Digitization and its Consequences for Creative-Industry Product and Labor Markets

Joel Waldfogel

University of Minnesota, NBER, and ZEW

May 15, 2020

Technological changes have sharply reduced the costs of creating, distributing, and promoting new creative products. This chapter explores the consequences of these changes for both creative product and labor markets. I apply the random long tail lens of Aguiar and Waldfogel (2018) to the product markets. Because new product success is unpredictable, falling costs can deliver products with high realized value, but which would not have been produced before, delivering substantial welfare benefits. I provide rough estimates of the welfare benefits of the growth in movies, television, and books. I have four basic findings. First, available data on movies, television, and books confirm existing findings for music that the random long tail is large compared with the conventional long tail: 9 times as large for books, 13 times as large for television, and 4 times as large for movies. Second and related, the absolute welfare benefit of new creative products is substantial. Third, available evidence on creative labor markets confirms increased activity evidence in product market creation data. Fourth, while total earnings of creative workers are rising, average earnings per worker are falling, although it is not clear how much of the decline in average earnings is simply compositional.

I am grateful for comments from participants in the NBER pre-conference and conference on Innovation and Entrepreneurship. In particular, I thank Gustavo Manso for discussant comments and the editors for additional guidance.

I. Introduction

Digitization has transformed many of the creative industries. Technological changes have sharply reduced the costs of creating, distributing, and promoting new products, with two broad consequences. First, there has been an explosion of new products – in movies, books, music, and television – with substantial welfare benefit for consumers. Second, because technological change has reduced the need for physical or financial capital for undertaking investment in new products, it has enabled individuals to bring new products to market largely by supplying their own labor to entrepreneurial creative projects. In this chapter I explore consequences of digitization for both consumers via the product market as well as entrepreneurial producers via their labor market activity.

A longstanding product market research tradition characterizes the effect of digitization on product markets generally, and markets for cultural goods in particular, through a "long tail" lens. The idea is that the Internet – and online retailing in particular – gives consumers access to a long tail of low-demand products not available at their local stores (Brynjolffson, et al 2003). This is an important insight about a large welfare benefit made possible by digitization that one might term a "long tail in consumption." Having access to, say, a million books at Amazon rather than, say, 50,000 titles at local store may deliver substantial welfare benefits to consumers.

The welfare benefits of digitization may be much larger, however. Digitization not only enables retailers to display products online without any "shelf space" constraints; digitization also reduces the costs of creating new varieties in the first place. For example, digitization has radically reduced the costs of production, distribution, and even promotion for books, music, movies, and television (Waldfogel, 2018 and cites therein). The numbers of new songs, books,

television shows, and movies brought annually to market have risen sharply. New song creation, for example, has more tripled.

Given the well-known unpredictability of product appeal at the time of investment, an increase in the volume of new product entry – a "long tail in production" – can have larger effects on welfare than the standard long tail. In the conventional long tail narrative, online retailing gives consumers access to large numbers of new products with insufficient appeal to have been stocked in local stores. All of the products whose availability is enabled by digitization are therefore less appealing (on average) than the lowest-selling product stocked offline. New products whose creation is made possible by digitization-induced cost reductions are different. Although such products had insufficient promise to justify their investment when costs were higher, because of unpredictability, these products can end up throughout the sales distribution and indeed, many turn out to be commercial successes. This approach parallels a view of entrepreneurship as experimentation explored in various studies.[1]

Aguiar and Waldfogel (2018) explore this mechanism explicitly using digitization of the recorded music industry as its context. Given the unpredictability of product success at the time of investment, they find that change in consumer surplus associated with the tripling of rate of new product introduction after digitization gives rise to a welfare benefit twenty times the size of the standard long tail. The music context is attractive because of the quality of data on the availability and sales of new products; but as a substantive matter, music sales are very highly concentrated in the top few percent of products. For a fuller sense of the effect of the welfare

[1] See, for example, Arrow (1969), Weitzmann (1979), Bergemann and Hege (2005), Manso (2011), and Kerr, Nanda, and Rhoder-Kropf (2014), for studies viewing entrepreneurship as experimentation. Ewens, Nanda, and Rhodes-Kropf (2018) study the effects of reduced costs of entrepreneurial experimentation on innovation in cloud computing.

benefits of this mechanism, it is of interest to revisit these sorts of calculations for books, movies, and television, three important cultural products whose sales concentration among top products – and predictability of sales success at release – may differ. That is the first goal of this chapter.

I also explore the implications of digitization for entrepreneurial creative labor markets. While digitization has lowered barriers to creating products available to broad audiences – and has therefore also enhanced entrepreneurial opportunities – the spread of digitization has also coincided with growing complaints from creators and intermediaries about earnings. This leads me to two broad questions. First, can I document evidence of new creative activity in various ongoing government databases confirming the growth in creative activity evident in product data? Second, what has happened to creators' earnings in the digital era?

I have four basic findings. First, available data on movies, television, and books confirm findings of Aguiar and Waldfogel (2018) for music that the random long tail is large compared with the conventional long tail. Second and related, the welfare benefit of new creative products is substantial. Third, available evidence on creative labor markets confirms increased activity evidence in product market creation data. Fourth, while total earnings of creative workers are rising, average earnings per worker are falling, although it is not clear how much of the decline in average earnings is simply compositional.

II. Theory

New technology enables individuals, or smaller-scale groups, without much costly capital to engage in creative entrepreneurship. The specific circumstances vary across creative products,

but the ability of individuals to create new products and bring them to market has increased across all of the creative industries.

Books provide an extreme example. Prior to digitization, an author needed to secure a contract with a major publisher in order to get a book created and brought to market. This was sufficiently difficult to prevent most would-be authors from attempting to create a book. With the advent of electronic self-publishing – in particular, with the appearance of Amazon's Kindle ecosystem – any author could create a text and make it available to millions of potential readers, without the permission or investment from the traditional gatekeepers (Waldfogel and Reimers, 2015).

Music is similar in the extent to which digitization enables individual entrepreneurial product creation. Prior to digitization, artists sought investments from record labels. Without record deals, an artist might perform on a small scale, but there was no real chance of finding a large audience. Digitization changed this radically. First, digitization allowed individuals to produce music using inexpensive hardware and software. Garageband software, for example, available on Apple computers and even iPhones, provides the functionality of a recording studio. Even more important, digital distribution – first via iTunes and more recently via streaming services – breaks the bottlenecks of both promotion and distribution. The resulting increase in creativity is evidenced by the fact that Spotify added nearly a million songs to its system in 2017; essentially anyone can create music and make it available to a wide audience.

Digitization has had similar effects on movie and video production. First, digital photography has reduced the cost of literally producing content. Second, and more important, digital distribution has eliminated distribution bottlenecks. A few decades ago, broadcast television could accommodate about 10 new series per year; and even today, movie theaters in

the US can accommodate about 250 films given that many are released on substantial numbers of screens. But the possibility of watching films and serials directly over the Internet allows for the creation of a great deal more content. The past few years have seen the creation of thousands of new movies per year, as well as literally hundreds of new television series.

While digitization has reduced costs for video production and distribution, it is worth noting that these media remain more expensive than music or books. Music and books can be created by individuals or small groups. Video typically requires a larger number of participants, depending on the subject matter.

A second feature worthy of note is that, particularly in movies, there is a bifurcation between small-scale new products whose success is difficult to predict and larger-scale products, often derivative of prior works, that are both expensive and less risky. Even as the movie industry, broadly construed, has created a large and growing number of new works, most of them small-scale, the traditional major studio players in Hollywood have continued to invest substantial sums in large-scale movies, often sequels to previous movies (see Benner and Waldfogel, 2020).

We would expect the technological changes above to do two things. First, they would facilitate the participation of more potential creators. That is, they would allow greater participation in the entrepreneurial creative labor force. Second, they would make additional products available to consumers. These outcomes would provide greater competition in the product market as well as some possible benefit to consumers.

The workings of both mechanisms depend on the sorts of products facilitated by the easing of entry barriers. If the additional products are unappealing to consumers, then they

would neither divert demand from existing products, nor would they provide much benefit to consumers. On the other hand, if the additional products included some products that consumers found appealing, the the relaxation of entry constraints would both provide competition for existing creative products – and their producers – as well as delivering benefits to consumers.

One well-known feature of creative products is the unpredictability of their appeal to consumers. It is well known that most new creative products fail (Caves, 2000; Vogel, 2014). William Goldman summarized this succinctly with his description of Hollywood executives ability to predict which movies would succeed, with the saying that "nobody knows anything." If this is correct, then a technological change that facilitates broad participation and many new products would be expected to deliver some products of value to consumers and therefore some consequential competition for other producers.

There is substantial evidence that this mechanism operates, the most corroborative of which is that large and growing shares of the successful products since digitization are products which entered the market with low ex ante promise. These include books originally release via self-publishing, music from independent record labels, and movies from independent producers. For example, over a tenth of the USA Today weekly top 150 bestselling books in 2012 began their commercial lives as self-published works. In the romance category, the share was over 40 percent (Waldfogel and Reimers, 2015). Similar evidence exists for music, movies, and television (Waldfogel, 2018).

