Journal of Economic Dynamics & Control 141 (2022) 104395

Contents lists available at ScienceDirect

Journal of Economic Dynamics & Control

journal homepage: www.elsevier.com/locate/jedc

Endogenous growth under multiple uses of data☆

Lin William Cong[a,*], Wenshi Wei[b], Danxia Xie[b], Longtian Zhang[c]

[a] SC Johnson College of Business, Cornell University, US
[b] Institute of Economics, Tsinghua University, China
[c] School of International Trade and Economics, Central University of Finance and Economics, China

ARTICLE INFO

Article history:
Available online 26 April 2022

Keywords:
Big data
Endogenous growth
Innovation
Nonrivalry
Privacy

ABSTRACT

We model a dynamic data economy with fully endogenous growth where agents generate data from consumption and share them with innovation and production firms. Different from other productive factors such as labor or capital, data are nonrival not only among firms but also in their uses across sectors, which affect both the level and growth of economic outputs. Despite the vertical nonrivalry, the innovation sector dominates the production sector in data usage and contribution to growth because (i) innovations are cumulative and benefit from data that are durable and dynamically nonrival; and (ii) innovations "desensitize" raw data into knowledge when entering production, which allays consumers' privacy concerns. Data uses in both sectors interact in generating allocative distortions and an apparent substitutability due to labor's rival usage across sectors and complementarity with data. Consequently, growth rates diverge under a social planner and a decentralized equilibrium, which is novel in the literature and has policy implications. Specifically, consumers' failure to fully internalize knowledge spillover while bearing privacy costs, combined with firms' market power, leads to an underpricing of data and inefficient data supply, causing underemployment in the innovation sector and suboptimal long-run growth. Improving data usage efficiency is ineffective in mitigating the underutilization of data, but interventions in the data market and direct subsidies hold promises.

1. Introduction

With the proliferation of Internet-based businesses and digital platforms, data are widely recognized as an important factor in the productive sectors and the long-run growth of any modern economy.[1] Because extant studies adopt the framework of semi-endogenous growth and focus on the use of data in prediction and production, one can neither quantify the

☆ All authors contribute equally to this work. Cong acknowledges financial support from the Ewing Marion Kauffman Foundation (RG-201805-4237); Xie acknowledges financial support from the National Natural Science Fundation of China (71973076); Zhang acknowledges financial support from the Academic Development Fundation in School of International Trade and Economics. We thank Shota Ichihashi and Liyan Yang for helpful discussions and comments, as well as the conference participants at the 2021 Conference on Markets and Economies with Information Frictions and the reviewers at the 2022 Western Finance Association Annual Conference. Yao Hou, Jinglei Huang, Desheng Ma, Ramtin Salamat, Daisy Shu assisted in proofreading the paper. All remaining errors are our own.

* Corresponding author.
E-mail addresses: will.cong@cornell.edu (L.W. Cong), wei_wenshi@163.com (W. Wei), xiedanxia@tsinghua.edu.cn (D. Xie), zhanglongtian@cufe.edu.cn (L. Zhang).

[1] Big data research has saved over 100 billion Euros for Europe and it has reduced medical care cost of the United States by 8% or 300 billion dollars every year (Manyika et al., 2011); Cong et al. (2021a) survey business applications of alternative data.

https://doi.org/10.1016/j.jedc.2022.104395
0165-1889/

aggregate level of data usage nor study how multiple uses of data interact to impact economic growth. Yet, in practice, data are used both in production and in innovation (e.g., in the R&D sector), and the amount of data usage directly drives economic growth. For example, while data-intensive industries process large quantity of data to directly improve their products and services (e.g., self-driving cars and VR technologies), it is still the case that universities, industry research institutes, and open-source initiatives advance the frontiers of fundamental data science (e.g., Google TensorFlow and GitHub), which adds to our understanding of general purpose technologies, such as AI and automation.

In this paper, we offer the first analysis on the endogenous growth of a data economy with multiple data uses, explicitly modeling the hallmark characteristics of data usage in innovation. This includes the endogenous supply of data, the dynamic accumulation of knowledge, and the nonrival data usage in other sectors. Agents in our model produce data through their consumption activities and sell data to innovators and production firms while being averse to potential privacy violations and data abuse. Meanwhile, they also supply labor to both the innovation sector and production sectors. Unlike labor, data are nonrival not only horizontally (Ichihashi, 2020; Jones and Tonetti, 2020) or over time (Cong et al., 2021b), but also vertically across sectors: Data usage in production does not limit their usage in innovation. That said, data uses across sectors interact in nontrivial ways to affect consumers' data contribution, privacy costs, and labor allocation across sectors; thus, affecting the long-run growth of the economy. Moreover, consumers endogenously contribute data, balancing the privacy costs incurred in their uses. We characterize the equilibria on the balanced growth path (BGP) and compare the importance of the multiple uses of data for economic growth.

The larger quantity of data the consumers share, the more severe the creative destruction is with the emergence of new varieties and firms. Because consumers who are only paid by incumbent firms fail to get fully compensated for the knowledge spillover benefits of their data contribution in the creation of future firms, the contemporaneous supply of data is suboptimally low. The growth rate is then endogenously lower in a decentralized economy than that in a social planner's solution. Meanwhile, the suboptimal usage of data lowers the productivity of labor more in the innovation sector than that in the production sector, which distorts the allocation of labor towards the production sector. This further slows economic growth because knowledge is only accumulated in the innovation sector. Market power of production firms, which underprices consumers' data contribution, aggravates the inefficiency, leading to underemployment in the innovation sector and overemployment in the production sector.

We further derive the long-run usage of data and decompose their contribution for the two sectors. The innovation sector dominates in terms of data usage and the contribution of data to growth for two reasons: (i) Data are dynamically nonrival and add to knowledge creation that is cumulative (also highlighted in Cong et al., 2021b). (ii) Innovations have a "desensitizing" effect on data usage in the production sector, i.e., knowledge generated from data usage in the innovation sector can be repeatedly utilized in the future and will not bring any additional privacy costs. When the same data used in the innovation sector are used again in the production sector, consumers' privacy costs are greater. But, if the data generate knowledge in the innovation sector (by expanding the innovation possibility frontier), which then enters the production sector, the privacy costs do not increase because the accumulated knowledge does not reveal private information.

Emerging studies on the data economy typically focus on how data add to contemporaneous production (Jones and Tonetti, 2020) and prediction (Farboodi and Veldkamp, 2021) without emphasizing long-run growth. While Cong et al. (2021b) introduces data usage in the innovation sector with knowledge accumulation and dynamic data nonrivalry, all of these models adopt a semi-endogenous growth framework and do not link the aggregate level of data usage to growth; nor do they offer insights on the multiple uses of data. In this paper, we use a fully endogenous model to isolate the impact of data on the economy from that of population growth. Our model generates differential growth rates and highlights the inefficiencies of a decentralized equilibrium relative to a social planner's solution. Unlike Jones and Tonetti (2020) and Cong et al. (2021b), which feature an overuse of data either in the innovation sector or in the production sector, our model reveals a general underutilization of data. Also, while both our paper and Farboodi and Veldkamp (2021) show that the contribution of data to growth is bounded, the latter relies on an informativeness bound in terms of making predictions from data, which we highlight a complementary channel of privacy concerns.

Our paper therefore adds to the large literature on economic growth. After Romer (1990)'s seminal study introducing innovation into the economy to generate long-run growth, Jones (1995) adds the knowledge spillover effect to derive a semi-endogenous result: The long-run growth rate now depends on the growth rate of population instead of the population level. Subsequent studies, such as Stokey (1998) and Jones (2016), also adopt semi-endogenous models, but they leave much to be desired since zero population growth in these models leads to zero growth of the economy, which is counterfactual. Our endogenous growth model circumvents this issue and highlights the distinguishing features of data.

This paper also contributes to the literature on the economics of data, information, and privacy in general. Previous studies, such as Hirshleifer (1971), Admati and Pfleiderer (1990), and Murphy (1996), focus on social values, sales, and property rights of information, while recent studies such as Akçura and Srinivasan (2005), Casadesus-Masanell and Hervas-Drane (2015), and Fainmesser et al., 2021 strive to connect digital information with privacy issues. More recently, Cong and Mayer (2022) analyze data reinforcement effects and platform competition in order to evaluate various policies concerning data privacy and sharing. Meanwhile, other studies, such as Easley et al. (2019), Jones and Tonetti (2020), and Ichihashi (2020, 2021a), highlight the nonrivalry of data horizontally and discuss competitions among data platforms or intermediaries. In this respect, Cong et al. (2021b) emphasizes the dynamic nonrivalry of data and their role in knowledge accumulation. We complement this perspective by incorporating the vertical and cross-sector nonrivalry of data, while allowing for data privacy concerns and multiple uses of data.

The remainder of this paper is organized as follows: Section 2 sets up the model and defines the equilibrium. Section 3 characterizes the social planner's solution, the decentralized economy, and the various usages of data, before drawing policy implications. Section 4 provides numerical results in terms of the contribution of data to growth through their multiple uses. It also further discusses the issue of misallocation and model robustness under alternative settings. Finally, Section 5 concludes.

2. The model

In this section, we first introduce agents in the economy, i.e., representative consumers, incumbent firms, innovation sector, and data intermediary. We then define the equilibrium.

