar to when firms own data, except now the household chooses how much data to share. Consumers license the same data in two ways when selling it: they sell data on variety i with a license that allows firm i to use it and, separately, they sell data on variety i with a license that allows it to be bundled and sold to all other firms. Because data can be sold in two ways, there are two different prices: data on variety i that will be used only by firm i sells at price p^a_{st}, while data on variety i that can be bundled and sold to any firm sells at price p^b_{st}. The representative household solves

$$U_0 = \max_{\{c_{it},x_{it},\tilde{x}_{it}\}} \int_0^\infty e^{-\tilde{\rho}t} L_0 u(c_t, x_{it}, \tilde{x}_{it}) dt \tag{47}$$

$$\text{s.t.} \quad c_t = \left(\int_0^{N_t} c_{it}^{\frac{\sigma-1}{\sigma}} di \right)^{\frac{\sigma}{\sigma-1}} \tag{48}$$

$$\begin{aligned} \dot{a}_t &= (r_t - g_L)a_t + w_t - \int_0^{N_t} p_{it} c_{it} di + \int_0^{N_t} x_{it} p^a_{st} c_{it} di + \int_0^{N_t} \tilde{x}_{it} p^b_{st} c_{it} di \\ &= (r_t - g_L)a_t + w_t - \int_0^{N_t} q_{it} c_{it} dt \end{aligned} \tag{49}$$

where $q_{it} := p_{it} - x_{it} p^a_{st} - \tilde{x}_{it} p^b_{st}$ is the effective price of consumption, taking into account that the fractions x_{it} and $\tilde{x}_{it}$ of each good consumed generate income when the associated data is sold.

Firm Problem. Each incumbent firm chooses how much data to buy, and two types of data are available for purchase: data from the firm's own variety (D_{ait}) and data from other varieties that the firm may buy (D_{bit}). Each firm sees the downward-sloping demand for its variety (aggregating the FOC from the Household Problem):

$$q_{it} = \left(\frac{c_t}{c_{it}}\right)^{\frac{1}{\sigma}} = \left(\frac{Y_t}{Y_{it}}\right)^{\frac{1}{\sigma}} = p_{it} - x_{it}p^a_{st} - \tilde{x}_{it}p^b_{st} \tag{50}$$

so that

$$p_{it} = \left(\frac{Y_t}{Y_{it}}\right)^{\frac{1}{\sigma}} + x_{it}p^a_{st} + \tilde{x}_{it}p^b_{st}. \tag{51}$$

Letting V_{it} denote the market value of firm i, the incumbent firm problem is:

$$\begin{aligned} r_t V_{it} = \max_{L_{it}, D_{ait}, D_{bit}} & \left[\left(\frac{Y_t}{Y_{it}}\right)^{\frac{1}{\sigma}} + x_{it}p^a_{st} + \tilde{x}_{it}p^b_{st}\right] Y_{it} - w_t L_{it} - p_{at}D_{ait} - p_b D_{bit} \\ & + \dot{V}_{it} - \delta(\tilde{x}_{it})V_{it} \qquad (52) \\ \text{s.t.} \quad & Y_{it} = D^{\eta}_{it} L_{it} \\ & D_{it} = \alpha D_{ait} + (1-\alpha)D_{bit} \qquad (53) \\ & D_{ait} \geq 0, \;\; D_{bit} \geq 0 \end{aligned}$$

Firms no longer face a simple constant elasticity demand curve because the effective price that consumers pay is different from the price that firms receive (because consumers sell data). From the perspective of the firm, D_{ait} and D_{bit} are perfect substitutes: scaled appropriately, the firm is indifferent between buying its own data versus an appropriately-sized bundle of other firms' data. This fact will help pin down the relative price of the two kinds of data.

Data Intermediary Problem. Because we have two types of data, we now introduce two different data intermediaries: one handles the sale of "own" data and the other handles the bundle. Each is modeled as earlier, i.e., as a monopolist who is constrained by free entry into data intermediation.

Taking the price p^a_{st} of data purchased from consumers as given, the data interme-

diary for own data solves the following problem at each date t:

$$\max_{\{p_{ait},D^a_{cit}\}} \int_0^{N_t} p_{ait}D_{ait}di - \int_0^{N_t} p^a_{st}D^a_{cit}di \tag{54}$$

s.t.

$$D_{ait} \leq D^a_{cit} \quad \forall i \tag{55}$$

$$p_{ait} \leq p^*_{ait} \tag{56}$$

subject to the demand curve $p_{ait}(D_{ait})$ from the Firm Problem above, where p^*_{ait} is the limit price associated with the zero profit condition that comes from free entry.

Similarly, taking the price p^b_{st} of data purchased from consumers as given, the data intermediary for bundled data solves

$$\max_{\{p_{bit},D^b_{cit}\}} \int_0^{N_t} p_{bit}D_{bit}di - \int_0^{N_t} p^b_{st}D^b_{cit}di \tag{57}$$

s.t.

$$D_{bit} \leq B_t = \left(N_t^{-\frac{1}{\epsilon}} \int_0^{N_t} (D^b_{cit})^{\frac{\epsilon-1}{\epsilon}} di\right)^{\frac{\epsilon}{\epsilon-1}} \quad \forall i \tag{58}$$

$$p_{bit} \leq p^*_{bit} \tag{59}$$

subject to the demand curve $p_{bit}(D_{bit})$ from the Firm Problem above, where p^*_{bit} is the limit price associated with the zero profit condition that comes from free entry.

The two data intermediaries are monopolists who choose the prices p_{ait} and p_{bit} of data as well as how much data to buy from consumers of each variety and type, taking the prices p^a_{st} and p^b_{st} as given. From the standpoint of the consumer, one unit of consumption generates one unit of data and data from all varieties sell at the same price, while each type of license may sell at a different price.

The constraints on the data intermediary problems are critical. Equation (55) says that the data intermediary can sell firm i's *own* data only up to the amount of type i's data that has been purchased. In contrast, equation (58) recognizes that data from all varieties can be bundled together and resold to each individual firm.

We assume free entry into the data intermediary sector at zero cost. This constrains the prices p_a and p_b that the data intermediaries can charge and implies that the monopolist earns zero profits. This condition together with the fact that the two types of

data are perfect substitutes in the firm production function pin down the prices.