Evidence that the random long tail mechanism operates does not directly indicate the size of the welfare benefit. The quantification of the welfare benefit is the task undertaken for music in Aguiar and Waldfogel (2018) and which we continue below for other creative products.

a. Products

An important research stream in digitization characterizes the benefit of the Internet through the lens of the "long tail." The idea is that online retailing gives consumers access to a larger number of products than they could obtain from their local retailers. The idea is summarized simply in a diagram showing the cumulative share of sales on the vertical axis and the cumulative share of products on the horizontal.

If all products sold equally well, the cumulative sales would be a straight, 45 degree line. In reality, of course, some products sell more than others, so the top *x* percent of products tends to account for more than *x* percent of sales. As a result, realistic cumulative sales curves initially rise more steeply than the 45 degree line.

The cumulative sales diagram is useful for illustrating the traditional long tail idea. Suppose that traditional brick and mortar stores carry a share, say $\frac{1}{3}$, of the total extant products, as in Figure 1. Then in the absence of online sales, consumers will have access to this share $\frac{1}{3}$, and sales will be at the quantity $q(\frac{1}{3})$. Online retailing gives consumers access to the remaining share $(1 - \frac{1}{3})$ of products, and sales in the presence of online retailing are q(1). Hence, the benefit from the additional sales relates to this difference, $\Delta = [q(1) - q(\frac{1}{3})]$. This is the basis for standard estimates of the benefit of online retailing for consumers (Brynjolfsson et al 2003).[2]

The random long tail idea is different. The idea is not simply that digitization gives consumers access to more extant products. Rather, the idea is that digitization, by reducing the

[2] See also Quan and Williams (2018), who document that terrestrial retailers adapt their assortments to local tastes, so that analysis along the lines of Figure 1 should be done separately by geography.

costs of bringing new products to market, allows the creation of more new products than would otherwise have been brought to market. The predictability of new product quality adds an important element to the story. If products' appeal to consumers were completely predictable at the time of investment, then while a reduction in cost would give rise to additional new products, all of those products would be "worse" than the previous cost threshold. For ease of comparison with the previous example, consider a cost reduction that triples entry (from $\frac{1}{3}$ to 1). Under the old cost threshold, entry occurred out to $\frac{1}{3}$, with associated sales of $q(\frac{1}{3})$. With lower costs – and perfect predictability – more entry occurs, but all of the products have lower realized sales than the products entering with higher costs. Hence, the additional entry – out to 1 – raises total sales to q(1). The benefit of additional entry with perfect predictability is formally equivalent to the traditional long tail benefit. Here, it is $\Delta = [q(1) - q(\frac{1}{3})]$.

It is well known that new product success is very unpredictable in media industries (Caves, 2000). Goldman (2012) colorfully declared that "nobody knows anything" about which potential Hollywood projects would find favor in the marketplace. Taken literally, the idea that nobody knows anything means that technological change giving rise to a growth in the number of products would bring forth products that are as good, on average, as existing products. In that extreme case – and putting aside substitutability across products - the growth in sales with a growth in products would lie along the 45 degree line, at least in expectation. A tripling in the number of product would then give rise to a tripling in sales and a tripling in the surplus associated with new production. It is useful to compare the welfare gain from new products under the "nobody knows" scenario with the standard long tail, in Figure 1.

The term Δ_C represents the standard long tail benefits (of additional/online products, all of which are "worse" than existing/local products), while the term Δ_R represents the "random long tail" benefits of additional products that are as good, on average, as existing products.

While it is easy to come to the conclusion that product success is not perfectly predictable, the polar opposite – that "nobody knows anything" - is a strong assumption that is probably not correct. The crucial point to understand, however, is that the degree of predictability determines the extent to which the additional products made possible by digitization add to welfare. If predictability were perfect, then the additional products would have benefits similar to standard long tail benefits. The lower the degree of predictability, the larger the benefit of new products. This analysis further points to the degree of predictability as a key determinant of the welfare benefits of new entry. Accordingly, the main empirical task of the product market part of this chapter is to use available if imperfect data on movies and books to assess the predictability of product success and the consequent size of the welfare benefit from new products, both absolutely and in comparison with traditional long tail approaches to measurement. That is, we will attempt to estimate Δ_C and Δ_R.

To be clear about the task, suppose we can observe the realized sales for a set of N products after an innovation that allows for additional entry. Order these from the top-selling (q_1) to the bottom selling (q_N), and suppose that absent the innovation, only the share N_0/N of the eventual products would have been produced, where $N_0<N$. Define $Q = \sum_{i=1}^{N} q_i$, and define $Q_o = \sum_{i=1}^{N_0} q_i$. Then the standard long tail benefit of the additional (N-N_0) products is $Q - Q_0$.

To quantify the random long tail benefit, we need to determine which N_0 of the N entering products would have entered absent the innovation. We do this by developing a

prediction of the realized sales of each product, based on information known at the time of investment decisions. Define the sequence of sales, ordered according to predicted sales, as $q_1', q_2', \ldots, q_N'$, where the predicted sales for q_k' exceeds the predicted sales for q_{k+1}', although the realized sales need not decline monotonically. That is, the ordering of products will differ from the ordering based on realized sales if there is imperfect predictability. Absent digitization, the N_0 products brought to market are the N_0 products with highest predicted sales. Output in the absence of digitization is given by $Q_0' = \sum_{i=1}^{N_0} q_i'$, and the welfare benefit of digitization is summarized by $Q - Q_0'$. The greater the predictability, the smaller the benefit of new products.

In particular, I seek to quantify the relative size of the "long tail in production" relative to the "long tail in consumption" for books, television, movies, alongside the quantification for music. Doing this requires two things. First, I need to know the amount by which the entry of new products has increased. Second, I need to calculate the share of sales attributable to the new products.

b. Entrepreneurial Creative Labor Markets

Digitization facilitates entry into the creative product market. A substantial input into production – the predominant input for books and music – is creative labor. Hence, we expect digitization to have consequences for the entrepreneurial creative labor market. It is possible that new modes of consumption, for example audio and video streaming, have expanded the market, raising demand for creative inputs enough for an increase in activity to be accompanied by higher earnings. It is also possible, however, that earnings would fall in the face of more competition. (It is worth noting here that average creative earnings, as opposed to earnings per hour, might also fall as more people are allowed to participate in create entrepreneurial labor markets on a part-time basis).

Since digitization, many artists have raised concerns about artist and intermediary earnings. Former RIAA head Cary Sherman raised concerns about the adequacy of streaming revenues, particularly at YouTube: "But it's harder and harder for more musicians to make a living. Because the revenue that they're getting from streaming isn't keeping pace with the revenue that they used to be able to earn. We're trying to get to a point where the streaming ecosystem works for everybody." [3] Entertainment executive Irving Azoff echoed Sherman's concerns in a tweet stating that "YouTube's below market rates are a threat to artists' livelihood."[4] Producer Kabir Seghal wrote, "Streaming services that we all use like Spotify and Apple Music offer great convenience to fans. But artists are getting a raw deal. The simple truth is musicians need to be paid more for their content."[5] Musician and business school professor David Lowery has written, "My song got played on Pandora 1 million times and all I got was $16.89, less than what I make from a single T-shirt sale."[6] Lowery continues, "… streaming flattens and commoditizes the spin. So you just have one price for every spin of a song across the entire spectrum, whether it's some kind of avant-garde classical work or whether it's a Miley Cyrus song. So that will work if you have lots and lots of spins. But it won't work if you have just a few spins. So what that will do is push out — and you already see that happening — it will push out any sort of niche or, you know … Specialty genres."[7]

[3] https://www.recode.net/2016/4/11/11586030/youtube-google-dmca-riaa-cary-sherman
[4] https://www.digitalmusicnews.com/2018/05/23/youtube-music-threat-artist-livelihood/
[5] https://www.cnbc.com/2018/01/26/how-spotify-apple-music-can-pay-musicians-more-commentary.html
[6] https://thetrichordist.com/2013/06/24/my-song-got-played-on-pandora-1-million-times-and-all-i-got-was-16-89-less-than-what-i-make-from-a-single-t-shirt-sale/
[7] https://www.salon.com/2014/08/31/david_lowery_heres_how_pandora_is_destroying_musicians/

Rights holder concerns are not limited to the music industry. An Author's Guild Survey released in early 2019 describes a "crisis of epic proportions for American authors, particular for literary writers."[8]

Below I seek to add to this discussion some information about official measures of labor market activity – numbers of people working in creative activities – as well as measures of earnings.

III. Data

We need two broad kinds of data for exploring implications of digitization. First, we need data on the product markets. Second, we need data on creative labor markets. Both kinds of data are challenging to obtain; but some useful data are available. We describe them below.

a. Product market data

The ideal data for measuring the welfare consequences of new products consist of three elements. First, we need a measure of the sales of each product in the market. Second, we need relevant variables for predicting the success of products, and these variables need to be known to agents at the time that investment decisions are made. Finally, we need to know the effect of the innovation on the number of products brought to market (i.e. N_0 vs N). These are all somewhat challenging to obtain, and I rely on different sources for different products.

i. Books

[8] https://www.authorsguild.org/industry-advocacy/six-takeaways-from-the-authors-guild-2018-authors-income-survey/

Rather than the entire distribution of sales, I observe the sales ranks for the top 150 best-sellers, by week. These data are drawn from the USA Today Bestseller list, which I have available weekly from 1993-2016. For each entry on the list, I observe the author, title, genre, publisher, and original release date. I have 20,264 distinct titles from 8,239 distinct authors.

