An Empirical Study of the Race, Ethnicity, Gender, and Age of Copyright Registrants

*Robert Brauneis & Dotan Oliar**

Abstract

Who is the author in copyright law? Knowing who our copyright system currently incentivizes to create which works is a necessary precondition for any effective copyright reform, yet copyright scholarship has thus far treated authors only through a priori conceptual analysis. This Article explores the author empirically.

Do those who self-identify as blacks (a U.S. Census category) register more music than members of other races per capita? Are Jewish authors particularly productive in registering literary works? What percentage of registrations represents works created by women? Which works tend to be registered by older authors? This Article provides answers to these questions—which happen to be yes, very likely, one-third, and literary works—and to many more by statistically analyzing the records of all fifteen million works registered with the Copyright Office from 1978 through 2012. It characterizes the modern-day American author along the axes of race and ethnicity, gender, and age.

The Article spells out the implications for copyright theory, policy, law, and reform. Copyright theory must explicitly account for the mechanism by which the copyright carrot induces authors of different demographics to create different types of works. This mechanism appears to contain substantial situated components—including social, cultural, and gender-related characteristics—that the major theories of copyright law that assume author uniformity do not acknowledge.

* Professor of Law and Co-Director of the Intellectual Property Law Program, The George Washington University Law School, and Member, Managing Board, Munich Intellectual Property Law Center; Professor of Law, University of Virginia School of Law. For valuable comments and discussions, we thank Michael Birnhack, Chris Buccafusco, Josh Fischman, Kristelia Garcia, Michael Gilbert, Alon Harel, Jerome Krief, Bobbi Kwall, Lydia Loren, Neil Netanel, Ariel Porat, Zvi Rosen, Rich Schragger, Micah Schwartzman, and participants in the 2015 Works in Progress Intellectual Property Colloquium, the 2015 Christopher A. Meyer Memorial Lecture, and intellectual property workshops at Berkeley, Lewis & Clark, Loyola Los Angeles, New York University, San Diego, St. John's, and Tel Aviv law schools. For access to and information about the Copyright Office Electronic Catalog, we thank Maria A. Pallante, Gail Sonneman, and many other members of the staff of the United States Copyright Office.

Table of Contents

INTRODUCTION

Who is the author in copyright law? Interpreting Congress's constitutional power to grant copyrights to "authors" for their "writings,"[1] the Supreme Court construed "author" to simply mean "one who completes a work of science or literature."[2] Unfortunately, today, more than 130 years later, we still do not know much about the author beyond this abstract and perfunctory statement.[3] But before determining whether our copyright system needs to change, and if so in what

1 U.S. CONST. art. I, § 8, cl. 8 ("The Congress shall have Power . . . To promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries . . . .").

2 Burrow-Giles Lithographic Co. v. Sarony, 111 U.S. 53, 57–58 (1884).

3 *See* Shyamkrishna Balganesh, *The Folklore and Symbolism of Authorship in American Copyright Law*, 54 HOUS. L. REV. 403, 405 (2016) (identifying "a problem that confronts modern American copyright jurisprudence to this day, despite the putative prominence of the author and authorship therein: the complete absence of a legal definition/account of the author, and of authorship"); Christopher Buccafusco, *A Theory of Copyright Authorship*, 102 VA. L. REV. 1229, 1230 (2016) ("Copyright jurisprudence did not begin with a theory of authorship, and it has not worked one out.").

way and how, lawmakers must first understand how the system currently works. This necessitates knowing more about the central figure in copyright law: the author.[4]

In this Article, we do not wish to join the scholarship that has thus far engaged in an a priori exploration of the author, whether conceptually,[5] ideologically,[6] theoretically,[7] historically,[8] or semiotically.[9] Rather, we believe that there is much to be gained from finding out who the author is empirically. We want to know who actually creates the books, articles, songs, movies, plays, art, and software that are the bedrock of American education, science, culture, and entertainment. What is the race, ethnicity, gender, and age of the authors of those works? Which authors are benefitting from our copyright system? Which authors are induced by the copyright carrot, and what are they induced to create?

We approach these questions by examining a hitherto untapped data source: the United States Copyright Office Electronic Catalog ("Catalog"). For the first time, through its Academic Partnership Program, the Copyright Office has provided us a full copy of the Catalog as it stood in late 2014. We expended much time and effort to clean

---

4 *See* Balganesh, *supra* note 3, at 404 ("Authorship is the real sine qua non of copyright law."); Oren Bracha, *The Ideology of Authorship Revisited: Authors, Markets, and Liberal Values in Early American Copyright*, 118 YALE L.J. 186, 186 (2008) ("The concept of the author is deemed to be central to copyright law."); Carys J. Craig, *Reconstructing the Author-Self: Some Feminist Lessons for Copyright Law*, 15 AM. U. J. GENDER SOC. POL'Y & L. 207, 209 (2007) (recognizing "the centrality of the concept of authorship to the operation and application of copyright law"); Jeanne C. Fromer, *Expressive Incentives in Intellectual Property*, 98 VA. L. REV. 1745, 1802 (2012) (emphasizing "how important the author is in copyright law"); Jane C. Ginsburg, *The Concept of Authorship in Comparative Copyright Law*, 52 DEPAUL L. REV. 1063, 1068 (2003) ("Much of copyright law in the United States and abroad makes sense only if one recognizes the centrality of the author, the human creator of the work.").

5 *See, e.g.*, Ginsburg, *supra* note 4.

6 *See, e.g.*, Bracha, *supra* note 4.

7 *See, e.g.*, Balganesh, *supra* note 3; Buccafusco, *supra* note 3; Tim Wu, *On Copyright's Authorship Policy*, 2008 U. CHI. LEGAL F. 335 (suggesting that copyright law should vest rights in authors to induce new types of creative works and new channels of distribution).

8 *See, e.g.*, MARK ROSE, AUTHORS AND OWNERS: THE INVENTION OF COPYRIGHT (1993); DAVID SAUNDERS, AUTHORSHIP AND COPYRIGHT (1992).