2.1. Representative consumers

Homogeneous agents (consumers) live in a continuous-time economy. We assume that the size of the agent population is a constant L in order to highlight endogenous growth—economic growth that is not driven by population growth. As argued in Jones and Tonetti (2020) and Cong et al. (2021b), each consumer produces data as a by-product of consumption and chooses the quantity of data to sell for profit, while incurring disutility due to concerns about privacy violations and data breaches.[2] Each consumer's instantaneous utility is:

$$u(c(t), d(t)) = \ln c(t) - \frac{\kappa d(t)^2}{2}, \quad \text{where} \quad c(t) \equiv \left(\int_0^{N(t)} c(v,t)^{\frac{\gamma-1}{\gamma}} \mathrm{d}v \right)^{\frac{\gamma}{\gamma-1}} \tag{1}$$

is the consumption index, i.e., the CES aggregate of the consumption of the differentiated varieties. $\gamma > 1$ is the elasticity of substitution, $c(v,t)$ is the consumption level of variety v at time t (which costs $p(v,t)$), $N(t)$ is the total number of varieties of products, and $d(t)$ is the quantity of data shared or sold. Note that $\kappa \in (0,1)$ captures the extent of privacy concerns, such as expected losses from leakage, the extra costs when a platform uses consumer data for price discrimination, the hacking of a centralized database, or the discomfort of existing under a Big Brother's constant surveillance. The disutility increases and is convex with respect to the quantity of data sold.

Each consumer supplies one unit of labor inelastically in exchange for an endogenous wage $w(t)$. They hold assets $a(t)$ that earn returns at an interest rate $r(t)$, and they sell data $d(t)$ at an endogenous price $p_d(t)$. The consumers' utility maximization problem is then:

$$\max_{\{c(v,t),d(t)\}} \int_0^\infty e^{-\rho t} u(c(t), d(t)) \mathrm{d}t, \tag{2}$$

subject to

$$\dot{a}(t) = r(t)a(t) + w(t) + p_d(t)d(t) - \int_0^{N(t)} p(v,t)c(v,t)\mathrm{d}v$$

and

$$d(t) \le g(c(t)). \tag{3}$$

Here, ρ captures the consumers' time preference. Constraint (3) dictates that consumers cannot supply more data than what they generate. $g(c(t))$ is an exogenous general function for data generating process.

Without loss of generality, we normalize the following price index to one:

$$P(t) \equiv \left(\int_0^{N(t)} p(v,t)^{1-\gamma} \mathrm{d}v \right)^{\frac{1}{\gamma-1}} = 1. \tag{4}$$

Then, the Euler equations for consumption and data sales from Hamiltonian are simply:

$$\frac{\dot{c}(t)}{c(t)} = r(t) - \rho \tag{5}$$

and

$$\frac{\dot{p}_d(t)}{p_d(t)} - \frac{\dot{d}(t)}{d(t)} = r(t) - \rho.$$

[2] For an empirical quantification of the value of privacy, see Tang (2019).

2.2. Incumbent firms

Each incumbent firm owns a patent and produces a distinct variety $v \in [0, N(t)]$ through buying data sets $D(v,t)$ from a data intermediary (whom we will introduce shortly) and hiring labor $L_E(v,t)$. Each firm takes the price of data sets and the wage of labor as given. The nonrival nature of data implies that, in contrast with other productive inputs, such as labor or capital, the duplication and repeated usage of data incur negligible costs. Consequently, per capita output no longer depends on per capita data usage but rather on the total quantity of data bought by the firms.

All of the incumbent firms constitute the production sector. Similar to the consumption index shown in (1), we define the aggregate output $Y(t)$ as follows:

$$Y(t) \equiv \left(\int_0^{N(t)} Y(v,t)^{\frac{\gamma-1}{\gamma}} \mathrm{d}v \right)^{\frac{\gamma}{\gamma-1}}. \tag{6}$$

With the FOCs from the consumers' problem and the price index shown in (4), a firm producing variety v takes the following demand function as given:

$$p(v,t) = \left(\frac{c(t)}{c(v,t)} \right)^{\frac{1}{\gamma}} = \left(\frac{Y(t)}{Y(v,t)} \right)^{\frac{1}{\gamma}}.$$

We denote the market value of the incumbent firm v as $V(v,t)$. The profit maximization then becomes:

$$r(t)V(v,t) = \max_{\{L_E(v,t), D(v,t)\}} \left(\frac{Y(t)}{Y(v,t)} \right)^{\frac{1}{\gamma}} Y(v,t) - w(t)L_E(v,t) - q_d^{prod}(v,t)D(v,t) + \dot{V}(v,t), \tag{7}$$

where the price of data sets faced by incumbent firms is denoted as $q_d^{prod}(v,t)$. In this optimization problem, the production function capturing the nonrival nature of data in production is defined as follows:

$$Y(v,t) = L_E(v,t)D(v,t)^{\eta}, \tag{8}$$

where parameter $\eta \in (0,1)$ captures the importance of data in production. We assume that data have diminishing marginal returns, consistent with the theoretical foundation in Farboodi and Veldkamp (2021) and the empirical evidence or calibration in Sun et al. (2017) and Jones and Tonetti (2020).

With firms and products symmetrically entering the economy, FOCs of incumbent firms are:

$$\left(1 - \frac{1}{\gamma}\right) \left(\frac{Y(t)}{Y(v,t)} \right)^{\frac{1}{\gamma}} \frac{Y(v,t)}{L_E(v,t)} = w(t) \tag{9}$$

and

$$\left(1 - \frac{1}{\gamma}\right) \eta \left(\frac{Y(t)}{Y(v,t)} \right)^{\frac{1}{\gamma}} \frac{Y(v,t)}{D(v,t)} = q_d^{prod}(v,t). \tag{10}$$

Since the price, outputs, and profits are equal for all incumbent firms, we simplify notations by writing $V(t) = V(v,t)$ and $D(t) = D(v,t)$ for all v and t in the remainder of this paper.

2.3. Innovation sector

In contrast with Jones and Tonetti (2020), potential entrants, also known as the innovation sector, can invent new varieties by using data and employing R&D labor, as in Cong et al. (2021b). Specifically, we set the innovation sector production function and the dynamics of the innovation frontier as:

$$\dot{N}(t) = \varepsilon N(t) L_R(t)^{1-\xi} D(t)^{\xi}, \tag{11}$$

where $L_R(t)$ is labor employed in the innovation sector, $\varepsilon > 0$ captures the innovation efficiency, and $\xi \in (0,1)$ is the relative importance of data $D(t)$. To generate endogenous growth, we adopt the coefficient for the knowledge spillover effect as in Romer (1990) (with the exponent of N being 1). Also, we assume a Cobb-Douglas combination of bundles of data and labor described in (11) which captures a competitive innovation process while maintaining free-entry conditions.

The potential entrants take the price of data, $q_d^{inno}(t)$, as given and maximize their expected profit by choosing the quantity of data to buy and the units of labor to employ:

$$\max_{\{L_R(t), D(t)\}} \dot{N}(t)V(t) - w(t)L_R(t) - q_d^{inno}(t)D(t), \tag{12}$$

where $D(t)$ denotes the data potential entrants purchase, which reflects the idea that all agents can utilize the same data. The FOCs with respect to labor and data usage are then:

$$(1-\xi)\varepsilon N(t) L_R(t)^{-\xi} D(t)^{\xi} V(t) = w(t) \tag{13}$$

and

$$\xi \varepsilon N(t) L_R(t)^{1-\xi} D(t)^{\xi-1} V(t) = q_d^{inno}(t). \tag{14}$$

2.4. Data intermediary

Following Jones and Tonetti (2020), we introduce a data intermediary industry with competitive entry. One data intermediary collects data from consumers at a given price $p_d(t)$ and integrates data into a data set $D(t)$. The intermediary maximizes profits by choosing the quantity of data to buy from consumers and the discriminatory price at which it sells data sets to incumbent firms and potential entrants, respectively. The data intermediary makes zero profit in the presence of free entry. Thus, we characterize data intermediation through a cost minimization problem and a zero-profit condition. The intermediary solves:

$$\min_{\{d(t)\}} p_d(t)d(t)L, \tag{15}$$

subject to

$$D(t) \leq Ld(t). \tag{16}$$

And the zero profit condition is:

$$\int_0^{N(t)} q_d^{prod}(v,t)D(t)\mathrm{d}v + q_d^{inno}(t)D(t) = p_d(t)d(t)L, \tag{17}$$

where $q_d^{prod}(v,t)$ is a discriminatory price subject to the demand curve $q_d^{prod}(D(t))$ in (10), and $q_d^{inno}(t)$ is subject to the demand curve $q_d^{inno}(D(t))$ in (14). Equation (16) and (17) together represent the idea that, with data nonrivalry, the intermediary can profit by combining all consumers' data and selling them to production firms and entrant innovators simultaneously.

2.5. Equilibrium definition

An equilibrium in a decentralized economy consists of quantities $\{c(t), Y(t), N(t), a(t), d(t), D(t), L_R(t)\}$, $\{c(v,t), Y(v,t), L_E(v,t)\}$, and prices $\{p_d(t), q_d^{inno}(t), w(t), r(t), V(t)\}$, $\{p(v,t), q_d^{prod}(v,t)\}$, where $v \in [0, N(t)]$, such that:

(i) $\{c(v,t), c(t), a(t), d(t)\}$ maximize consumers' utility in (2), $\{Y(v,t), Y(t), L_E(v,t), D(t), p(v,t), V(t)\}$ maximize the discounted value of incumbent firms in (7), $\{L_R(t), D(t)\}$ maximize the expected return of new entrants in (12), and $\{q_d^{prod}(v,t), q_d^{inno}(t)\}$ minimize the cost of data intermediary in (15) and (17).

(ii) $\{w(t)\}$ clears the labor market with $\int_0^{N(t)} L_E(v,t)\mathrm{d}v + L_R(t) = L$, $\{r(t)\}$ clears the asset market with $a(t)L = N(t)V(t)$, and $\{p_d(t)\}$ clears the data market with $d(t)L = D(t)$. $\{N(t)\}$ follows the R&D production function in (11).

3. Data uses for endogenous growth

We characterize the equilibria under the optimal allocation and in a decentralized economy, respectively, to identify inefficiencies due to knowledge accumulation externality and market power. We then contrast multiple uses of data with the singular use of data in either the production or innovation sector, before drawing implications for regulation and policy intervention.