These data fall short of the ideal in two respects. First, I do not observe the full distribution of sales across all releases. Rather, I observe only those making the top 150 in at least one week of the year. Second, I do not observe sales quantities. Rather, I observe only sales ranks. I transform sales ranks into quantities using the rough approximation that sales are proportional to the reciprocal of the rank.[9] I then sum these (1/rank) terms across all weeks for which a title enters the bestseller list. This gives me an estimate of total sales. The estimate is deficient in two ways, both that the estimated sales are only approximations to the true values and that I attribute no sales to the titles in weeks when they don't appear in the top 150. Still, the resulting "sales" estimates allow me to calculate a scalar total sales quantity per title.

I have no direct way to deal with the problem that I observe only the head of the sales distribution except to amend my empirical exercise. Rather than studying the predictability of product success among all released titles, I study the predictability of success among those achieving top-150 status in at least one week. Given the evidence, cited above, that many works with low ex ante promise become best sellers, I can be confident that the head of the sales distribution contains diversity of works according to their ex ante promise. Because I have bestseller lists back to 1993, I am able to construct author-specific past sales measures, which I

[9] This is an approach common in the analysis of rank data. See, for example, Chevalier and Goolsbee (2003).

can use to help predict the success of the current release. Other variables potentially relevant to predicting product success include genre and publisher.

b. Movies

I observe all US-released movies, 1980-2016. The movie data fall short of the ideal in one major respect. While I would like to observe the full distribution of revenue across movies, the only revenue data that are systematically available are box office revenues. These are important for movies in wide release, but this measure misses much of the revenue for movies made possible by digitization, which are generally distributed mainly – and sometimes exclusively – outside of theaters (see Benner and Waldfogel, 2020).

What I use instead is a measure of interest that I can obtain for every movie, the number of IMDb users rating each movie. This measure is highly correlated with box office revenue for titles where box office revenue is available, providing some support for its use as a sales proxy. IMDb provides a great deal of information that is potentially relevant to the prediction of movie success (again, measured by the number IMDb ratings). These variables include the production budget, the genre, the identities and past success of the major actors, and the production company. My effective movie database contains 34,279 movies.

c. Television Data

My television data are also drawn from IMDb. I use have information on 16,159 television series produced between 1948 and 2016. I include those with a reported rating on IMDb, which therefore have at least five persons rating the show. As with movies, I use the number of persons rating the show as a measure of its success. I use the following variables for predicting success. I have the show's classification into one of 52 genres and its three most important cast members.

I calculate each cast member's experience as the number of series they had appeared in prior to the current series.

d. Labor Market Data

Ideally, I would have data on time spent on, and earnings derived from, new creative products. That way, I could measure both time spent making creative products, as well as both the overall earnings of those involved and the return to such activities, e.g. the earnings per hour of effort. What I actually have, while substantial, falls short of the idea. I have household surveys as well as data from tax returns indicating how many people filing a Schedule C as a nonemployer working in creative activities.

The household survey providing information on employment by occupation is the American Community Survey (ACS). The main purpose of the ACS is to provide "annual (or multi-year average) estimates of selected social, economic, and housing characteristics of the population for many geographic areas and subpopulations."[10] The ACS is based on surveys of 3 million addresses per year. The ACS asks respondents their occupations and their incomes and contains sampling weights that allow for the creation of population estimates. Table 1 lists the relevant creative occupations in the ACS.[11]

A second government data source of interest covers "nonemployer establishments." These data, from tax records, provide another possible glimpse into creators' labor force activity. Self-employed individuals with business income are required to complete a Schedule C. In filling out this form, the individual also indicates their industry. The Internal Revenue Service

[10] https://www.census.gov/topics/income-poverty/poverty/guidance/data-sources/acs-vs-cps.html
[11] The Current Population Survey (CPS) has a similar approach but much smaller coverage. Efforts to detect evidence of an increase in creative activity among individuals in creative occupations were unsuccessful with the CPS.

maintains statistics on nonemployer establishments with Schedule C filings of $1,000 or more. Industries relevant to the creation of books, music, movies, and television include those listed in Table 2.

"Nonemployer Statistics (NES) is an annual series that provides subnational economic data for businesses that have no paid employees and are subject to federal income tax. The data consist of the number of businesses and total receipts by industry. Most nonemployers are self-employed individuals operating unincorporated businesses (known as sole proprietorships), which may or may not be the owner's principal source of income. Statistics are available on businesses that have no paid employment or payroll, are subject to federal income taxes, and have receipts of $1,000 or more."[12] While these data are technically available at the industry level, the nonemployer "establishments" are generally self-employed individuals.

IV. Results: Welfare Benefits of New Products

A natural way to quantify the welfare benefit of new products is to estimate a utility-theory consistent demand model that allows calculation of consumer surplus as a function of the products in the choice set. Aguiar and Waldfogel (2018) present such an approach, while also documenting that the size of the random long tail in relation to the conventional long tail is well summarized with a simple calculation. That simple calculation is the ratio of the share of sales accounted for by the ex ante long tail to the share of sales in the ex post long tail.

Accordingly, I estimate the welfare benefit of digitization by ascertaining which of recent products only exist because of digitization. To do this, I attempt to determine which among a

[12] https://www.census.gov/programs-surveys/nonemployer-statistics/about.html

set of recent products, had modest ex ante probabilities of success. I assume that, say, *x* percent products would not have come to market absent digitization. I then ask what share of current sales are accounted for by the products which would have been created without digitization. Finally, I compare this "random long tail" in production with something analogous to the standard long tail, the share of sales accounted for by the lowest-selling *x* percent of new products.

Doing this requires two steps. First, I need to determine which among a crop of recent products would not have been produced but for digitization. For this purpose I predict product success using information available at the time of entry. I assume that the products with low ex ante probabilities of success (the "ex ante losers") would have come to market without digitization. I then quantify the share of sales accounted for by the ex ante losers, which I view as a rough estimate of the welfare gain from digitization.

a. Predicting Ex Ante Product Success

I am interested in predictions of product success, as opposed to explanation. Hence, I use predictive tools suited to this purpose. In particular, I use cross-validated LASSO regressions. For each of the three products – books, movies, and television series – I regress the log of my "sales" measure on interactions of the explanatory variables described above. I allows the cross validation procedure to choose the penalty parameter that minimizes out-of-sample mean squared error.

To predict the success of individual books and movies, I regress measures of "sales" for an entering cohort of products on various explanatory variables and interactions. For books these include: interactions of publisher, genre, publication year, and authors' prior sales, for a

total of 179 possible explanatory variables. From these, the LASSO procedure selects 146 for inclusion. For movies, these include interactions of genre, budget, and year for a total of 102 explanatory variables. LASSO includes 85 of these variables. For television series these include 191 possible variables, and LASSO selects only 31. The resulting models explain different shares of the variation across products. The R-squared for movies is 0.57, while it's 0.21 for books, and 0.11 for television shows. Table 3 summarizes. It is interesting that the movie industry, which inspired the phrase, "nobody knows anything," has the highest share of variance explained by the regression. The lower R-squared values for the other products suggests higher random long tail benefits for those products, relative to the conventional long tail.

b. Welfare Effects

The sales predictions above ($\hat{q}_i$) allow us to order products according to ex ante promise. Then given the number of products that would have been produced but for the innovation that reduced the cost threshold, we can calculate the realized sales that the chosen products would have delivered. The top panels of Figures 2-4 report these results via comparisons between the cumulative sales distributions ordered according to realized vs predicted sales, for each of the three products for particular recent years (2016 for books and movies and 2015 for television). The smooth, upper lines show the cumulative sales in decreasing order according to realized sales. The lower, jagged line shows the cumulative realized sales but ordered according to expected sales. By construction, both lines begin at the origin and terminate in the sale cumulative sales. But they diverge between the extremes according because of imperfect prediction.

Patterns differ fairly substantially among books, movies, and television series. First, realized sales are far more concentrated for movies and television shows than for books. We see this in the initial steepness of the realized sales for movies and television series. The gini coefficients bear out the comparison: 0.935 for television and 0.938 for movies, compared with 0.806 for books. This means that the conventional long tail is larger for books than for the others. Second, movie success is far more predictable than television or book success. We see this in the proximity of the jagged line – sales ordered by ex ante promise – to the smooth one for movies.

What do these patterns mean for the welfare benefits of digitization? We have two measures of interest, both of which depend on the number of new products which would have been produced absent digitization. First, we can quantify the random long tail in relation to the conventional long tail (Δ_R/Δ_C). Second, we can measure the share of total sales attributable to products made possible by digitization.

Consider first the bottom panel of Figure 2, for movies. The downward-sloping line shows the share of total sales accounted for by the new products made possible by digitization. The vertical line at 250 reflects the idea that the movie industry produced roughly 250 movies per year prior to digitization. At *N*=250, the welfare gain – measured as additional revenue – is about 10 percent of revenue.[13] How large is this in absolute terms? As Table 4 shows, US box office revenue in 2016 was $11.4 billion. As of the early 2000s, box office revenue accounted for 17.9 percent of overall Hollywood revenue. This suggests that total US movie industry

[13] This 10 percent is the difference between the total revenue from all products and the value of the ex ante line at N=250, divided by total revenue.

domestic revenue is on the order of $63 billion. Hence, the share of revenue attributable to products that exist only because of digitization is ten percent of $63 billion, or about $6.3 billion.