9 *See, e.g.*, Peter Jaszi, *Toward a Theory of Copyright: The Metamorphoses of "Authorship,"* 1991 DUKE L.J. 455 (deconstructing the concept of authorship using modern literary theory); Martha Woodmansee, *On the Author Effect: Recovering Collectivity*, 10 CARDOZO ARTS & ENT. L.J. 279 (1992) (critiquing the modern view of the author as an individual who is the sole source of a work); *see also* ROLAND BARTHES, *The Death of the Author*, *in* IMAGE MUSIC TEXT 142 (Stephen Heath trans., 1977) (elevating the reader's role, relative to the author's, in assigning meaning to the text); MICHEL FOUCAULT, *What Is an Author?*, *in* LANGUAGE, COUNTER-MEMORY, PRACTICE 113 (Donald F. Bouchard ed., trans. & Sherry Simon trans., 1977) (exploring the socially constructed relations between the author, reader, text, and meaning).

and organize the data on copyright registrations, which include the name and birth year of the author, the type of the registered work, its title, and its dates of registration, creation, and publication.

Our empirical analysis focuses on three variables that are not in the Copyright Office's data, but that we generate: authors' race and ethnicity, gender, and age. We are able to calculate authors' ages by subtracting their birth year from the year in which they created their works. Establishing authors' gender is not as simple. While it is easy to guess the likely gender of John and Jane, what about Pat or Terry? To answer this question, we use probabilities drawn from the gender distribution of first names released by the U.S. Census Bureau, in this case from the 1990 Census. Similarly, we determine authors' probabilistic race and ethnicity using last name data released by the U.S. Census Bureau from the 2000 Census.

Relying on Census statistics involves the risk that the gender distribution of authors' first names, or the racial distribution of their last names, might be different than those in the general population, such that the statistics reported may not be accurate. This risk is not substantial in the gender context, because the vast majority of first names are exclusively male or female, or virtually so. In contrast, many popular last names are more evenly distributed among races and ethnicities. We therefore estimate authors' race using two methods. First, as a benchmark, we use the racial and ethnic distribution of last names in the general population. Second, we use regression analysis. We explain why the results reached under our first method likely underestimate the true racial and ethnic registration disparities. Qualitatively, however, the two estimates are consistent with one another as indicators of over- and underrepresentation of certain demographics among authors.

Part I provides basic information about the Catalog and the subset of registration records that we analyze in this Article. Part II analyzes authors' race and ethnicity. Authors of different races differ in the rate and type of works registered. For example, black authors tend to register music at rates significantly higher, and Hispanic authors tend to register all works at rates significantly lower, than those of authors of all other races and ethnicities. Last names that Jewish sources suggest are often borne by those who self-identify as Jewish are associated with a high per-capita rate of registrations, particularly of textual works. Part III analyzes authors' gender. Among other things, we find that two-thirds of authors are male, but the gender gap in registration differs across types of works. We also find that men and

women show a strong within-group bias in choosing co-authors. Part IV focuses on authors' age. It shows that the average age of authors has increased over time, on par with the general population age trend. Different works tend to be created by authors of different age profiles: musical works tend to be created by authors who are on average ten years younger than those who create literary works. The production of music is also much more age-concentrated than that of literature. All aforementioned registration patterns have not been time-invariant: while authorial participation has shown signs of greater diversity over time, this trend has neither been linear nor universal.

Part V details policy implications. Our findings suggest a need for fundamental revision of copyright theory. The past decades have seen a blossoming of theories of copyright law and authorship. Due to the paucity of empirical data, it was hard to affirm or refute any of them. Because theories are evaluated by their ability to explain known data and predict future ones, it is striking that none of the existing theories of copyright law predicted the patterns discovered—that authors of different races and ethnicities, genders, and ages tend to create different types of works and at different rates—and only a few are consistent with them. Copyright theory—which tends to view the author in an abstract, uniform, ahistorical, and individualistic manner[10]—needs to account for the mechanism by which copyright entitlements induce particular authors to choose which works to create and at what rate. Our findings suggest that this mechanism contains important situating components, including social, cultural, and biological characteristics.

I. THE DATASET

Since January 1, 1978—the effective date of our current Copyright Act[11]—the Copyright Office has kept its records digitally. The records have thus far been accessible to the public only by means of an online search page that is suitable for researching rights in a particular title but not for conducting statistical analyses of millions of records.[12] For the first time, the Copyright Office, through its Aca-

10 *See* Dan L. Burk, *Copyright and Feminism in Digital Media*, 14 AM. U. J. GENDER SOC. POL'Y & L. 519, 546 (2006) ("The author is thus envisioned as a discrete and solitary individual, separate from both the community that consumes the work and from the relational network of shared understandings and cultural images within which the work arises.").

11 *See* Act of Oct. 19, 1976, Pub. L. No. 94-553, 90 Stat. 2541 (codified as amended in 17 U.S.C.).

12 Dotan Oliar was involved in a project that created a computer program to systematically download five years' worth of registration data, from 2008 through 2012. *See* Dotan Oliar,

demic Partnership Program, has provided us a full copy of the Catalog as it stood in late 2014. We expended much work to clean the data, reverse-engineer Office recordkeeping protocols that changed over time, and, importantly, convert the data from the Library of Congress's unique Machine-Readable Cataloging ("MARC") archival format to a customary columns-and-rows dataset structure. Conducting these steps—a laborious and time-consuming task—made it possible for us to analyze the data statistically. In the academic spirit of openness, and to facilitate third-party follow-up research, we plan to release the dataset with accompanying documentation.

