Manhattan Transfer: Productivity effects of agglomeration in American authorship

Lukas Kuld*, Sara Mitchell*, and Christiane Hellmanzik*

July 21, 2021

Abstract

We investigate quantity and quality effects of agglomeration in the careers of American authors. We combine novel yearly data on publications and work location of 471 eminent authors with US Census data to provide a complete picture of industry concentration and agglomeration economies from 1850-2000. We find that, on aggregate, an author has 40% higher odds of publishing while living in New York City. The effect size increases with industry concentration but declines with industry maturity and technological progress after WWII. Taking relocation of working-age authors to New York City as an event study, we see a significant immediate increase in publications after arriving. In comparison, the penalty of moving away from the city is mild. Works published while an author lived in New York City were more likely to achieve critical acclaim and are more likely to have lasting influence in terms of present-day popularity.

JEL-codes: N30 • N90 • R11 • Z11

Keywords: Agglomeration economies, urban history, geographic clustering, productivity, literature, creativity

*Department of Economics, TU Dortmund, Vogelpothsweg 87, 44227 Dortmund, Germany. Corresponding author:lukas.kuld@tu-dortmund.de

The authors thank Maria Fleming, Seán McGuane, Vivien Gaida, Leyla Ruzdija, and Jens Kirsten for their excellent research assistance. We also thank John O'Hagan for his support and supervision of the first stage of the data collection process, and the Department of Economics at Trinity College Dublin for their financial support of the first stage of data collection.

1 Introduction

During the last decade of the 19th century, the economics of American literary writing entered a period of rapid transformation. Fast growing demand from an expanding base of educated readers, falling costs due to new printing technologies, and legal changes that eased the trade of literary texts induced the transformation of the publishing trade into a modern industry during the first half of the 20th century. In this paper, we investigate the existence and magnitude of agglomeration economies in 150 years of American literary production observing both the rise and relative decline of New York as a literary cluster.[1] We show how New York City offered significant productivity benefits to authors residing there during these early stages of the industry's professionalisation path. In turn, authors responded to these benefits by moving there.

To investigate industry concentration and agglomeration effects, we use data from historical US Censuses and build a novel panel dataset, which includes the complete careers of 473 prominent authors of American literature obtained from encyclopedias. We also combine several sources to measure the quantity and quality of publications produced by these authors. These data allow us to provide econometric and narrative evidence to draw a complete picture of the industrial development and the evolution of agglomeration effects in American literary production from 1850 to 2000.

We use the author panel to identify the existence and magnitude of agglomeration economies. In a fixed effects panel framework, we estimate the difference of the publication propensity while being located in New York City and elsewhere for a working-age author to assess evidence for overall agglomeration economies. As a first step to address endogeneity concerns, we condition our publication estimates on age and year, as well as past publications and migration. On aggregate, we find that an author has 40% higher odds of publishing during the periods of her life when she lives in New York City than when she lives elsewhere. However, we find significant heterogeneity in this effect by age and time, with younger authors benefiting most and agglomeration benefits being strongly correlated to the industry's evolution. These patterns confirm our theoretical expectations about agglomeration economies. In this context, we also address endogeneity concerns in the overall estimate.

First, we observe that the individual benefit of moving to New York is much larger for an author than the penalty of moving away. As the advantages accrued in New York seem partly spatially transferable, we focus on the gains for working age authors moving to New York in an event study. We model these moves in two frameworks: a panel regression

[1] In this paper, we use "New York" to refer to New York City unless otherwise specified.

framework with yearly dummies and a quasi-experimental matching framework for a comparison with publication activity of similar authors who never move to New York. Reassuringly, our estimates do not show an increase in publication activity relative to their peers before their move to New York. Therefore, relocation does not appear to follow an upswing in publications, which would have been indicative of positive selection. Instead, we observe a strong and immediate increase in publications only after arriving in New York City.

Second, we explore how productivity gains correlate with New York City's industrial dominance and, therefore, expected agglomeration economies. We show this by estimating the publication differential by decade over our long time horizon. We find that the publication differential between New York City and other locations rises and declines with the city's industry share and age. The agglomeration benefits associated with residence in New York City diminish with industry maturity. By the 1950s, we see no significant evidence for localised benefits on publication quantity.

Third, we explore the effect of agglomeration on quality of published works. This lets us glimpse into the black box of agglomeration mechanisms and provides additional evidence on the causality of the New York effect. Arguably, authors have less choice over the timing of their peak quality than their career more generally. We provide evidence that works published while an author lived in New York City were more likely to achieve critical acclaim at time of publication and are more likely to have lasting influence in terms of present-day popularity. We do not observe similar localised benefits for contemporaneous bestsellers suggesting that New York's agglomeration effects are not associated with immediate mass market success. Thus, co-location with other writers appears to be more important for success within literary circles and peers than with the wider public.

This paper adds to the rich economic literature using individual industries to understand agglomeration, such as the Saxenian [1996] study on semiconductor producers in Silicon Valley and Boston's Route 128, the Arzaghi and Henderson [2008] study of advertising agencies in Manhattan, the Buenstorf and Klepper [2009] study of the Akron tyre cluster, and Cabral et al. [2018] study of the auto industry in Detroit, among others. However, these studies using relatively modern data are only able to analyse agglomeration over the short-run. Our unique datasets and empirical setting allows us to to investigate agglomeration under vastly changing circumstances over a period of 150 years. Linking New York City's changing position within the publishing industry and as a cultural centre to authors' productivity, we are able to provide a dynamic assessment of agglomeration economies from a fledgling industry to a fully mature industry.

Furthermore, the empirical setting allows us to provide unique insights into productivity

effects of agglomeration using historical data. As Hanlon and Heblich [2020] note, direct measures of productivity are often not available in historical studies, so many studies rely on city size, employment levels, or patents as a measure of agglomeration forces. Moreover, many papers are not able to assess agglomeration benefits in the medium to long-run because time series data on firm activity are typically too short to observe the rise and fall of the industry.

Our unique data provide several advantages in these respects. We are able to explore the impact of agglomeration on a direct measure of individual productivity: literary publications. Our measure of productivity allows us to assess agglomeration benefits over a very long time horizon — longer than possible for many other industries — and across several quality metrics spanning short- and long-run success. To this end, we utilise measures of long-run critical acclaim (via inclusion in *Kindlers Literatur Lexikon*), contemporary commercial success (via inclusion in *Publishers Weekly* bestseller lists), and current day popularity (via the number of online ratings on *Goodreads.com*).[2]

This paper also adds to a growing literature on the economic history of the arts and, more specifically, to agglomeration effects in various artistic fields, such as British and Irish literature [Mitchell, 2019], classical music composers [Borowiecki, 2013], and visual artists [Hellmanzik, 2010]. In contrast to these studies, we can link the observed individual gains to quantitative as well as qualitative measures over a very long time horizon. Moreover, our empirical setting allows us to link the rise and decline of agglomeration benefits to industry maturity.

Similar to the Borowiecki and Dahl [2021] descriptive analysis of the broader creative industries in the US, we start by linking these industry developments and broader economic changes to the emergence and decay of literary clusters. In contrast, as our main contribution, we identify the impact of these literary clusters on the individual-level quantity and quality of produced work.

The rest of the paper is structured as follows. The next section provides an overview of the empirical setting and sets a conceptual framework for agglomeration economies in literary writing. Section 3 describes the datasets and the data collection methodology, and 4 provides preliminary evidence on patterns of agglomeration in American literature. In Section 5, we set up our strategy to identify individual agglomeration benefits. The results of this exercise are discussed in Section 6. Finally, Section 7 concludes.

[2]For a descriptive overview of American authorship based on the *Britannica* authors and authors in *Publishers Weekly* bestseller lists, see O'Hagan [2021] and Ceulemans et al. [2020], respectively.

2 Empirical setting and conceptual framework

2.1 Development of the American literary marketplace

In the years after American independence, British literature remained central to American book culture, with most books in the US being imported from London, Dublin, and Edinburgh [Green, 2010]. Until the last decades of the 19th century, competition with reprints of massively popular and commercially proven British authors, such as Charles Dickens and Sir Walter Scott, kept the incentives to publish American authors low and, therefore, limited the profitability of American authorship. During this time, we see the development of an American book manufacturing and publishing industry without much benefit to American literary authors [Green, 2010; West III, 1989].[3] Before international copyright was recognised in 1891, American "writers could be had cheap compared with paper, type, founders, and compositors" [Jaszi and Woodmansee, 2009, p. 91].[4] As a result, there were few financial (and reputational) incentives for American authors until the mid- to late-19th century, and those American authors that were successful often had individual or institutional patrons [Jaszi and Woodmansee, 2009; West III, 1989].

During this time, publishing houses rushed to publish the newest novel as soon as copies arrived on the ships from Britain.[5] New York City's port was closer and more geographically convenient than other major port cities along the eastern seaboard, allowing New York-based publishers to receive new imported books before publishers in other port cities and easily distribute them throughout the United States.[6] Due in part to these natural advantages, New York City became the leading port in terms of book imports and the

[3]Green [2010] notes that only 20 novels of more than 100 pages written by American authors were published in the 1790s, and only 25 novels by American authors were published in the next decade. At that time, most of the novels read by Americans were imported volumes borrowed from circulating libraries.

[4]See Jaszi and Woodmansee [2009] and West III [1989] for a detailed discussion of the evolution of authorship in the United States during the 19th century and how it both impacted and was impacted by the development of copyright laws.

[5]Until international copyright was recognised in 1891, publishing houses operated under an informally-enforced "courtesy of the trade" in which the first publishing house or printer to announce a foreign book, usually via newspaper, had the right to print it and no other printer would print the same book. For works by popular authors, publishing houses printed copies as quickly as possible without waiting to see who announced the book first [Green, 2010; West III, 1989].

[6]As Glaeser [2005] explains, New York City had a deep harbour that was free from sea ice and was relatively easy for ships to enter and leave, and New York City was close to Hudson River which served as an important distribution point inland. New York City was also a geographically convenient distribution point for the book market because it had relatively easy access to other sea ports in the United States. Later, the construction of the Erie Canal "not only opened an easy path to the upper part of the West from New York but helped make the land through which it passed one of the most active book markets in the nation" [Green, 2010, p. 119]. See Glaeser [2005] and Green [2010] more detailed discussions on the competitive advantages of New York City and their impact on both the city and the publishing industry.

centre of the publishing industry by the mid-1800s [Green, 2010; Glaeser, 2005].[7] Around a third of the bookbinders in the United States were located in New York City in the mid 19th century, as seen in Figure 1. By 1860, New York City was the centre of the printing industry and had the largest concentration of writing-related occupations, such as editors and reporters.

It was around this time — mostly after New York City became established as the centre of the printing and publishing industry — that authorship finally developed as an economically viable activity. The development of a highly concentrated and competitive printing and publishing industry facilitated the development of the magazine and advertising industries. These provided important income streams for authors [Williams, 2007; West III, 1989]. At the same time, there was a growing "cultural patriotism that encouraged Americans to become authors" [Williams, 2007, p. 116]. This shift occurred relatively quickly: the first list of bestselling novels released by Publishers Weekly in 1895 included only included two American authors in the top 10 list; five years later, nine of the top ten novels were by American authors.

2.2 Agglomeration economies in literary production

Authors likely benefited from co-locating with publishing, and to a lesser degree printing, manufacturing, and other related industries. Gabe and Abel [2016] argue that co-agglomeration "facilitates movement among jobs and provides a constant market for skill ... workers benefit from locating in places with an abundance of jobs that require the knowledge they possess as well as jobs with similar knowledge requirements" (p. 2-3). Similarly, Haller and Heuermann [2020] assert that "denser labor markets insure workers against the risk of unemployment by providing a larger pool of potential job" (p. 273).

Thus, the co-agglomeration of industries such as advertising, editing and reporting, and magazines in New York City could have offered additional employment opportunities for authors [West III, 1989]. The thick labour market would have reduced the risk of engaging in a writing career as authors could be employed in a variety of industries requiring writing skills. This would provide authors with more employment opportunities and also allow them to signal their quality to printing houses or gain popularity. Moreover, the favourable market also could have facilitated better matches with employers that allowed authors to pursue multiple income streams at the same time or move from publishing in periodicals to negotiating royalty contracts [Williams, 2007].

[7] It is important to note that natural advantages cannot fully explain industrial concentration. For example, Ellison and Glaeser [1999] found that only around 20% of industrial concentration can be attributed to natural advantages.

Gabe and Abel [2016] argue that "in the case of labour pooling, sharing in the same knowledge as others in the local labour market allows workers to move between jobs more easily. In the case of knowledge spillovers, being around people with similar knowledge facilitates the sharing of information and ideas" (p. 3). Therefore, the pooling of artists, authors, publishers, literary critics, and other creative workers in New York City could facilitate the exchange of knowledge, ideas, and information. Indeed, the American author Henry James (1843-1916) noted:

> "The best things come...from the talents that are members of a group; every man works better when he has companions working in the same line, and yielding to the stimulus of suggestion, comparison, emulation. Great things have of course been done by solitary workers, but they have usually been done with the double the pains they would have cost if they had been produced in more genial circumstances." (James, 1909, p. 31)

Storper and Venables [2004] stress the importance of face-to-face contact for efficient communication, solving incentive problems, facilitating learning, and providing motivation. An author living in the periphery may have difficulty establishing meaningful social connections with publishers or others in critical gate-keeping positions, such as literary critics. Moving to New York City would increase physical proximity to gatekeepers and therefore increase the probability of having the face-to-face interactions necessary for building trust, developing professional relationships, and facilitating knowledge spillovers.

Before the rise of intermediaries, the author-publisher relationship was effectively a two-sided matching market[8] and, in its taste aspect, was not too dissimilar to the Roth and Sotomayor [1992] marriage market model. The pooling of industries related to authorship would have reduced search costs and resulted in higher quality matches between authors and publishers and other relevant industry gatekeepers [Helsley and Strange, 1990]. In addition, a dense network of publishers and other suppliers in New York City could reduce authors' cost of realising their ideas [Helsley and Strange, 2002]. This may have been particularly beneficial for authors in the early stage of their career, as this dense network could increase their likelihood of being represented by a publisher.

Indeed, there is evidence that this was the case for some authors. For example, the New York Times [1901] reported that Upton Sinclair wrote the novel *Springtime and Harvest:*

[8]West III [1989] describes the search process for a publisher: "What kind of publisher is best for a serious literary author? [...] Authors and publishers often come together by chance; sometimes an author is dissatisfied with a previous house, or the new house is rumoured to be the 'right' kind of firm. Most of the matches are imperfect, but occasionally an author does find an entirely compatible publisher." (p. 25).