We can do a similar calculation for television. The bottom panel of Figure 3 two things. First, prior to digitization, there were roughly 100 new shows per year. Second, the Figure's downward-sloping line shows that roughly half of television industry "sales" are attributable to products beyond the first 100, those made possible by digitization. Television industry revenue is difficult to calculate, since some of television of broadcast on ad-supported networks, while other television is distributed via subscriptions (e.g. HBO or Netflix). We can get a rough sense of the order of magnitude of the industry from annual production costs. These came to $37 billion in the US for 2013. On the logic that production occurs in the expectation of revenue in excess of production costs, the production expenditures would provide an underestimate of aggregate revenue. Half of the $37 billion would be $18.5 billion.

Books are slightly more complicated in that we don't observe the entire population of new works. To perform the analogous calculation on books, we need to know the number of the bestsellers, rather than total works, that would have existed absent digitization. This is difficult to say for sure. Since the mid-2010s about 10 percent of bestsellers were works that came to market as self-published books. It is difficult beyond that to say what share of bestsellers only came to market because of digitization. The bottom panel of Figure 4 has a vertical line 1500, as if 1500 of the bestsellers would have existed absent digitization. Under that assumption, about 10 percent of the sales of bestsellers would be for books made possible by digitization. US book sales were about $26 billion in 2016, so books made possible by digitization account for about $2.6 billion of this.

And how large is the random long tail relative to the conventional long tail? Evaluated at the vertical lines in the bottom panels of Figures 2-4, the ratio Δ_R/Δ_C - which was roughly 20 for music in Aguiar and Waldfogel (2018) - is 3.83 for movies, 12.89 for television, and 8.62 for books. Here, too, the random long tail is much larger than its conventional counterpart.

V. Results: Labor Market Outcomes

We know that the numbers of new products have risen sharply, in books, music, television, and movies. The creation of these products requires some activity by people, which might appear in labor market statistics. That is, the product creation documented above reflects entrepreneurial labor market activity by creative individuals. The resulting products, as we have seen, have varying degrees of success. Moreover, the existence of a large number of new products provides competition for other products, with possible consequences for the returns to creating new products. We explore these questions below in turn. The questions here have clear parallels to research on whether entrepreneurship pays. Some important examples include Hamilton (2000) and Moskowitz Vissing-Jørgensen (2002), who find that entrepreneurship does not pay, and Manso (2016), who finds that it does, when option value is properly measured.

1. Can We See Digitization-Enabled Creative Activity in the Government Data?

Our first question is a mundane but important one: do the available data sources, the American Community Survey and the IRS nonemployer statistics, reflect the activity underlying the increase in the number of creative products created? Before turning to this question, we can

make an easier ask of these data sources: do they indicate the growth in drivers apparently working for Uber and Lyft? Uber's revenue grew from $0.1 billion in 2013 to $6.5 billion in 2016 and reached $11.3 billion in 2018. The growth has been rapid and abrupt, and rides require drivers, so it should be possible to see evidence of this new digitization-enabled activity in data.

Among the occupations in the ACS is the category of "taxi driver and chauffeurs." Figure 5 shows the number of people reporting that they work in this occupation in the ACS. The figure rises slowly from about 400,000 to 500,000 between 2000 and 2013. Between 2013, and 2017, the figure rises by another 300,000, topping 800,000 in 2017. This coincides well with the rapid growth in ridesharing apps, particularly Uber, documented in Hall and Krueger (2016).

The nonemployer statistics provide similar corroboration. Figure 6 shows the number of nonemployer establishments NAICS code 4853 ("taxi and limousine services") rising from about 100,000 in the late 1990s to about 200,000 in 2013. By 2016, the number was about 700,000. At least for occupations with abrupt growth, the ACS and IRS statistics corroborate what one expects for underlying activity.

With Figure 7, we turn to numbers of individuals working in creative occupations in the ACS. The four relevant occupations continuously available using the 2010 occupation classifications include actors, producers, and directors; musicians, singers, and related workers; writers and authors; and photographers. All show substantial growth over the period 2000-2016. Actors grow from 200,000 to nearly 300,000. Musicians grow from 200,000 to almost 280,000. Writers and authors grow from under 200,000 in 2000 to over 300,000 in 2016, and there is a

jump in 2012, which coincides with the Kindle era at Amazon.[14] Photographers grow from 150,000 to nearly 250,000.

Figure 8 shows aggregate earnings in each category from the ACS. Despite fluctuations, aggregate earnings rise in all but the photography category. Figure 9 shows what happened to real average earnings in each of these categories. While all fluctuate year to year, there are clear downward trends. As the number of people working in these occupations have risen, the average earnings per worker has declined.

Figure 10 documents the evolution of creative occupation employment according to the IRS nonemployer statistics. Here the relevant categories are independent artists, writers, and performers (NAICS 7115), sound recordings (NAICS 5122), motion pictures (NAICS 5121), and publishing except internet (NAICS 511). The first – and broad – category grows steadily and sharply over the digital era, from about 425,000 in 1997 to about 850,000 in 2016. Sound recording and motion picture nonemployer establishments also grow, but by much smaller absolute amounts. Publishing grows quickly from 1997 to about 2004, then holds steady.

Digitization's enablement of creative work has no discrete date as clear as, say, the arrival of Uber. Hence, it is difficult to say whether the broad growth of individuals filing Schedule C's for nonemployer establishments in creative industries is specifically caused by digitization.

The IRS data are nevertheless potentially useful for documenting the evolution of both total self-employment earnings in these occupations, as well as the average earnings per filer. Figure 11 aggregates the four NAICS codes together. The top panel shows the substantial

[14] December 2011 saw the peak search volume on the term "Amazon Kindle" according to Google Trends. See https://trends.google.com/trends/explore?date=all&geo=US&q=%2Fm%2F03d068f .

growth in individuals across these categories, from about half a million to a million. The second panel shows that the total earnings have risen from about $16 to $24 billion. The third panel shows that the average earnings have fallen from $30,000 in 1997 to about $24,000 in 2009 and have remained at that level in real terms to 2016.

The tax return-based figures appear to confirm much of what's evident in the ACS data. First, there is quite substantial growth in the number of establishments (individuals) creating works for money. This provides evidence that they large outpouring of new works is generating income for the individuals creating it. The IRS data also show that the per capita business income of those individuals with this income is falling, by roughly 10 percent in the largest category and by much more in the more specific categories.

Even if the data are relatively clear, much remains unanswered. That is, while the government data do reflect the activity manifesting itself as a growth in new products, it is not clear that the reduction in average earnings reflects falling returns to creative entrepreneurship, as opposed to a changing mix of people involved in the activities.

Figure 12 provides suggestive evidence that composition – and the influx of new workers – explains the decline in average earnings over time. The figure presents the 90^{th} , 50^{th} , and 10^{th} percentiles of the ACS log earnings distributions, by category. At the top and the middle of the distributions, earnings are stable over time. Earnings at the bottom of the distribution, by contrast, fall substantially.

Conclusion

Digitization has changed the conditions surrounding the production of creative products. Less capital is required, so not only has there been more entry; there has also been a shift of new product creation outside of traditional firms. To put this another way, digitization has enabled viable creative entrepreneurship that would have been difficult earlier. The results of these changes include substantial benefits to consumers, in the form of products accounting for substantial shares of sales that would not have existed without digitization. These products are made available because many more would-be creators are able to bring new products to market; and as with ridesharing drivers, we can see this activity in government data. Activity is rising, as are total earnings of creative workers; but average earnings are falling, particularly at the bottom of the earnings distribution. It is difficult to draw more nuanced conclusions about returns with existing data; but it seems a topic fruitful for additional research.

References

Abraham, Katharine G., John C. Haltiwanger, Kristin Sandusky, and James R. Spletzer. 2017. "Measuring the Gig Economy: Current Knowledge and Open Issues." March 2.

Aguiar, L. and Waldfogel, J., 2018. Quality predictability and the welfare benefits from new products: Evidence from the digitization of recorded music. *Journal of Political Economy*, *126*(2), pp.492-524.

Arrow, K. 1969. Classificatory notes on the production and diffusion of knowledge. American Economic Review 59:29–35.Benner, MJ, Waldfogel, J. Changing the channel: Digitization and the rise of "middle tail" strategies. *Strat. Mgmt. J*. 2020; 1– 24. https://doi.org/10.1002/smj.3130

Bergemann, D. and Hege, U., 2005. The financing of innovation: Learning and stopping. *RAND Journal of Economics*, pp.719-752.Brynjolfsson, E., Hu, Y. and Smith, M.D., 2003. Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers. *Management Science*, *49*(11), pp.1580-1596.

Caves, R.E., 2000. *Creative industries: Contracts between art and commerce* (No. 20). Harvard University Press.

Chevalier, J. and Goolsbee, A., 2003. Measuring prices and price competition online: Amazon. com and Barnes and Noble.com. *Quantitative marketing and Economics*, *1*(2), pp.203-222.

Cuntz, A. and Miller, A.L., *Unpacking predictors of income and income satisfaction for artists* (Vol. 50). WIPO.

Cuntz, A., 2018. *Creators' Income Situation in the Digital Age* (No. 755). LIS Cross-National Data Center in Luxembourg.