3.1. Optimal allocation

For simplicity, we first discuss the multiple uses of data under the optimal allocation. We denote the fraction of labor hired in the innovation sector by $l_R(t)$ and the fraction in the incumbent firm v by $l_E(v,t)$, then the labor market clearing becomes $\int_0^{N(t)} l_E(v,t)\mathrm{d}v + l_R(t) = l_E(t) + l_R(t) = 1$. The planner needs to determine the labor allocation and the quantity of data that consumers contribute, solving:

$$\max_{\{l_E(t), d(t)\}} \int_0^{\infty} e^{-\rho t} L\left(\ln c(t) - \frac{\kappa d(t)^2}{2}\right)\mathrm{d}t,$$

such that

$$\begin{aligned} c(t) &= Y(t)/L, \\ Y(t) &= N(t)^{\frac{1}{\gamma-1}} l_E(t) L^{1+\eta} d(t)^{\eta}, \\ \dot{N}(t) &= \varepsilon l_R(t)^{1-\xi} L d(t)^{\xi} N(t), \\ 1 &= l_E(t) + l_R(t), \\ d(t) &\leq g(c(t)) \equiv \bar{d}(t). \end{aligned} \tag{18}$$

Note that in (18), we denote the maximum data available by $\bar{d}(t)$.

The Hamiltonian for the optimal allocation is:

$$\mathcal{H}(l_E(t), d(t), N(t), \lambda(t)) = \ln\left[N(t)^{\frac{1}{\gamma-1}} l_E(t) L^\eta d^\eta\right] - \frac{\kappa d(t)^2}{2} + \varepsilon\lambda(t) l_R^{1-\xi} L d(t)^\xi N(t).$$

Then, the FOC with respect to $d(t)$ is:

$$\begin{aligned}\frac{\partial \mathcal{H}}{\partial d(t)} &= -\kappa d(t) + \frac{\eta}{d(t)} + \lambda(t)\varepsilon\xi l_R(t)^{1-\xi} L d(t)^{\xi-1} N(t) \\ &= \underbrace{-\kappa d(t)}_{\text{Marginal disutility}} + \underbrace{\frac{\eta}{d(t)}}_{\text{Marginal production}} + \underbrace{\lambda(t)\xi\frac{\dot{N}(t)}{d(t)}}_{\text{Marginal innovation}}.\end{aligned} \tag{19}$$

In contrast with Jones and Tonetti (2020) as well as Cong et al. (2021b), the planner in our model seeks to share data since data facilitate both production and innovation.

Due to the upper limit of data shown by the constraint (18), for some period t', the maximum data available for the social planner $\bar{d}(t') = g(c(t'))$ may not be large enough to satisfy $\kappa\bar{d}(t')^2 - \eta > 0$, which is the sum of the first two terms in (19). In these corner solutions, the social planner opts to share all of the data available, and the quantity of shared data increases synchronously with consumption (according to $\bar{d}(t) = g(c(t))$). Then, the increasing quantity of data in both the production sector and the innovation sector interact with one another to fuel economic growth, which is clear when we express $Y(t)$ in terms of growth rate and derive the following results:

$$\begin{aligned}\frac{\dot{Y}(t)}{Y(t)} &= \frac{\gamma}{\gamma-1}\frac{\dot{N}(t)}{N(t)} + \frac{\dot{Y}(v,t)}{Y(v,t)} \\ &= \underbrace{\frac{1}{\gamma-1}\varepsilon l_R(t)^{1-\xi} L d(t)^\xi}_{\text{Growth from innovation sector}} + \underbrace{\frac{\dot{l}_E(t)}{l_E(t)} + \eta\frac{\dot{d}(t)}{d(t)}}_{\text{Growth from production sector}}.\end{aligned} \tag{20}$$

When $\kappa\bar{d}(t)^2 - \eta > 0$, an interior optimal quantity of data $d_s(t)$ balances the marginal disutility and the sum of marginal production and marginal innovation. Along the BGP, the social planner does not share unlimited data, but keeps a constant level of data d_s in the long run. We derive the next proposition in Appendix A.1.

Proposition 3.1. *When $\kappa d^2 - \eta > 0$, along the BGP, the optimal balanced growth rate of varieties g_{Ns} and the optimal data contribution per capita d_s exist and are uniquely determined by two equations:*

$$g_{Ns}(d) = \frac{(\gamma-1)\rho}{\xi}(\kappa d^2 - \eta) \tag{21}$$

and

$$g_{Ns}(d) = \varepsilon L \underbrace{\left[\frac{\frac{1-\xi}{\xi}\left(\kappa d^2 - \eta\right)}{1 + \frac{1-\xi}{\xi}\left(\kappa d^2 - \eta\right)}\right]^{1-\xi}}_{\text{Changes in labor allocation caused by data}} d^\xi. \tag{22}$$

Equation (21) characterizes the relationship between data disutility and growth rate. The social planner requires higher growth rates as the compensation for the higher disutility caused by a greater usage of data. Equation (22) reflects a technology constraint, which is the growth rate that can be achieved through each unit of data produced. These two equations together pin down economic growth and per capita data usage along the BGP, which are shown in Fig. 1. Importantly, endogenous long-run growth occurs even when the population experiences no growth—a desirable and realistic feature that is unattainable under semi-endogenous growth models.

Moreover, along the BGP, the growth rate of the aggregate output $Y(t)$ is

$$\frac{\dot{Y}(t)}{Y(t)} = \frac{\gamma}{\gamma-1}\frac{\dot{N}(t)}{N(t)} + \frac{\dot{Y}(v,t)}{Y(v,t)} = \underbrace{\frac{1}{\gamma-1}\varepsilon l_{Rs}^{1-\xi} L d_s^\xi}_{\text{Growth from innovation sector}}, \tag{23}$$

the aggregate output $Y(t)$ at time t is

$$Y(t) = N(t)^{\frac{1}{\gamma-1}}(1 - l_{Rs}) d_s^\eta L^{1+\eta}, \tag{24}$$

and the fraction of labor employed in the innovation sector is

$$l_{Rs} = \frac{\frac{1-\xi}{\xi}(\kappa d_s^2 - \eta)}{1 + \frac{1-\xi}{\xi}(\kappa d_s^2 - \eta)}. \tag{25}$$

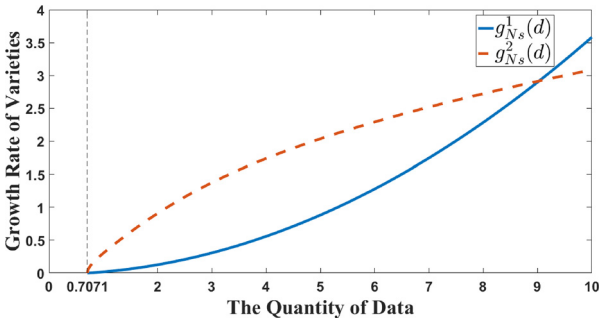


Fig. 1. Steady states in the optimal allocation in (21) and (22). Note: This figure shows the steady states in the optimal allocation. The solid line corresponds to (21) and the dashed line, (22). The position marked by the vertical line is the lower bound of the domain of d. When the solid line is above the dashed line, no more data are shared. The intersection is the value of g_{Ns} and d_s along the BGP. The first state, where $g_{Ns} = 0$ and $d = (\eta/\kappa)^{1/2} = 0.7071$, can be ruled out in the long run, because the FOC with respect to d is $\partial\mathcal{H}/\partial d(t) > 0$ in (19), and the social planner can share data until the quantity of data reaches the intersection where $\partial\mathcal{H}/\partial d(t) = 0$.

Equation (23) reveals that the growth of the aggregate output is determined by the growth of varieties in the long run. The usage of data in the innovation sector plays an important role in economic growth, whose contribution is denoted by the parameter ξ. Compared with (20), where the maximum quantity of data available is not sufficiently large, the growth from the production sector in (23) is zero in the long run. The reason is that increasing privacy costs dominate and the quantity of data input is limited. Growth from production needs the quantity of data to become greater, but this means that consumers will face larger privacy costs. Because of the diminishing returns in terms of data, the output contribution from directly using data makes it difficult to counteract the exponential growth of privacy costs. As a result, the quantity of data input is constant, implying that privacy costs are limited and growth from the production sector alone cannot be sustained.

That said, the innovation sector transforms data into knowledge, creating "desensitized" data, such as blueprints, algorithm codes, or aggregate economic patterns. A firm may innovate incrementally by observing previous data-based knowledge repeatedly without incurring any additional privacy costs. As such, the "desensitized" data can be utilized in every future period and the data supplied at a constant level in each period can promote endogenous economic growth in the long run due to the "dynamic nonrivalry", as described in Cong et al. (2021b). As we will see for the decentralized economy in Section 3.2, knowledge represented by "desensitized" data still suffers from insufficient data sharing, since consumers cannot internalize the benefits of knowledge spillovers.

The results presented in Jones and Tonetti (2020) are based on an assumption that privacy costs are measured by the proportion of shared data rather than the total quantity of data. As a result, the corresponding disutility caused by data usage is kept at a constant level while the total quantity of shared data keeps growing. In semi-endogenous growth models, the population growth can impact the dynamics of data usage because a greater quantity of data is contributed in the aggregate without increasing privacy costs per capita. Thus, in the long run, they note that the quantity of data each person shares will decrease while we find that it remains fixed in our model. Finally, in both Cong et al. (2021b) and our paper, economic growth is sustainable, which contrasts with Xie (2017), who finds that innovations having to do with disutility stall the endogenous growth of an economy.