A Romance while visiting Lake Massawippi in rural Canada, and he returned to New York City in order to find a publisher. The book was rejected by several publishers before Sinclair decided to self-publish under the imprint "Sinclair Press". Upton encountered similar difficulties when trying to publish his novel *The Jungle* [Blum, 2020] due to its shocking content. After several rejections from publishers, Sinclair considered self-publishing again, but ultimately the novel was published by the New York City-based publisher Doubleday, Page & Company.

Furthermore, it may have been more important for authors pursuing critical acclaim to be located in a literary cluster than those pursuing commercial success. Literary historians often make the distinction between the anti-popular and original *romantic genius* who is in conflict with the market and the derivative *popular writer* who caters to the commercial market [Williams, 2007]. Bourdieu [1985] describes these as the *field of large-scale cultural production* (the creation of cultural goods for non-producers of cultural goods or 'the public at large') and the *field of restricted production* (cultural goods created for an audience of other producers of cultural goods). He notes that the *field of restricted production* develops its own evaluation criteria that breaks away from that of the general public.[9] Therefore, authors seeking critical acclaim may have benefited from the geographic clustering of literary and artistic elites by learning about up and coming literary styles and readers' tastes through a network of peer writers. This environment may not have been as beneficial for authors pursuing commercial success because, by design, the tastes and preferences of highly localised literary elite may not have been appealing to the masses.

2.3 Core-periphery dynamics

It is important to note that agglomeration effects not only reflect the advantages associated with the concentration of economic activity in New York City but also the disadvantages or penalty associated with living elsewhere, for example, due to a lack of a local publisher. As an extreme, in an idealised cluster study, the cluster is completely isolated from the periphery. Those located in the cluster have perfect access (and no limits to access) to the benefits within the city, and all who live outside the city have no access to these benefits. Therefore, individual benefits from agglomeration arise from differences in access to ideas, publishers, critics, and so on. However, in practice, the effect's spatial bound and decay is less clear cut and likely to change with technological development. Therefore, the disadvantage of being located in the periphery can soften through, for example, advances in transportation and ICT technology as much as through an expansion

[9]See also Bourdieu [1969], Becker [1974], among others.

of the publishing industry. As spatial imbalances either decline or lose importance, the agglomeration benefits associated with the cluster will also decline.

Several technological shocks in the 20th century resulted in a decline in the concentration of the printing and publishing industries in New York City. For example, the rise of automobiles, road freight and telecommunication reduced the need for high-density cities. Indeed, all major cities experienced some degree of population decline during the period from the 1930s to the 1990s, although this decline was less dramatic in New York City compared to all other major cities except Los Angeles [Glaeser, 2005]. These technological shocks did not necessarily decrease the advantages of living in New York City so much as they decrease the *dis*advantages of being located elsewhere.

There were several major changes in the industries related to publishing and writing that further shifted this balance. Technological advancements dramatically reduced the cost of printing, allowing for the expansion of smaller printing presses for books, magazine, newspapers, etc. across the United States [Winship, 2009]. The development of a national book trade system further facilitated a rapid expansion in the number of publications.[10] Thus, it became easier to publish outside New York City. The film industry began to emerge in Hollywood from the 1940s, expanding the number of professional opportunities for some writers outside New York City. Around the same time, there was an increased professionalisation of the publishing industry with agents becoming more important as brokers to publishing houses. These literary agents had become an established intermediary between authors and publishing houses due to the growth in subsidiary rights and more complex contracts [Kaestle and Radway, 2009; West III, 1989]. Thus, this period was characterised by a combination of factors that both decreased the advantages of being in New York City and decreased *dis*advantages of being located elsewhere.

Therefore, we expect to see large growth in the concentration of authors in New York City around the turn of the 20th century, after the publishing, printing, and bookbinding industries in New York City were already established. We also expect to see agglomeration effects emerge from this time. Authors living in New York City would benefit from the agglomeration of economic activity. At the same time, there was an almost complete lack of access to the market for those located elsewhere. We expect to see a decay of the agglomeration effects from the 1940s when there is a notable decrease in the disadvantages of being located elsewhere. We also expect to see a differential effect for commercially successful and critically acclaimed works, with authors writing for a network of peers receiving the greatest gains.

[10] Winship [2009] notes that the US population more than doubled between 1880 and 1916 while the number of new book titles and editions increased sixfold. Luey [2009] observes that "title production increased doubly fivefold between the late 1940s and 1990" (p. 35).

3 The data

This paper utilises several historic and contemporary data sources to identify trends in industry concentration and estimate agglomeration effects. First, we use cross-sectional population data from historical US Censuses in order to identify trends in industry concentration and major literary clusters. We also use these data to compare the demographics of New York-based authors (by far the largest literary cluster) to authors based elsewhere to provide preliminary evidence on whether the systematic self-selection of certain demographics into New York exists. Second, we use a purpose-built panel dataset with the lifetime biographic, location, and publication quantity data of 473 prominent authors associated with American literature. Third, three different measures of publication quality, namely current-day popularity via the number of ratings on *Goodreads.com*, critical acclaim via inclusion in the *Kindler Literatur Lexikon*, and commercial success via inclusion in *Publishers Weekly* top 10 bestseller lists. We use the combined panel and publication quality data in the regression analysis to identify the existence and magnitude of agglomeration effects associated with New York City.

3.1 Population and industry concentration: *US Census* data

We utilise the IPUMS USA full count Census data from 1850-1940, 1% samples for 1950, two combined metro 1% samples for 1970, and 5% microsamples for 1960 and 1980-2000.[11] We identify a total of 47,619 responses for the 1950 basis occupation (OCC1950) category *006 Authors*.[12] We examine high skilled occupations (professional and technical occupations as well as professors and instructors), which we define as the 1950 basis occupation category codes 000 to 099. We also examine several complementary occupational groups including artists, editor and reporters, bookbinders, and occupations relating to the publishing industry such as pressmen, printers, and apprentices in the printing trade.[13] For each group, we utilise data on the total population, location, and demographic charac-

[11] More specifically, we utilise the Ruggles et al. [2020] IPUMS USA data. We use 1% microsamples for 1950 and 1970 because 5% samples are not available for these Census waves.

[12] Before 1960, the Censuses were conducted by US Marshalls or, from 1880, specially trained census-takers or enumerators via in-person interview [US Census Bureau, 2020]. The Census enumerators were instructed to ask for the interviewee to describe their occupation and ask follow-up questions for clarification as needed. The enumerator recorded the occupation as a string that was later classified into an occupational category. In the historic Censuses, the most frequent strings that were classified as the occupation *author* include: writer (543) author (496) authoress (173) waiter (29) poet (22) literature (21) literary (18) literary writer (13) auther (12) dramatist (11) autheress (9) playwright (9) story writer (9).

[13] Artists are identified by the occupational code *004 Artists and art teachers*, editors and reporters are identified by *036 Editors and reporters*, and bookbinders are identified by *502 Bookbinders*. Occupations related to the publishing industry include *575 Pressmen and plate printers, printing* and *614 Apprentices, printing trades*.

teristics such as age, gender, ethnicity, location of birth (foreign-born or native-born), education level, and marital status.

We use these historic US Census data to examine patterns of industry concentration and identify major literary clusters. We also use these data to compare the demographics of New York City-based authors to determine whether they are systematically different from authors based elsewhere, and we use these data as a baseline to determine whether the demographics of the panel of prominent authors differ systematically from the general author population.

3.2 Author sample and publication quantity: *Britannica* data

The purpose-built panel dataset on author productivity and residence as used in this paper was generated following the methodology of Mitchell [2019].[14] An initial list of prominent authors was generated from the Encyclopaedia Britannica [2016], and detailed lifetime biographic and publication data were collected from the authors' biographies in Encyclopaedia Britannica [2016], The Literary Encyclopaedia [2016], and Literature Online [2016]. For this study, only those authors were considered which had at least one literary work, be it prose, poetry or a play. Thus, we exclude authors whose publications were limited to any other type of writing activity be it journalism, criticism, or song writing. For each author, we collected general biographic data including year of birth, year of death, ethnicity, and gender, as well as annual data on age, number of publications, and location. The data on location were brought into a panel structure using the time of movement as the beginning and end point of any spell spent at a particular location.[15] In total, the panel data set includes 471 prominent authors associated with American literature who were born between 1800 and 1949. We only use the working age population of authors, defined as ages 18 to 65, for our analyses.

For these authors (henceforth referred to as the *Britannica* authors), we extracted information on the number of works published in each year and categorised them as novels, plays and poems. The encyclopedias often only included a list of "notable publications", so the number of publications was collated from lists of publications across all three sources. While the names of publications were not collected, the lists of publications

[14]The data collection for the American author research was conducted at and funded by the Department of Economics at Trinity College Dublin under the supervision of John O'Hagan, with assistance from Sara Mitchell, in summer 2016 by Maria Fleming and in summer 2018 by Seán McGuane. The data were merged, cleaned, and extended by Lukas Kuld and Sara Mitchell in Spring 2019. Additional data was collected at the Chair of Urban, Regional, and International Economics at TU Dortmund in summer 2019 by Vivien Gaida, Leyla Ruzdija, and Jens Kirsten under the supervision of Christiane Hellmanzik and Sara Mitchell.

[15]Short-term travels were recorded; however, they are not considered in this study.

were cross-referenced to ensure that publications were not double-counted. This resulted in a 4,258 publications over the years 1826 to 2013.[16] As the list of publications for the *Britannica* author dataset was compiled from non-comprehensive lists across three sources, we combine the *Britannica* publications with the publications listed in *Goodreads* to ensure completeness. (See next section for a discussion of the *Goodreads* data.) Thus, the output measure indicates whether an author's published work is mentioned in either the *Britannica* list or the *Goodreads* list in a given year.

3.3 Publication quality: *Goodreads.com*, *Publishers Weekly* and *Kindler* data

We also gathered information on publication quality to complement the data on quantity of publications. To analyse agglomeration effects along different time horizons and in different types of markets we utilise three different measures of quality: a measure of present-day popularity, a measure of market success at the time of publication, and a measure of critical acclaim by contemporary expert opinion.

We link publication data from Goodreads.com [2020] for 352 of the *Britannica* authors. For these authors, we retrieve data on 1,388 publications, including the year of first publication and ratings on Goodreads.com. On Goodreads.com, users can list and discuss literature and rate each publication from one to five stars. We focus on how many ratings a publication receives as an indication of how widely known and read a publication is currently or, more precisely, since the launch of Goodreads.com 2006. We use these data as a measure of present-day market success. As Bourrier and Thelwall [2020] note, the most popular books on *Goodreads.com* generally fall into two categories: recent 'popular' literature and older 'classic' (canonical) literature as, for instance, read in secondary schools. Therefore, present-day market success partly reflects critical acclaim and canonisation. We refer to these data as the *Goodreads* data.

In order to assess novels' historical success, we collected the top ten bestselling novels in the United States in each given year from 1895 to 2011 as determined by *Publishers Weekly*. We scraped these *Publishers Weekly* lists of bestselling novels from Wikipedia, where the lists are organised by year [Wikipedia Contributors, 2019]. We then linked data on 164 bestselling novels to 79 authors in the *Britannica* dataset. Inclusion in the *Publishers Weekly* list reflects contemporaneous market success. We refer to these data as the *Publishers Weekly* data.

[16] See O'Hagan [2021] for a descriptive analysis of the dataset, author migration patterns, and productivity over the lifecycle.

Finally, we obtain data from Kindlers Literatur Lexikon [2019], a comprehensive German encyclopedia on literature that aims to only cover the most important works in all languages in cultural history. As such, it is a valuable source of experts' opinion on the quality of authors' works.[17] These encyclopedia entries are primarily about individual publications which are considered canonical by the editorial team. It is highly selective, and many authors only have a single work included in the encyclopedia.[18] We linked data on 226 publications included in the *Kindler encyclopedia* to 171 authors in the *Britannica* data set. Around half of these authors only have one publication listed in *Kindlers Literatur Lexikon*. As this variable is backward-looking in nature and not based on market success, we consider it a complementary measure of literary quality and reflective of a very long-run, ex-post agglomeration effect. We refer to these data as the *Kindler* data.

Summary statistics for the full dataset on publications are provided in Table 1. This table includes the time period covered, the total publications, the range of publication years and the total number of authors included in each dataset (for the *Goodreads*, *Publishers Weekly* and *Kindler* datasets, this is the number of authors in the respective dataset which could be linked to the *Britannica* panel). It is important to note that we were not able to link publications based on title, as these data were not collected during creation of the *Britannica* author dataset. Therefore, we rely on the assumption that editorial teams of the respective encyclopedias and online databases maintain accurate records of the year of publication.

4 Descriptive patterns of literary agglomeration

As a first step in assessing agglomeration effects in literary production, we conduct a descriptive analysis of the development and drivers of literary clusters. We analyse the places of residency of authors across time to identify and quantify the major literary clusters. We then explore the degree to which demographic and industry characteristics of New York City (the only major literary cluster) can explain author concentration. We then compare the characteristics of authors in New York to authors living elsewhere as a simple check for evidence of systematic self-selection of specific groups of authors into New York City. Last, we show how the in-migration of early career authors has contributed to New York's dominance.

[17] See Heinz Ludwig Arnold [2009] for more information on the aims of the encyclopedia, the curatorial process, and the contributors.

[18] We received permission to collect data from the Kindlers Literatur Lexikon web portal through personal correspondence on 5 December 2017. The data collection process via web scraping was conducted in early 2018.

4.1 Author location and the rise of New York as literary cluster

To assess agglomeration effects in literary production, we first quantify the main clusters by mapping the places of residency of authors across time. The results for all authors in the US census and the *Britannica* authors (aged 18 to 65) are presented in Figure 2. Until 1990, New York is the single most important location for authors in general with a peak in agglomeration with more than a third of authors in the first half of the 20th century. Consistent with historical accounts as well as theory mentioned in Section 2, most writing is scattered along the East coast with New England, Boston and Chicago in particular being important albeit much smaller centres. The West Coast, Los Angeles in particular, gains in importance as a place of residence for American writers from 1920 onward, while New York's importance wanes from 1950. The decline in the concentration of authors in New York City is not due to a decline in authorship — the total number of authors in New York City continues to increase — but rather a marked increase in the number of authors located elsewhere.