Ewens, M., Nanda, R. and Rhodes-Kropf, M., 2018. Cost of experimentation and the evolution of venture capital. *Journal of Financial Economics*, *128*(3), pp.422-442.Goldman, W., 2012. *Adventures in the screen trade*. Hachette UK.

Hall, Jonathan V. and Alan B. Krueger. 2016. An Analysis of the Labor Market for Uber's Driver-Partners in the United States. NBER Working Paper 22843.

Hamilton, B.H., 2000. Does entrepreneurship pay? An empirical analysis of the returns to self-employment. *Journal of Political economy*, *108*(3), pp.604-631.

Jackson, Emilie, Adam Looney, and Shanthi Ramnath. 2017. "The Rise of Alternative Work Arrangements: Evidence and Implications for Tax Filing and Benefit Coverage. Office of Tax Analysis Working Paper 114, January.

Katz, Lawrence F. and Alan B. Krueger. 2016. "The Rise and Nature of Alternative Work Arrangements in the United States, 1995-2015" NBER Working Paper 22667

Katz, Lawrence F. and Alan B. Krueger. 2019. "Understanding Trends in Alternative Work Arrangements in the United States." NBER Working Paper 25425.

Kerr, W., R. Nanda, and M. Rhoder-Kropf. 2014. Entrepreneurship as experimentation. Journal of Economic Perspectives 28:25–48.

Manso, G., 2011. Motivating innovation. *The Journal of Finance*, *66*(5), pp.1823-1860.

Manso, G., 2016. Experimentation and the Returns to Entrepreneurship. *The Review of Financial Studies*, *29*(9), pp.2319-2340.

Moskowitz, Tobias, J., and Annette Vissing-Jørgensen. 2002. "The Returns to Entrepreneurial Investment: A Private Equity Premium Puzzle?" *American Economic Review*, 92 (4): 745-778.

Quan, T.W. and Williams, K.R., 2018. Product variety, across-market demand heterogeneity, and the value of online retail. *The RAND Journal of Economics*, *49*(4), pp.877-913.

Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. IPUMS *USA: Version 8.0* [dataset]. Minneapolis, MN: **IPUMS**, 2018. https://doi.org/10.18128/D010.V8.0

Vogel, H.L., 2014. *Entertainment industry economics: A guide for financial analysis*. Cambridge University Press.

Waldfogel, J., 2016. Cinematic explosion: New products, unpredictabilty and realized quality in the digital era. *The Journal of Industrial Economics*, *64*(4), pp.755-772.

Waldfogel, J., 2017. The random long tail and the golden age of television. *Innovation Policy and the Economy*, *17*(1), pp.1-25.

Waldfogel, J., 2018. *Digital Renaissance: What Data and Economics Tell Us about the Future of Popular Culture*. Princeton University Press.

Waldfogel, Joel, 2017. How Digitization Has Created a Golden Age of Music, Movies, Books, and Television. *Journal of Economic Perspectives*, *31*(3), pp.195-214.

Waldfogel, J. and Reimers, I., 2015. Storming the gatekeepers: Digital disintermediation in the market for books. *Information economics and policy*, *31*, pp.47-58.

Weitzman, M.L., 1979. Optimal search for the best alternative. *Econometrica: Journal of the Econometric Society*, pp.641-654.

Table 1: ACS creative occupations (2010 definition), plus taxi and limo

occupation
Artists and Related Workers
Actors, Producers, and Directors
Musicians, Singers, and Related Workers
Entertainers and Performers, Sports and…
Editors, News Analysts, Reporters, and…
Writers and Authors
Media and Communication Workers, nec
Broadcast and Sound Engineering Technic…
Photographers
Television, Video, and Motion Picture C…
Taxi Drivers and Chauffeurs

Table 2: Codes for schedule C and therefore for nonemployer statistics

NAICS code	Name	2016 establishments
711510	Independent artists, writers, & performers	849,176
511000	Publishing industries (except Internet)	72,348
512100	Motion picture & video industries (except video rental)	83,331
512200	Sound recording industries	25,206

Notes: from 2018 Instructions for Schedule C, Principal Business or Professional Activity Codes, p C-17, at https://www.irs.gov/pub/irs-pdf/i1040sc.pdf. From page C-3: "Enter on line B the six-digit code from the Principal Business or Professional Activity Codes chart at the end of these instructions."

Table 3: Product success prediction

	Television	Movies	Books
# possible variables	191	102	179
# chosen by LASSO	31	85	146
R2 out of sample	0.110	0.5721	0.2151

Note: For each product I run a LASSO model relating log sales or its proxy to potential predictors, including past measures of author or actor success, genre, etc.

Table 4: Revenue, products absent digitization, and Δ_R/Δ_C

	US Revenue	**Products absent digitization**	$\boldsymbol{\Delta_R/\Delta_C}$
Books	$26.27 b (2016)	1500	8.62
Television	$37 billion (2013)	100	12.89
Movies	$63 billion = $11.4/0.179 (2016)	250	3.83

Notes: book revenue (https://www.statista.com/statistics/271931/revenue-of-the-us-book-publishing-industry/). Movie (https://www.latimes.com/business/hollywood/la-fi-ct-mpaa-annual-report-20180404-story.html) - US box office only. For box office as a share of total revenue, see http://www.edwardjayepstein.com/table2.htm . Box office = 17.9 percent. Television production revenue (https://www.statista.com/statistics/293450/revenue-of-television-production-in-the-us/).

Table 5: Disintermediating industries: those in which nonestablishment growth exceeds employment growth, 1999-2016 (Creative industries in bold)

NAICS code	industry name	growth in nonemployer establishments, 1999-2016	growth in CBP employment, 1999-2016	nonemployer establishments, 2016	CBP employment, 2016
812	Personal and Laundry Services	1,208,604	179,902	2,720,918	1,441,285
531	Real Estate	1,026,738	337,006	2,595,577	1,563,001
485	Transit and Ground Passenger Transportation	716,632	145,970	869,052	515,992
711	**Performing Arts, Spectator Sports, and Related Industries**	**610,364**	**173,370**	**1,221,596**	**503,751**
7115	**Independent Artists, Writers, and Performers**	**367,394**	**9,752**	**849,176**	**46,638**
484	Truck Transportation	217,607	76,420	587,038	1,460,598
811	Repair and Maintenance	128,181	-53,455	747,224	1,265,012
492	Couriers and Messengers	70,242	33,578	197,355	611,946
115	Support Activities for Agriculture and Forestry	26,013	-762	112,936	97,574
5122	**Sound Recording Industries**	**14,239**	**-265**	**25,206**	**22,940**
511	**Publishing Industries (except Internet)**	**9,211**	**-88,098**	**72,348**	**916,599**
339	Miscellaneous Manufacturing	8,972	-193,183	72,089	541,059
325	Chemical Manufacturing	8,043	-119,583	14,302	766,771
336	Transportation Equipment Manufacturing	7,198	-401,944	10,769	1,504,272
221	Utilities	5,266	-28,145	19,613	638,917
332	Fabricated Metal Product Manufacturing	5,020	-382,218	38,222	1,406,266
481	Air Transportation	4,233	-116,398	20,585	466,440
442	Furniture and Home Furnishings Stores	3,018	-72,091	40,015	453,251
315	Apparel Manufacturing	2,176	-478,117	26,412	96,791
337	Furniture and Related Product Manufacturing	2,046	-254,251	18,119	368,902
532	Rental and Leasing Services	1,445	-109,511	79,373	512,405
314	Textile Product Mills	1,355	-108,958	3,706	113,013
334	Computer and Electronic Product Manufacturing	1,266	-828,790	9,461	786,387
316	Leather and Allied Product Manufacturing	1,198	-48,288	5,246	25,678
313	Textile Mills	1,056	-260,334	2,020	101,952
331	Primary Metal Manufacturing	622	-221,750	4,069	375,873
324	Petroleum and Coal Products Manufacturing	441	-4,356	1,568	104,748
483	Water Transportation	342	-6,712	6,645	65,132
333	Machinery Manufacturing	95	-367,476	15,487	1,030,750

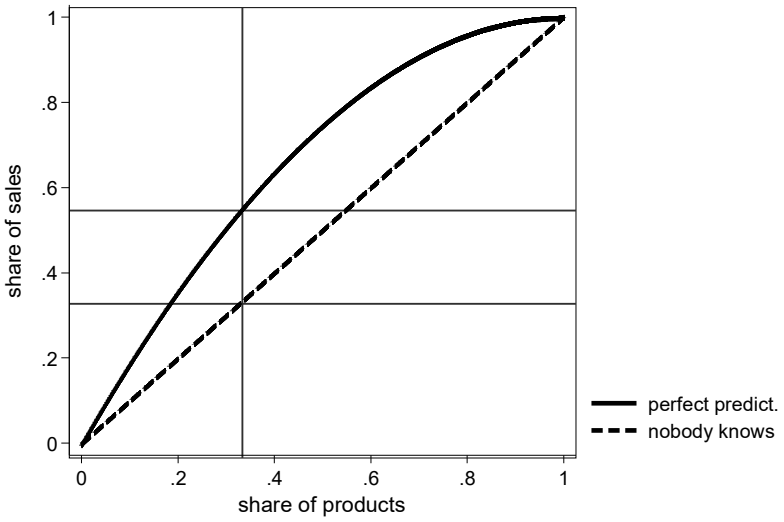

1
.8
.6
.4
.2
0
share of sales
0
.2
.4
.6
.8
1
share of products
perfect predict.
nobody knows