Equation (24) reveals that the usage of data in the production sector determines the current period's productivity $Y(t)$ and its contribution is denoted by the parameter η. Then, it is natural that parameter η can also affect the growth rate shown in Proposition 3.1, since η affects the quantity of data in the equilibrium, and it then further affects labor allocation and growth rate through the innovation sector. However, this indirect impact on economic growth is based on the utilization of data in the innovation sector, which is crucial in the analysis conducted here. We extend our model in Section 3.3 in order to further elaborate on this point.

Several comparative statistics reveal that labor allocations under multiple data usage differ significantly from those in studies such as Jones and Tonetti (2020) and Cong et al. (2021b). From (25), the fraction of labor employed in the innovation sector is determined by the following three factors: the contribution of data in production, η, the contribution of data in innovation, ξ, and the quantity of data, d_s. In addition, from Proposition 3.1, d_s is determined by η and ξ endogenously. As a result, ξ and η affect the labor allocation through the following two channels:

The first channel reflects a direct effect. As argued in Jones and Tonetti (2020), in the production function (8), we let the labor contribution in the production sector be 1 and assume that the labor contribution is independent of the data contribution which is denoted by η, in order to capture the increasing returns to scale for incumbent firms. Along the BGP, the demand curve for labor from the point of view of the social planner is

$$\kappa d_s^2 = \eta + \frac{\xi}{1-\xi}\frac{l_R(t)}{l_E(t)}.$$

We treat the quantity of data d_s as given, and consider the direct effect of η. According to the above equation, with the increase of η, the social planner needs less labor in the innovation sector to compensate for privacy costs. Thus, η does not

cause the substitution of labor but rather encourages the planner to devote more labor to production. Meanwhile, we use the settings described in Cong et al. (2021b) and assume constant returns to scale in innovation: The contributions of data and labor are ξ and $(1-\xi)$, respectively. Thus, the increase of ξ substitutes labor in the innovation sector, which is similar to the automation as discussed in Acemoglu and Restrepo (2018) and Aghion et al. (2019). The difference in returns to scale helps us capture the differences in data utilization between incumbents in the production sector and potential entrants of the innovation sector.

In addition to this direct effect, an indirect effect can be seen through (19), which can be transformed into:

$$\underbrace{\eta}_{\text{Compensation from production}} + \underbrace{\xi\lambda(t)\dot{N}(t)}_{\text{Compensation from innovation}} = \underbrace{\kappa d_s^2}_{\text{Disutility}}. \tag{26}$$

The increases of the two parameters η and ξ on the left hand side can increase the quantity of data d_s, which means that the planner can tolerate greater privacy costs. Then, in (25), l_{Rs} should increase to compensate for the additional privacy costs, which are denoted in the term $(\kappa {d_s}^2 - \eta)$. This indirect effect counteracts the direct effect, rendering the overall impact parameter dependent.

3.2. Decentralized economy

In the decentralized economy, the Hamiltonian for consumers' problem is:

$$\mathcal{H}(c(v,t),d(t),a(t),\mu(t)) = \ln c(t) - \frac{\kappa d(t)^2}{2} + \mu(t)\begin{bmatrix} r(t)a(t) + w(t) + p_d(t)d(t) \\ -\int_0^{N(t)} p(v,t)c(v,t)\mathrm{d}v \end{bmatrix}. \tag{27}$$

With $\mu(t) = 1/c(t)$, the FOC for consumers problem with respect to $d(t)$ is:

$$\frac{\partial \mathcal{H}}{\partial d(t)} = -\kappa d(t)^2 + \underbrace{\left(1-\frac{1}{\gamma}\right)\eta}_{\text{Compensation from production sector}} + \underbrace{\frac{\xi\dot{N}(t)V(t)}{N(t)^{\frac{\gamma}{\gamma-1}}Y(v,t)}}_{\text{Compensation from innovation sector}} = 0. \tag{28}$$

Here, consumers' privacy costs are compensated by the products from the production sector and assets from the innovation sector. As we have discussed in the optimal allocation, along the BGP, we have $d(t) \to d_c$ and $g_{Yc} \equiv \dot{Y}(t)/Y(t) \to g_{Nc}/(\gamma-1)$, so the function of data in productivity determines the output $Y(t)$, and the function in innovation determines the growth rate of varieties g_{Nc}. Focusing on the interior solution, i.e., $d_c^2 > (1-1/\gamma)\eta/\kappa$, we have:

Proposition 3.2. *In a decentralized economy along the BGP, the balanced growth rate of varieties g_{Nc} and data per capita d_c are determined by the following two equations:*

$$g_{Nc}(d) = \rho\frac{\kappa d^2 - \left(1-\frac{1}{\gamma}\right)\eta}{\xi\Gamma - \left[\kappa d^2 - \left(1-\frac{1}{\gamma}\right)\eta\right]} \tag{29}$$

and

$$g_{Nc}(d) = \varepsilon L\left[\frac{\frac{1-\xi}{\xi}\frac{\kappa d^2-\left(1-\frac{1}{\gamma}\right)\eta}{1-\frac{1}{\gamma}}}{1+\frac{1-\xi}{\xi}\frac{\kappa d^2-\left(1-\frac{1}{\gamma}\right)\eta}{1-\frac{1}{\gamma}}}\right]^{1-\xi} d^{\xi}, \tag{30}$$

where

$$\Gamma \equiv \left[1-\left(1-\frac{1}{\gamma}\right)(1+\eta)\right].$$

Moreover, if $1-(1-1/\gamma)(1+\eta) > 0$, then $\kappa d^2 - (1-1/\gamma)\eta \in (0, \xi\Gamma)$.

The parameter Γ here can be viewed as the share of profit owned by the incumbent firms and we assume $\Gamma > 0$. In the above two equations, (29) describes the relationship between the exponential disutility and the required growth from the view of consumers, while (30) describes the relationship between data demand and achievable growth subject to the innovation possibility frontier from the view of potential entrants. Fig. 2 shows the above two equations and their intersection. When data can be used for both production and innovation, their allocations are distorted by the monopoly markup in the production sector, as well as the higher required rate of return compared with the optimal allocation. Generally, we can observe the following three channels of inefficiencies from the above two equations in a decentralized economy:

The first inefficiency concerns distortions in the price of data and is similar to that discussed in Jones and Tonetti (2020), which is reflected by the term $(1-1/\gamma)\eta$ in (29). Compared with the optimal allocation, consumers are compensated by fewer products from the monopolists, which is equal to $(1-1/\gamma)\eta$. The reason for this is that the price is inefficiently low since the monopolistic incumbent firms require a markup $(1-1/\gamma)$. With the same privacy costs, consumers are compensated by less output. As is shown in (28), this distortion pushes consumers to share a smaller quantity of data while other conditions remain unchanged.

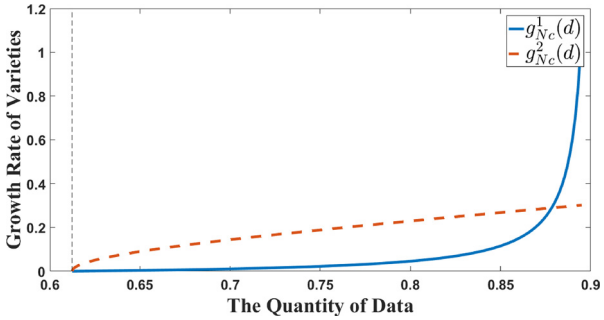


Fig. 2. Steady states in the decentralized economy in (29) and (30). Note: This figure shows the steady states in the decentralized economy. The solid line corresponds to (29) and the dashed line, (30). The intersection after the vertical line is the value of g_{Nc} and d_c along the BGP. Because of this inefficiency, (29) has a larger convexity than (21) in Fig. 1.

The second inefficiency comes from the insufficient growth rate to compensate for the disutility of consumers, which is reflected by the term $-[\kappa d^2-(1-1/\gamma)\eta]$ in (29). The required growth rate of varieties increases faster than that under conditions of optimal allocation (see the denominator in (21) and (29)), which means that, with the same privacy concerns, consumers require a higher growth rate to compensate for the privacy concerns. The reason for this lies in the insufficient compensation to consumers due to knowledge spillover effects. As a result, the value of innovations is lower in competitive equilibrium than that in the optimal allocation. Formally, along the BGP and under condition of optimal allocation, the disutility compensated by new innovations is equal to

$$\kappa d_s^2-\eta=\lambda(t)\xi\dot{N}(t)=\underbrace{\frac{\xi}{\rho(\gamma-1)}\frac{1}{N(t)}}_{\text{Privacy compensation of each time of innovation}}\dot{N}(t).$$

Similarly, in a decentralized economy, the disutility compensated by new innovations is

$$\begin{aligned}\kappa d_c^2(t)-\left(1-\frac{1}{\gamma}\right)\eta&=\xi\frac{V(t)}{N(t)^{\frac{1}{\gamma-1}}Y(v,t)}\frac{\dot{N}(t)}{N(t)}=\frac{\xi\Gamma}{r(t)-\frac{\dot{V}(t)}{V(t)}}\frac{\dot{N}(t)}{N(t)}\\&=\underbrace{\frac{\xi\Gamma}{\rho+g_{Nc}}\frac{1}{N(t)}}_{\text{Privacy compensation of each time of innovation}}\dot{N}(t).\end{aligned}\tag{31}$$

As seen in (31), when consumers share more data, the innovation sector can use it to generate dynamically nonrival "desensitized" data and increase the growth rate of varieties g_N. Through the "desensitizing" effect of the innovation sector, consumers' privacy concerns from such dynamically nonrival data are alleviated in terms of future reuse, while, at the same time, they cannot get the corresponding benefits. This process increases the required rate of return $(r(t)-\dot{V}(t)/V(t))$ to $(\rho+g_N)$, which diminishes the value of innovation (and the value of each asset, both of which are denoted by $V(t)$). Therefore, the greater the quantity of data consumers share, the lower the value that each time of innovation has, and the less compensation consumers get. This inefficiency pushes consumers to share a smaller quantity of data under the condition of the same growth rate, given other conditions unchanged.