5.2 Equilibrium when Consumers Own Data

An equilibrium in which consumers own data consists of quantities $\{c_t, Y_t, c_{it}, x_{it}, \tilde{x}_{it}, a_t, Y_{it}, L_{it}, D_{it}, D_{ait}, D_{bit}, D^a_{cit}, D^b_{cit}, B_t, N_t, L_{pt}, L_{et}\}$ and prices $\{q_{it}, p_{it}, p_{ait}, p_{bit}, p^a_{st}, p^b_{st}, w_t, r_t, V_{it}\}$ such that

1. $\{c_t, c_{it}, x_{it}, \tilde{x}_{it}, a_t\}$ solve the Household Problem
2. $\{L_{it}, Y_{it}, p_{it}, D_{ait}, D_{bit}, D_{it}, V_{it}\}$ solve the Firm Problem
3. (q_{it}) The effective consumer price is $q_{it} = p_{it} - x_{it}p^a_{st} - \tilde{x}_{it}p^b_{st}$
4. D^a_{cit}, D^b_{cit}, B_t, p_{ait}, and p_{bit} solve the Data Intermediary Problem subject to the constraint that there is free entry into this sector, so it makes zero profits
5. p^a_{st} clears the data market so that supply equals demand: $D^a_{cit} = x_{it}c_{it}L_t$
6. p^b_{st} clears the data market so that supply equals demand: $D^b_{cit} = \tilde{x}_{it}c_{it}L_t$
7. (L_{et}) Free entry into producing a new variety leads to zero profits (including the entrant's share of the rents from creative destruction): $\chi w_t = V_{it} + \frac{\int_0^{N_t} \delta(\tilde{x}_{it})V_{it}\, di}{\dot{N}_t}$
8. Definition of L_{pt}: $L_{pt} = \int_0^{N_t} L_{it} di$
9. w_t clears the labor market: $L_{pt} + L_{et} = L_t$
10. r_t clears the asset market: $a_t = \int_0^{N_t} V_{it} di / L_t$
11. N_t follows its law of motion: $\dot{N}_t = \frac{1}{\chi}(L_t - L_{pt})$
12. $Y_t := c_t L_t$ denotes aggregate GDP

5.3 Understanding the Equilibrium when Consumers Own Data

While Section 7 will discuss the "big picture" findings related to this allocation, it is worth pausing here for a bit to highlight some of the smaller results related to the allocation when consumers own data in this environment.

First, the "perfect substitutes" character of data in equation (53) means that in equilibrium,

$$p_{at} = \frac{\alpha}{1-\alpha} p_{bt} \tag{60}$$

where we've dropped the "i" subscript because of symmetry. At any other price ratio, firms would buy only one type of data and not the other. Similarly, the consumer prices for each type of data satisfy

$$p^a_{st} = p_{at} \text{ and } p^b_{st} = N_t p_{bt}. \tag{61}$$

Second, consider the inequality constraints in the Data Intermediary's problem. In equilibrium, the data intermediary will sell any data that it buys. Moreover, because of nonrivalry, data can be bought once and sold multiple times. This means that both inequality constraints will bind. First, $D_{ait} = D^a_{cit} = x_{it} Y_{it}$; that is, all data of variety i that the data intermediary purchases will be sold back to firm i. Second, $D_{bit} = B_t = N D^b_{cit} = N \tilde{x}_{it} Y_{it}$ (using symmetry); that is, *all* data that the data intermediary buys will be sold to all varieties as bundled data.

Notice what this implies about the data the firm uses in making variety i. From equation (53),

$$\begin{aligned} D_{it} &= \alpha D_{ait} + (1-\alpha) D_{bit} \\ &= [\alpha x_{it} + (1-\alpha)\tilde{x}_{it} N_t] Y_{it}. \end{aligned} \tag{62}$$

Finally, this expression can be substituted into the production function for a variety and manipulated to yield

$$Y_{it} = [(\alpha x_{it} + (1-\alpha)\tilde{x}_{it} N_t)^\eta L_{it}]^{\frac{1}{1-\eta}}. \tag{63}$$

The now-familiar increasing returns associated with data shows up in the $1/(1-\eta)$ exponent. Additionally, the term $\alpha x_{it} + (1-\alpha)\tilde{x}_{it} N_t$ reminds us that when all data is shared, this overall data term scales with the measure of varieties, N_t, which will ultimately generate the scale effect associated with data nonrivalry.

6. Outlaw Data Sharing

The final allocation that we consider is motivated by recent concerns over data privacy. In the world in which firms own data, suppose the government, in an effort to protect privacy, limits the use of data. In particular, it mandates that

$$\tilde{x}_{it} = 0$$
$$x_{it} \leq \bar{x} \in (0, 1].$$

That is, firms are not allowed to sell their data to any third parties: $\tilde{x}_{it} = 0$. Moreover, the government may restrict firms to use less than 100 percent of their own-variety data, parameterized by $x_{it} = \bar{x}$. We require $\bar{x} > 0$ in our setting — otherwise output of each firm would be zero because data is an essential input to production.

With this determination of $\tilde{x}_{it}$ and x_{it}, the rest of the equilibrium looks exactly like the firms-own-data case, so we will not repeat that setup here. Instead, we turn next to comparing the equilibrium outcomes across these different allocations.

7. Key Insights from Comparing the Different Allocations

This section delivers the payoff from the preparation we've made in the previous sections: we see how the different allocation mechanisms we've studied lead to different outcomes. We compare the allocations on the balanced growth path for the social planner (*sp*), when consumers own data (*c*), when firms own data (*f*), and when the government outlaws data sharing (*os* for "outlaw sharing"). When firms restrict the sale of data to limit their exposure to creative destruction, what are the consequences? When consumers own data and can sell it, is the allocation optimal? What if data sharing is banned out of a concern for privacy?

Privacy and Data Sharing. The steady-state fraction of data that is shared with other firms is given by[2]

$$\tilde{x}_{sp} = \left(\frac{1}{\tilde{\kappa}} \cdot \frac{\eta}{1-\eta} \right)^{1/2} \tag{64}$$

[2]We assume $\frac{\epsilon}{\epsilon-1} > \sigma\eta$ and $\frac{\epsilon}{\epsilon-1} > \frac{3}{2}\sigma\eta - \frac{1}{2}\eta$ so that $0 < \Gamma < 2$ holds in equation (66).

$$\tilde{x}_c = \left(\frac{1}{\tilde{\kappa}} \cdot \frac{\eta}{1-\eta} \cdot \frac{\sigma-1}{\sigma}\right)^{1/2} \tag{65}$$

$$\tilde{x}_f = \left(\frac{\Gamma\rho}{(2-\Gamma)\delta_0}\right)^{1/2} \text{ where } \Gamma := \frac{\eta(\sigma-1)}{\frac{\epsilon}{\epsilon-1} - \sigma\eta} \tag{66}$$

$$\tilde{x}_{os} = 0. \tag{67}$$

Interestingly, even when consumers own and sell their own data, the equilibrium allocation features inefficiently low data sharing because of the $\frac{\sigma-1}{\sigma} < 1$ term in equation (65). The equilibrium price of data that consumers receive in exchange for selling is influenced by this same factor:

$$p^b_{st} = \frac{\eta}{1-\eta} \cdot \frac{\sigma-1}{\sigma} \cdot \frac{1}{\tilde{x}_c} \left(\psi_c L_t\right)^{\frac{1}{\sigma-1}}.$$

Recall that $\frac{\sigma}{\sigma-1}$ is the standard monopoly markup, so the intuition is that the monopoly markup distortion leads data to sell for a price that is inefficiently low, causing consumers to share too little data.