Figure 1

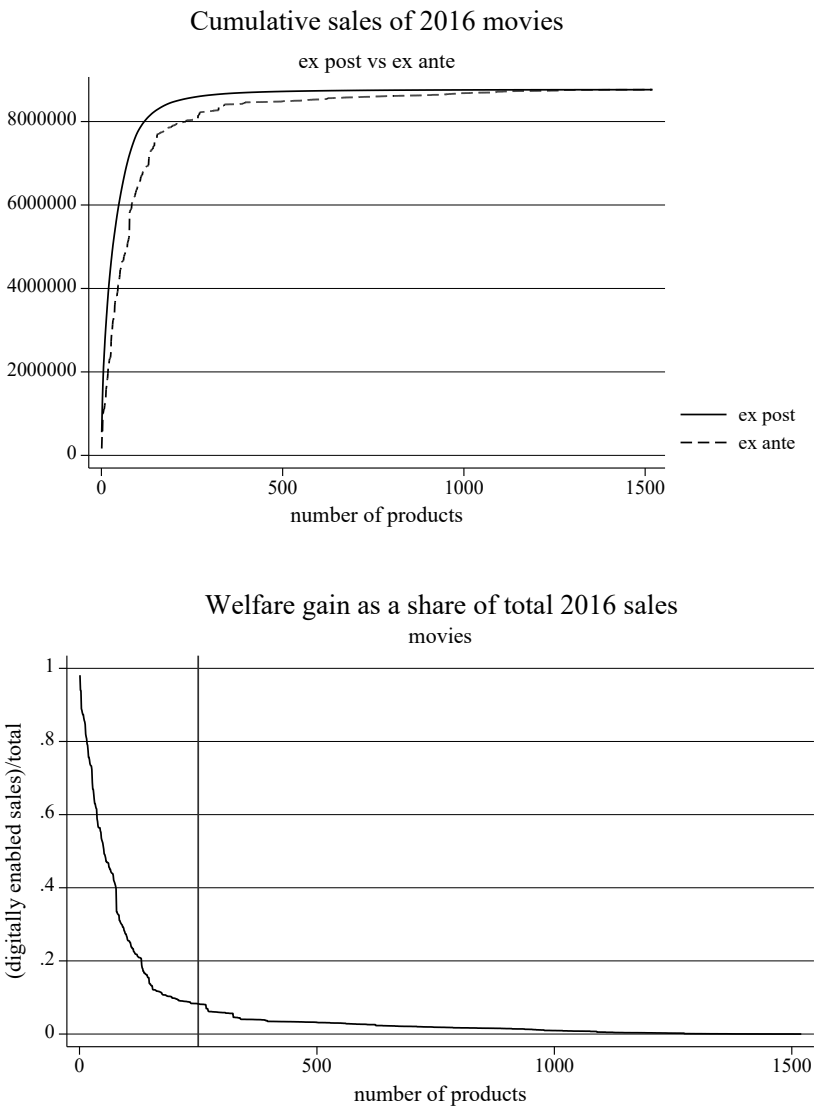
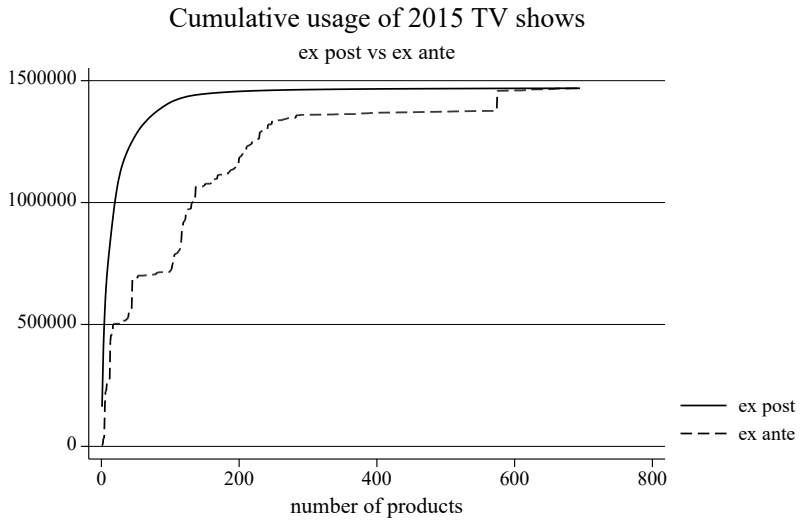

Cumulative sales of 2016 movies
ex post vs ex ante
8000000
6000000
4000000
2000000
0
0
500
1000
1500
number of products
ex post
ex ante
Welfare gain as a share of total 2016 sales
movies
(digitally enabled sales)/total
1
.8
.6
.4
.2
0
0
500
1000
1500
number of products

Figure 2


Cumulative usage of 2015 TV shows
ex post vs ex ante
1500000
1000000
500000
0
0
200
400
600
800
number of products
ex post
ex ante

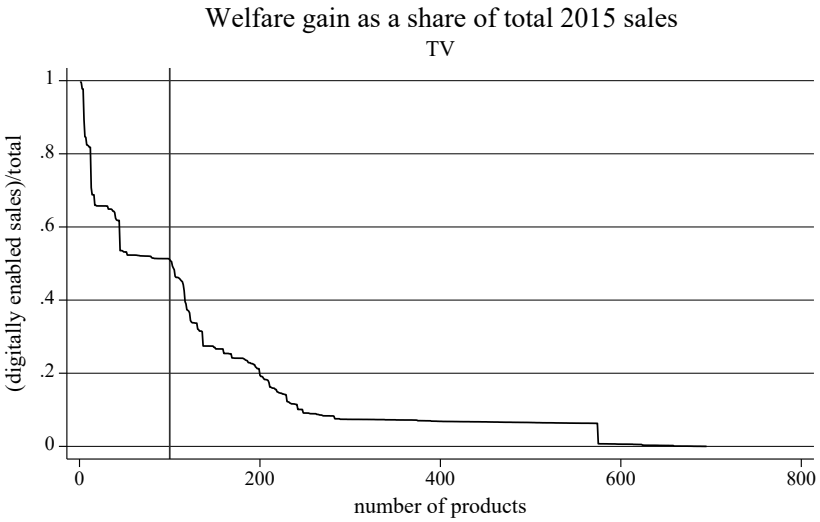

Welfare gain as a share of total 2015 sales
TV
(digitally enabled sales)/total
1
.8
.6
.4
.2
0
0
200
400
600
800
number of products

Figure 3

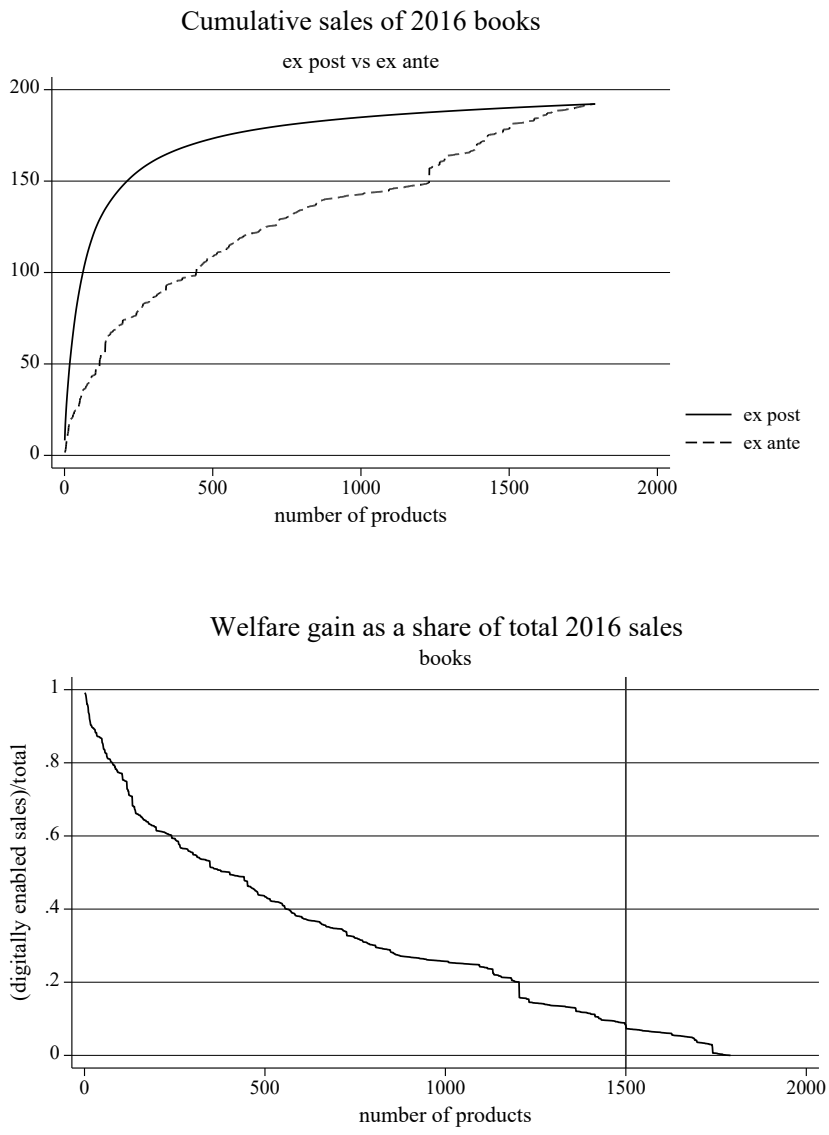

Cumulative sales of 2016 books
ex post vs ex ante
200
150
100
50
0
0
500
1000
1500
2000
number of products
ex post
ex ante
Welfare gain as a share of total 2016 sales
books
(digitally enabled sales)/total
1
.8
.6
.4
.2
0
0
500
1000
1500
2000
number of products