The concentration of authors in New York City before World War II is striking, even when compared to the concentration of workers in other occupations associated with book production such as bookbinders, publishers, and printers, other writing-related occupations such as editors and reporters, and other artists. As seen in Figure 3, the concentration of authors is most similar in its pattern to that of artists, although artists' concentration in New York City remained consistently lower than that of authors.[19] It is important to note that, as expected, author concentration follows other industry concentration. That is, the publishing and printing markets are established in New York City first, and then authors begin to concentrate there, in particular working age *Britannica* authors, that is, eminent literary authors.

Interestingly, the concentration of *Britannica* authors across major literary clusters appears to be slower to respond to changes in the publishing industry than the overall author population. In 1850, Boston was the largest cluster of Britannica authors, even though the city had already waned in importance in the publishing industry and the broader author population had largely relocated to New York City. For these prominent authors, New York City did not overtake Boston as the largest literary cluster until more than a decade later, and Boston remained a close second until the 1890s. By 1900, however, the concentration of prominent authors in New York City surpassed that of the general author population. New York City's position as the largest cluster of prominent authors peaked around 1925 and remained steady until just before the 1950s, during which time 45-50%

[19]The impact of the spatial constraints faced by artists' in form of necessary access to physical capital (e.g., an art studio, canvasses, paintbrushes, paint, and other supplies) for spatial concentration, while not clear, could have suggested a higher concentration than authors.

of all prominent authors lived in New York City. In contrast, the city's importance for the wider population of writers began to wane from 1930.

4.2 Demographics and industry concentration as drivers of agglomeration

We explore the degree to which demographic and industry characteristics can explain author concentration using a simple prediction procedure. We first fit a negative binomial regression model using the the subsection of data which includes all US counties except the seven counties that compose the New York City agglomeration.[20] Specifically, we estimate the number of authors in a given county based on log industry employment in related industries (printing, the publishing industry, the paper industry, bookbinders, artists, and editors and reporters), the total population (in logs), the share of skilled workers, the share of women, the share of people born abroad, and the share of population that is white. We then use the parameters estimated to predict the expected number of authors in New York City based on the actual industry and demographic data of the seven New York agglomeration counties. Last, we divide the predicted values for New York City by the numbers predicted for all other counties to get the predicted New York share.[21]

The results, illustrated in Figure 4, indicate that the actual share of authors located in New York City is notably higher than the predicted share. That is, the concentration of authors is greater than we would expect given the observable economic and social opportunities present in the city. This trend is particularly notable from just after 1900 until around 1940. Puga [2010] notes that "substantial localization or spatial concentration of economic activity may be seen as a sign of agglomeration economies" (p. 204) as per the Starrett [1978] spatial impossibility theory. Thus, we argue that the exceptional degree of concentration of authors in New York City is due, in part, to a combination of natural advantages and agglomeration economies.

In later years, the concentration of authors and related industries in New York City consistently declined starting around 1950. It is important to note that this relative decrease of New York City is driven by a strong US-wide growth of authors and employment in related industries. Therefore, we observe a catch-up of other regions and not an absolute

[20] We use a negative binomial regression model because the dependent variable is overdispersed count data, as suggested by [Cameron and Trivedi, 2001].

[21] That is we first estimate $log(\mathbb{E}(y_{i \notin NYC,t})) = x_{i \notin NYC,t}\beta$, with $x_{i \notin NYC,t}$ the vector of covariates for county i (outside NYC) in year t, and $y_{i,t}$ its number of authors. We then use the estimated $\hat{\beta}$ vector to predict the share of authors in the New York City agglomeration: $\sum_{i \in NYC} \hat{y_{i,t}} = \sum_{i \in NYC} exp(X_{it}\hat{\beta})$. Last, we calculate the predicted share as $\sum_{i \in NYC} \hat{y_{i,t}} / \sum_{i \notin NYC} \hat{y_{i,t}}$.

decline for most of these industries. Nevertheless, agglomeration effects accrue through the difference between agglomeration and periphery, and this difference is therefore the focus of our analysis.

4.3 Characteristics of authors in New York and elsewhere

We compare the demographic characteristics of authors in New York City to the general author population to determine whether there are systematic differences between these groups. New York-based authors do not appear to be an exceptional group, as seen in the US Census data in Figure 5.[22] For the majority of period from 1850 to 2000, authors living in New York City are broadly similar to authors located elsewhere with respect to mean age, the share of white authors, the share of female authors, the share of authors that never married (except after 1950), and the share of authors with at least some college education. Thus, authors in New York City do not appear to be systematically different to the general author population with respect to demographics. There are two notable exceptions, first, during the period of the highest concentration of authors in New York City (1900-1930), the average New York City author was around five years younger than the average author located elsewhere. Second, throughout the sample period, the average author in New York City is more likely to have been born abroad than authors located elsewhere, reflecting New York City's history as a major hub of immigration.

This pattern is also discernible for the *Britannica* sample of eminent authors, as seen in 6. For most of the time period, the *Britannica* authors in New York City and those living elsewhere are broadly similar in terms of age, ethnicity, gender, and being born abroad. (Data on marital status and higher education is not available for the *Britannica* sample). This is important in terms of identifying a causal effect of living in New York City on productivity, as systematic differences between these groups could indicate self-selection of certain groups into New York City which might bias the estimates. Systematic differences between these samples could also indicate sample selection bias; for example, there could be oversampling of prominent authors if New York-based authors are more likely to be considered worthy of inclusion in encyclopedias. We can also use Figures 5 and 6 to compare the *Britannica* sample of prominent authors to the wider population of authors. In general, the *Britannica* authors tend to be younger, more ethnically diverse from around 1900, less likely to be born abroad before 1975, and less likely to be female before 1925 compared to authors in the US Census.

Authors who move to New York City tend to do so before age 35, as illustrated in Figure 7 (histogram on right y-axis). This is the typical age range for labour migration (see e.g.

[22] All summary statistics are calculated using the sample of working-aged authors (ages 18-65).

Stark and Bloom [1985]), suggesting that authors move to New York City for professional reasons. In fact, human capital theory would suggest a higher pay-off and lower relocation costs for young authors [Becker, 1964]. This age structure is therefore indicative of market opportunities in New York City under the assumption that authors are rational actors.

The average moving age to New York City by *Britannica* authors remains relatively constant for most of the sample period (see the local average line in Figure 7), suggesting that the pattern of moving to the city during the early career stages is stable over time. The frequency of moves shows that a large part of the concentration of authors in New York City is due to in-migration and not due to a large share of authors being born in the city. The number of author relocations to New York over time closely aligns with the industry's trend in concentration, that is, the number of moves increases until the 1930s and decreases thereafter (see histogram on top x-axis of Figure 7). At no point in time would New York City have been the observed literary cluster purely with native New Yorkers.

5 Empirical Strategy

As our baseline, we determine the existence and magnitude of agglomeration effects in our data by estimating the difference of the publication propensity while being located in New York City and elsewhere for a working-age author. We condition the estimate of the productivity effect of living in New York City on the author, year, and age. Specifically, we use a binomial Logit fixed effect model to estimate the following equation:

$$g(\mathbb{E}(y_{it})) = \beta NYC_{it} + \gamma_i + \tau_t + \alpha_{it} \tag{1}$$

where y_{it} is a binary indicator for a publication by author i in year t, NYC_{it} is a binary indicator for whether author i lives in New York City in year t, γ_i and τ_t are fixed effects for author and year, and α_{it} is a fixed effect for the age of author i in year t. g is the Logit function used as link. We limit our estimation to authors living in the United States at a known location and in the age of 18 to 65. Standard errors are clustered at author level.

This approach is similar to other literature on agglomeration economies. Although, many papers regress local wages on a local characteristic, such as employment density, under the assumption that workers are paid the value of their marginal product in competitive labour markets (See Combes and Gobillon [2015] and Ciccone and Hall [1996], among others). However, this empirical setting provides the unique opportunity to utilise a

direct measure of individual productivity rather than relying on wage proxies.[23]

5.1 Identification and heterogeneity of effect

Due to the use of individual fixed effects, identification of the New York City coefficient β in Equation 1 comes from individual authors' moves to and from New York City.[24] These moves are unlikely to be exogenous to an author's time-invariant characteristics (e.g., natural ability, personality) or time-variant characteristics (e.g, career stage, martial status).[25] Therefore, we explore whether the estimated effect reflects a causal relationship from being in New York to increased productivity as measured by the probability of publishing or whether it is in fact the result of confounding factors, such as social reasons which have successful authors move to New York City.

In Equation 1, we condition our estimate on the individual author, year, and age. Therefore, we address constant productivity differences, be it between authors, between different ages or a general publication slump, for instance, during the Great Depression. Thus, identification is not threatened by the possibility that New York City-based authors are constantly more (or less) productive than authors based elsewhere. Still, we note that authors living in New York City are not more productive overall (see Section 6.1).

Instead, our main concern is that the decision of moving to New York is correlated with career stage in a way that is not captured by age fixed effects (e.g., if authors begin their career at different ages). We provide three counterfactuals to investigate the causality of productivity changes after an author moves to New York City: their own productivity before relocation, their peers' productivity elsewhere before and after relocation, and New York moves at times when industry concentration and maturity predict lower agglomeration economies.

First, we set up the move of working age authors to New York as an event study. This allows us to explore their productivity trajectory before and after relocation. In addition, we can compare the productivity pattern to the productivity of their peers who never

[23]This approach also relates to the literature on agglomeration and innovation, which use patents as a direct measure of individual productivity. (See Carlino and Kerr [2015] for an overview.) The literary market is characterised by highly differentiated products, each of which can be considered an incremental innovation. The use of literary publications as a measure of innovation has a unique advantage over traditional measures of innovation such as patents: many inventions are patented but relatively few are commercialised and introduced into the market; whereas, we only observe publications that are commercialised and introduced into the market.

[24]Any year of any author contributes indirectly to the identification via age and year fixed effects. However, this contribution is limited as the omission of authors who never live in New York City in the regressions shows: The estimates remain almost unchanged, which indicates that these authors are not fundamentally different in their age and year characteristics.

[25]See for instance Baum-Snow and Ferreira [2015] and Combes and Gobillon [2015].

move to New York as a comparison group via a quasi-experimental matching procedure. If the conditional (age, author, year) productivity levels after moving to New York City are an expression of the author's career that would have played out in a similar way without any New York agglomeration economies, we expect to observe an acceleration in publications before moving, especially in comparison to peers located elsewhere. Otherwise, the move can still reflect a career step but the agglomeration economies in New York might have been a necessary condition for the success of this endeavour.

Second, we utilise our long time horizon and observed shift in New York City's dominance. If authors move to New York at a similar career stage (and possibly with similar motivations) throughout our sample period but their observed gains change in a way consistent with the expected agglomeration economies associated with industry concentration and maturity, then we can more confidently attribute the observed productivity gains to being in New York City instead of an expression of factors correlated with the decision to move.

In the last part of our empirical analysis, we provide evidence on the effect on quality. This approach addresses within-author endogeneity, as authors arguably have less choice over the timing of their peak quality than when to start their career. If the observed productivity gains were driven by the fact that moving to New York City purely coincides with the decision to start a career, it is unlikely that we would also observe these patterns of quality gains. Therefore, we can more confidently attribute these productivity and quality gains to agglomeration economies in New York City.

5.2 Moving to New York City as an event study

We study the productivity impact of moving to New York City in two set-ups: a regression framework and a direct comparison of publication means of authors who move to New York City to the publication means of a matched control group. In order to compare productivity before and after the move to New York City, we use an approach similar to that which de la Roca and Puga [2017] use to estimate dynamic agglomeration effects.[26] Specifically, we re-estimate our Equation 1 using dummies that indicate periods relative to an author's relocation to New York City, for instance a dummy indicating that the author will move to New York City in two years.[27] We estimate the productivity for the following periods: more than 5 years before, 15+ years after moving to New York City, and either individual years or five-year blocks in between. We show estimates for two

[26] Also see Glaeser and Mare [2001] and Combes and Gobillon [2015] for more information on dynamic agglomeration economies.

[27] If an author never lives in New York City, we still include her in our regression but all her time dummies will be zero throughout.

reference periods: the years immediately before the move and the years more than five years before, that is we estimate the productivity relative to this time period. Our main interest in this exercise is in showing a clear change in the publication quantity after moving to New York City. By contrast, an increased publication quantity before moving to New York could indicate reverse causality, that is an uptick in publications leading to a relocation to New York City.

In addition to the described regression framework, we set up a quasi-experimental matching design. We match all authors around their move to New York with authors who never lived in New York by age, year (± 1 year), ethnicity, and gender (coarsened exact matching, see for instance Iacus et al., 2012). We then directly compare the yearly sample means of the *New York City movers* group to that of the *never New York City* group for our variable of interest: the number of publications per year.

This approach provides three important results beyond the fixed effects regression framework. First, it explicitly states and limits the comparison group. We know that both groups of authors have the same age, gender, ethnicity, and time composition. Second, it allows to non-parametrically gauge the trajectory of the yearly productivity of treated (*New York City-movers*) and control group around the time of their move. For instance, we can distinguish between a positive New York City estimate due to higher long-term productivity and a shorter 'head start' effect immediately after relocation. Finally, matching without restricting the comparison on the prior publication record provides further evidence whether positive sorting into New York City exists. It also addresses the potential issue caused by time-variant unobservable individual characteristics to the extent that the matched characteristics are the same for the two groups.

5.3 Moves to and from New York by prior publication record

Thus far, we have only discussed a general New York City effect without determining whether this effect is being driven by moves to and away from New York City. Similarly, it is possible that this effect is driven by a subset of authors; for example, established authors (or novice authors) may receive the greatest gains from residing in New York City. Therefore, we investigate the heterogeneity of the New York City agglomeration effect through, first, separate estimates for moves to and from New York City and, second, by differentiating authors by their prior publications. Specifically, we estimate variations of Equation 2.

$$
\begin{aligned}
g(\mathbb{E}(y_{i,t})) =& \beta_1 NYC_{i,t} + \beta_2 NYC_{i,\tau \leq t} + \beta_3 y_{i,\tau<t} \\
&+ \beta_4 NYC_{i,t} \times y_{i,\tau<t} + \beta_5 NYC_{i,\tau \leq t} \times y_{i,\tau<t} \\
&+ \gamma_i + \tau_t + \alpha_{it}
\end{aligned} \tag{2}
$$

NYC_t is a binary indicator for whether the author lives in New York City in time t. $NYC_{i,\tau \leq t}$ is an indicator variable which denotes whether author i lives in New York City in year t or has lived there at some point τ before t ($NYC_{i,\tau \leq t} = 1 - \prod_{\tau \leq t}(1 - NYC_{i,\tau})$). Because NYC_t represents a subset of $NYC_{\tau \leq t}$, the coefficient for NYC_t provides the difference between the two dummy variables when taken together. That is, β_1 indicates the penalty (or gains) associated with leaving New York, while β_2 gives an estimate of the productivity differential between the time periods before and after moving first to New York City (both periods restricted to working age, set as 18-65).