The last inefficiency comes from different allocations of labor, which is reflected by the term

$$\left[\frac{\frac{1-\xi}{\xi}\frac{\kappa d^2-\left(1-\frac{1}{\gamma}\right)\eta}{1-\frac{1}{\gamma}}}{1+\frac{1-\xi}{\xi}\frac{\kappa d^2-\left(1-\frac{1}{\gamma}\right)\eta}{1-\frac{1}{\gamma}}}\right]^{1-\xi}$$

in (30). In the decentralized economy, labor is allocated as

$$\left(\frac{l_R}{l_E}\right)_c=\frac{1-\xi}{\xi}\frac{\kappa d_c^2-\left(1-\frac{1}{\gamma}\right)\eta}{1-\frac{1}{\gamma}}.\tag{32}$$

Since the monopolistic markup of incumbent firms distorts the equilibrium allocation, the production sector underemploys labor, which is captured by the denominator in (32). However, when data are used in multiple channels, an insufficient quantity of data also affects the labor allocation, which is captured by the corresponding numerator. This data shortage may trap the decentralized economy in a low-growth regime with underemployment in the innovation sector. Section 4 quantitatively investigates g_{Nc} and d_c relative to the optimal allocations.

From (32), the fraction of labor employed in the innovation sector is also determined by a direct effect and an indirect one, such as those that we discussed previously. The demand function for labor from incumbent firms can be rewritten as

$$\left(1-\frac{1}{\gamma}\right)N(t)^{\frac{\gamma}{\gamma-1}}(d_c(\eta,\xi)L)^{\eta}=w(t).$$

Treating $d_c(\eta,\xi)$ as given, we can see that the demand for labor from incumbent firms increases with η because of increasing returns to scale, which reflects the direct effect of η. As for the indirect effect of η, the demand for labor from incumbent firms increases with $d_c(\eta,\xi)$. At the same time, the demand function for labor from the innovation sector is

$$(1-\xi)\varepsilon N(t)L_R(t)^{-\xi}(d_c(\eta,\xi)L)^{\xi}V(t)=w(t).$$

When considering the direct effect of ξ, we can see that the demand for labor from the innovation sector decreases with ξ and increases with $d_c(\eta,\xi)$.

Similar patterns in terms of comparative statistics are manifested in the optimal allocation as well, but the relative magnitude of the direct and indirect effects plays a nontrivial role. Section 4 discusses some further results through numerical examples. We can note that the market power in the production sector should lead to underemployment in production (e.g., Cong et al., 2021b). Yet, we find the opposite.

3.3. Allocations with a singular usage of data

We now investigate the two uses of data separately in order to understand their contribution to economic growth. First, we set $\xi=0$, i.e., the social planner can only use data in the production sector. Juxtaposing the results with those in Section 3.1 reveals the contribution of data in the innovation sector and its complementarity with data usage in production. The problem for the social planner now becomes:

$$\max_{\{l_E(t),d(t)\}}\int_0^{\infty}e^{-\rho t}L\left(\ln c(t)-\frac{\kappa d(t)^2}{2}\right)\mathrm{d}t,$$

subject to

$$\begin{aligned}
c(t)&=Y(t)/L,\\
Y(t)&=N(t)^{\frac{1}{\gamma-1}}l_E(t)L^{1+\eta}d(t)^{\eta},\\
\dot{N}(t)&=\varepsilon l_R(t)LN(t),\\
1&=l_E(t)+l_R(t),\\
d(t)&\le g(c(t))\equiv\bar{d}(t).
\end{aligned}$$

The Hamiltonian for the optimal allocation is:

$$\mathcal{H}(l_E(t),d(t),N(t),\lambda(t))=\ln\left[N(t)^{\frac{1}{\gamma-1}}l_E(t)L^{\eta}d^{\eta}\right]-\frac{\kappa d(t)^2}{2}+\varepsilon\lambda(t)l_RLN(t).$$

Proposition 3.3. *When data are only used in the production sector, along the BGP under conditions of optimal allocation, the per capita usage of data d' is*

$$d'=\left(\frac{\eta}{\kappa}\right)^{\frac{1}{2}},\tag{33}$$

the growth rate of varieties is

$$g_N'=\varepsilon L-(\gamma-1)\rho,$$

the aggregate output at time t is

$$Y'(t)=\frac{\rho(\gamma-1)}{\varepsilon}\left(\frac{\eta}{\kappa}\right)^{\frac{\eta}{2}}L^{1+\eta}N'(t)^{\frac{1}{\gamma-1}},$$

and the fraction of labor employed in the R&D is

$$l_R'=1-\frac{\rho(\gamma-1)}{\varepsilon L}.$$

Appendix A.4 contains the derivations of these expressions. Shutting down the usage of data in the innovation sector helps us better understand why the quantity of shared data becomes constant in the long run. If we consider the data usage shown in (33), it is a constant. Due to diminishing marginal returns ($\eta<1$), the social planner does not share all the available data and instead chooses a fixed quantity. In addition, when we compare the results with Section 3.1, we get:

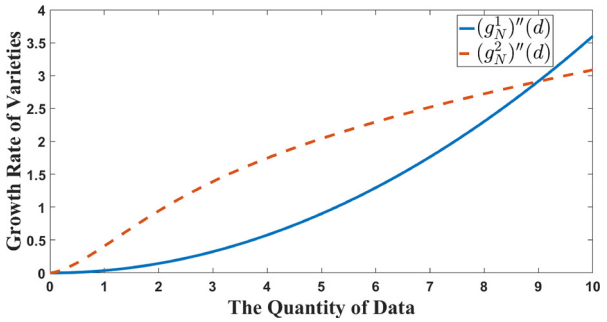


Fig. 3. Steady states when data only enter the innovation sector in (34) and (35). Note: The solid line corresponds to (34) and the dashed line, (35). The intersection gives the value of g_N'' and d'' along the BGP.

Corollary. When data are used only in production, data usage (e.g., captured by κ and η) does not affect the growth rate of the economy in the long run but only the level of GDP $Y'(t)$ during each period. Moreover, the per capita usage of data is less than that when data are used in innovation.

Next, we set $\eta = 0$ to analyze the case in which the social planner can only use data in the innovation sector. The problem for the social planner now becomes:

$$\max_{\{l_E(t),d(t)\}} \int_0^\infty e^{-\rho t} L\left(\ln c(t) - \frac{\kappa d(t)^2}{2}\right) \mathrm{d}t,$$

subject to

$$\begin{aligned} c(t) &= Y(t)/L, \\ Y(t) &= N(t)^{\frac{1}{\gamma-1}} l_E(t) L, \\ \dot{N}(t) &= \varepsilon l_R(t)^{1-\xi} L d(t)^{\xi} N(t), \\ 1 &= l_E(t) + l_R(t), \\ d(t) &\le g(c(t)) \equiv \bar{d}(t). \end{aligned}$$

The Hamiltonian for the optimal allocation is then:

$$\mathcal{H}(l_E(t), d(t), N(t), \lambda(t)) = \ln\left[N(t)^{\frac{1}{\gamma-1}} l_E(t)\right] - \frac{\kappa d(t)^2}{2} + \varepsilon \lambda(t) l_R^{1-\xi} d^{\xi} L N(t).$$

Proposition 3.4. *When data are only used in the innovation sector, along the BGP under conditions of optimal allocation, the balanced growth rate of varieties g_N'' and data per capita d'' are determined by the following two equations:*

$$(g_N^1)''(d) = \frac{(\gamma-1)\rho}{\xi} \kappa d^2, \tag{34}$$

and

$$(g_N^2)''(d) = \varepsilon L \left[\frac{\frac{1-\xi}{\xi}\kappa d^2}{1 + \frac{1-\xi}{\xi}\kappa d^2}\right]^{1-\xi} d^{\xi}. \tag{35}$$

Also, the fraction of labor employed in the innovation sector is

$$l_R'' = \frac{\frac{1-\xi}{\xi}\kappa (d'')^2}{1 + \frac{1-\xi}{\xi}\kappa (d'')^2}.$$

The aggregate output at time t is:

$$Y''(t) = N(t)^{\frac{1}{\gamma-1}} (1 - l_R'') L.$$

Appendix A.5 contains the derivations of these expressions, and Fig. 3 represents the above two equations and their intersection. When η is relatively small (e.g., 0.03 to 0.10, as estimated in Jones and Tonetti, 2020), the growth rates for the varieties and the quantity of data are similar to those in Proposition 3.1.

Comparing the two singular uses of data, we see that the role of data is much more important in the innovation sector because of the "desensitizing" effect and the accumulation of knowledge, both of which are absent in the production sector. However, when the social planner uses data only in the innovation sector, the economy suffers losses in the production sector. This is intuitive since a nonrival use of data in the production sector (without requiring additional data contributed by consumers) can further increase production outputs.

Table 1
Baseline Parameters for Quantitative Analyses.

Parameter	Description	Value	Source
η	Contribution of data in production sector	0.1	Standard
ξ	Contribution of data in innovation sector	0.5	Discretionary
κ	Risk aversion to privacy concerns	0.2	Discretionary
γ	Elasticity of substitution between varieties	4	Standard
ρ	Rate of time preference	0.03	Standard
L	Population level	1	Standard
ε	Innovation efficiency	1	Standard

3.4. Policy interventions and further discussion

In a decentralized economy, the output level and the value of innovation are both lower than what they would be under conditions of an optimal allocations. In other words, consumers choose to share a socially inefficient quantity of data. To push decentralized allocations to the optimal level, we propose the following policy interventions to alleviate distortions in the levels of data usage, growth rates, and labor allocations:

Proposition 3.5. *To restore socially efficient outcomes, a government can institute revenue subsidies in the production and the innovation sectors with the following rates:*

$$s_p = \frac{1}{\gamma - 1} \quad \textit{and} \quad s_n = \frac{\rho + g_{Ns}}{\rho\Gamma(\gamma - 1)(1 + s_p)} - 1.$$

Here, s_p is the production subsidy, which is the conventional markup correction to increase the output of monopolists, and s_n is the innovation subsidy to encourage innovation.