In Parts II through IV below, we empirically characterize copyright demographics as they are reflected in the 14,598,621 original valid monograph registrations for the years 1978–2012 that were included in the Catalog as of September 30, 2014.[13]

A. Original Valid Monograph Registrations, 1978–2012

The Catalog contains records of various Copyright Office transactions that the Office keeps as part of its administration of the copyright system. Those transactions include copyright registrations and preregistrations, mask work registrations, document recordations, and mandatory deposits of published works. The Catalog currently contains records dating back to January 1, 1978, and new records are added to the Catalog on a daily basis.[14]

The Catalog as we received it contained over twenty-seven million records. We focus on a portion thereof—about 54%—that we call original valid monograph ("OVM") registration records. This subset narrows down the records of interest pursuant to the following criteria and reasons:

Monographs. Monographs are not serials, serials being works published in a series (such as magazines) that usually contain a collection of contributions by multiple authors. We exclude serials because their registration records contain thin authorship information that ap-

---

Nathaniel Pattison & K. Ross Powell, *Copyright Registrations: Who, What, When, Where, and Why*, 92 TEX. L. REV. 2211, 2219–20 (2014).

13 The most recently altered record in the version of the Catalog that we are using, CSN0107839, was last modified on September 30, 2014, at 17:07.17 (as recorded in field 005 of the MARC record).

14 The records in the Catalog are currently maintained in the MARC format for bibliographic records. For additional details on the history of the Catalog, see Robert Brauneis & Dotan Oliar, From the Copyright Office Catalog to the Original Valid Monograph Registration Datasets: Some History and Technical Details, http://www.robertbrauneis.net/registeringauthors/OnlineAppendixI.pdf [https://perma.cc/XK7E-FFCU] [hereinafter Online Appendix I].

plies only to the compilation as a whole, and contain no information about the types of work included.[15]

Original. Original registrations are those making an initial claim of copyright. We therefore exclude supplementary and renewal registrations. The former correct earlier-filed registrations, and including them would amount to double counting.[16] Renewal registrations were filed to lengthen the term of copyright or enhance the set of exclusive rights in works that obtained federal copyright before 1978.[17] We exclude them because they are not informative as to authorship patterns in our post-1978 period of interest.[18]

Valid. Finally, we only consider registrations that were valid as of the time our snapshot of the Catalog was taken. Although a registration record is created when an application is granted, it can later be cancelled for various reasons.[19] When this happens, the record is simply marked cancelled, rather than being removed from the Catalog.[20] Cancelled registrations do not represent valid legal claims and were excluded for that reason. Moreover, including them would often

---

15 *See generally* 17 U.S.C. §§ 101, 103 (2012) (defining and establishing copyright in compilations).

16 The 9/2014 Catalog contains 67,064 records of supplementary registrations relating to monographs, 67,035 of which are still valid. For graphic representation of the categories of monograph registrations, see Robert Brauneis & Dotan Oliar, Additional Tables and Charts 1 tbl.2, http://www.robertbrauneis.net/registeringauthors/OnlineAppendixII.pdf [https://perma.cc/PRR2-FX6V] [hereinafter Online Appendix II]. Under Copyright Office practice, if a second record is created while the content of the original registration is left unchanged, cross-references between the original and supplementary records are added. If there were a substantially larger number of supplementary registrations, we would have to figure out how to integrate the corrections and additional information that they contain into the original registrations, because the record of an original registration that has been the subject of a supplemental filing is incorrect or incomplete. However, less than one half of one percent of original registrations have been the subject of supplemental registrations. Therefore, for most statistical purposes, the supplemental registrations will make little difference, and we have decided not to undertake the difficult and time-consuming task of reading over 67,000 supplemental registrations and determining how the original registrations should be altered in light of those supplemental filings.

17 *See* U.S. Copyright Office, Compendium of U.S. Copyright Office Practices §§ 2102–2109 (3d ed. 2014).

18 Until 1992, renewals had to be filed to obtain copyright protection beyond the initial twenty-eight-year term; until the end of 2005, there remained some residual benefits to filing them. In the 9/2014 Catalog there are 730,401 records of renewal registrations for works that originally gained federal copyright before 1978. For graphic representation of these renewal registrations, see Online Appendix II, *supra* note 16, at 1 tbl.2.

19 *See* 37 C.F.R. § 201.7 (2016).

20 Of the 15,313,668 original registration records in the 9/2014 Catalog, 50,570 records represent cancelled registrations. (Similarly, 29 records of supplementary registrations represent cancelled registrations, and 384 records of renewal registrations represent cancelled registrations.) For graphic representation of these figures, see Online Appendix II, *supra* note 16, at 1 tbl.2.

amount to double counting, as many works claimed in cancelled registrations end up being re-registered.[21]

We have further excluded registrations that had blanks or invalid values in critical fields.[22] We have also decided to consider only registrations dated from January 1, 1978, through December 31, 2012. Because processing applications in the Copyright Office takes time, statistics drawn concerning registrations in 2013 and 2014 would be incomplete.[23] Applying these criteria left us with 14,598,621 records. It is those records that we analyze below.

B. The Basic Information in OVM Registration Records

Registration records systematically include three types of information[24]: information about the registration itself, about the work registered, and about the work's authors and claimants.

1. Information About the Registration. The most important datum in this category is the effective date of the registration, which all registrations have. This is the date that the Copyright Office received a valid application, deposit, and payment.[25] Because this is an objective and verifiable date, and because the Catalog begins with registrations with effective dates on or after January 1, 1978, we use it to organize our data by full years.

2. Information About the Registered Work. The most important information for our analyses are the following:

a. The Work's Year of Creation. Over 99% of registrations indicate the year in which the work was created.[26] Creation year is inferior

---

21 Subtracting the 50,570 records of cancelled original monograph registrations from the total of 15,313,668, we arrive at a total of 15,263,098 original valid monograph registration records. Three of those records contained no usable information and were therefore not included in the dataset we have generated.

22 Those exclusions of an additional 590 records, detailed in Online Table 3, leave the dataset with 15,262,519 records. *See* Online Appendix II, *supra* note 16, at 2 tbl.3.

23 There were 663,884 original valid monograph registration records with registration dates in 2013 and 2014 excluded.

24 Registration records can contain various other types of information, such as information about the deposit submitted with the registration application, initials identifying the Copyright Office staff member who prepared the registration, and so on, but we decided that this additional information was either irrelevant to our purposes or entered too inconsistently to be of use. Beginning in 2008, the Catalog has included additional information, such as mailing addresses associated with claimants and authors, and their citizenship.