Similarly, $y_{i,\tau<t}$ refers to author i having published at some point τ before t ($y_{i,\tau<t} = 1 - \prod_{\tau<t}(1 - y_{i,\tau})$). $NYC_{i,\tau \leq t} \times y_{i,\tau<t}$ and $NYC_{i,t} \times y_{i,\tau<t}$ are interaction terms indicating whether an author lives in New York City in year t or has lived there at some point, and having published before, respectively. The parameters, their definitions, and their interpretation in full specification are summarised in Table 2.

5.4 Agglomeration effects over time and by industry concentration

So far, we have ignored the time dimension in our data beyond controlling for yearly changes in overall productivity. As a next step, we explicitly model the development of New York City agglomeration benefits over time. As discussed in Section 2, we observe a notable shift in New York City's industrial dominance over the sample period. We also observe that the age (and thus likely the career stage) at which authors move to New York stays relatively constant over time in Figure 7. Therefore, if the observed gains associated with living in New York change in a way consistent with the expected agglomeration economies associated with industry concentration and maturity, then we can more confidently attribute the observed productivity gains to being in New York City instead of an expression of factors correlated with the decision to move.

First, we investigate how the magnitude of the New York City effect varies by decade by estimating the following equation.

$$g(\mathbb{E}(y_{it})) = \beta_k NYC_{it} \times \mathbb{1}_{t \in D_k} + \gamma_i + \tau_t + \alpha_{it} + \alpha_{it} \times \mathbb{1}_{t \in D_k} \tag{3}$$

where $\mathbb{1}_{t \in D_k}$ is a dummy indicating that year t falls within decade k and NYC_{it} indicates whether author i lives in New York City in year t. Subsequently, β_k is the decade specific effect of being located in New York City in year t. y_{it} is a dummy for a publication by author i in year t, γ_i, τ_t, and α_{it} are fixed effects for author, year, and age, respectively. As a further precaution, we interact the age effect with decade dummies to control for the case that age-specific productivity changes over time.

Next we repeat the decade estimation while distinguishing moves to and from New York City. This is crucial as some decades might see more moves to New York, while others see authors leaving. We will see that most of the productivity differential comes from moving to New York City, so this could, potentially, invalidate the comparison by decade. Therefore, we re-estimate Equation 2 using decade dummies to interact with dummies for living in or having lived in New York City. The former then indicates the penalty of moving away, while the latter gives an estimate of the productivity differential between the time period before moving first to New York to later years, conditional on the age-specific productivity (and both age periods restricted to working age: 18 to 65). In particular, we estimate the following Equation 4.

$$g(\mathbb{E}(y_{i,t})) = \beta_k NYC_{i,t} \times \mathbb{1}_{t \in D_k} + \beta_k NYC_{i,\tau \leq t} \times \mathbb{1}_{t \in D_k} + \gamma_i + \tau_t + \alpha_{it} + \alpha_{it} \times \mathbb{1}_{t \in D_k} \tag{4}$$

Last, we directly estimate a central hypothesis of this paper: the concurrence of individual productivity gains and industry concentration in New York City. To this end, we interact the New York dummy first with the share of American authors, and second with the difference of this share over five years. Specifically, we estimate Equation 5.

$$g(\mathbb{E}(y_{it})) = \beta_1 NYC_{it} + \beta_2 NYC_{it} \times Share_t + \gamma_i + \tau_t + \alpha_{it} \tag{5}$$

This equivalent to Equation 1, with the addition of $Share_t$, defined either as the share of authors in year t residing in New York City[28] or its change compared to $t - 5$.

[28] $|\{i : i \in NYC \wedge year = t\}|/|\{i : year = t\}|$

5.5 Evidence on commercial and critical success

In the last part of our empirical analysis, we provide evidence on the effect on quality. We provide two sets of estimates. First, we link residence in New York City to writing a bestseller or critically acclaimed work. Second, we link present-day popularity on the website *Goodreads* to residence in New York City. As an aside, our investigation into quality effects also addresses within-author endogeneity, as authors arguably have less choice over the timing of their peak quality than when to start their career. If the observed productivity gains were driven by the fact that moving to New York City purely coincides with the decision to start a career, it is unlikely that we would also observe these patterns of quality gains. Therefore, we can more confidently attribute these productivity and quality gains to agglomeration economies in New York City.

First, we estimate the following regression using publications in the *Kindlers Literatur Lexikon* as reflection of critically acclaim, long-rum eminence and recognition by experts and, separately, in *Publishers Weekly* to determine the probability of having commercial success at the time, respectively.

$$g(\mathbb{E}(y_{it})) = \beta_1 NYC_{i,t} + \beta_2 NYC_{i,\tau \leq t} + \gamma_i + \tau_t + \alpha_{it} \tag{6}$$

As in Section 5.3, NYC_t is a binary indicator for whether the author lives in New York City in time t. $NYC_{\tau \leq t}$ is a binary indicator for whether the author has lived in New York City at any point in their lives prior or equal to year t. When taken together with $NYC_{\tau \leq t}$, the coefficient for NYC_t provides the difference between the two dummy variables; that is, β_1 indicates the penalty of leaving New York.

Second, we re-estimate Equation Equation 6 using ratings on *Goodreads*. We use a Logit regression to estimate the probability of publishing more or less popular books (using quartiles for the number of online ratings) depending on whether the author was living in New York City at the time of publication. The publication probability regressions are equivalent to the estimation described by Equation 6. Last, we directly estimate the number of ratings received on *Goodreads* using a quasi-Poisson regression. For this analysis, we exclude years without a publication; that is, we focus on ratings per book as opposed to publication quantity.

6 Results

6.1 New York City agglomeration effect

The estimates of Equation 1, presented in Table 3, indicate that the probability of an author publishing is greater during the periods of her life when she lives in New York City than when she lives elsewhere. Column (1) shows that an author has around 40 percent higher odds of publishing while living in New York City, conditional on her overall publication probability, and the overall year and age specific publication probability. (The coefficients are presented in log-odds units.)[29] Columns (2) and (3) show estimates from alternative empirical models, that is a Probit and linear (probability) model. Column (4) presents a more conservative Logit estimate in which the outcome variable is limited to publications listed on *Goodreads* only. We provide the average marginal effect for all specifications at the bottom of the table to allow a direct comparison. This shows that the average marginal effect is very similar between the different statistical specifications though smaller for *Goodreads* publications. The latter is expected as the frequency of these publications is lower, which translates a similar relative effect into a smaller absolute effect.

Between the age of 18 and 65, *Britannica* authors publish on average every four years. Based on this average, the New York City effect estimated in Column (1) would imply an increase in the publication probability of around nine percentage points or a publication every three years after a move to New York.[30] However, we see that the average marginal effect in a Logit, Probit or OLS specification is only four to five percentage points. This could indicate that the New York City effect is dominated by authors who published relatively few works. For these authors a small absolute increase in publication probability would result in a larger proportional increase.

As a first robustness check of the overall New York City estimate, we investigate whether this differential purely reflects correlated effects associated with recent moving activity or living in a major city. These are in particular, a variable indicating whether the author lives in another major city (see cities in Figure 2), whether the author has moved to a new place in the last five years, and the cumulative number of moves up to the current year. The results of this exercise are presented in Column (4). We find that the inclusion of

[29] A coefficient of 0.35 corresponds to an odd-ratio of $exp(0.35) = 1.42$ or a roughly 40 percent higher chance of publishing in year t.

[30] The mean probability of a sample author outside NYC publishing is 0.22. The estimated New York coefficient corresponds to a 42 percent increase in the probability of publishing from 22 percent to 31 percent ($1.42 * 22\% + 22\% = 31.24\%$) or an increase of nine percentage points. Therefore, the New York City Effect is associated with an increase of nine percentage points

these additional mobility-specific variables does not change our New York City estimate, though we do find that authors with a high number of moves given their age also have a higher probability of publishing.

Our results could be inflated if there is positive selection into New York City, that is overall more productive authors having a higher probability of moving to New York City. In a first simple check for such a selection effect, we compare the means of the individual fixed effects estimated for authors who live at least partly in New York City to those of authors who consistently live elsewhere. As seen in the last row of Table 3 (NYC - other mean FEs), we find that authors who never live in New York City have a higher average publication productivity though their average productivity might be different once we take into account the positive impact of New York City and the effect of age and year. Therefore, we obtain some reassurance that there is no positive selection.

These findings are consistent with the findings of other studies on agglomeration economies in cultural production. Hellmanzik [2010] finds evidence of substantial premiums for paintings produced in the artistic clusters of Paris and New York. Borowiecki [2013] and Borowiecki [2015] observe agglomeration economies in classical music production. More closely related, Mitchell [2019] finds that there is a significant productivity gain for London-based writers. Consistent with findings here, Mitchell [2019] does not observe evidence of agglomeration effects associated with living in minor literary clusters and does not observe evidence of positive selection of authors into London.

6.2 Moving to New York City as an event study

Since identification comes from moves to New York, we then zoom in on the yearly productivity of authors who move to New York City in order to assess the trend in productivity around the time of the move to New York City. This is important for our understanding of the agglomeration mechanisms at work as well as for understanding possible selection mechanisms. We consider all moves of working age authors to New York City as an event study; that is, we analyse the productivity of these authors around their relocation to New York City in two frameworks: a panel regression framework with yearly dummies around the time of changing place of residence and a matching framework for a comparison between New York-based authors with similar authors who never move to New York.

6.2.1 Pre- and post-move to New York City

In Figure 8 we show the estimates of relocating to New York City based on dummies indicating the time to or since an author's first working age move to New York City. On the left, we show yearly odds-ratios with the period more than five years before the move as base (reference period). We see that in any year after the move to New York City authors are significantly more likely to publish compared to the time period more than five years before the move. The point estimates range from 2.5 to five times higher odds. Before the move to New York City, authors have significantly higher odds than the reference period only during one year, the second last before the move. Crucially, we do not observe an upward trend in the publication odds over the last five years before the move but a clear jump in productivity immediately after the move to New York which is sustained through several years after the initial relocation.

On the right, in Figure 8b, we take the last five years before relocation as the base and compare the publication propensity during these years to other five-year periods. We observe that the publication propensity is significantly higher for all time periods after relocation with around twice the odds of publishing. On the other hand, authors are more productive during the five years before relocation compared to years before. In contrast to the discussion in the last paragraph, this difference is significant if we consider pooled time periods. Since comparing individual years brings a lot of variability, these pooled time periods also give a better understanding of the potential range of the effect.

For context, we observe in Figure 7 that authors tend to move to New York when they are relatively young (typically before age 30). Therefore, these authors would be between ages 18-25 in the period $t < -5$ and we might expect considerable variability in output as some authors might not have begun to publish yet. For the interpretation of the estimates, it is important to note that these are conditional on age, author, and year. Therefore, we do not see direct publication patterns but differences relative to the author's overall publication propensity and their peer's propensity at the same age.

As described, the primary aim of this exercise is to determine whether there is a distinct jump or kink in the productivity trajectory after the move (and thus the gains can be more confidently attributed to residence in New York) or whether authors accelerated their productivity in the years immediately before the move to New York (suggesting that the estimated gains may reflect correlated effects). However, this exercise also allows us to analyse the dynamics of the agglomeration effect. We find that the gains remain relatively constant for the first several years after relocation and then begin to decay ten or more years after moving to New York City. This suggests that the New York City agglomeration effect on quantity is primarily driven by static agglomeration economies.

That is, we observe evidence that the productivity gains are associated with exposure to agglomeration forces that can be accessed and exploited immediately and do not observe evidence of a dynamic learning effect. This is in contrast to the de la Roca and Puga [2017] finding that Spanish workers learn by working big cities, as reflected by trends in individual earnings. However, this finding is similar to that of Mitchell [2019] for London-based authors.

This pattern does not exclude that authors could learn in other ways that impact the type of work they produce (e.g., transitioning from short stories to novels or plays) or the quality of the work they produce (e.g., writing a bestseller or critically acclaimed work). It is also possible that moving to New York City helped authors match with publishers that understand and respect the authors' artistic perspectives and offer a multi-book contract. It was not necessary to live in New York City to have access to the literary market or to find a publisher, but living in New York City and having access to a dense network of publishers likely made it easier. Once a good match has been made, an author's productivity might increase as the author no longer has to search for a new publisher for each respective publication therefore reducing search costs and increasing time available for writing. The rate of production will increase after this match has been made, but the rate of production would not necessarily continue to increase over time, especially if there are additional limits to publication quantity, such as author-side limitations in the writing process.

6.2.2 Quasi-experimental matching of movers and non-movers

Next, we move from an analysis of the relative impact of relocation on productivity to a direct comparison of publication means. We compare the mean productivity of authors who move to New York City with authors who never live in New York. To this end, we run a quasi-experimental approach matching authors who move to New York between 20 and 60 with authors who never live in New York by age, year (± 1), ethnicity, and gender as described in Section 2 and the respective 95% confidence intervals. Figure 9 illustrates the results.

We see that authors who move to New York City do so on average at an age corresponding to their, and their comparison group's, career start. Both groups have a clear upwards trajectory in their yearly publication propensity. However, the New York City *movers* group shows significantly higher yearly publication means than their comparison group during most of the early years in New York but not before their move there. Before, and long after the move, the difference between both groups is insignificant or, in fact, the comparison group's average is higher.

These findings provide further evidence that the observed New York City effect presented above is not driven by model choice. In addition, because we match authors by age and time, the effect is also not driven by some type of sample imbalance. For example, there are relatively few authors in New York City after World War II, but they are more productive than earlier cohorts even at later stages of their careers. This kind of sample imbalance might bias the panel fixed effects estimates, but the results of our matching exercise show that results are not driven by this kind of sample imbalance. Last, we do not observe evidence that highly prolific authors systematically self-select into New York City.