These equations can be contrasted with data sharing when firms own data, as in equation (66). First, the utility cost associated with privacy $\tilde{\kappa}$ does not enter the firm solution, as firms do not inherently care about privacy. Second, $\tilde{x}_f$ depends on δ_0, capturing the crucial role of creative destruction — which does not enter the planner or consumer solutions for $\tilde{x}$. As we will see in our numerical examples, reasonable values for δ_0 mean that creative destruction concerns are first-order for firms, so they share little data with other firms and $\tilde{x}_f$ is small. They therefore inadvertently deliver privacy benefits to consumers. But as we will see, this aversion to sharing has other consequences. An extreme version of this allocation is the one that outlaws data sharing entirely, so that $\tilde{x}_{os} = 0$.

The privacy considerations that involve only firm i and consumption of variety i are similar. In particular,

$$x_{sp} = \frac{\alpha}{1-\alpha}\frac{\tilde{\kappa}}{\kappa} \cdot \tilde{x}_{sp} \tag{68}$$

$$x_c = \frac{\alpha}{1-\alpha}\frac{\tilde{\kappa}}{\kappa} \cdot \tilde{x}_c \tag{69}$$

$$x_f = 1 \tag{70}$$

$$x_{os} = \bar{x} \in (0, 1]. \tag{71}$$

These equations show that when firms own data, they overuse it. That is, firms set $x_f = 1$, while the social planner and consumers take into account the privacy costs associated with κ and generally choose less direct sharing of data, $x < 1$.

Firm Size. Because of symmetry, firm size L_{it} equals the ratio of production employment to varieties, L_{pt}/N_t. This quantity plays an important role in all the allocations and is denoted by the parameter ν:

$$L_{it}^{alloc} = \left(\frac{L_{pt}}{N_t}\right)^{alloc} = \nu_{alloc}, \quad \text{for } alloc \in \{sp, c, f, os\} \tag{72}$$

where

$$\nu_{sp} := \chi\rho \cdot \frac{\sigma - 1}{1 - \eta} \tag{73}$$

$$\nu_c := \chi g_L \cdot \frac{\rho + \delta(\tilde{x}_c)}{g_L + \delta(\tilde{x}_c)} \cdot \frac{\sigma - 1}{1 - \sigma\eta} \tag{74}$$

$$\nu_f := \chi g_L \cdot \frac{\rho + \delta(\tilde{x}_f)}{g_L + \delta(\tilde{x}_f)} \cdot \frac{\sigma - 1}{1 - \sigma\eta\frac{\epsilon-1}{\epsilon}} \tag{75}$$

$$\nu_{os} := \chi\rho \cdot \frac{\sigma - 1}{1 - \sigma\eta}. \tag{76}$$

For all allocations, firm size as measured by employees is constant. This is because the entry cost technology is such that a fixed number of workers can create a new variety. Several economic forces determine firm size. First, notice how similar ν_{sp} and ν_{os} are. That is, steady-state firm size in the allocation with no data sharing features a firm size that looks superficially similar to the optimal firm size. Both are increasing in χ (the entry cost) and ρ (the rate of time preference). Higher values of these parameters deter entry, and since the two uses for labor are entry and production, this increases labor used in production.

The only difference between the two expressions is that the optimal firm size depends on $1 - \eta$ where the equilibrium firm size depends on $1 - \sigma\eta$. This difference is subtle and important to understand, as this same difference plays an important role throughout the allocations. To understand this difference, we rewrite the optimal allo-

cation as

$$\left(\frac{L_{pt}}{N_t}\right)^{sp} = \nu_{sp} = Const \cdot \frac{1/(1-\eta)}{1/(\sigma-1)}. \tag{77}$$

The left-hand side of this expression is the ratio of production labor to variety, and variety is closely related to entry. The right-hand side is the ratio of two elasticities. The numerator, $1/(1-\eta)$, is the degree of increasing returns to scale at the firm level that results from the nonrivalry of data. The denominator, $1/(\sigma-1)$, is the degree of increasing returns to scale associated with the love of variety. Perhaps not surprisingly, the planner makes the ratio of production labor to variety proportional to the ratio of these two elasticities, which capture the social value of production labor and entry.

In contrast, consider the equilibrium allocation when data sharing is outlawed. Flipping the numerator and denominator, equation (76) can be expressed as

$$\left(\frac{N_t}{L_{pt}}\right)^{os} = \frac{1}{\nu_{os}} = Const \cdot \frac{1-\sigma\eta}{\sigma-1}. \tag{78}$$

As shown in Appendix equation (**??**), this expression derives from the free entry condition for firms, i.e., $\chi w_t = V_{it}$ (since there is no creative destruction in the outlaw-sharing equilibrium). The value of a firm is the present discounted value of future profits. The number of firms in the economy, N_t, depends on profits relative to entry costs. Aggregate profits as a share of GDP equals $(1-\sigma\eta)/\sigma \cdot 1/(1-\eta)$, while aggregate payments to production labor as a share of GDP equals $(\sigma-1)/\sigma \cdot 1/(1-\eta)$. Equation (78) says that equilibrium variety is proportional to this ratio. And the inverse of this expression gives ν_{os}.

Now look back at equations (73) and (76), and recall that ν equals the employment of firms for a given allocation. These equations imply that firm employment is larger in the equilibrium with no data sharing than in the optimal allocation since $\sigma > 1$. This occurs because of the profit share term. Intuitively, the equilibrium allocation creates varieties based on profits, while the social planner creates varieties based on the full social surplus. Because profits are less than social surplus — the standard appropriability problem — the outlaw-sharing equilibrium features too few firms. The flip side is that firms in equilibrium are inefficiently large.

We will discuss the equations for ν_c and ν_f after considering the number of firms

and varieties, next.