Figure 4

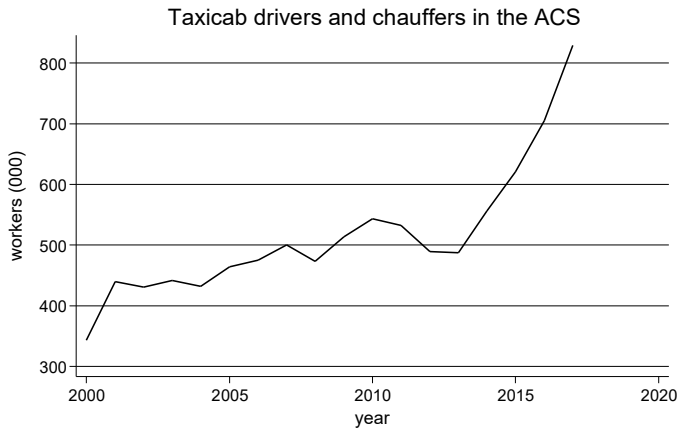


Figure 5

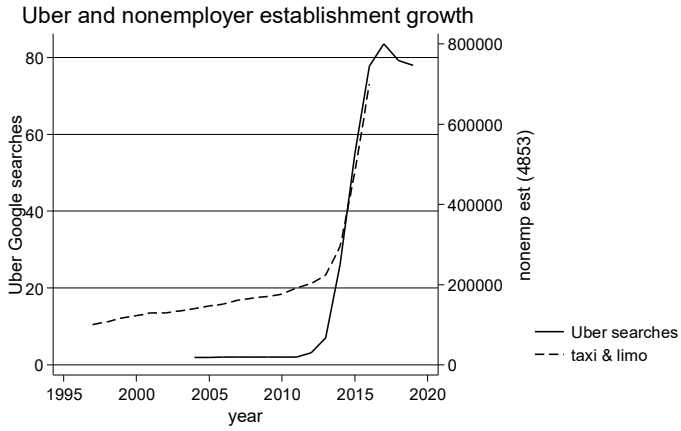


Figure 6

Using occ2010

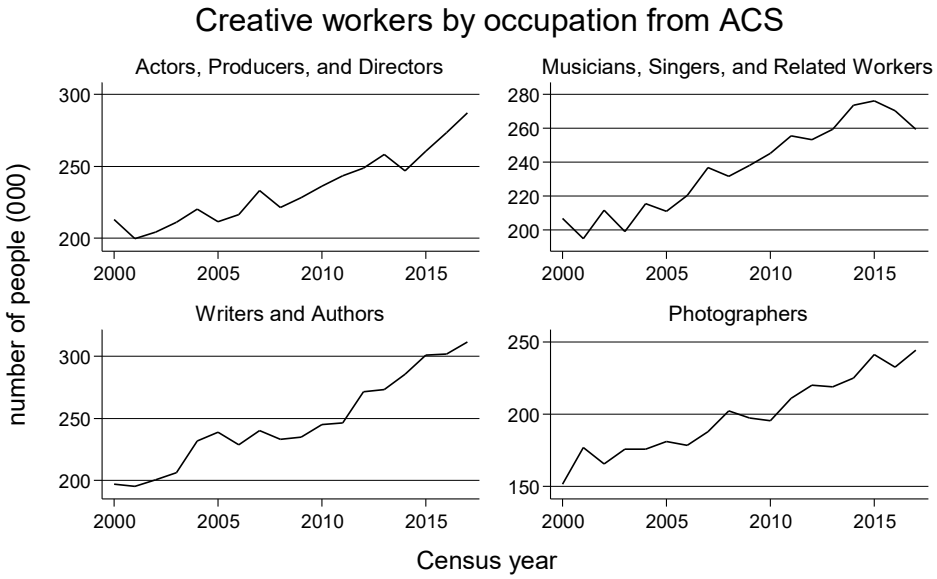


Figure 7

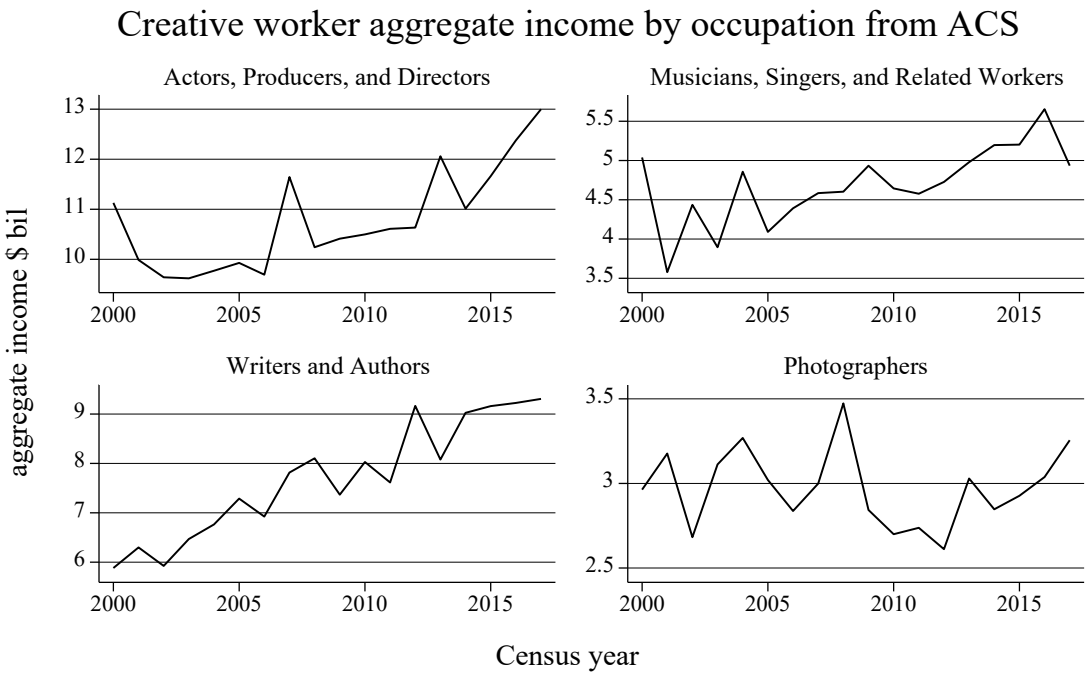


Figure 8

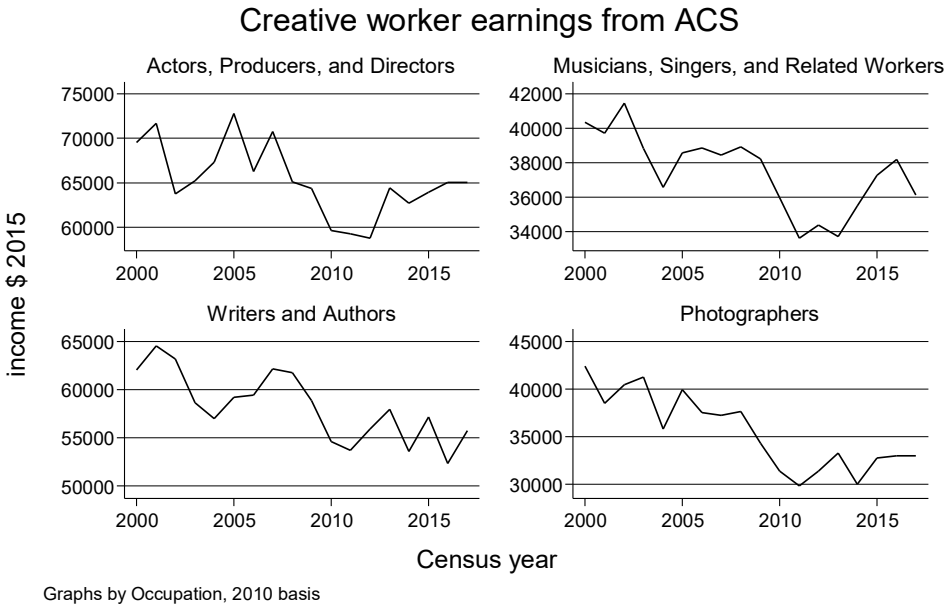

Creative worker earnings from ACS
Actors, Producers, and Directors
Musicians, Singers, and Related Workers
Writers and Authors
Photographers
income $ 2015
Census year
Graphs by Occupation, 2010 basis

Figure 9

Figure 10: Nonemployer establishments related to books, music, movies, and television