25 The effective date of "registration is the day on which an application, deposit, and fee, which are later determined by the Register of Copyrights or by a court of competent jurisdiction to be acceptable for registration, have all been received in the Copyright Office." 17 U.S.C. § 410(d) (2012).

26 Some 104,091 registrations (about 0.72%) of the 14,472,367 registrations we focus on

as a running variable not only because some registrations do not have one, but also because it is self-reported by registrants. We further have no way of knowing (as we do in the case of registration years) that we have the complete set of works created in a particular year, as works can be registered at any time after their creation.[27]

b. Type of work. Each registration contains a two-letter code that identifies the work as predominantly belonging to one of eleven categories that are listed in Table 1 below.[28] Some categories are broad and cover a large number of registrations, while others are narrower and cover a comparatively small number of registrations.

For most of our analyses, we omit the three smallest type-of-work categories—"Map," "Sound Recording and Text," and "Multimedia Kit"—that together represent less than one percent of all registrations. Excluding them enables us to concentrate on the more consequential categories and to construct more legible charts and tables. We have also decided to combine three music-related categories, namely, "Musical Work," "Musical Work/Sound Recording," and "Sound Recording." Delving into them, one of us has revealed changing patterns in how music is created and registered.[29] Nevertheless, all three relate to the production of commercially distributed music, lead-

---

under the six major categories of work below do not have creation year data. In a little over 100,000 records, the creation year field is blank; in about 400 others, it was likely mistakenly entered, because it is either before 1500 or after 2014.

27 Registered works' creation and registration dates are quite close: 56.87%, 85.58%, 93.93%, 96.34%, and 98.29% of registered works were registered within zero, one, three, five, and ten years of creation, respectively. The mean difference in our dataset between the year of registration and the year of creation is 1.1 years. All numbers above were calculated after omitting 9129 registrations with a registration year earlier than their creation year, which are erroneous.

28 Type-of-work categories have always been meant to represent the predominant type into which a work submitted for registration falls, recognizing that works sometimes cross categories, and that a registration will cover all aspects of the work registered that have been created by the author or authors named in the application. A work fixed in a book, for example, may be primarily a literary work, but may also contain some illustrations that would qualify as pictorial works. *See, e.g.*, U.S. COPYRIGHT OFFICE, *supra* note 17, § 609.2(C) ("If the work contains more than one type of authorship, the applicant should select the type of work or the paper application that corresponds to the predominant form of authorship in that work."). Some of the categories of works listed in § 102 of the Copyright Act themselves recognize the hybrid character of many works in that category; for example, § 102(a)(2) defines one category as "musical works, including any accompanying words"; § 102(a)(3) defines another category as "dramatic works, including any accompanying music." *See* 17 U.S.C. § 102 (2012).

29 *See* Robert Brauneis, *Musical Work Copyright for the Era of Digital Sound Technology: Looking Beyond Composition and Performance*, 17 TUL. J. TECH. & INTELL. PROP. 1, 28–31 (2014).

ing us to believe that combining them is appropriate for present purposes.

As a result, when we analyze data in terms of types of works, we will be using six categories. As Table 1 shows, we will refer to them by single-word abbreviations, namely, "Text," "Music," "Art," "Movies," "Drama," and "Software."[30]

TABLE 1. TYPE-OF-WORK CATEGORIES

Categories in Registrations	Our Abbreviations	Number of OVM Registrations, 1978–2012	Percentage of OVM Registrations	Percentage of Our Six-Category Scheme
Non-Dramatic Literary Work	Text	5,462,210	37.42%	37.74%
Musical Work		3,926,918	26.90%	
Musical Work/ Sound Recording		623,835	4.27%	
Sound Recording		362,813	2.49%	
Music Combined	Music	4,913,566	33.66%	33.95%
Visual Material	Art	2,519,555	17.26%	17.41%
Motion Picture	Movies	747,262	5.11%	5.16%
Dramatic Work or Choreography	Drama	527,900	3.61%	3.65%
Computer Program	Software	301,874	2.07%	2.09%
Map		48,027	0.33%	
Sound Recording/ Text		42,154	0.29%	
Multimedia Kit		36,073	0.25%	

c. Publication Status and Publication Date. Each registration record notes whether the concerned work or works were published at the time of registration. If they were, a date of publication is usually also included. A little over half of all works were registered as published.[31]

3. Author and Claimant Information. Each record contains information about the work's authors and claimants, and whether each is an individual or a corporate entity. Below, we analyze the

30 A more complicated table in an online appendix shows the relationship between the categories we are using and other schemes for categorizing works of authorship, including the eight categories in § 102(a) of the Copyright Act. *See* Online Appendix II, *supra* note 16, at 3 tbl.4. It shows, among other things, that "Movies" includes all audiovisual works, that "Art" includes all pictorial, graphic, and sculptural works, and that "Drama" includes choreography and any music that might accompany a dramatic work.

31 Overall, 7,863,069 registrations, or about 54%, are for published works, while 6,735,551 registrations, or about 46%, are for unpublished works.

demographics of individuals. We parsed individuals' names to first and last, in order to use them in connection with our gender and race analyses, respectively. Sometimes, records would list both a noncopyright author (who may be an employee, or whose name may appear on the deposit copy) and the author for copyright purposes (such as an employer for hire). Our analyses attempt to count as authors only those who are authors for copyright purposes. We further employed various counting rules to deal with the use of pseudonyms, "doing business as" names, and other alternate names, which are detailed in Online Appendix I.[32]

II. Race and Ethnicity

A. Methodology: Inferring Race and Ethnicity from Last Names

Registration records do not specify individual authors' race or ethnicity, so we use their last names as a proxy. Luckily, almost all registrations by individual authors include their last names. In developing statistics on race and ethnicity we rely on information elicited from the 2000 U.S. Census regarding the racial and ethnic distribution of people with particular last names.[33] Under federal policy, the Census Bureau asked people to self-identify as members of one or more of six races—white, black, American Indian or Alaska Native, Asian, Native Hawaiian or Other Pacific Islander, and "Some Other Race."[34] In addition, it asked them to separately note whether they are "Spanish, Hispanic, or Latino," which it regards as their ethnicity, rather than race.[35]

Based on the answers it received, the Census Bureau provides the probability that holders of various last names are either of Hispanic ethnicity, regardless of race, or are, alternatively, non-Hispanic and fall into one of five mutually exclusive racial categories: white only, black only, Native American or Alaska Native only, Asian or Other

---

32 *See* Online Appendix I, *supra* note 14.