Number of Firms and Varieties. The effect of the appropriability problem on the measure of varieties can be seen more directly in our next set of equations. The number of firms (varieties) in an allocation is proportional to the labor force:

$$N_t^{alloc} = \psi_{alloc} L_t \quad \text{where} \quad \psi_{alloc} := \frac{1}{\chi g_L + \nu_{alloc}}. \tag{79}$$

Notice that the last half of the denominator of the ψ expression is just the ν term itself. For g_L small, variety is basically inversely proportional to firm size, verifying the statements we just made about firm size and variety in providing the intuition above.

Next, it is now worth comparing firm size and variety between the equilibrium in which consumers own data and the outlaw-sharing equilibrium in which firms own data. Equations (74) and (76) show that firm sizes differ in these two allocations only because of creative destruction, which enters in two ways. In the numerator of (74), there is a $\rho + \delta(\tilde{x}_c)$ term. This captures the extent to which creative destruction raises the effective rate at which firms discount future profits. In the denominator, however, there is an additional term involving $\delta(\tilde{x}_c)$. This term captures the rents from destroyed firms as they flow to new entrants — business stealing — essentially raising the return to entry. If $\rho = g_L$, then these two terms cancel and creative destruction does not influence firm size and variety creation.

A similar effect impacts firm size and the number of firms in the equilibrium when firms own data and can legally buy and sell it, as seen in equation (75). However, in that allocation, data sharing is typically lower than when consumers own data, implying that creative destruction is also lower, limiting the role of this term.

GDP and Economic Growth. The key finding of the paper is how data sharing influences living standards and economic growth. The next set of equations shows aggregate GDP in the various allocations. For the allocations that feature some data sharing, the equation for aggregate output is

$$Y_t^{alloc} = \left[\nu_{alloc}(1-\alpha)^{\eta} \tilde{x}_{alloc}^{\eta}\right]^{\frac{1}{1-\eta}} \left(\psi_{alloc} L_t\right)^{1+\frac{1}{\sigma-1}+\frac{\eta}{1-\eta}} \quad \text{for } alloc \in \{sp, c, f\} \tag{80}$$

There are essentially three key terms in this expression, and all have nice intuitions. First, ν_{alloc} captures the size of each individual firm, and it is raised to the power $1/(1-\eta)$ because of the increasing returns to scale at the firm level associated with data. Second, the term $(1-\alpha)x_{alloc}$ captures data. In particular, recall (e.g., from equation (31)) that

$$D_{it} = [\alpha x_{it} + (1-\alpha)\tilde{x}_{it}N_t]Y_{it} = N_t\left[\frac{\alpha x_{it}}{N_t} + (1-\alpha)\tilde{x}_{it}\right]Y_{it}. \tag{81}$$

As N_t grows large, the "own sharing" term $\alpha x_{it}/N_t$ disappears, and data is ultimately proportional to $(1-\alpha)\tilde{x}_{alloc}$. This is raised to the power η because of the usual D_i^{η} term in the production function for output, and it is further raised to the power $1/(1-\eta)$ because of the feedback effect through Y_{it}. Finally, the last term in equation (80) is $N_t = \psi_{alloc}L_t$ raised to the power $1 + \frac{1}{\sigma-1} + \frac{\eta}{1-\eta}$. This exponent captures the overall degree of increasing returns to scale in the economy: $1/(\sigma-1)$ comes from the standard variety effect associated with the nonrivalry of ideas while $\eta/(1-\eta)$ comes from the extra degree of increasing returns associated with the nonrivalry of data. This last effect enters directly because of the N_t term associated with data sharing in (81) that we just discussed.

This solution for GDP when there is some data sharing can be contrasted with the solution when data sharing is outlawed:

$$Y_t^{os} = [\nu_{os}\alpha^{\eta}x_{os}^{\alpha}]^{\frac{1}{1-\eta}}\left(\psi_{os}L_t\right)^{1+\frac{1}{\sigma-1}}. \tag{82}$$

Two main differences stand out. The first is related to the ν and ψ terms and the differences in the allocations in these two economies. But the second is perhaps surprising and potentially even more important: there is a fundamental difference in the role of scale between the allocations that involve data sharing and the outlaw-sharing equilibrium. In the allocations that involve data sharing, the exponent on L_t is $1 + \frac{1}{\sigma-1} + \frac{\eta}{1-\eta}$, while in the outlaw-sharing equilibrium, the additional returns associated with data sharing $\frac{\eta}{1-\eta}$ are absent. The reason for this can be seen directly in equation (81) above: when $\tilde{x} = 0$, the additional scale term associated with $(1-\alpha)\tilde{x}N_t$ disappears and the amount of data just depends on αx_{os}. That is, firms learn only from their own production and not from the N_t other firms in the economy.

The results for per capita income illustrate this even more clearly. In this econ-

omy, consumption per person equals output per person, Y_t/L_t. Dividing the equations above by L_t gives

$$c_t^{alloc} \propto L_t^{\frac{1}{\sigma-1}+\frac{\eta}{1-\eta}} \qquad \text{for } alloc \in \{sp, c, f\} \tag{83}$$

$$c_t^{os} \propto L_t^{\frac{1}{\sigma-1}}. \tag{84}$$

The exponents in these equations denote the degree of increasing returns to scale in the economies. When data is not shared, the degree of increasing returns equals the familiar $\frac{1}{\sigma-1}$, which is the standard increasing returns associated with the nonrivalry of ideas and the love of variety in a Romer / Dixit-Stiglitz environment. Without data sharing, that is the end of the story. However, when data is shared across firms, there is the additional scale effect of $\frac{\eta}{1-\eta}$.

The importance of this effect can be seen by taking logs and derivatives of these equations to obtain the growth rate of income and consumption per person along a balanced growth path:

$$g_c^{alloc} = \left(\frac{1}{\sigma-1} + \frac{\eta}{1-\eta}\right) g_L \qquad \text{for } alloc \in \{sp, c, f\} \tag{85}$$

$$g_c^{os} = \left(\frac{1}{\sigma-1}\right) g_L. \tag{86}$$

Even though this is a semi-endogenous growth setup in which standard policies have level effects but not growth effects, we see that data sharing is different. The allocations that involve data sharing feature faster long-run rates of economic growth.

Notice that the nature of data sharing matters for this result. If every firm shares with 10 others, then this looks like the "outlaw sharing" equilibrium because the number of firms benefiting from the data does not grow with the economy. Conversely, if all firms share their data with one quarter of the other firms, then this looks like the sharing economy: the number of firms benefiting from data increases as the economy grows larger.