Alternatively, the government can subsidize firms for purchasing data and employing labor with the following rates:

$$s_{d1} = \frac{1}{\gamma}, \qquad s_{d2} = 1 - \frac{\rho(\gamma - 1)\Gamma}{\rho + g_{Ns}}, \quad \textit{and} \quad s_l = 1 - \frac{1 - s_{d2}}{1 - \frac{1}{\gamma}}.$$

Here, s_{d1} and s_{d2} are the data subsidies for the production sector and the innovation sector, respectively, and s_l is the labor subsidy for the innovation sector.

4. Quantitative analysis

In this section, we simulate the growth of our data economy to quantify misallocation in a decentralized economy and to compare the various uses of data. Furthermore, we explore some alternative specifications of the model to demonstrate the robustness of our findings.

4.1. Contribution of multiple uses of data to growth

First, we build on Section 3 in examining the relative contribution of different uses of data. For simplicity, we focus on optimal allocations. In Section 3.1, data are used in both the production and innovation sectors, while in Section 3.3, data are used either in the production or innovation sectors but not in both. We compare the results derived in these three subsections. Table 1 summarizes the model parameters.[3] Our quantitative analyses mainly focus on the following three key variables: the growth rate of varieties g_N, the quantity of data being shared d, and the fraction of labor employed in the innovation sector l_R. We do not discuss the time-varying output level $Y(t)$ (which is affected by the variety level $N(t)$), since that would tend to make our analyses more confusing.

Table 2 briefly summarizes our baseline numerical results. The values for the three key variables shown in this table differ dramatically across different scenarios. The optimal allocation with multiple uses of data realizes the highest growth rate for varieties with the largest quantity of data usage. Compared with situations in which data are only used in the innovation sector, sharing data with the production sector may increase the quantity of data and the growth rate. This is due to the fact that, when the production sector can use data simultaneously, consumers can receive additional compensation from selling data to the production sector captured by η. The additional compensation increases the quantity of data in the economy. Thus, through vertical nonrivalry, the innovation sector can use more data and, thus, the growth rate increases. This reflects the fact that the two sectors are complementary in terms of the growth of the economy, where one sector's data usage can benefit the other's. In addition, the quantity of data when data are only used for the purposes of the innovation sector is higher than that when data are only used in the production sector. The reason lies in that when data enter the innovation

[3] Some parameters related to data cannot be determined from the existing literature, e.g., η, κ, and ξ. As a result, we illustrate our theoretical framework within a rational range of values for these parameters and choose standard values for the other parameters.

Table 2
Results for Quantitative Analysis.

Model	g_N	d	l_R
Social Planner	2.9160	9.0277	0.9419
Decentralized Economy	0.2897	0.8783	0.0956
Only in Production (SP)	0.9100	0.7071	0.9100
Only in Innovation (SP)	2.9097	8.9903	0.9417

Notes: The table reports three key variables for different allocations using the parameter values in Table 1. g_N is BGP level of the growth rate of varieties, d is BGP level of the quantity of data, and l_R is BGP level of R&D employment. Compared with single usages in the last two lines, multiple usages in the social planner have higher g_N, and the quantity of data is not a simple addition of the singular usage numbers.

sector, they become "desensitized" as knowledge accumulates and, thus, alleviate privacy concerns. This "desensitization" makes data usage in innovation more economical compared with that in production and brings about higher growth rates.

Figure 4 shows allocations under the conditions of different values of η. Against the backdrop of multiple data uses, the quantity of data and the growth rate are strictly larger than in situations when data are only used either in the production or innovation sectors. Here, the higher contribution of data in the production sector, which is captured by η, can increase the growth rate for varieties g_N, while other parameters remain unchanged. Intuitively, the planner is willing to share a greater quantity of data with more compensation from the production sector. With vertical nonrivalry, the additional data can also be used in the innovation sector, meaning that the growth rate increases simultaneously. Recalling the discussion of (25) and (26), if the quantity of data is sufficient and increases faster than η, the planner will choose to devote more labor to the innovation sector, which is shown in Fig. 4.

Figure 5 represents the allocations under different values of ξ. Unlike the role of η shown in Fig. 4, higher values of ξ more significantly increase the growth rate and the quantity of shared data. Meanwhile, the changing patterns of the quantity of data and labor employment are similar both with multiple data uses and with singular data usage in the innovation sector. However, when data are only used in production, the value of ξ does not significantly influence the three key variables compared with the other two cases.

Figure 6 shows allocations under different values of κ. When data are only used in the production sector, increasing κ reduces the quantity of data but has no impact on economic growth. In contrast, κ has a broader impact when data have multiple uses or only have singular usage in the innovation sector. In these two cases, as a household's risk aversion to privacy concerns becomes greater, the planner reduces the extent of data sharing and the fraction of employment in the innovation sector, further affecting the growth rate.

4.2. Misallocation in equilibrium

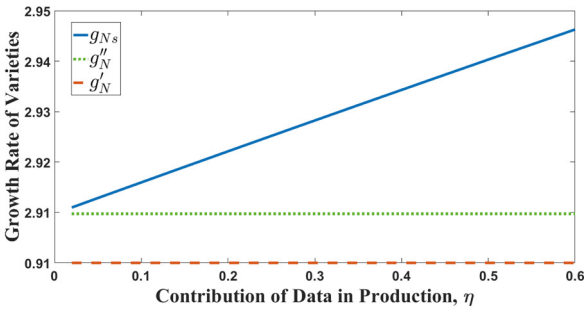
We next discuss allocations in the four different models in Section 3 to characterize misallocation in a decentralized economy. According to Table 2, in a decentralized economy, due to market distortions, the quantity of data usage is only slightly higher than that when data are only used for production. Compared with the optimal allocation, the underutilization of data is obvious. In the long run, the decentralized economy suffers from a growth trap of insufficient varieties and innovation. Fig. 7 reveals that this growth trap becomes more serious as κ increases.

Furthermore, we simulate a wide range of values for η and ξ to find out whether inefficiency can be mitigated by the improvement in data utilization efficiency. In Fig. 8, with the increase of η, the fraction of labor employed in the innovation sector and the growth rate of varieties decrease. Although the quantity of data has increased, it is still too limited to attract labor to the innovation sector. The increase in η shows an effect of attracting labor to the production sector. This process of labor allocation is the opposite of the optimal allocation shown in Fig. 4(c); that is, the increase in the quantity of data has little effect on economic growth under these circumstances. Fig. 9 reveals a substitution effect for ξ on labor, which differs from Fig. 5(c), since the quantity of data has not increased significantly (see (32)). In addition, in Fig. 9, the maximum value of ξ is only 0.9. There is still a big gap from the optimal distribution because of an insufficient compensation for data sharing (discussed in Section 3.2) and the loss of labor in the innovation sector.

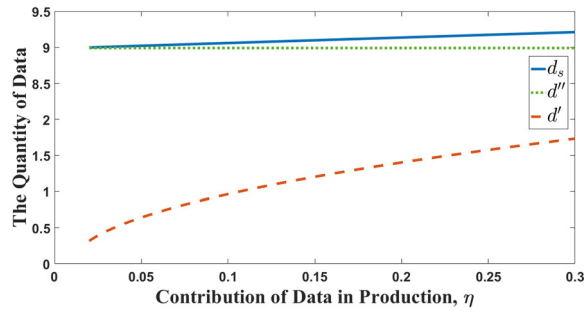
4.3. Alternative specifications

Following both Jones and Tonetti (2020) and Cong et al. (2021b), we set the production sector to have increasing returns and the innovation sector to have constant returns. However, one may question whether the production sector should also have constant returns to scale, e.g., $Y(v,t) = L_E(v,t)^{1-\theta} D(t)^{\theta}$, where $L_E(v,t)$ and θ are the labor employed and the data contributed in the production sector. Here, both the direct and indirect effects of θ increase labor employment in the innovation sector. Thus, we show in Appendix A.7 and Figure A.1-A.3 that our key findings remain robust.

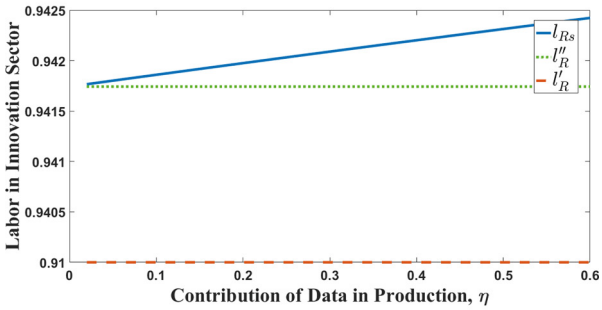
One may also anticipate that privacy concerns become greater as the number of firms using data increases, since their repeated usage across firms increases the likelihood of leakage and hacking. We can allay this concern by allowing the disutility term to be $\int_0^{N(t)} [\kappa d(v,t)^2/2] \mathrm{d}v$. Along the BGP, the quantity of data decreases and the growth rate of varieties



(a) Growth rate of varieties



(b) Quantity of data



(c) Labor allocation

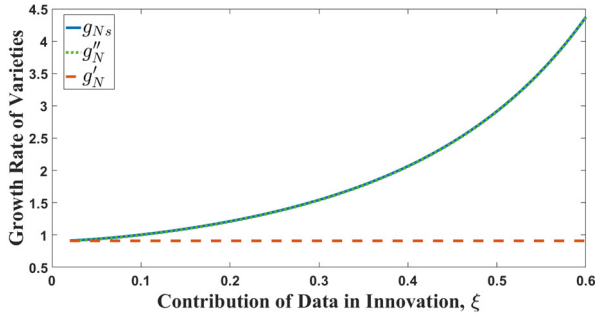
Fig. 4. Equilibrium outcomes for different η. Note: The figure shows equilibrium outcomes for various usages of data and different η. The solid line represents the optimal allocation, the dashed line represents the economy where data only enter the production sector, and the dotted line represents the economy where data only enter the innovation sector. η does not impact g_N', l_R' and all variables where data only enter the innovation sector. Other parameters are given in Table 1.

gradually declines towards zero, as shown in Appendix A.8. In this case, innovation leads to an increasing number of firms, while it leads to greater privacy costs, which impedes further data sharing. Meanwhile, a decrease of the quantity of data sharing in turn pushes innovation to become insufficient over time. Note that in reality, with multiple uses of data, it is difficult for consumers or the government to distinguish the specific uses.