33 *See* U.S. Census Bureau, Frequently Occurring Surnames from the Census 2000, http://www.census.gov/topics/population/genealogy/data/2000_surnames.html (last updated Sept. 15, 2014) (download File A, File B, and Technical Documentation: Demographic Aspects of Surnames – Census 2000) (containing information on the probability that individuals with particular last names belong to one of six racial or ethnic categories).

34 *See* Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity, 62 Fed. Reg. 58,782 (Oct. 30, 1997), https://www.gpo.gov/fdsys/pkg/FR-1997-10-30/pdf/97-28653.pdf; Elizabeth M. Grieco & Rachel C. Cassidy, U.S. Census Bureau, Overview of Race and Hispanic Origin 2000: Census 2000 Brief (2001), https://www.census.gov/prod/2001pubs/c2kbr01-1.pdf.

35 Grieco & Cassidy, *supra* note 34, at 1–2.

Pacific Islander only, or two or more races. Although four of the categories are thus properly prefaced by "non-Hispanic"—non-Hispanic white, non-Hispanic black, non-Hispanic native American or Alaskan, and non-Hispanic Asian or Pacific Islander—we will, for the sake of convenience, refer to each of those categories without the "non-Hispanic" prefix. Thus, for example, it should be understood that every reference in this Article to "white" is really a reference to "non-Hispanic white," and every reference to "black" is really a reference to "non-Hispanic black." We would like to emphasize that in conducting race and ethnicity statistics, we have not taken any decision as to which races and ethnicities exist, how they should be denoted, and which individuals belong in which group. Rather, our statistics reflect a list of races and ethnicities defined and named by the government, into which Census respondents self-selected.

Relying on this governmental six-category taxonomy, we were able to assign probabilities of race or ethnicity to the vast majority of individual authors of registered works. The Census data provides probabilities for 151,671 last names.[36] Our dataset contains 10,425,336 registrations of works that were created by individual authors. Of those, 1,092,026 registrations did not contain a last name that appeared in the Census list of most common surnames, and therefore do not feature in our statistics on race. Our statistics build on the probable race or ethnicity of the individual authors of the remaining 9,333,310 registered works.

Last names are rarely determinative of their bearer's race. True, among the most common surnames, some are overwhelmingly held by people who self-identify as Hispanic or as Asian or Pacific Islander. Garcia, Rodriguez, Martinez, Lopez, and Gonzalez are all among the twenty-five most popular last names in the United States, and over 90% of people bearing those last names identified as Hispanic.[37] Nguyen, Tran, and Patel are among the 200 most popular last names, and over 90% of people bearing those last names identified as Asian or Pacific Islander.[38] By contrast, however, those who self-identify as white or as black tend to share surnames more evenly. For example, the five most popular last names in the United States are Smith (73% white, 22% black); Johnson (62% white, 34% black); Williams (49%

36 *See* U.S. CENSUS BUREAU, *supra* note 33.

37 DAVID L. WORD ET AL., DEMOGRAPHIC ASPECTS OF SURNAMES FROM CENSUS 2000, at 4, https://www2.census.gov/topics/genealogy/2000surnames/surnames.pdf (last visited Jan. 2, 2018).

38 U.S. CENSUS BUREAU, *supra* note 33.

white, 47% black); Brown (61% white, 35% black); and Jones (58% white, 38% black).[39] When, as shorthand, we make statements about the race or ethnicity of a certain cross-section of authors, we are referring to the average of the probable race or ethnicity of individuals in that cross-section.[40]

B. Main Findings

1. Overrepresentation of White Authors

Many people believe that the United States has slowly moved towards greater equality and opportunity, and, as a corollary, that the differences between races and ethnicities across various social and economic metrics are on the decline. Yet between 1978 and 2012, the proportion of white authors reflected in copyright registrations, compared to their proportion in the population, has grown. In 1980, whites accounted for about 79.6% of the general population in the United States.[41] In that year, they accounted for 79.47% of copyright registrations—almost exactly equal to their proportion of the general population. Since 1980, the percentage of whites in the U.S. population has been decreasing. It dropped to 75.6% by 1990,[42] 69.1% by 2000,[43] and 63.7% by 2010.[44] While the percentage of white authors represented in copyright registrations has also been dropping, it has not dropped nearly as much. It dropped to 77.41% in 1990; 75.19% in 2000; and 73.96% in 2010. Thus, as of 2010, white authors were producing 116% of the registrations they would be if they were producing at a rate

---

39 *Id.*

40 For the purposes of statistically analyzing race, we have excluded works that have no individual authors, such as works created by corporations, as these have no race. The probability that a work was authored by a particular race has been calculated as the average of that particular race among the work's individual authors for whom we have last name statistics. Race statistics, such as for a category of works or for a year, have been calculated as the average of the relevant works' probabilistic racial or ethnic authorship.

41 *See* U.S. Census Bureau, 1980 Census of Population: General Population Characteristics 1-52 tbl.49 (1983), https://www2.census.gov/prod2/decennial/documents/1980/1980censusofpopu8011u_bw.pdf (noting that there were 180,256,366 non-Hispanic whites out of a total of 226,545,805 persons in the United States).