In an economy in which firms do not share data, firms learn only from their own production. Because the entry cost is a fixed number of units of labor, the number of firms is directly proportional to the amount of labor in the economy. But this is just another way of saying that firm size is invariant to the overall population of the economy:

a bigger economy has more firms but not larger firms. This means that in the outlaw-sharing economy, there is no additional data benefit to having a larger economy, so the growth rate does not incorporate a boost from the increasing returns associated with the nonrivalry of data. Contrast this with an economy in which data is shared. In that case, the amount of data that each firm can learn from *is* an increasing function of the size of the economy. Therefore, the scale of the economy and the increasing returns associated with the nonrivalry of data interact.[3]

Data and Firm Production. This difference in the returns to scale shows up throughout the allocations. This can be seen, for example, in the comparisons of data used by each firm and aggregate data use:

$$D_{it}^{alloc} = [\nu_{alloc}(1-\alpha)\tilde{x}_{alloc}\psi_{alloc}L_t]^{\frac{1}{1-\eta}} \quad \text{for } alloc \in \{sp, c, f\} \tag{87}$$

$$D_{it}^{os} = [\nu_{os}\alpha x_{os}]^{\frac{1}{1-\eta}} \tag{88}$$

and

$$D_t^{alloc} = ND_i = [\nu_{alloc}(1-\alpha)\tilde{x}_{alloc}]^{\frac{1}{1-\eta}}(\psi_{alloc}L_t)^{1+\frac{1}{1-\eta}} \quad \text{for } alloc \in \{sp, c, f\} \tag{89}$$

$$D_t^{os} = [\nu_{os}\alpha x_{os}]^{\frac{1}{1-\eta}}\psi_{os}L_t. \tag{90}$$

The scale difference also shows up in firm production. Recall from the discussion about firm size measured by employment at the start of this section that firm employment is invariant to the size of the economy. What we see next, however, is that firm production is not invariant to the size of the economy when data is shared. In that case, firm production grows with the overall size of the economy because of the nonrivalry of data:

$$Y_{it}^{alloc} = \left[\nu_{alloc}(1-\alpha)^{\eta}\tilde{x}_{alloc}^{\eta}\right]^{\frac{1}{1-\eta}}(\psi_{alloc}L_t)^{\frac{\eta}{1-\eta}} \quad \text{for } alloc \in \{sp, c, f\} \tag{91}$$

$$Y_{it}^{os} = [\nu_{os}\alpha^{\eta}x_{os}^{\eta}]^{\frac{1}{1-\eta}}. \tag{92}$$

[3]Notice that this finding is robust to specifying the entry cost differently. For example, if the entry cost is such that the number of firms is $N = L^{\beta}$, then firm size will be $\frac{L}{N} = L^{1-\beta}$ and firm data will grow in proportion. Notice that β could be less than one or greater than one: it is possible that firm size is decreasing in the overall scale of the economy if varieties are easy to create. Contrast that with the data sharing case, in which each firm benefits from all data in the economy: $D_i = NY_i = N \cdot \frac{L}{N} = L$. That is, regardless of β, the full scale effect is passed through.

Wages, Profits, and Pricing. In the equilibrium allocations, i.e., $alloc \in \{c, f, os\}$, the factor income share of production labor and profits in aggregate GDP add to one and are given by

$$\left(\frac{w_t L_{pt}}{Y_t}\right)^c = \left(\frac{w_t L_{pt}}{Y_t}\right)^{os} = \frac{\sigma-1}{\sigma(1-\eta)}, \qquad \left(\frac{w_t L_{pt}}{Y_t}\right)^f = \frac{\sigma-1}{\sigma(1-\eta\frac{\epsilon-1}{\epsilon})} \tag{93}$$

$$\left(\frac{N_t \pi_t}{Y_t}\right)^c = \left(\frac{N_t \pi_t}{Y_t}\right)^{os} = \frac{1-\sigma\eta}{\sigma(1-\eta)}, \qquad \left(\frac{N_t \pi_t}{Y_t}\right)^f = \frac{1-\sigma\eta}{\sigma(1-\eta\frac{\epsilon-1}{\epsilon})}. \tag{94}$$

By comparison, recall from equation (17) that the aggregate production function for the economy is

$$Y_t = N_t^{\frac{1}{\sigma-1}} D_{it}^{\eta} L_{pt}. \tag{95}$$

Therefore, the marginal product of production labor multiplied by L_{pt} as a share of GDP from the social planner's perspective is equal to one. That is, as is standard in models with varieties, labor is underpaid relative to its social marginal product so that the economy can provide some profits to incentivize the creation of new varieties.

It is also interesting to compare the monopoly markup and pricing in the different equilibrium allocations. The price of a variety is

$$q_{it}^c = N_t^{\frac{1}{\sigma-1}} = (\psi_c L_t)^{\frac{1}{\sigma-1}} \tag{96}$$

$$p_{it}^c = \left(1 + \eta \cdot \frac{\sigma-1}{\sigma(1-\eta)}\right) N_t^{\frac{1}{\sigma-1}} = \left(1 + \eta \cdot \frac{\sigma-1}{\sigma(1-\eta)}\right) (\psi_c L_t)^{\frac{1}{\sigma-1}} \tag{97}$$

$$p_{it}^f = N_t^{\frac{1}{\sigma-1}} = (\psi_f L_t)^{\frac{1}{\sigma-1}} \tag{98}$$

$$p_{it}^{os} = N_t^{\frac{1}{\sigma-1}} = (\psi_{os} L_t)^{\frac{1}{\sigma-1}}. \tag{99}$$

Two points are worth noting. First, the effective price paid by consumers (i.e., incorporating the fact that they can sell their data) in the consumers-own-data allocation — q_{it}^c — and the actual price paid by consumers in the other allocations — p_{it}^f, p_{it}^{os} — are both equal to $N_t^{\frac{1}{\sigma-1}}$. Of course, N_t will differ across these allocations, but the point is that the consumer prices are both the same function of the number of firms. Moreover, there is no "markup" term that shows up in this expression. This is a feature of the exogenous labor supply in our environment. One way or the other, labor can only be used to

produce goods and so the monopoly markup does not result in a misallocation of labor. This is true even though firms internalize that they have increasing returns because of the learning-by-doing associated with data. This is something we will explore more fully in a future version on the paper.