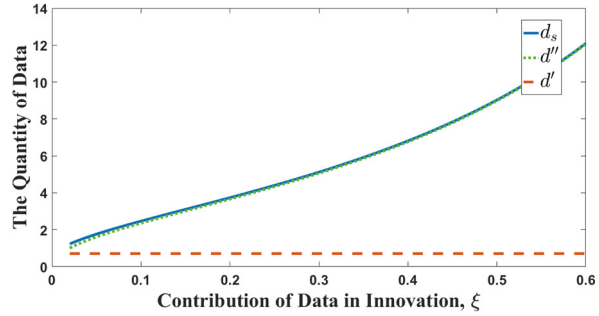
Similarly, privacy costs may increase with multiple uses of data in multiple sectors. Specifically, each consumer's instantaneous utility function can be specified as:

$$\max_{\{l_E(t),d(t)\}} \int_0^\infty e^{-\rho t} L\left[\ln c(t) - (1+\alpha)\frac{\kappa d(t)^2}{2}\right] \mathrm{d}t,$$

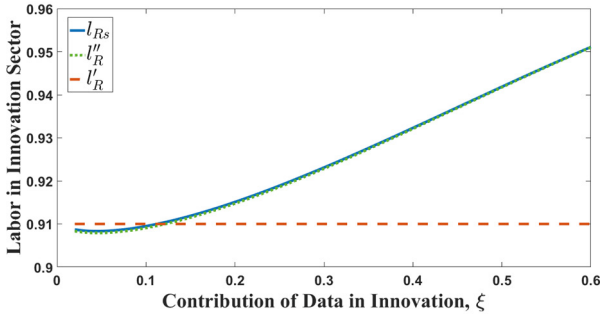
where $\alpha \geq 0$ denotes the additional privacy concerns due to the multiple uses of data. In Appendix A.9, Figure A.4 and Table A.1 show the equilibrium and the comparative statistics with respect to α (or, equivalently, with respect to κ). When α is close to zero, the model reduces to the baseline with multiple uses of data as discussed in Section 3.1. The multiple uses bring about higher compensation for privacy costs and, consequently, a greater quantity of data usage, as we discuss in Section 4.1. When α increases, sharing data with the production sector may decrease the quantity of data and the growth



(a) Growth rate of varieties

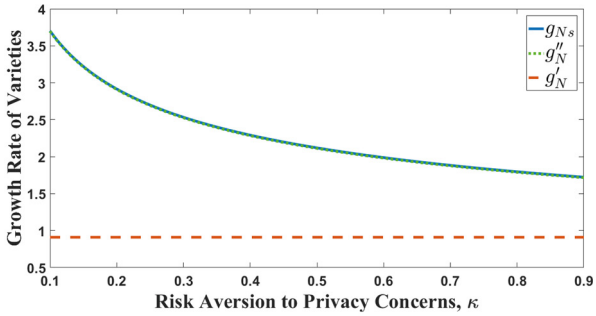


(b) Quantity of data

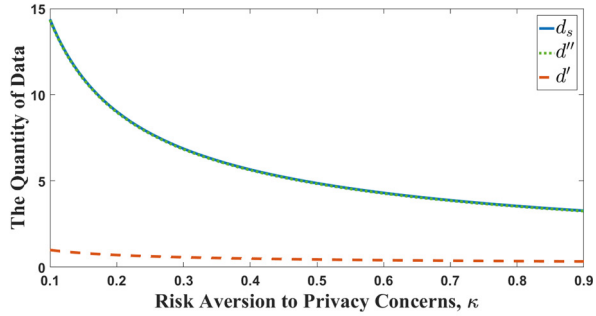


(c) Labor allocation

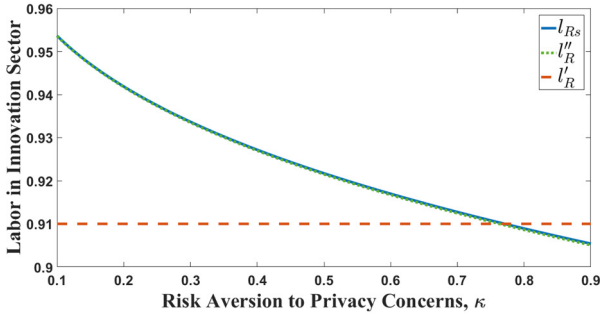
Fig. 5. Equilibrium outcomes for different ξ. Note: The figure shows equilibrium outcomes for various usages of data and different ξ. The impact of ξ on the multiple uses and only innovation use is similar. For example, ξ significantly increases d_s and g_{Ns}, but it also causes the outflow of labor from the production sector.



(a) Growth rate of varieties

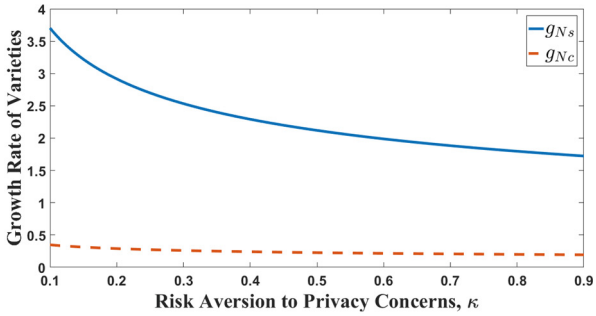


(b) Quantity of data

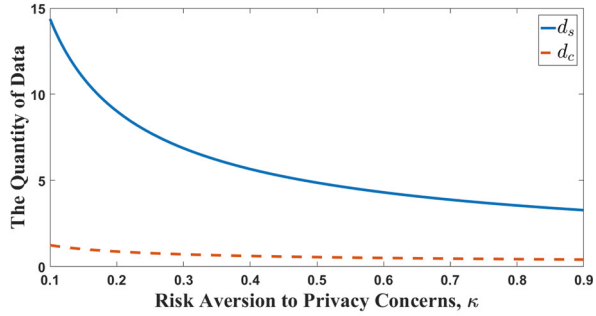


(c) Labor allocation

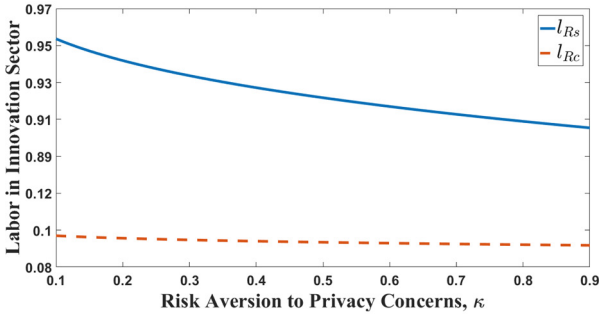
Fig. 6. Equilibrium outcomes for different κ. Note: The figure shows equilibrium outcomes for various usages of data and different κ. When data only enter the production sector, increasing κ reduces d' but have no impact on economic growth g'_N. In optimal allocation, which is represented by the solid line, κ affects three variables.



(a) Growth rate of varieties

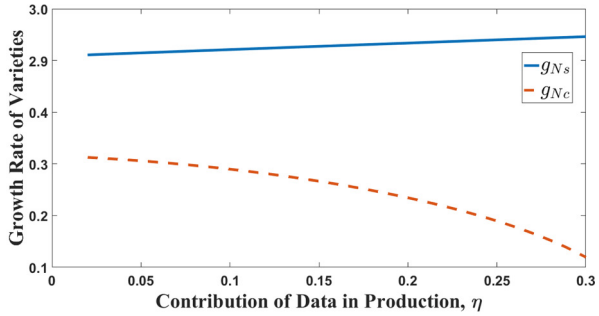


(b) Quantity of data

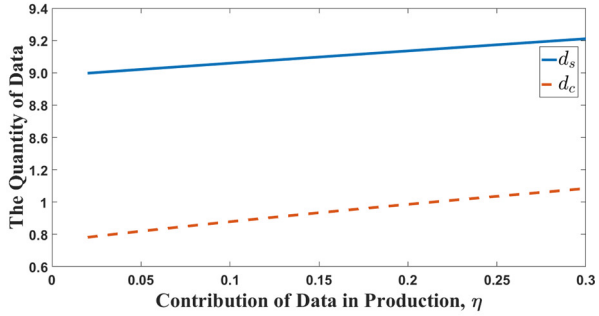


(c) Labor allocation

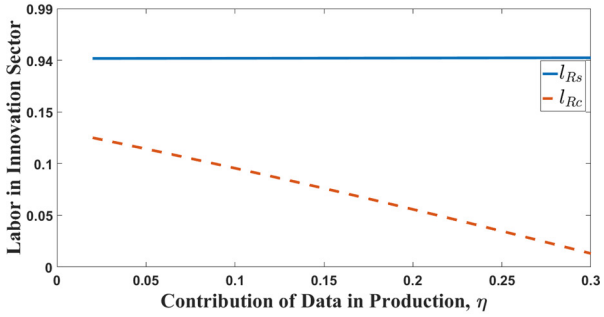
Fig. 7. Equilibrium outcomes in the optimal allocation and decentralized economy for different κ. Note: The figure shows three key variables along the BGP for different κ. Orange dashed lines represent the decentralized economy and for different κ, and blue solid lines represent the optimal allocation. d_c, g_{Nc} and l_{Rc} display the negative correlation with κ.