42 *See* U.S. Census Bureau, 1990 Census of Population: General Population Characteristics 3 tbl.3 (1992), http://www2.census.gov/library/publications/decennial/1990/cp-1/cp-1-1.pdf (noting that there were 188,128,296 non-Hispanic whites out of a total of 248,709,873 persons in the United States).

43 *See* U.S. Census Bureau, Profiles of General Demographic Characteristics: 2000 Census of Population and Housing, at 1 tbl.DP-1 (2001), http://www2.census.gov/census_2000/datasets/demographic_profile/0_United_States/2kh00.pdf.

44 *See* U.S. Census Bureau, 2010 Census Briefs: The White Population: 2010, at 3 tbl.1 (2011), https://www.census.gov/prod/cen2010/briefs/c2010br-05.pdf.

equal to their proportion of the general population, which was the rate at which they were producing registrations in 1980, three decades earlier.

Why are white authors now overrepresented in copyright registrations, when they were not at the beginning of our study period? Part of the explanation may be age. The white population is relatively older than the population of other racial and ethnic groups,[45] and in particular has a smaller percentage of its population that is under twenty-five years of age,[46] a segment of the population that produces very few copyright registrations.[47] It is also possible that our methodology underestimates white authors before 2000 because it allocates to last names the population distribution as of 2000, whereas whites comprised a larger percentage of the population between 1978 and 1999 than in 2000 (although a smaller percentage between 2001 and 2012), which may suggest that whites were somewhat overrepresented in 1980 and 1990 (and not so overrepresented after 2000). Finally, some of the increase in overrepresentation may be the reciprocal of an increase in underrepresentation of Hispanic authors, which may have its own causes, and which we will now turn to discuss.

2. *Extraordinary Underrepresentation of Hispanic Authors*

In 1980, Hispanics constituted 6.4% of the U.S. population,[48] but Hispanic authors contributed only 4.45% of copyright registrations. Thus, Hispanic authors were producing only 69.5% of the registrations that they would if they were producing at a rate equal to their proportion of the population. Since 1980, the Hispanic population in the United States has grown considerably to 9.0% in 1990,[49] 12.5% in 2000,[50] and 16.3% in 2010.[51] By contrast, Hispanic authorship has grown at a slower pace to 5.3% in 1990, 6.8% in 2000, and 7.27% in 2010. Thus, as of 2010, Hispanic authors were producing only 44.6% of the registrations that they would be if they were producing at a rate equal to their proportion of the general U.S. population. That is by far

---

45 *See* U.S. CENSUS BUREAU, CENSUS 2000 SUMMARY FILE: RACE AND HISPANIC OR LATINO ORIGIN BY AGE AND SEX FOR THE UNITED STATES: 2000, at tbl.8 (2002), https://www.census.gov/population/www/cen2000/briefs/phc-t8/tables/tab08.pdf.

46 *See id.*

47 *See infra* Table 6 (showing registration rates by age groups).

48 U.S. CENSUS BUREAU, PERCENT HISPANIC OF THE U.S. POPULATION: 1970 TO 2050, at 5 (2012), https://www.census.gov/newsroom/cspan/hispanic/2012.06.22_cspan_hispanics_5.pdf.

49 U.S. CENSUS BUREAU, *supra* note 42, at 3 tbl.3.

50 U.S. CENSUS BUREAU, *supra* note 43, at 1 tbl.DP-1.

51 U.S. CENSUS BUREAU, 2010 CENSUS BRIEFS: THE HISPANIC POPULATION: 2010, at 3 tbl.1 (2001), https://www.census.gov/prod/cen2010/briefs/c2010br-04.pdf.

the largest underrepresentation of any racial or ethnic group. As mentioned above, in 2010 whites were at 116% (73.96/63.7). To round out the figures, blacks were at 120% (15.11/12.60); Asian or Pacific Islanders at 83% (4.05/4.9); American Indian/Alaskan Natives at 77% (0.7/0.9); and people of two or more races at 62% (1.8/2.9).

What can explain the striking and growing underrepresentation of Hispanic authors? The relatively young age of the Hispanic population can explain a small part of the difference. In 2000, Hispanics constituted 12.5% of the total U.S. population and 11.2% of the U.S. population between ages 25 and 64.[52] Yet in that year, Hispanics still only produced 6.8% of copyright registrations. Thus, Hispanic registration rates stand at 54.4% or 58.6% of what one would expect them to be based on Hispanics' share of the relevant comparison group. These numbers are even lower for 2010, where the Hispanic registration rate is at 44.6% (relative to population share) and 49.8% (relative to the 25–64 cohort).

A somewhat larger portion of the difference may be explained by the considerable portion of Hispanics that are unauthorized immigrants, a group that is counted in the Census.[53] Of the 50.5 million American Hispanics in 2010,[54] approximately 8 million were unauthorized immigrants,[55] comprising the majority group of all unauthorized immigrants, estimated at about 11 million in total.[56] It seems quite likely that unauthorized immigrants produce copyright registrations at a rate far less than the general population; even if they are producing works of authorship, most would likely be uncomfortable with submitting a registration application to the federal government on which they must state, among other things, their citizenship and home address. If about 16% of Hispanics living in the United States are unau-

52 *See* U.S. Census Bureau, *supra* note 45.

53 *See Congressional Apportionment: Frequently Asked Questions*, U.S. Census Bureau, https://www.census.gov/population/apportionment/about/faq.html#Q16 [https://perma.cc/5E4P-WHWE] (last updated Feb. 4, 2013) (noting that undocumented alien residents are included in the Census); *Foreign Born*, U.S. Census Bureau, http://www.census.gov/topics/population/foreign-born/about.html [https://perma.cc/UN75-QW7S] (last updated July 6, 2016) (noting that "unauthorized migrants are implicitly included in Census Bureau estimates").

54 *See* U.S. Census Bureau, *supra* note 51, at 2.