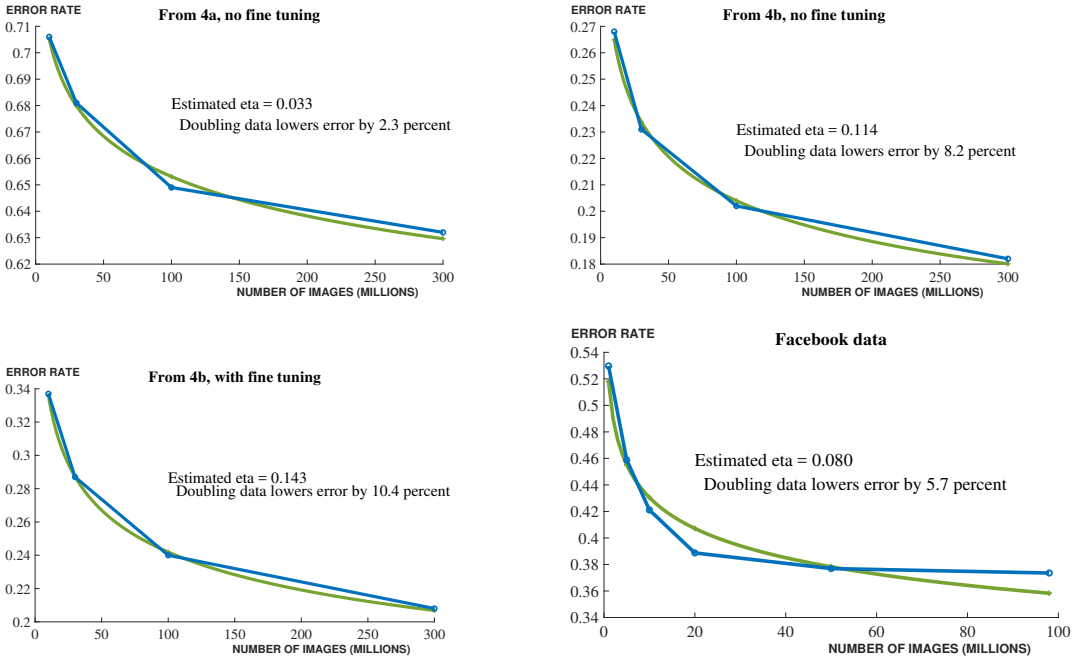
Second, notice that the price that firms receive for their sales in the consumers-own-data equilibrium, p_{it}^c, does involve a markup term given by $1 + \eta \cdot \frac{\sigma-1}{\sigma(1-\eta)}$. If $\eta = 0$, this term would drop out. Instead, it captures the fact that firms know that consumers can sell their data. Therefore, firms charge an additional markup over marginal cost to capture this revenue.

8. Numerical Example

We now provide a numerical example to illustrate the forces in the model. This should not be viewed as a formal calibration that can be compared quantitatively to facts about the U.S. economy. For example, our model assumes that all firms benefit from data equally and that data shared by each firm is equally useful. In this sense, the model might more naturally be compared to a particular industry, such as medical care or autonomous cars. Nevertheless, we find it useful to think about how large the various forces in the model might possibly be. To answer this question, we need to specify a value for η and values for the other parameters.

How Large is η? Sun, Shrivastava, Singh and Gupta (2017) study how the error rate in image recognition applications of machine learning changes with the number of images in the learning sample. They examine four different approaches with a number of images that ranges from 10 million to 300 million. While their paper does not report estimates of an elasticity that corresponds to η, it is relatively straightforward to compute such an estimate from their results. In particular, if we assume that the error rate is proportional to $M^{-\eta}$ where M is the number of images and that productivity equals the inverse of the error rate, then we can compute an estimate of η. Using their data together with a related exercise from Facebook from Joulin, van der Maaten, Jabri and Vasilache (2015) we obtain 5 different estimates of η, ranging from 0.033 to 0.143, with

Figure 2: Estimating η from Image Recognition Algorithms



Note: The parameter η is estimated from regressing the log of the error rate $1-mAP$ on the log number of images using data from Sun, Shrivastava, Singh and Gupta (2017) in the first three panels and from Joulin, van der Maaten, Jabri and Vasilache (2015) in the last panel. A fifth estimate from Figure 4a of Sun, Shrivastava, Singh and Gupta (2017) with fine tuning is omitted but yields an estimate of $\eta = 0.040$. The data are plotted in blue while the fitted log-linear curve is shown in green.

a mean of 0.082, as shown in Figure 2.[4] At this mean value, a doubling of the amount of data leads the error rate to fall by 5.9 percent. Notably, the power function fits well and there is no tendency (at least in the Google study) for the error rate to flatten at a high number of images. Furthermore, as data proliferates, firms will use and develop algorithms that make even better use of more data. Obviously, it would be valuable to use a broader set of applications in order to estimate η in different contexts.

Nevertheless, taking the midpoint of $\eta = 0.08$ as our baseline implies $\eta/(1-\eta) = 0.087$. Multiplying by a value of g_L of 2% implies that the steady state growth rate with data sharing would be around 0.17 percentage points higher than without data sharing.

[4]We are grateful to Abhinav Shrivastava and Chen Sun for providing the data points from their paper and the Facebook paper that we used to estimate η and for help interpreting the "mAP" metric.

Other parameters. Other parameter values for our example are reported in Table 2. We consider an elasticity of substitution of 5 implying that the degree of increasing returns in the economy is $\frac{1}{\sigma-1} = 0.25$ when there is no data sharing, rising to $\frac{1}{\sigma-1} + \frac{\eta}{1-\eta} = 0.34$ in the presence of data sharing. Population growth in advanced economies is around 1 percent per year, but the growth rate of R&D labor is closer to 4 percent; as a compromise, we choose $g_L = 0.02$. Combined with the returns to scale, this implies steady-state growth rates of consumption per person of 0.5 percent when data is not shared and 0.67 percent when data is shared across firms. Of course these are lower than what we see in advanced economies, but our model omits quality improvements within firms/varieties, so we probably should not match a higher growth rate.

Regarding the privacy cost parameters, κ and $\tilde{\kappa}$, Athey, Catalini and Tucker (2017) show that people express concerns about privacy but are willing to share once incentivized, even by a relatively small reward: a majority of MIT students in their survey were willing to share the email addresses of three close friends in exchange for a free pizza. Nevertheless, we give an important role to privacy. We set $\tilde{\kappa} = 0.20$, implying that having all of one's data shared with all firms is equivalent to a reduction in consumption of 10 percent.

We set $L_0 = 100$, corresponding to a workforce of around 100 million people; labor units are therefore millions of people. We set the rate of time preference to 3 percent (it must be larger than g_L). Entry requires $\chi = 0.01$ workers; because labor units are in millions of people, this corresponds to 10,000 people, and with an R&D share of the population of around 1 percent, this would mean 100 researchers. If all a variety's data is shared, this increases the rate of creative destruction by $\delta_0/2$, which we calibrate to 20 percent; absent any other death, this corresponds to an expected lifetime of 5 years. Finally, we set the weight on "own data" to $\alpha = 0.5$; this parameter plays very little role in our results.

8.1 Consumption-Equivalent Welfare

Growth rates are only one of the differences in our allocations. Consumers also care about the level of consumption and about privacy considerations. We use a consumption-equivalent welfare measure to summarize these differences.