6.3 Moves to and from New York by prior publication record

We now investigate the heterogeneity of the New York City effect by moves to and from New York and by prior publication record. Column (1) in Table 4 shows that the New York effect is significantly higher for unpublished authors doubling the probability of publishing when living in New York City ($exp(0.75) = 2.12$). Consistently, the point estimate for published authors gives a 30 percent increase in publication probability, slightly lower than the overall estimate in Table 3. Hence, moving to New York City is more beneficial for novice authors than established authors. This is consistent with the hypothesis that a dense network of publishers and other industry gatekeepers could have reduced search costs and resulted in, first, an increased likelihood of an author being represented by a publisher and, second, higher quality matches between authors and publishers. However, due to the nature of the available data, we cannot empirically confirm this mechanism.

Next, we include a second New York dummy to distinguish the moving to New York City effect from a moving away effect. The new dummy indicates whether an author lives currently in New York City or has lived there previously: $NYC_{i,\tau\leq t} = 1 - \prod_{\tau\leq t}(1 - NYC_{i,\tau})$, see also Section 5. When we use both dummies in a regression, the $NYC_{i,\tau\leq t}$ parameter expresses the probability differential of publishing between the years before moving to New York and those after the move, which include years spent at different locations. (Therefore, authors who already live in New York at 18 or younger or not before they are 65 are no different from authors who never live there for this variable.) Correspondingly, the NYC_t parameter can now be interpreted as an interaction term that indicates the additional benefit of still being located in New York City, or put differently, the size of the penalty of moving away.

We see in Column (2) that the penalty of moving away from New York City is relatively mild and the overall New York effect is mostly an expression of authors moving there.

The publication probability after moving away, which is estimated at 90 percent of the New York City publication record, is not significantly different. In contrast, the initial move to New York is associated with an increase to a publication probability which is more than twice as high in any year afterwards ($exp(0.81) = 2.25$). However, if authors maximise their publication output and if, in addition, any move between locations inflicts further costs, such as social adaption, insecurity, and direct financial costs, it is reasonable to expect a positive benefit of any move on productivity. Even a low penalty of moving away, might, therefore indicate significant location effects in New York.

Last, we show in Column (3) that the initial move bonus is only slightly lower for already published authors. This differentiation is interesting as we are more likely to observe published authors moving away from New York than moving there. Therefore, our estimates for the two move directions could have also been expressions of different career stages. However, the estimates in this column confirm the interpretation of an initially positive productivity effect in New York City even after an author's first publication. At the same time, both the relocation bonus and the moving-away penalty decrease for published authors.

In sum, we find that the estimated New York City effect reflects productivity gains associated with moving to New York City rather than a penalty associated with moving away from New York. Therefore, in the following analyses, we focus primarily on moves to New York or being in New York rather than exploring moves away from New York in more detail. We also find that novice authors receive greater gains from moving to New York compared to established authors. This is further evidence against the possibility of positive selection into New York City.

6.4 Agglomeration effects over time and by industry concentration

We now analyse the development of the New York City agglomeration benefits over time and across different levels of industry concentration. New York City's industrial dominance changed dramatically over the sample period. We link these dynamics in this section to agglomeration benefits directly and via decade-level estimates for the period 1880-1990 in Figure 10. This figure illustrates the estimated odds-ratios and 95% confidence intervals that resulted from the estimation of Equation 3 in Section 5.

These results indicate statistically significant increase in the probability of publishing associated with living in New York City during the 1900s, 1910s, 1920s, and 1930s. This effect peaks during the 1900s and begins to wane from the 1910s. By the 1940s, authors

are no more productive during the periods when they live in New York City compared to when they live elsewhere. This is consistent with our hypothesis that the timing of agglomeration effects would coincide with major developments in the publishing industry, with agglomeration effects emerging around the turn of the 20th century and beginning to decay from the 1940s. Strikingly, this peak impact also coincides with the highest observed concentration of authors in New York City as recorded by the US Census.

One caveat in the interpretation of the estimates in Figure 10 is that the relative number of moves to and from New York City changes over time, and we have seen the diverging estimates for the two categories. We therefore repeat the decade estimation while distinguishing moves to and from New York City. This is crucial as some decades might see more moves to New York, while others see authors leaving. If the benefit of moving to New York is larger than the penalty for moving away, decades with more in-migration would mechanically see higher estimates for a New York agglomeration effect.

Therefore, we re-estimate Equation 2 using again decade dummies to interact with the dummies for living in or having lived in New York City. The obtained estimates in Figure 11 confirm the decade pattern for moves to New York City. The moving away penalty is significantly different from zero only during the 1910s. We conclude that the estimates shown in Figure 10 are driven by moves to New York City but that this does not invalidate the comparison between decades as the time trends are the same for the overall effect and the effect of in-migration.

Last, we directly estimate a central hypothesis of this paper: the concurrence of individual productivity gains and industry concentration in New York City. We interact the New York City effect with either the share of authors in New York in a given year (Figure 12a) or the difference of the share over the last five years (Figure 12b). We therefore test whether the agglomeration effect on productivity is higher at times of high industry concentration, when many authors lived in New York City compared to the rest of the US. Or, in or second specification, whether the effect is higher when the industry concentration increases, that is when many authors recently moved to New York City. We compute industry concentration using US census data, so this statistic is computationally independent of the behaviour of the authors in the *Britannica* sample.

In sum, we observe in Figure 12 and Table 5 that the productivity gains associated with residing in New York City rise and fall in tandem with the shift in the industry concentration of New York City over time. Given that the average age of moves to and from New York remains constant over time (see Figure 7), age or career stage productivity effects around the time of a move to New York should remain constant, too. However, the productivity benefits change significantly over time and closely align with the pattern

in industry concentration. Therefore, we can more confidently attribute the estimated productivity gains to changing agglomeration patterns in New York City as opposed to confounding effects due to the decision to start a career.

6.5 Evidence on commercial and critical success

We now examine the effect of moving to New York City on quality of published works. This is interesting to better understand the mechanisms involved but also as a check on our quantity estimates presented earlier. Authors can strictly determine when to start their career but have little or no control over the timing of their peak quality. Table 6 shows estimates for the probability of publishing critically acclaimed works, included in the *Kindler* encyclopedia, and commercially successful publications, included in the *Publishers Weekly* bestselling lists. Column (1) estimates the probability of publishing a work which entered the *Kindler* encyclopedia. Column (3) shows estimates for having a book in the *Publishers Weekly* top ten in year t. Columns (2) and (4) show estimates for the general probability of publishing in year t for the respective subset of authors; that is, those who have at least one *Kindler* or *Publishers Weekly* entry.

The results show that moving to New York City has a positive and statistically significant association with the probability of publishing critically acclaimed works. The gains for critically acclaimed works are greater than for publications overall. In contrast, authors are no more or less likely to experience commercial success when living in New York City compared to living elsewhere (Columns (3-4)).

These findings with respect to quality support theoretical arguments for cluster benefits discussed in Section 2: that authors pursuing critical acclaim will receive greater gains from agglomeration than those pursing mass market success. These findings could indicate that living in New York facilitates better matches between authors and relevant industry gatekeepers. They could also indicate that the transmission of tacit knowledge of literary styles, norms, and evaluation criteria was necessary to achieve critical acclaim.

Next, we explore the longevity or lasting impact of works that were published after an author moved to New York as measured by contemporary popularity via the number of ratings on *Goodreads*. Here, we use the term "contemporary" to mean present-day popularity, not contemporaneous (at time of publication) popularity. The results in Table 7. Columns (1)-(4) show the heterogeneity of the effect the number of ratings of *Goodreads*. We estimate the probability of publishing a book whose number of ratings places it into a given quartile using a Logit regression. Column (5) shows estimates from a Poisson regression for the number of ratings on *Goodreads*. We restrict observations in

this regression to years with at least one publication on *Goodreads*. The terms NYC_t and $NYC_{\tau \leq t}$ have the same interpretation as in the previous regression.

We find that the magnitude of the effect of moving to New York increases with the contemporary popularity of the publication. That is, authors who publish living in or after having lived in New York are most likely to publish a book with a high degree of contemporary popularity (i.e., the highest quartile of number of ratings on *Goodreads*). In addition, the number of ratings per book estimated directly is also significantly higher after a move to New York City. Moving away from New York City has no statistically significant effect on the quantity of publications rated on *Goodreads*. However, individual works receive more ratings on *Goodreads* if the author has lived in New York City at the time of publication.

Therefore, these findings show the importance of having lived or living in New York City on contemporary popularity, that is agglomeration effects which materialise over the very long run. As Bourrier and Thelwall [2020] show, ratings on Goodreads for older publications mirror literature read in secondary and tertiary education. The impact measured here is, therefore, probably at least partly transmitted by critical acclaim. As such, these findings complement our previous findings on critical acclaim, albeit using a more indirect measure. However, these results also illustrate the impact on an interested readership. Similar to citations of scientific articles or patents, once a publication has been deemed important or canonised by relevant industry gatekeepers, it is more likely to endure.

These findings are also important by supporting the plausibility of our main effects on publication quantity. In particular, it addresses within-author endogeneity, as authors arguably have less choice over the timing of their peak quality than when to start their career. If the observed productivity gains were driven by the fact that moving to New York City purely coincides with the decision to start a career, it is unlikely that we would also observe these patterns of quality gains. Therefore, we can more confidently attribute these productivity and quality gains to agglomeration economies in New York City.

7 Conclusion

This paper shows how New York City offered significant productivity benefits to authors residing there during the early stages of the industry's professionalisation path. However, we also find how individual benefits and industry concentration in New York decline with industry growth and maturity. Given this strong growth of the publishing industry all over the United States and widely decreased costs of publication after World War II,

we attribute the decreasing publication quantity differential between New York and the rest of the county primarily to a catch-up by the periphery, the professionalisation of the publishing industry, and technological progress, rather than an absolute decline of literary circles and the publishing industry in New York City.

A number of factors made it easier and less costly to be located elsewhere particularly after World War II, in particular the shift in author-publisher relationship and rise of literary agents as intermediaries facilitated this development. This could have implications for other clusters of solitary, talent-based activity, as they could be more subject to fluctuations in economic and technological development than those requiring greater physical infrastructure or team-based activity.

With our data, we cannot give a final answer to the persistence of quality rather than quantity effects. We have shown that residence in New York did not just make it more likely to publish but to publish critically acclaimed works. From a theoretical perspective, it is probable that success with peers including literary critics and editors is largely helped by social or cultural and in turn geographic proximity. Lower publication costs would not directly remedy disadvantages through social and cultural distance. Therefore, the impact on quality should be more persistent than the impact on quantity. Although, this might change the more that social and professional interactions move online.

Due to the composition of our sample, our estimate cannot be seen as the average treatment effect for an average American but the effect on a talented and relatively young subgroup of the population. However, our primary research interest is in observing the evolution of agglomeration economies in a creative sector over a long time horizon. We focus less on the composition of the beneficiaries in this paper, but we acknowledge that literary agglomeration may have been particularly beneficial to specific demographics of authors.

Furthermore, we have highlighted the importance of location in a developing industry. Naturally, the strict geographic bounds of these agglomeration effects may lead to specific ethnic and socioeconomic groups being over-represented. Here, our estimates would imply that a random sample of authors is more likely to be representative of New York than the overall US population or demographically similar to historically mobile groups such as young men. In addition, social and cultural barriers as described often for literary circles would further limit diversity among successful authors. More research into the interaction of agglomeration economies and exclusionary practices could provide important insights.

References

Arzaghi, M. and Henderson, J. V. (2008). Networking off Madison Avenue. *The Review of Economic Studies*, 75(4):1011–1038.

Baum-Snow, N. and Ferreira, F. (2015). Causal inference in urban and regional economics. In *Handbook of Regional and Urban Economics*, volume 5, pages 3–68. Elsevier.

Becker, G. S. (1964). Human capital: A theoretical and empirical analysis, with special reference to education. *University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship*.

Becker, H. S. (1974). Art as collective action. *American Sociological Review*, pages 767–776.

Blum, D. (2020). "Gloom and Horror Unrelieved": Upton Sinclair's classic novel 'The Jungle' almost never saw the light of day. *Public Broadcasting Service*.

Borowiecki, K. J. (2013). Geographic clustering and productivity: An instrumental variable approach for classical composers. *Journal of Urban Economics*, 73(1):94–110.

Borowiecki, K. J. (2015). Agglomeration economies in classical music. *Papers in Regional Science*, 94(3):443–468.

Borowiecki, K. J. and Dahl, C. M. (2021). What makes an artist? The evolution and clustering of creative activity in the US since 1850. *Regional Science and Urban Economics*, 86:103614.

Bourdieu, P. (1969). Intellectual field and creative project. *Social Science Information*, 8(2):89–119.

Bourdieu, P. (1985). The market of symbolic goods. *Poetics*, 14(1-2):13–44.

Bourrier, K. and Thelwall, M. (2020). The Social Lives of Books: Reading Victorian Literature on Goodreads. *Journal of Cultural Analytics*, 1(1).

Buenstorf, G. and Klepper, S. (2009). Heritage and agglomeration: The Akron tyre cluster revisited. *The Economic Journal*, 119(537):705–733.

Cabral, L., Wang, Z., and Xu, D. Y. (2018). Competitors, complementors, parents and places: Explaining regional agglomeration in the US auto industry. *Review of Economic Dynamics*, 30:1–29.

Cameron, A. C. and Trivedi, P. K. (2001). Essentials of count data regression. *A companion to theoretical econometrics*, 331.

Carlino, G. and Kerr, W. R. (2015). Agglomeration and innovation. In *Handbook of Regional and Urban Economics*, volume 5, pages 349–404. Elsevier.

Ceulemans, C., Ginsburgh, V., Prieto-Rodriguez, J., and Weyers, S. (2020). One hundred years of solitude: Bestsellers in the United States, 1900-1999. *Poetics*, 79:101422.

Ciccone, A. and Hall, R. E. (1996). Productivity and the density of economic activity. *American Economic Review*, pages 54–70.

Combes, P.-P. and Gobillon, L. (2015). The empirics of agglomeration economies. In *Handbook of Regional and Urban Economics*, volume 5, pages 247–348. Elsevier.

de la Roca, J. and Puga, D. (2017). Learning by working in big cities. *The Review of Economic Studies*, 84(1):106–142.

Ellison, G. and Glaeser, E. L. (1999). The geographic concentration of industry: Does natural advantage explain agglomeration? *American Economic Review*, 89(2):311–316.

Encyclopaedia Britannica (2016). Encyclopaedia Britannica. Encyclopædia Britannica Inc. `www.britannica.com`.

Gabe, T. M. and Abel, J. R. (2016). Shared knowledge and the coagglomeration of occupations. *Regional Studies*, 50(8):1360–1373.