55 *See* Michael Hoefer, Nancy Rytina & Bryan C. Baker, U.S. Dep't of Homeland Sec., Estimates of the Unauthorized Immigrant Population Residing in the United States: January 2009, at 4 tbl.3 (2010), https://www.dhs.gov/xlibrary/assets/statistics/publications/ois_ill_pe_2009.pdf (estimating that about 8,150,000 unauthorized immigrants in the United States originated from the countries of Mexico, El Salvador, Guatemala, Honduras, and Ecuador).

56 *See id.* (estimating that about 10.8 million unauthorized immigrants were living in the United States in January 2009).

thorized immigrants—calculated as 8/50.5 million—and if they submitted no copyright registrations at all, that alone could reduce Hispanic author representation from 100% to 84%; but there is still a long way from 84% to a rate under 50% as calculated above.

3. Overrepresentation of Black Authors

The black population of the United States has remained relatively stable as a percentage of the total population, rising from 11.7% in 1980 to 12.6% in 2010.[57] Black authors have also contributed a relatively stable, slightly rising percentage of copyright registrations, from 14.22% in 1980 to 15.11% in 2010. Thus, black authors have been steadily overrepresented in copyright registrations—from 122% (14.22/11.7) in 1980 to 122% (14.73/12.1) in 1990, 118% (14.5/12.3) in 2000, and 120% (15.11/12.6) in 2010.

4. Authors of Different Races Tend to Create Different Works

Members of different races and ethnicities differ substantially in the types of works they tend to create. The strongest areas of registration by white authors have been dramatic works and software, while their weakest areas have been arts and music. Black authors have been the strongest in music and drama and weakest in software and art. Hispanics have been strongest in music and movies and weakest in software and text. Lastly, Asians and Pacific Islanders have been strongest in art and software, and weakest in music and drama. Table 2 presents registration patterns across race and creative areas:

57 *See* U.S. CENSUS BUREAU, A LOOK AT THE 1940 CENSUS 9, https://www.census.gov/newsroom/cspan/1940census/CSPAN_1940slides.pdf [https://perma.cc/23LJ-ESAB] (last visited Jan. 2, 2018).

TABLE 2. PERCENT OF REGISTRATIONS BY RACE AND WORK TYPES[58]

	Text	Music	Drama	Art	Movies	Software	All
White	77.77(3)	74.56(6)	77.82(2)	76.68(5)	76.96(4)	78.52(1)	76.21
Black	13.57(3)	16.07(1)	13.97(2)	12.57(5)	12.81(4)	12.06(6)	14.61
Hispanic	4.65(5)	7.42(1)	5.76(3)	5.65(4)	6.55(2)	4.46(6)	6.09
Asian / Pacific Islander	4.27(3)	1.86(6)	2.76(5)	5.63(1)	4.20(4)	5.54(2)	3.25
Native Am. / Alaskan	0.69	0.73	0.69	0.69	0.70	0.65	0.71
Two or more races	1.71	1.67	1.68	1.69	1.78	1.72	1.69

The strengths and weaknesses of white and Asian authors overlap somewhat: both are strong in software and weakest in music. Black and Hispanic authors' strengths and weaknesses also substantially overlap—both are strongest in music and weakest in software. And, as these similarities suggest, the relative strengths and weakness of the white/Asian group on the one hand, and the black/Hispanic group on the other, seem to be substantially opposite.

5. *Per-Capita Production of Copyright Registrations and the Extraordinary Representation of Jewish Authors*

Including this topic in this Part of the Article is not free from difficulty. First, there is no universally recognized definition of what Judaism is and who is Jewish, and we have no intention of providing or adopting one. Judaism is not a race—those who self-identify as Jewish can also self-identify as white, black, Asian, or Hispanic; the U.S. Census does not consider Judaism to be a race; and most Jews would cringe at the suggestion.[59] Although some view Judaism as a religion, it cannot be reduced to it, as many secular Jews do not follow religious practices yet have a strong Jewish identity. The topic is included here because Judaism likely has an ethnic element to it, though it has other components as well, including religion and culture.[60] Sec-

58 Superscripts designate a work type's rank per given race or ethnicity. Rankings were not added in the last two rows as there is little variation across work types.

59 While the Supreme Court has in one case considered Judaism to be a race, it emphasized that it was doing so only in the context of affording the protection of antidiscrimination laws to a congregation whose synagogue was painted with anti-Semitic slogans and symbols. *See* Shaare Tefila Congregation v. Cobb, 481 U.S. 615 (1987).

60 *See generally* RELIGION OR ETHNICITY? JEWISH IDENTITIES IN EVOLUTION (Zvi Gitelman ed., 2009) (exploring ethnic, religious, national, and cultural aspects of Judaism, among

ond, and relatedly, whatever the exact definition of Judaism might be, there is nothing in the data we have that tells us what percentage of those bearing particular last names self-identify as Jewish.

With these qualifications in mind, it appears that last names that reputable Jewish sources identify as being borne by many who self-identify as Jewish tend to be highly represented among top last names in terms of copyright registrations per-capita.[61] For the purposes of this Section, we started out with the 5003 most populous of the 151,671 last names in the 2000 U.S. Census data.[62] Using Copyright Office data, we calculated the number of copyright registrations under each last name, both in general and per work type. Cross-referencing these two data sources makes it possible to generate the following per-capita statistics:

---

others); ROBERTA ROSENTHAL KWALL, THE MYTH OF THE CULTURAL JEW (2015) (exploring various aspects of Jewish identity while emphasizing the importance of Jewish law and culture).

61 For these sources, see *infra* note 66.

62 We set out to limit our inquiry to the most populous 5000 last names, but ended up analyzing 5003 because four last names were tied for the 5000th place, having an equal number of bearers, 6435. This number thus marks the lower bound of bearers for the purposes of a last name's inclusion in Table 3's per-capita registration statistics.