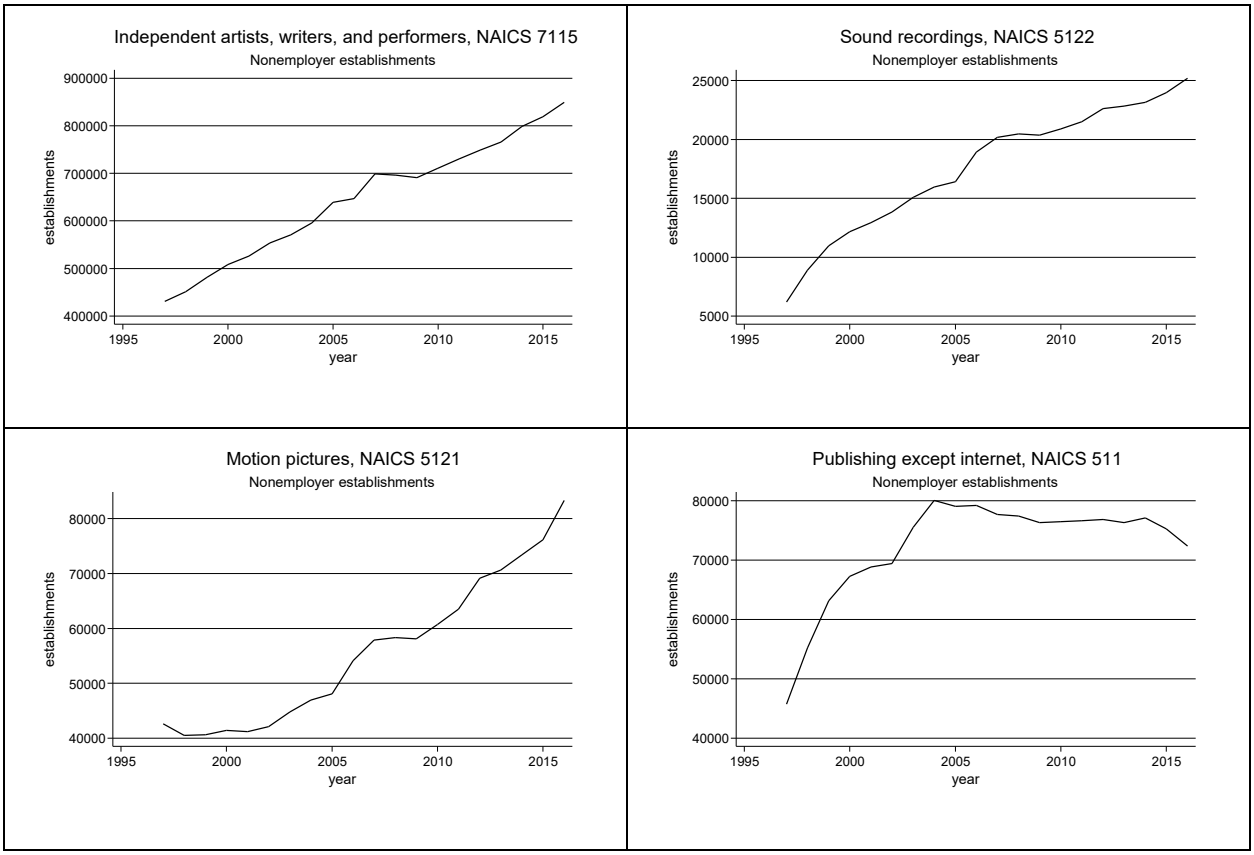
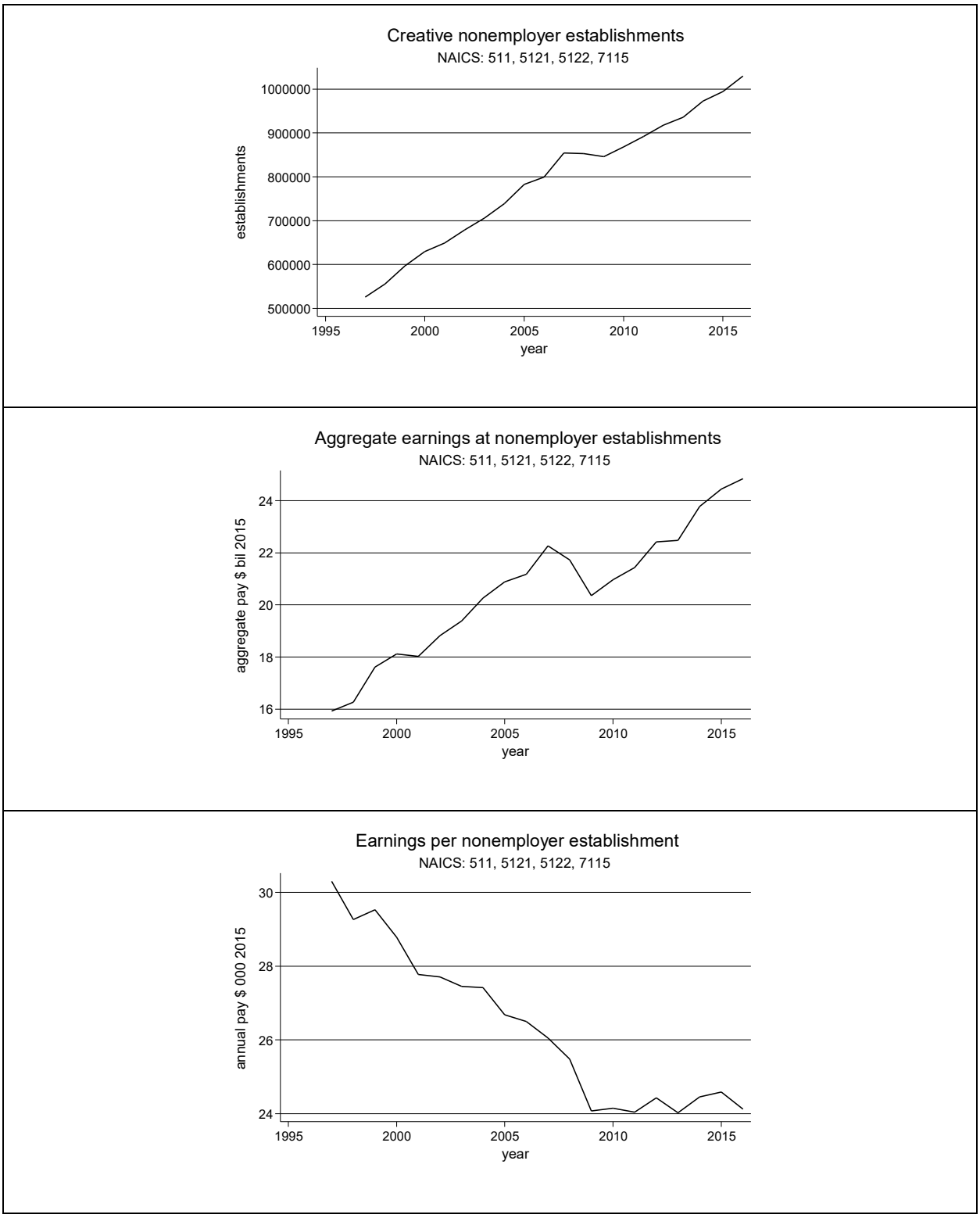
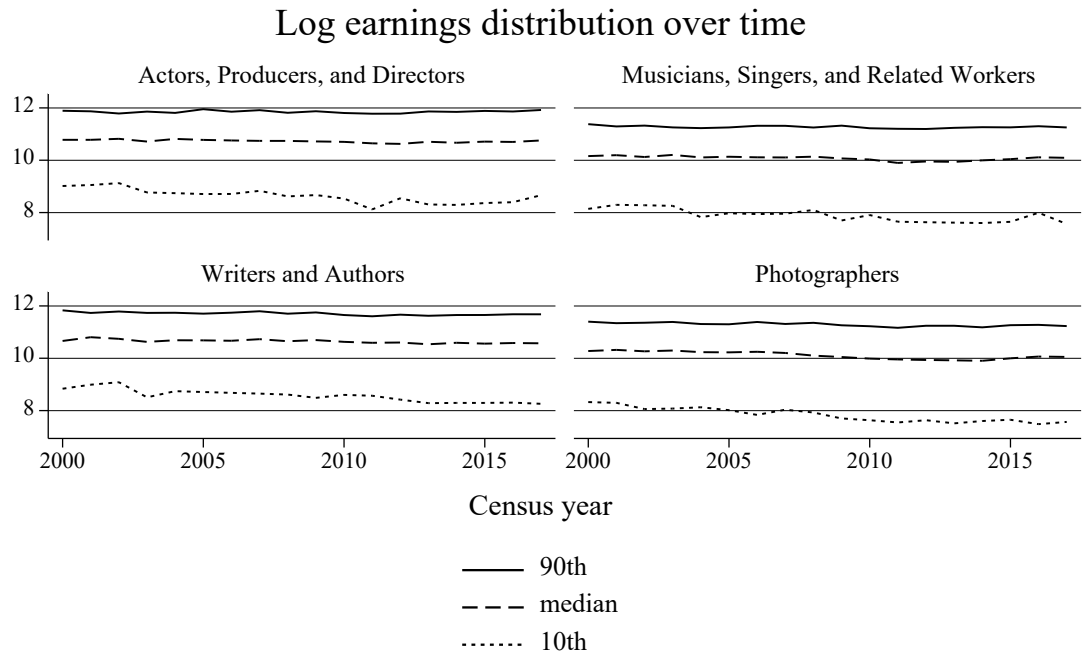


Figure 11: Aggregate and per capita earnings at creative nonemployer establishments





Graphs by Occupation, 2010 basis

Figure 12