Glaeser, E. L. (2005). Urban Colossus: Why is New York America's largest city? *Economic Policy Review*, page 7.

Glaeser, E. L. and Mare, D. C. (2001). Cities and skills. *Journal of Labor Economics*, 19(2):316–342.

Goodreads.com (2020). `https://www.goodreads.com`.

Green, J. N. (2010). The book trades in the new nation. In Gross, R. A. and Kelley, M., editors, *A history of the book in America: Volume 2: An extensive republic: Print, culture, and society in the new nation, 1790-1840*, pages 75–127. University of North Carolina Press.

Haller, P. and Heuermann, D. F. (2020). Opportunities and competition in thick labor markets: Evidence from plant closures. *Journal of Regional Science*, 60(2):273–295.

Hanlon, W. W. and Heblich, S. (2020). History and urban economics. *CEPR Discussion Papers*, (15303).

Heinz Ludwig Arnold (2009). Preface to the third, completely revised edition. In H.L. Arnold (Ed.), Kindlers Literatur Lexikon. J.B. Metzler. `https://www.derkindler.de/das-lexikon/vorwort-kindler-2009/`.

Hellmanzik, C. (2010). Location matters: Estimating cluster premiums for prominent modern artists. *European Economic Review*, 54(2):199–218.

Helsley, R. W. and Strange, W. C. (1990). Matching and agglomeration economies in a system of cities. *Regional Science and Urban Economics*, 20(2):189–212.

Helsley, R. W. and Strange, W. C. (2002). Innovation and input sharing. *Journal of Urban Economics*, 51(1):25–45.

Iacus, S. M., King, G., and Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, 20(1):1–24.

James, H. (1909). *Hawthorne*. Macmillan, London.

Jaszi, P. and Woodmansee, M. (2009). Copyright in transition. In Kaestle, C. F. and Radway, J. A., editors, *A history of the book in America, Volume 4: Print in motion: The expansion of printing and reading in the United States, 1880-1940*, pages 90–101. University of North Carolina Press.

Kaestle, C. F. and Radway, J. A. (2009). A framework for the history of publishing and reading in the United States, 1880-1940. In Kaestle, C. F. and Radway, J. A., editors, *A history of the book in America, Volume 4: Print in motion: The expansion of printing and reading in the United States, 1880-1940*, pages 7–21. University of North Carolina Press.

Kindlers Literatur Lexikon (2019). Kindlers Literatur Lexikon. J.B. Metzler. `http://www.kll-online.de`.

Literature Online (2016). Literature Online. ProQuest LLC. `www.lion.chadwyck.co.uk`.

Luey, B. (2009). The organization of the book publishing industry. In Nord, D. P., Rubin, J. S., and Schudson, M., editors, *A history of the book in America, Volume 5: The Enduring Book: Print Culture in Postwar America*, pages 29–54. University of North Carolina Press.

Mitchell, S. (2019). London calling? Agglomeration economies in literature since 1700. *Journal of Urban Economics*, 112:16–32.

New York Times (1901). A young author's romance of love; "Springtime and Harvest," the Story of Upton Sinclair's Life. He Wrote It and Then Eloped, and Now Wishes to Found a Library for the People. *The New York Times Archive*.

O'Hagan, J. (2021). Prominent US authors, born 1800 to 1950. *Poetics*, page 101589.

Puga, D. (2010). The magnitude and causes of agglomeration economies. *Journal of Regional Science*, 50(1):203–219.

Roth, A. E. and Sotomayor, M. (1992). Two-sided matching. *Handbook of game theory with economic applications*, 1:485–541.

Ruggles, S., Flood, S., Goeken, R., Grover, J., Meyer, E., Pacas, J., and Sobek, M. (2020). IPUMS USA: Version 10.0 [USA Full Count 1850-1940 and USA Samples 1850-2000]. IPUMS: Minneapolis, MN. https://doi.org/10.18128/D010.V10.0.

Saxenian, A. (1996). Inside-out: Regional networks and industrial adaptation in Silicon Valley and Route 128. *Cityscape*, pages 41–60.

Stark, O. and Bloom, D. E. (1985). The new economics of labor migration. *American Economic Review*, 75(2):173–178.

Starrett, D. (1978). Market allocations of location choice in a model with free mobility. *Journal of Economic Theory*, 17(1):21–37.

Storper, M. and Venables, A. J. (2004). Buzz: face-to-face contact and the urban economy. *Journal of Economic Geography*, 4(4):351–370.

The Literary Encyclopaedia (2016). The Literary Encyclopaedia. The Literary Dictionary Company Ltd. `www.litencyc.com`.

US Census Bureau (2020). Census instructions. `https://www.census.gov/history/www/through_the_decades/census_instructions/`.

West III, J. L. (1989). *American authors and the literary marketplace since 1900*. University of Pennsylvania Press.

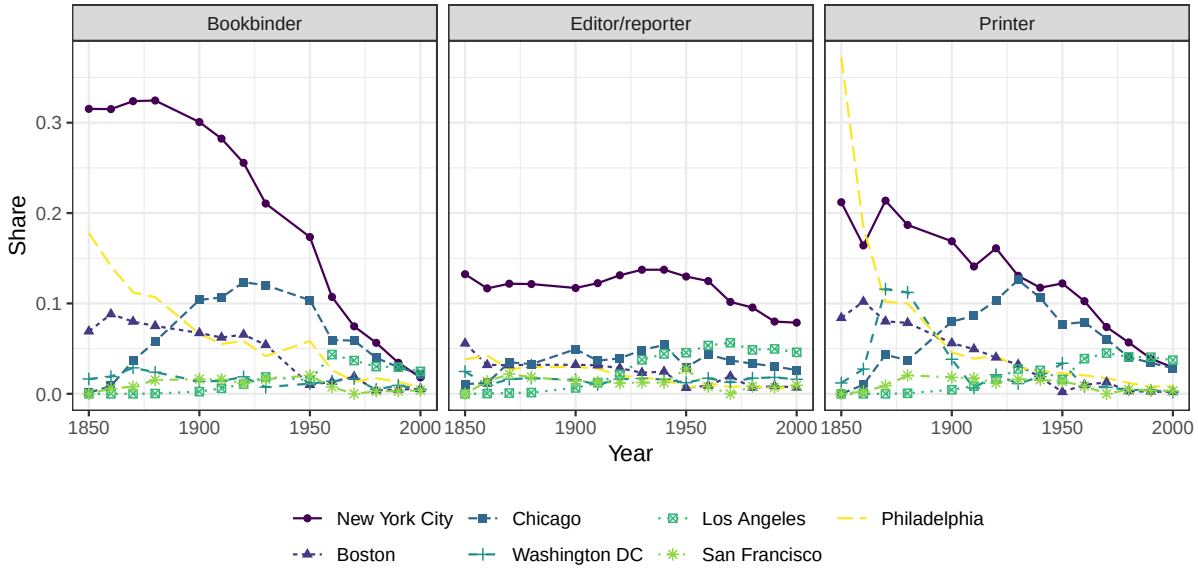
Wikipedia Contributors (2019). Publishers Weekly lists of bestselling novels in the United States. Wikipedia, The Free Encyclopedia. `https://en.wikipedia.org/w/index.php?title=Publishers_Weekly_lists_of_bestselling_novels_in_the_United_States&oldid=986386855`. Retrieved June 2019.

Williams, S. S. (2007). Authors and literary authorship. In Casper, S. E. and et al., editors, *A history of the book in America, Volume 3: The industrial book, 1840-1880*, pages 90–116. University of North Carolina Press.

Winship, M. (2009). The rise of a national book trade system in the United States. In Kaestle, C. F. and Radway, J. A., editors, *A history of the book in America, Volume 4: Print in motion: The expansion of printing and reading in the United States, 1880-1940*, pages 56–77. University of North Carolina Press.

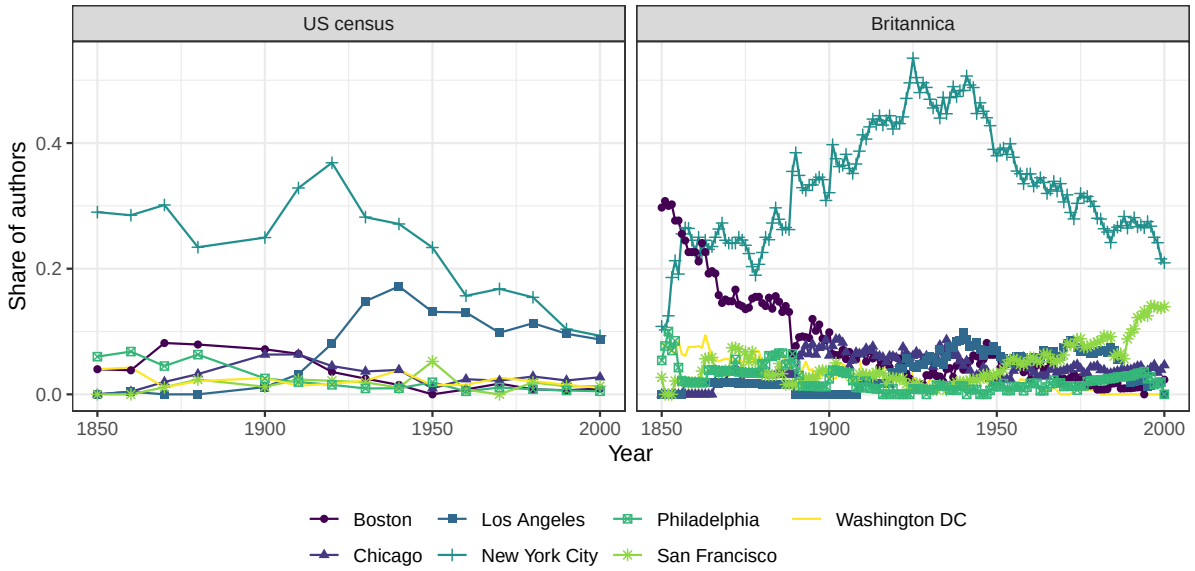
A Figures

Figure 1: Major cities for publishing-related industries



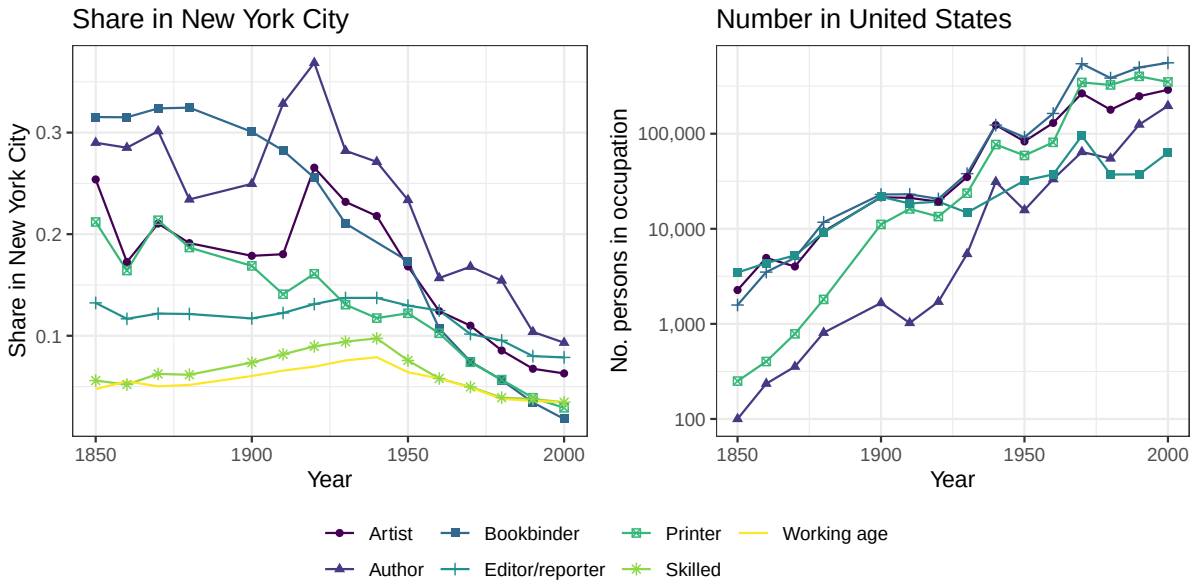
Source: Own calculation based on Ruggles et al. [2020] US census data (IPUMS) data.

Figure 2: Share of authors living in major cities



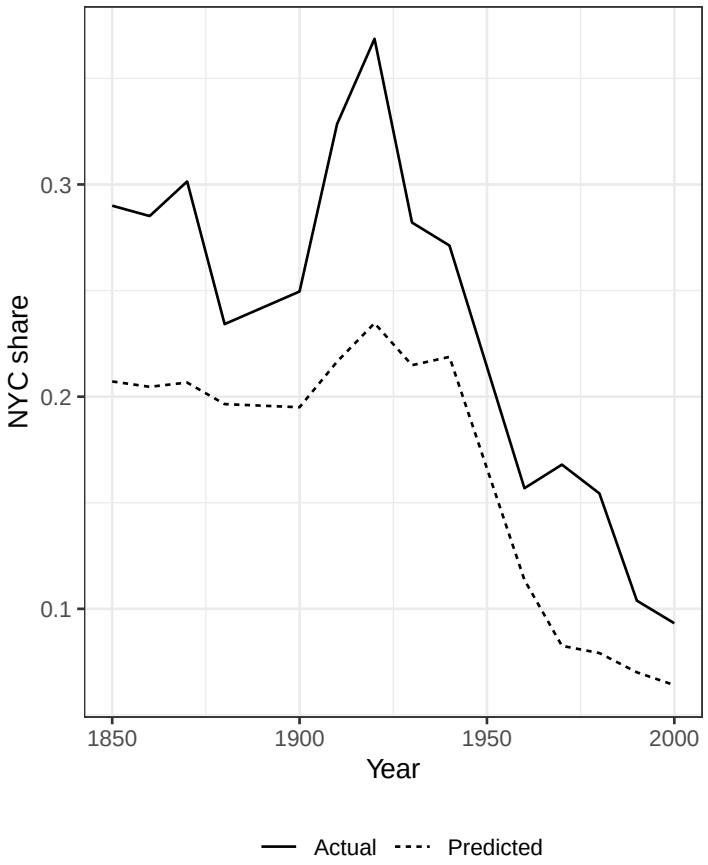
Source: Own calculation based on Ruggles et al. [2020] US census data (IPUMS) and *Britannica* data.