Table 2: Parameter Values

Description	Parameter	Value
Importance of data	η	0.08
Elasticity of substitution	σ	5
Weight on privacy	$\kappa = \tilde{\kappa}$	0.20
Population level	L_0	100
Population growth rate	g_L	0.02
Rate of time preference	ρ	0.03
Labor cost of entry	χ	0.01
Creative destruction	δ_0	0.4
Weight on own data	α	1/2
Use of own data in NS	$\bar{x}$	1

Note: Baseline parameter values for the numerical example.

Along a balanced growth path, welfare is given by

$$U_{ss}^{alloc} = \frac{1}{\tilde{\rho}} \left(\log c_0^{alloc} - \frac{\tilde{\kappa}}{2} \tilde{x}_{alloc}^2 + \frac{g_c^{alloc}}{\tilde{\rho}} \right).$$

Notice that the x_{it} "own privacy" term drops out because it is scaled by $1/N$; recall equation (2). Let $U_{ss}^{alloc}(\lambda)$ denote steady-state welfare when we perturb the allocation of consumption by some proportion λ:

$$U_{ss}^{alloc}(\lambda) = \frac{1}{\tilde{\rho}} \left(\log(\lambda c_0^{alloc}) - \frac{\kappa}{2} x_{alloc}^2 + \frac{g_c^{alloc}}{\tilde{\rho}} \right).$$

Then consumption equivalent welfare λ^{alloc} is the fraction by which consumption must be decreased in the optimal allocation to deliver the same welfare as in some other allocation:

$$U_{ss}^{sp}(\lambda^{alloc}) = U_{ss}^{alloc}(1).$$

Moreover, it is straightforward to see that this consumption equivalent welfare measure

is given by

$$\log \lambda^{alloc} = \underbrace{\log c_0^{alloc} - \log c_0^{sp}}_{\text{Level term}} - \underbrace{\frac{\tilde{\kappa}}{2}\left(\tilde{x}_{alloc}^2 - \tilde{x}_{sp}^2\right)}_{\text{Privacy term}} + \underbrace{\frac{g_c^{alloc} - g_c^{sp}}{\tilde{\rho}}}_{\text{Growth term}}. \tag{100}$$

That is, there is an additive decomposition of consumption-equivalent welfare into terms reflecting differences in the level of consumption, the extent of privacy, and the growth rate.

8.2 Results from the Numerical Example

The top panel of Table 3 shows summary statistics for the numerical example. The fraction of data that is shared differs dramatically across the allocations. The social planner chooses to share 66 percent of data, even taking privacy considerations into account. When consumers own data, they share less at 59 percent.[5] As discussed earlier, the reason for this difference is the monopoly markup $\frac{\sigma}{\sigma-1} = 1.2$ that leads the price at which consumers sell their data to be too low relative to what the planner would want. These "high sharing" allocations can be contrasted with the bottom two allocations. When firms own data, they distort the use of data in two ways. First, they use 100 percent of their "own" data, more than what consumers or the planner would desire. In this sense, firms do not satisfy the privacy concerns of consumers. Second, in terms of sharing with *other* firms, there is too little sharing relative to the planner: firms share only 16 percent of their data with other firms. The key factor in this decision is creative destruction. And of course, when data sharing is outlawed, the allocation features no data sharing.

The next two columns of the top panel show that firm size and the number of varieties differ across the allocations. When firms own data or when sharing is outlawed, the rate of creative destruction is low (see the last column). This has two countervailing effects. On the one hand, it raises the present value of profits, which tends to promote entry. On the other hand, it reduces the boost to entry associated with business stealing. When $\rho > g_L$ the business stealing effect dominates and higher rates of creative destruction lead to more entry and smaller firms. This can be seen in the top panel of

[5] In the planner and consumers-own-data allocations $x = \tilde{x}$ because we've set $\kappa = \tilde{\kappa}$ and $\alpha = 1/2$.

Table 3: Numerical Example

Summary Statistics:

Allocation	Data Sharing "own" x	Data Sharing "others" $\tilde{x}$	Firm size ν	Variety $N/L=\psi$	Consu-mption c	Growth g	Creative Destruct. δ
Social Planner	0.66	0.66	1304	665	18.6	0.67%	0.0870
Consumers Own Data	0.59	0.59	1482	594	18.3	0.67%	0.0696
Firms Own Data	1	0.16	1838	491	16.0	0.67%	0.0052
Outlaw Sharing	1	0	2000	455	7.3	0.50%	0

Consumption-Equivalent Welfare:

Allocation	Welfare λ	$\log\lambda$	Level term	Privacy term	Growth term
Optimal Allocation	1	0	..	..	..
Consumers Own Data	0.9886	-0.0115	-0.0202	0.0087	0.0000
Firms Own Data	0.8917	-0.1146	-0.1555	0.0409	0.0000
Outlaw Sharing	0.3429	-1.0703	-0.9399	0.0435	-0.1739

Note: The table reports statistics from our numerical example for the different allocations using the parameter values in Table 2. The top panel shows baseline statistics along a balanced growth path. Firm size is multiplied by 10^6 and therefore is measured in people. The bottom panel reports consumption equivalent welfare calculated according to equation (100). In particular, λ is the fraction by which consumption must be decreased in the optimal allocation to deliver the same welfare as in some alternative allocation.

Table 3, where the number of varieties is higher when consumers own data than in the two limited-sharing allocations. Similarly, firm size is smaller when consumers own data.

The outlaw-sharing equilibrium features a smaller scale effect, which shows up both in economic growth being slower and in the overall level of consumption being substantially lower.

The bottom panel of Table 3 shows the welfare decomposition using the baseline parameter values. The allocation in which data sharing is outlawed is stunningly inferior: consumption-equivalent welfare is only 1/3 that of the social planner. A small part of this is the growth rate differential, but the bulk comes from distortions to the level of consumption, most importantly the missing scale effect associated with data sharing. Because data is nonrival, there are large social gains associated with sharing data, even in the presence of privacy concerns. The optimal trade off between sharing and privacy still features some data sharing: efficiency requires that nonrival factors be shared to some extent. Laws that prohibit such sharing can have dramatic effects, reducing incomes by a factor of three or more.

One institution that appropriately balances these concerns is assigning ownership of data to consumers. Data sharing is close to that of the social planner and consumption-equivalent welfare falls short of optimal by just a percentage point or two. Consumers take their own privacy considerations into account but are incentivized by markets to sell their data broadly to a range of firms, leading them to nearly-optimal allocations.