(a) Growth rate of varieties

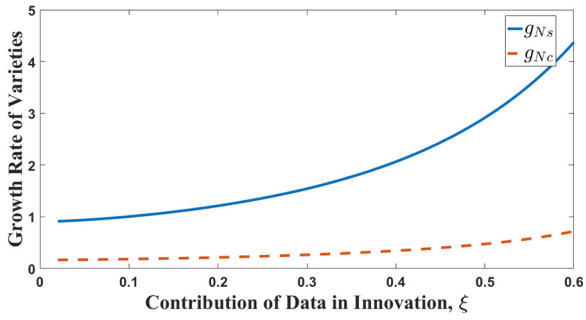


(b) Quantity of data

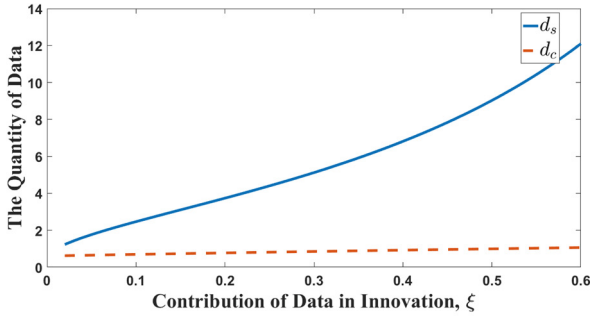


(c) Labor allocation

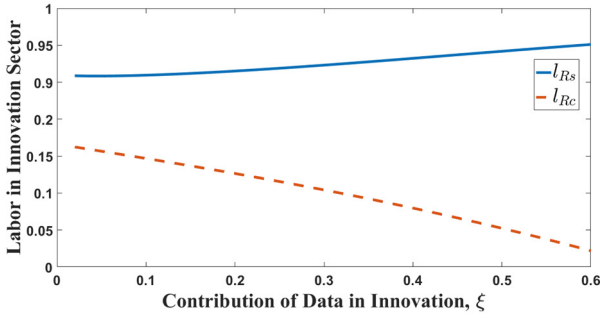
Fig. 8. Equilibrium outcomes in the optimal allocation and decentralized economy for different η. Note: The figure shows three key variables along the BGP for different η. Orange dashed lines represent the decentralized economy, and blue solid lines represent the optimal allocation. g_{Nc} and l_{Rc} display the negative correlation with η, which is opposite to g_{Ns} and l_{Rs} in Fig. 4(a) and 4(c).



(a) Growth rate of varieties



(b) Quantity of data



(c) Labor allocation

Fig. 9. Equilibrium outcomes in the optimal allocation and decentralized economy for different ξ. Note: Orange dashed lines represent the decentralized economy and for different ξ, and blue solid lines represent the optimal allocation. l_{Rc} displays the negative correlation with ξ, which is opposite to l_{Rs} in Fig. 5(c).

rate because the production sector cannot provide economic growth and the additional compensations coming from η are limited. As a result, sharing the same quantity of data as that shared for the innovation sector to the production sector may be accompanied by huge privacy costs. With vertical nonrivalry, higher privacy costs lead to a reduction in the quantity of data and the growth rate, and the results presented in Table 2 become reversed. In this case, the quantity of data used in the production sector should be limited in order to impede the increase in additional privacy costs.

5. Conclusion

The nonrival nature allows data to be employed in both the innovation and production sectors simultaneously. In this paper, we develop an endogenous growth model to understand the interactions and differences in data uses for these sectors. We show that consumers' privacy concerns provide bounds for the growth in the overall quantity of data used. Moreover, the usage of data in the innovation sector plays a more important role in economic growth than in the production sector, because (i) data are dynamically nonrival and add to knowledge accumulation, and (ii) innovations "desensitize" data, reducing consumers' privacy costs when knowledge enters the production sector instead of raw data.

Data uses in both sectors interact to generate spillover in allocative distortion and they exhibit an apparent substitutability due to labor's rivalry and complementarity with data. Consequently, growth rates under a social planner and a decentralized equilibrium differ, which is a novel result obtainable only in a model with fully endogenous growth. According to an optimal allocation, the majority of labor is employed in the innovation sector. For a decentralized equilibrium, however, consumers' failure to fully internalize knowledge spillover in the face of privacy concerns, combined with firms' market power, underprices data and inefficiently limits their supply, leading to underemployment in the innovation sector and a substantially lower data utilization and growth rate in the long run. Also, we discuss interventions in the data market and potential direct subsidies to mitigate the aforementioned inefficiencies.

We are at the dawn of research on data economy and growth. To offer clear insights in a tractable way, we have necessarily left interesting aspects of the economics of data for future research. For example, a lower value of κ, the exogenous parameter for consumers' disutility, facilitates faster growth. It can be endogenously influenced by technological innovations such as privacy-preserving computation (Cao et al., 2019; Hastings et al., 2020) and distributed ledgers (Chen et al., 2021; Cong and He, 2019). It is equally interesting to consider how regulatory policies, such as GDPR, may affect κ in the long run. Furthermore, consumers' cost of data contribution is modeled in reduced form here; active literature is beginning to provide microfoundations for privacy costs (e.g., Ichihashi, 2020; 2021b; Liu et al., 2020), including possibly negative data externalities exacerbating privacy concerns (Acemoglu et al., 2021; Ichihashi, 2021b).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jedc.2022.104395.

References

Acemoglu, D., Makhdoumi, A., Malekian, A., Ozdaglar, A., 2021. Too much data: prices and inefficiencies in data markets. American Economic Journal: Microeconomics, forthcoming.

Acemoglu, D., Restrepo, P., 2018. Automation and new tasks: the implications of the task content of production for labor demand. Journal of Economic Perspectives 33 (2), 3–30.

Admati, A.R., Pfleiderer, P., 1990. Direct and indirect sale of information. Econometrica 58 (4), 901–928.

Aghion, P., Jones, B.F., Jones, C.I., 2019. Artificial Intelligence and Economic Growth. University of Chicago Press.

Akçura, M.T., Srinivasan, K., 2005. Research note: customer intimacy and cross-selling strategy. Manage Sci 51 (6), 1007–1012.

Casadesus-Masanell, R., Hervas-Drane, A., 2015. Competing with privacy. Manage Sci 61 (1), 229–246.

Chen, L., Cong, L.W., Xiao, Y., 2021. A brief introduction to blockchain economics. In: Information for Efficient Decision Making: Big Data, Blockchain and Relevance. World Scientific, pp. 1–40.

Cong, L.W., He, Z., 2019. Blockchain disruption and smart contracts. Review of Financial Studies 32 (5), 1754–1797.

Cong, L.W., Li, B., Zhang, Q.T., 2021. Alternative data in fintech and business intelligence. In: The Palgrave Handbook of FinTech and Blockchain. Springer, pp. 217–242.

Cong, L.W., Mayer, S., 2022. Antitrust and user union in the era of digital platforms and big data. Working Paper.

Cong, L.W., Xie, D., Zhang, L., 2021. Knowledge accumulation, privacy, and growth in a data economy. Manage Sci 67 (10), 6480–6492.

Easley, D., Huang, S., Yang, L., Zhong, Z., 2019. The economics of data. Working Paper.

Hastings, M., Hemenway Falk, B., Tsoukalas, G., 2020. Privacy-preserving network analytics. Available at SSRN 3680000.

Cao, S., Cong, L. W., Yang, B., 2019. Financial reporting and blockchains: audit pricing, misstatements, and regulation. Available at SSRN 3248002.

Farboodi, M., Veldkamp, L., 2021. A growth model of the data economy. Working Paper 28427. National Bureau of Economic Research.

Fainmesser, I. P., Galeotti, A., Momot, R., 2021. Digital privacy. Available at SSRN 3459274.

Hirshleifer, J., 1971. The private and social value of information and the reward to inventive activity. American Economic Review 61 (4), 561–574.

Ichihashi, S., 2020. Online privacy and information disclosure by consumers. American Economic Review 110 (2), 569–595.

Ichihashi, S., 2021. Competing data intermediaries. Rand J Econ 52 (3), 515–537.

Ichihashi, S., 2021. The economics of data externalities. J Econ Theory 196, 105316.

Jones, C.I., 1995. R&D-based models of economic growth. Journal of Political Economy 103 (4), 759–784.

Jones, C.I., 2016. Life and growth. Journal of Political Economy 124 (2), 539–578.

Jones, C.I., Tonetti, C., 2020. Nonrivalry and the economics of data. American Economic Review 110 (9), 2819–2858.

Liu, Z., Sockin, M., Xiong, W., 2020. Data privacy and temptation. Working Paper 27653. National Bureau of Economic Research.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H., 2011. Big data: the next frontier for innovation, competition and productivity. Technical Report. Institute M. G. McKinsey & Company.

Murphy, R.S., 1996. Property rights in personal information: an economic defense of privacy. Georgetown Law Journal 84 (7), 2381–2417.

Romer, P.M., 1990. Endogenous technological change. Journal of Political Economy 98 (5), S71–S102.

Stokey, N.L., 1998. Are there limits to growth? Int Econ Rev 39 (1), 1–31.

Sun, C., Shrivastava, A., Singh, S., Gupta, A., 2017. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 843–852.

Tang, H., 2019. The Value of Privacy: Evidence from Online Borrowers. Working Paper. HEC, Paris.

Xie, D., 2017. Regulatory Growth Theory. Available at SSRN 3238773