Figure 3: Size and concentration of industry over time



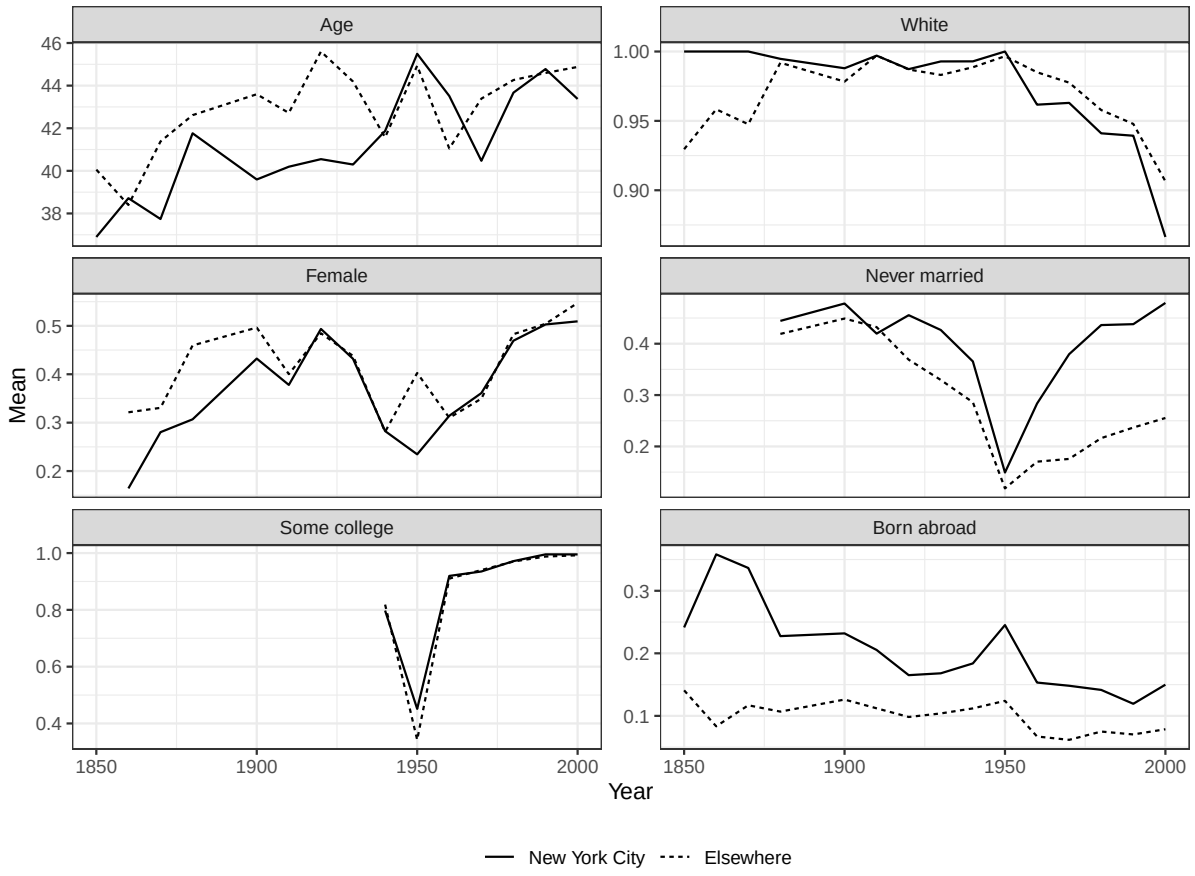
Notes: The figure on the left shows the share of the respective occupational group that resides in the New York City agglomeration. The figure on the right shows the total number of people employed of the respective occupational group in the US. Groups are classified by occupation or industry following the IPUMS OCC1950 and IND1950 classifications as described in Section 3. Data for bookbinders is not available in the 1940 Census.
Source: Own calculation based on Ruggles et al. [2020] US census data (IPUMS).

Figure 4: Author concentration in New York City



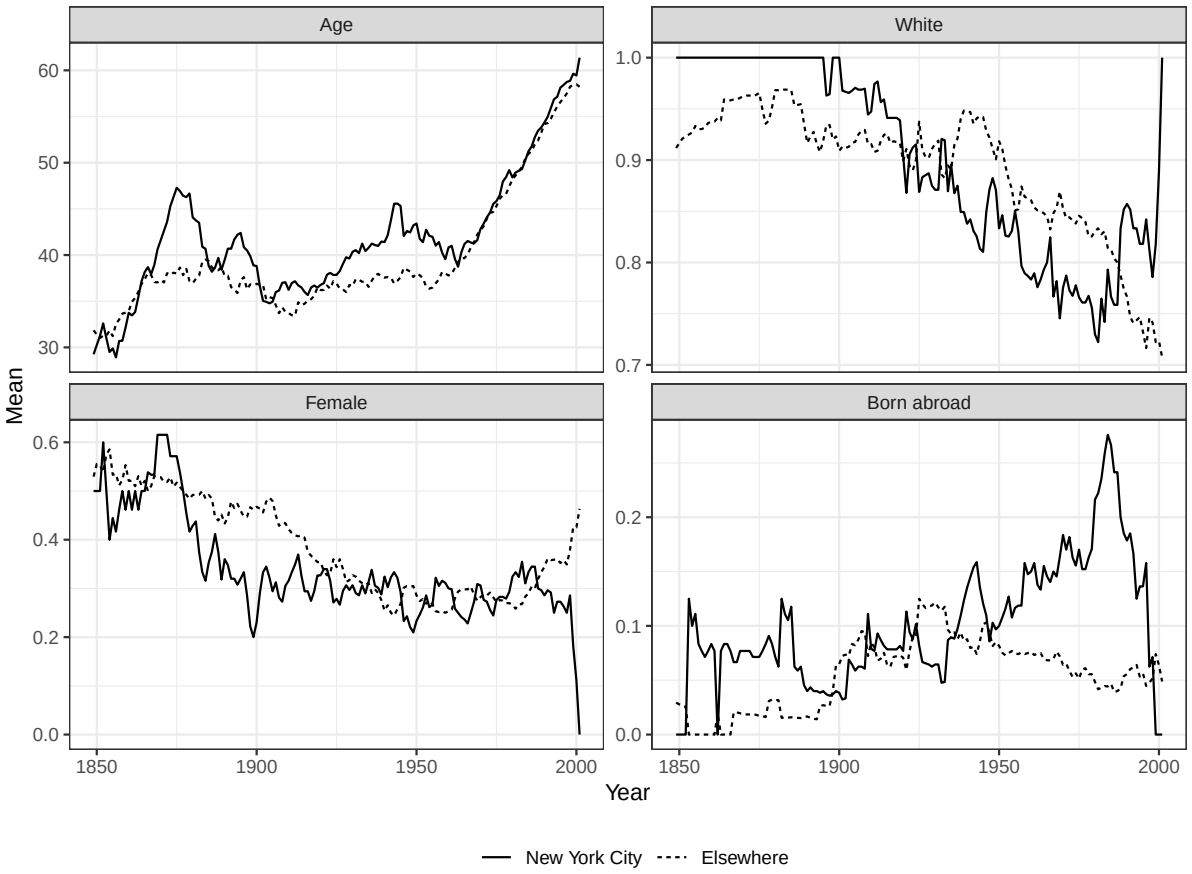
Source: Own calculations based on Ruggles et al. [2020] US census data (IPUMS) as described in Section 4.

Figure 5: Characteristics of *Census* authors in New York City and elsewhere



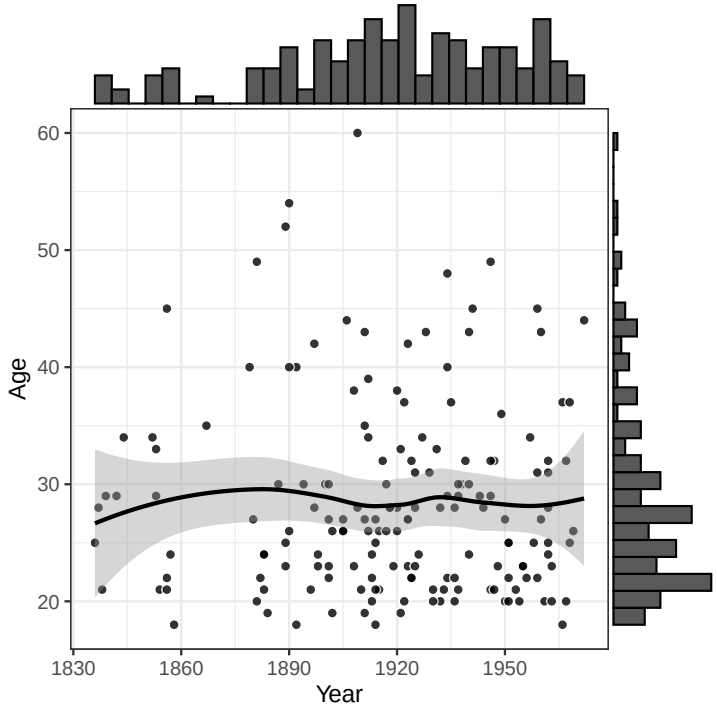
Source: Own calculation based on Ruggles et al. [2020] US census data (IPUMS).

Figure 6: Characteristics of *Britannica* authors in New York City and elsewhere



Source: Own calculation based on *Britannica* data.

Figure 7: Distribution of age and year for working age moves to New York City

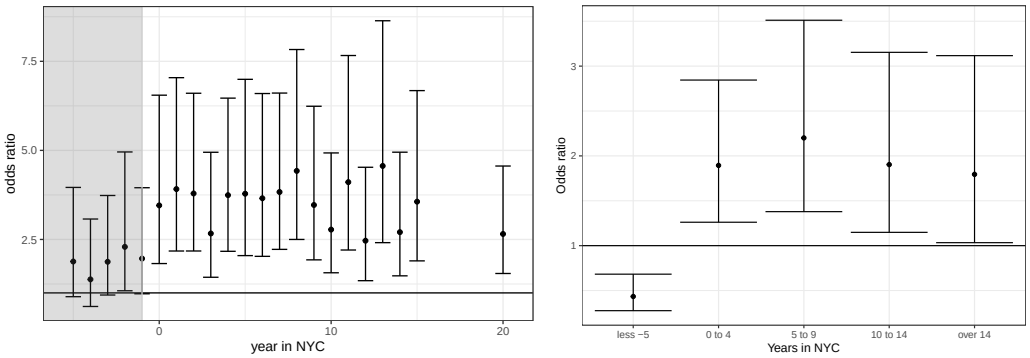


Notes: The plot shows the year and age at an author's first move to NYC. Moves are restricted to working age: 18-65. The trend is estimated using local regressions (LOESS with a span of 0.75).
Source: Own calculations based on *Britannica* data.

Figure 8: Probability of publishing pre- and post-move to NYC

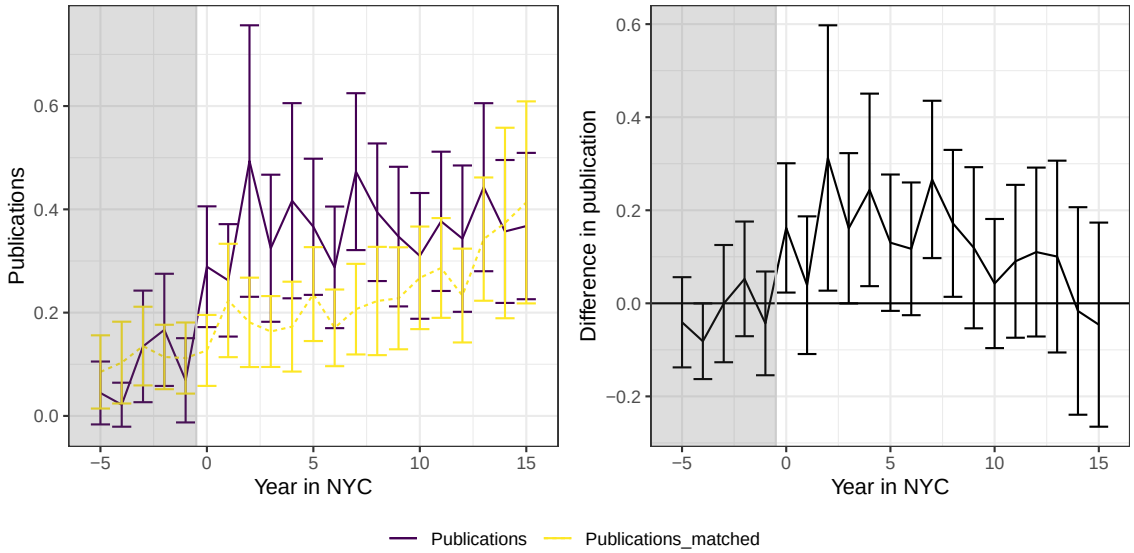
(a) 1-year periods

(b) 5-year periods



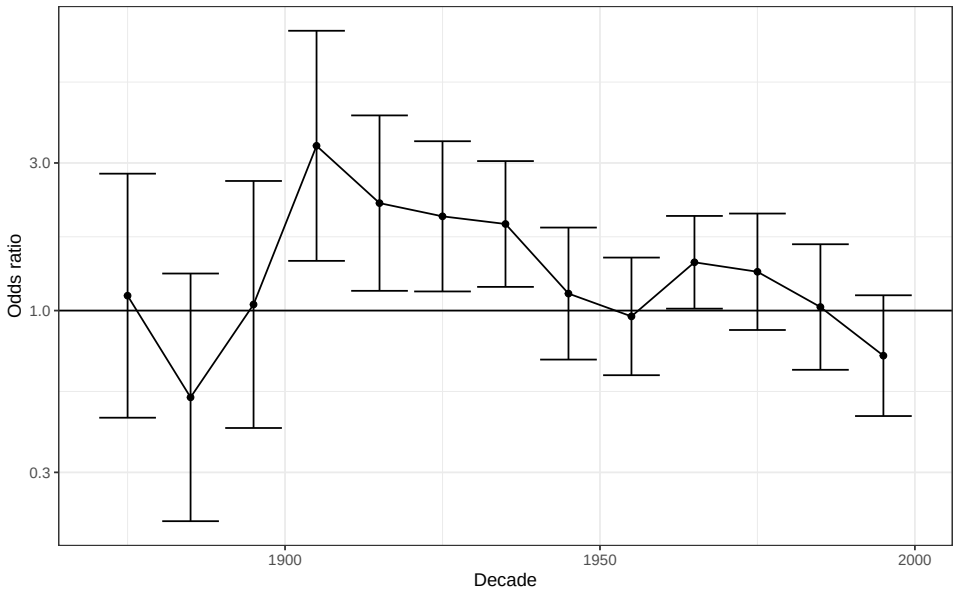
Source: Own calculations based on the estimation of Equation 1 with 1-year and 5-year dummies, respectively, as described in Section 5. The reference period for the 1-year estimates is the period 5+ years before moving to New York City. The reference period for the 5-year estimates is the period 5 years before moving to New York City. The graphs illustrate the estimated odds ratios and the respective exponentiated 95% confidence intervals.