In contrast, when firms own data, concerns about creative destruction sharply limit the amount of data they sell to other firms. While this generates some privacy benefits, equal to about 4 percent of consumption, the social loss from nonrival data not being used by other firms is much larger. Equilibrium welfare is just 89% of optimal when firms own data, compared to 99% of optimal when consumers own data. Failing to appropriately take advantage of the nonrivalry of data leads consumption to be lower by more than 15 percent along the balanced growth path, even in this example in which there are sharply diminishing returns to additional data.

9. Discussion

Implications for IO. Several issues related to antitrust and IO are raised by this framework. First, because firms see increasing returns to scale associated with data and, perhaps more importantly, because of the nonrivalry of data, firms in this economy would like to merge into a single economy-wide firm. Our paper provides a concept of a firm as the boundary of data sharing and the nonrivalry of data may create strong pressures to increase scale.

Second, data may serve as a barrier to entry. A natural concern about the limited-sharing allocations is that as a firm accumulates data, this may make it harder for other firms to enter. In our framework, this force appears somewhat mechanically through the dependence of the rate of creative destruction $\delta(\tilde{x})$ on the amount of data sharing. It would be interesting in future research to consider this force more explicitly, say, in a quality ladder model.

The Boundaries of Data Diffusion: Firms and Countries. At the beginning of the paper, we noted that both ideas and data are nonrival. Both can be expressed as bit strings, and it is natural to wonder about the differences between them. For example, while ideas give rise to increasing returns and people create ideas, growth theory does not typically suggest that Luxembourg and Hong Kong should be much poorer than Germany and China because of their relatively small size. Instead, the view is that ideas diffuse across countries, at least eventually and in general, so that the relevant scale is the scale of the global market of connected countries rather than that of any individual economy.

Data may be different. For example, it seems much easier to monitor and limit the spread of data than to limit the spread of ideas. Perhaps this is because ideas, in order to be useful, need to be embodied inside people in the form of human capital (which makes it inherently hard to keep from spreading). In contrast, data can be encrypted and tightly controlled.

This raises an interesting question about whether the quantity of data that an organization has access to can serve as an important productivity advantage. This could apply to firms or even to countries. For example, the Chinese economy is large. Could access to the inherently larger quantities of data associated with a large population

provide an advantage. Lee (2018) suggests "China has more data than the US — way more. Data is what makes AI go. A very good scientist with a ton of data will beat a super scientist with a modest amount of data." Similarly, a government that places a lower weight on consumer privacy might induce more data sharing, leading to a higher level of GDP (but perhaps lower welfare). Or, in an industry context with trade, could this difference lead to firms (e.g., in China) having a distinct productivity advantage in data-intensive products?

10. Conclusion

The economics of data raises many important questions. Privacy concerns have appropriately received a great deal of attention recently. Our framework supports this: when firms own data, they may overuse it and not adequately respect consumer privacy.

But another important consideration arises from the nonrivalry of data. Because data does not get depleted, there are large social gains to allocations in which the same data is used by multiple firms simultaneously.

An analogy may be helpful. Because capital is rival, each firm must have its own buliding, each worker needs her own desk and computer, and each warehouse needs its own collection of forklifts. But if capital were nonrival, it would be as if every worker in the economy could use the *entire* aggregate stock of capital at the same time. Clearly this would produce tremendous economic gains. This is what is possible with data. Because data is nonrival, it is technologically feasible for all medical data to be used by each health researcher and for all driving data to be used by every machine learning algorithm. Obviously there may be incentive reasons why it is inefficient to have all data shared with all firms. But the equilibrium in which firms own data and sharply limit its use by other firms may also be inefficient. Our numerical examples suggest that these costs can be large.

Government restrictions that, out of a concern for privacy, outlaw data sharing entirely may be particularly harmful. Instead, our analysis suggests that giving the data property rights to consumers can lead to allocations that are close to optimal. Consumers appropriately balance their concerns for privacy against the economic gains that come from selling data to all interested parties.

References

Acquisti, Alessandro, Curtis Taylor, and Liad Wagman, "The Economics of Privacy," *Journal of Economic Literature*, June 2016, *54* (2), 442–92.

Arrieta Ibarra, Imanol, Leonard Goff, Diego Jimenez Hernandez, Jaron Lanier, and E. Glen Weyl, "Should We Treat Data as Labor? Moving Beyond "Free"," *American Economic Association Papers and Proceedings*, 2018, pp. 38–42.

Athey, Susan, Christian Catalini, and Catherine Tucker, "The Digital Privacy Paradox: Small Money, Small Costs, Small Talk," Working Paper 23488, National Bureau of Economic Research June 2017.

Bajari, Patrick, Victor Chernozhukov, Ali Hortasu, and Junichi Suzuki, "The Impact of Big Data on Firm Performance: An Empirical Investigation," Working Paper 24334, National Bureau of Economic Research February 2018.

Begenau, Juliane, Maryam Farboodi, and Laura Veldkamp, "Big Data in Finance and the Growth of Large Firms," 2017. NYU manuscript.

Benassy, Jean-Pascal, "Taste for Variety and Optimum Production Patterns in Monopolistic Competition," *Economics Letters*, 1996, *52* (1), 41–47.

Bollard, Albert, Peter J. Klenow, and Huiyu Li, "Entry Costs Rise with Development," 2016. Stanford University manuscript.

Chiou, Lesley and Catherine Tucker, "Search Engines and Data Retention: Implications for Privacy and Antitrust," Working Paper 23815, National Bureau of Economic Research September 2017.

Farboodi, Maryam and Laura Veldkamp, "Long Run Growth of Financial Technology," NBER Working Papers 23457, National Bureau of Economic Research, Inc May 2017.

Gentzkow, Matthew, Bryan T. Kelly, and Matt Taddy, "Text as Data," Working Paper 23276, National Bureau of Economic Research March 2017.

Joulin, Armand, Laurens van der Maaten, Allan Jabri, and Nicolas Vasilache, "Learning Visual Features from Large Weakly Supervised Data," *CoRR*, 2015, *abs/1511.02251*.

Lee, Kai-Fu, "Tech companies should stop pretending AI won't destroy jobs," *MIT Technology Review*, February 21 2018.

Romer, Paul M., "Endogenous Technological Change," *Journal of Political Economy*, October 1990, *98* (5), S71–S102.

Sun, Chen, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta, "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era," *CoRR*, 2017, *abs/1707.02968*.

Varian, Hal, "Artificial Intelligence, Economics, and Industrial Organization," in Ajay K. Agrawal, Joshua Gans, and Avi Goldfarb, eds., *The Economics of Artificial Intelligence: An Agenda*, University of Chicago Press, 2018.

A. Appendix

To be completed.