Figure 9: Productivity of NYC-movers and matched authors elsewhere



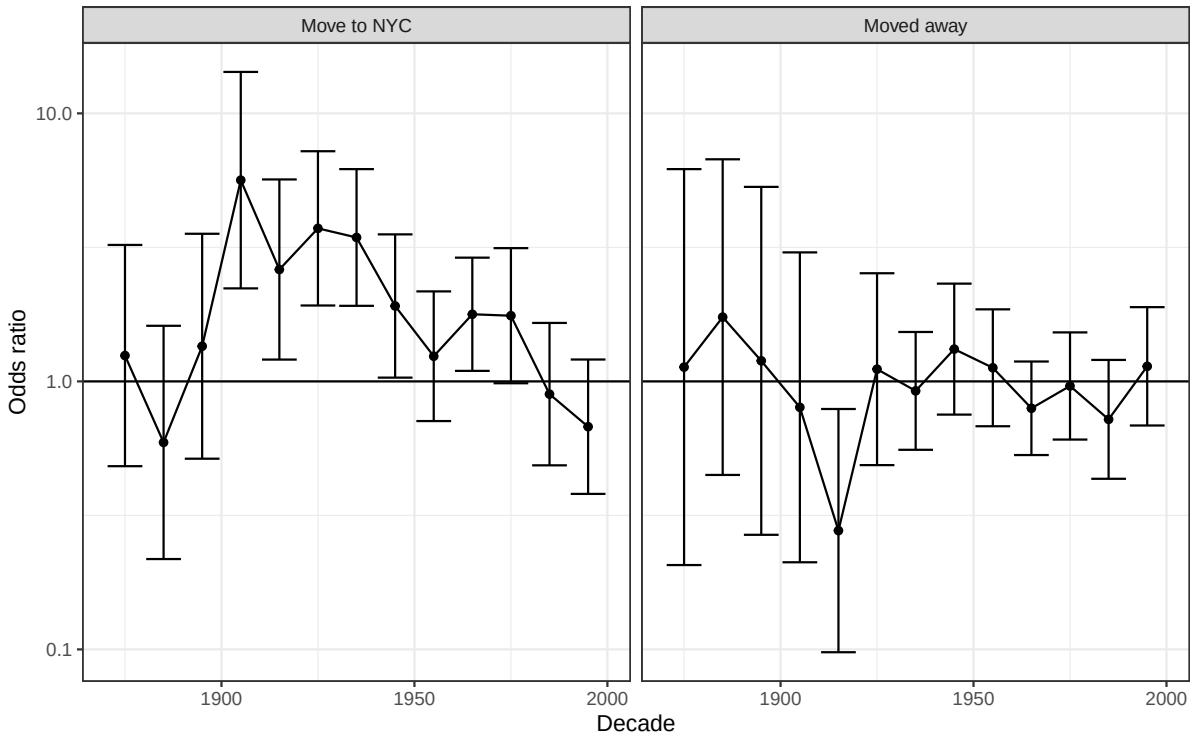
Source: Own calculations based on the quasi-experimental matching method described in Section 5. Authors are matched on age, year (± 1 years), ethnicity, and gender. The graph shows the sample means of publications for NYC-movers and matched authors elsewhere. The bars represent 95% confidence intervals.

Figure 10: New York City effect by decade



Source: Own calculations based on the estimation of Equation 3 as described in Section 5. The bars represent 95% confidence intervals.

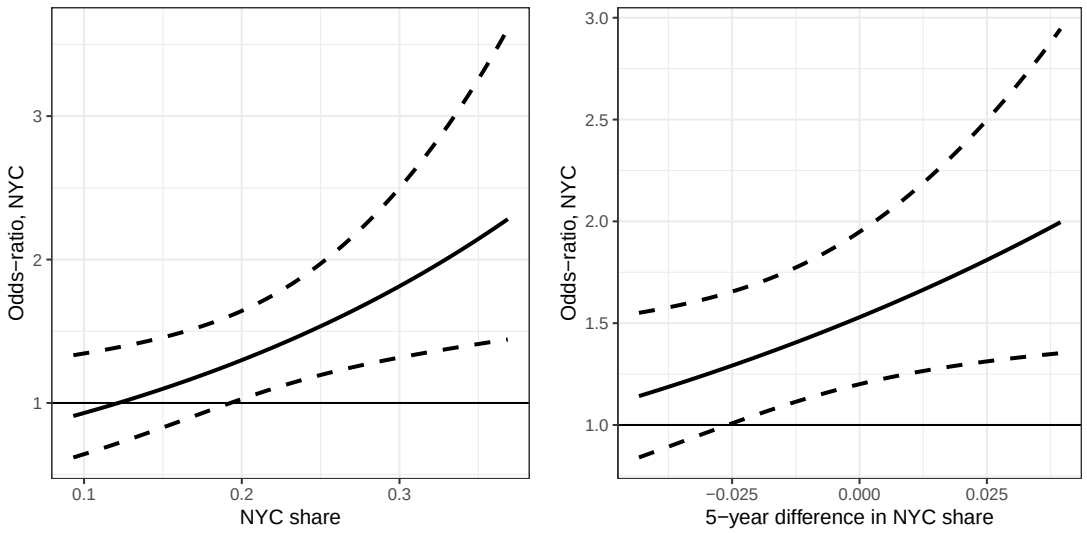
Figure 11: Effect of moving to and away from New York City, by decade



Source: Own calculations based on the estimation of Equation 3 with the addition of dummy variables for moves to and away from New York City, as described in Section 5. The bars represent 95% confidence intervals.

Figure 12: Effect of author concentration in New York City

(a) Author concentration in NYC (b) 5-year difference in concentration



Own calculation based on the estimation of Equation 5 as described in Section 5. The dashed lines represent 95% confidence intervals.

B Tables

Table 1: Summary Statistics for author datasets

	Britannica	Goodreads	Kindler	Publishers Weekly	Census
Authors	471	352	171	79	47619
NYC authors	269	206	102	53	8645
% years in NYC	56	55	55	57	n/a
Female authors	159	106	47	28	n/a
Years covered	1820 - 2014	1821 - 2014	1820 - 2013	1852 - 2014	1850 - 2000
Publications	4258	1388	226	164	n/a
Publication years	1826 - 2013	1827 - 2013	1824 - 1998	1895 - 2011	n/a

Notes: All statistics refer to yearly observations of authors with a known location in the US and aged 18 to 65.
Source: Own calculations based on data collection process described in Section 3.

Table 2: Summary of Equation 2 parameters and their interpretations

Parameter	Variable	Definition	Interpretation in full specification
β_1	NYC_t	Author lives in NYC in year t	Penalty of moving away/benefit of currently living in NYC
β_2	$NYC_{\tau \leq t}$	$1 - \prod \tau \leq t(1 - NYC_\tau)$	Benefit of moving to NYC
β_3	$y_{\tau<t}$	$1 - \prod \tau < t(1 - y_\tau)$	Effect of having previously published
β_4	$NYC_t \times y_{\tau<t}$	Product of two dummies	Additional effect of currently living in NYC for published authors
β_5	$NYC_{\tau \leq t} \times y_{\tau<t}$	Product of two dummies	Additional effect of moving to NYC for published authors

Notes: The table provides details to the specification shown in Equation 2 and Table 4. The interpretation refers to the full specification including all variables in the regression.

Table 3: Probability of publishing while living in New York City

	Logit	Probit	OLS	Logit	Logit
	(1)	(2)	(3)	(4)	(5)
	All	All	All	Goodreads	All
NYC	0.35**	0.21**	0.04**	0.40*	0.40***
	(0.12)	(0.07)	(0.02)	(0.20)	(0.10)
Major literary city					0.07
					(0.10)
Any move$_{t-5,\ldots,t-1}$					0.03
					(0.06)
Cumulative moves					0.10***
					(0.03)
N	15420	15420	15865	10569	15420
Author FEs	430	430	470	306	430
Year FEs	151	151	151	145	151
Age FEs	48	48	48	46	48
Deviance	14394.07	14399.32		6892.44	14376.82
R^2 (full model)			0.18		
NYC (avg. marg. eff.)	0.05***	0.05***	0.04**	0.03*	0.06***
	(0.01)	(0.01)	(0.02)	(0.01)	(0.02)
NYC - other mean FEs	-0.071	-0.032	-0.001	-0.398	-0.177

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Notes: The columns show estimates from Logit, Probit, and OLS regressions with a binary dependent variable indicating a publication by an author in year t. Standard errors are clustered on the author level and reported in parentheses. Coefficients are in log-odds units. The last row reports the difference in the means of the fixed effects for authors who lived in NYC and does who never lived there.

Table 4: Heterogeneity of NYC effect by moves to and from NYC and prior publication record

	(1)	(2)	(3)
NYC_t	0.75^{***}	0.11	0.39
	(0.17)	(0.13)	(0.33)
$\text{y}_{\tau<t}$	1.17^{***}		1.26^{***}
	(0.12)		(0.16)
$\text{NYC}_t \times \text{y}_{\tau<t}$	-0.47^{**}		-0.31
	(0.18)		(0.34)
$\text{NYC}_{\tau\leq t}$		0.81^{***}	1.00^{**}
		(0.21)	(0.37)
$\text{y}_{\tau<t} \times \text{NYC}_{\tau\leq t}$			-0.30
			(0.30)
N	16155	16155	11456
Author FEs	435	435	313
Age FEs	48	48	48
Year FEs	186	186	182
Deviance	14820.67	14932.27	10244.68

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Notes: The table reports estimates from Logit regressions for the probability of publishing in year t. Standard errors are clustered on the author level and reported in parentheses. Coefficients are in log-odds units. NYC_t is a binary indicator for whether the author lives in New York City in time t. $\text{y}_{\tau<t}$ is a binary indicator for whether the author has published in a year before t. $NYC_{i,\tau\leq t}$ is binary indicator for whether the author lives in New York City in time t or has lived there in a year before t, that is $NYC_{i,\tau\leq t} = 1 - \prod_{\tau\leq t}(1 - NYC_{i,\tau})$.

Table 5: Effect of industry concentration in NYC on probability of publishing

	(1) Share in NYC	(2) Δ Share in NYC
NYC_t	−0.41	0.42
	(0.30)	(0.12)
NYC_t*NYC $Share_t$	3.34**	
	(1.30)	
NYC_t*Δ NYC $Share_t$		6.75**
		(3.15)
N	15420	15206
Author FEs	430	427
Age FEs	48	48
Year FEs	151	146
Deviance	14382.66	14241.99

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

The table reports estimates from Logit regressions for the probability of publishing in year t. Standard errors are clustered on the author level and reported in parentheses. Coefficients are in log-odds units. NYC_t is a binary indicator for whether the author lives in New York City in time t. NYC $Share_t$ is the share of American authors living in NYC in year t. NYC $Share_t$ Standard errors are clustered on the author level and reported in parentheses. Δ NYC Share is the change of this share over the last five years up to year t.

Table 6: New York effect for bestselling and critically acclaimed works

	Kindler Authors		Pub. Weekly Authors	
	(1)	(2)	(3)	(4)
	Kindler works	All works	PW works	All works
$\text{NYC}_{\tau \leq t}$	1.50^{*}	1.15^{**}	0.35	1.11
	(0.68)	(0.38)	(1.93)	(0.95)
NYC_{t}	0.37	0.02	-0.07	-0.21
	(0.37)	(0.21)	(0.63)	(0.55)
N	3084	5484	1351	1303
Author FEs	131	154	56	52
Age FEs	39	48	38	38
Year FEs	108	151	82	82
Deviance	1251.63	5231.04	776.23	1328.69

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Notes: The table shows estimates for the probability of publishing in year t from using Logit regressions. Standard errors are clustered on the author level and reported in parentheses. Coefficients are in log-odds units. Column (1) only uses works included in *Kindler*. Column (2) uses all works by authors discussed in *Kindler*. Column (3) only uses works listed in *Publishers Weekly* bestsellers lists. Column (4) includes all works by authors included the *Publishers Weekly* bestsellers lists. Column (1) estimates the probability of publishing an work discussed in the *Kindler* encyclopedia. Columns (2) and (4) show estimates for the general probability of publishing in year t. Column (3) shows estimates for having a book in the *Publishers Weekly* top ten in year t. NYC_t is a binary indicator for whether the author lives in New York City in time t. $NYC_{\tau \leq t}$ is a binary indicator for whether the author has lived in New York City at any point in their lives prior or equal to year t.

Table 7: Effect of being in New York at time of publication on contemporary popularity

	No. online ratings on Goodreads				(5)
	(1)	(2)	(3)	(4)	No. of
	4th quartile	3rd quartile	2nd quartile	1st quartile	ratings
$\text{NYC}_{\tau \leq t}$	-0.64	0.90	1.82^{*}	13.51^{***}	6.23^{**}
	(0.58)	(0.71)	(0.89)	(2.77)	(2.29)
NYC_{t}	$0.62^{\cdot}$	0.18	0.10	-0.06	1.03^{**}
	(0.37)	(0.31)	(0.37)	(0.46)	(0.34)
N	4369	4569	4121	3211	1318
Author FEs	137	152	145	114	304
Age FEs	43	43	41	43	46
Year FEs	122	107	103	108	145
Deviance	1990.37	2085.28	1955.80	1612.67	8826962.70

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Notes: The table show estimates from Logit (1-4) and Poisson (5) regressions. Standard errors are clustered on the author level and reported in parentheses. Coefficients for Logit regressions are in log-odds units. Columns (1-4) refer to the probability of publishing a book in a category based on the number of ratings on *Goodreads* from the fewest ratings (4th quartile) to publications with the most ratings. Column (5) shows estimates from a Poisson regressions for the number of ratings on *Goodreads* with observations in this regression restricted to years with at least on publication on *Goodreads*. Standard errors are all clustered on the author level. NYC_t is a binary indicator for whether the author lives in New York City in time t. $NYC_{\tau \leq t}$ is a binary indicator for whether the author has lived in New York City at any point in their lives prior or equal to year t.