TABLE 3. TOP AUTHOR SURNAMES FOR PRODUCTION OF COPYRIGHT REGISTRATIONS PER CAPITA

	All	Music	Text	Art	Drama	Movies	Software
1	Brent	Bach	Hubbard	Loomis	Segal	Correll	Gass
2	Loomis	Diamond	Lerner	Brent	Dickens	Dobson	Hubert
3	Lerner	Shapiro	Epstein	Ahn	Frankel	Schiller	Doucette
4	Segal	Baptiste	Greenberg	Boynton	Freedman	Palacio	Furr
5	Bach	Wayne	Siegel	Wimberly	Brody	Loza	Szabo
6	Gottlieb	Muhammad	Freedman	Su	Wilde	To	Kiefer
7	Shapiro	Seals	Eisenberg	Hummel	Cohen	Cosby	Adler
8	Levin	Gold	Adler	Ennis	Gottlieb	Mackenzie	Booher
9	Greenberg	Gaither	Bernstein	Hillman	Levin	Mancuso	Peterman
10	Steinberg	Segal	Fishman	Pan	Eisenberg	Tomlin	Alford
11	Weinberg	Berlin	Gottlieb	Chang	DeStefano	Fuchs	Wyman
12	Ahn	Holiday	Frankel	Moss	Goldman	Burrows	Christman
13	Bernstein	Kaye	Levin	Rockwell	Weinstein	Jerome	Forsythe
14	Eisenberg	Macleod	Segal	Clough	Bernstein	Lyman	Tsai
15	Epstein	Steinberg	Horowitz	Rigsby	Calabrese	Landon	Kirsch
16	Levine	Silver	Lieberman	Lung	Israel	Shapiro	Baer
17	Freedman	Bernstein	Kaye	Healy	Martins	Frankel	Simpkins
18	Adler	Lerner	Levine	Weinberg	Kaplan	Bloom	Freedman
19	Kaye	Pinson	Brody	Keane	Shapiro	Grossman	Rao
20	Siegel	Wainwright	Shapiro	Chiu	Katz	Hartwell	Hutchings
21	Gold	Levine	Weinberg	Giordano	Kahn	Levi	Feldman
22	Diamond	Conte	Silverman	Haskell	Epstein	Jankowski	Gottlieb
23	Frankel	Paxton	Rosen	Pak	Friedman	Foley	Doty
24	Brody	Mandel	Kahn	Burch	Stern	Ackerman	Mandel
25	Horowitz	Richman	Kaplan	Rinehart	Goldberg	Rubino	Tennant

It is worth noting that the content of Table 3 is affected by the fact that it is limited to the most populous 5003 last names. That limitation was intended to guard against two concerns. On the one hand, the more last names one includes, the greater the possibility that high-ranked last names will represent prolific individual authors with a relatively uncommon last name. We have not entirely avoided that issue in Table 3. Its top last name for textual work registrations per capita is Hubbard. There are 13,837 registrations under that last name in our database, of which at least 10,341 are attributed to a single person, L. Ron Hubbard, the prolific writer and founder of Scientology.[63] However, if we considered more last names, the table would include more names representing prolific individual authors. For example, if we

63 The number of registrations attributed to L. Ron Hubbard above likely underestimates the true one. Other registrations likely belong to him—as they may have, for example, in addition to his last name, his birth and death year, or his first name in full, Lafayette, but do not contain L. Ron or Lafayette Ron and thus were not counted.

considered the most populous 10,000 last names, the last name Disney would occupy the top spot both overall and in the art category. On the other hand, the fewer last names one includes in per-capita statistics, the greater the danger of losing valuable information and excluding from the analysis cross-sections that are not very populous. While the 5003 cutoff represents our judgment regarding a reasonable balance between these opposing concerns, varying it would vary the contents of Table 3.

How can one tell whether Jewish authors are overrepresented in Table 3? As there is no clear definition of who is Jewish, estimates of the Jewish population in the United States vary. For present purposes, that rate is likely not greater than 3.3% (and probably lower), comprising those who self-identify as Jewish (about 2.2%) and those with Jewish ancestry (an additional 1.1%).[64] If all members of society produced copyright registration at equal rates, one would expect the rate of last names that reputable Jewish sources identify as being borne by many who self-identify as Jewish in Table 3 (or in any randomly selected list of last names) to be no more than about 1 in 33.[65]

Table 3 suggests that Jewish authors produce copyright registrations at a rate that greatly exceeds their proportion in the population. The first column notes the last names with the highest rate of per-capita registrations for all works. It seems that at least nineteen of the twenty-five last names in that column, or over three-quarters of the last names, represent copyright registrations by authors bearing last names that Jewish sources self-identify as borne by many Jewish people.[66] Even if only half of the bearers of these last names self-describe as Jewish, that would still amount to a substantial overrepresentation.

---

64 *See, e.g.*, *A Portrait of Jewish Americans*, PEW. RES. CTR. (Oct. 1, 2013), http://www.pewforum.org/2013/10/01/jewish-american-beliefs-attitudes-culture-survey/ [https://perma.cc/UG9P-UAEQ] (finding that 2.2% of the U.S. population self-describe as Jewish, and that an additional 1.1% have a Jewish parent or were raised Jewish).

65 One caveat is that the proportion of the Jewish population among members of the most populous 5003 last names might be greater (or smaller) than that in the general population. As the overall rate of what might be considered Jewish last names in the first column of Table 3 is likely greater than 75%, for this to be a valid alternative explanation one would need that rate to similarly hold among the most populous 5003 last names, which is clearly not the case based on casual observation. Further, based on five 25-surname random samples of the most populous 5003 last names, comprising of those ranked 976–1000, 1976–2000, 2976–3000, 3976–4000, and 4976–5000, the rate appears to be less than about 6.5%, using an inclusive criterion.

66 *See, e.g.*, BENZION C. KAGANOFF, A DICTIONARY OF JEWISH LAST NAMES AND THEIR HISTORY (2d ed. 1996) (tracing the origin of about 4000 common Jewish last names); *The Memi De-Shalit Database of Jewish Family Names*, BEIT HATFUTSOT https://www.bh.org.il/databases/family-names/jewish-family-names-introduction/ (last visited Feb. 26, 2018) (containing an open, searchable database of Jewish last names).

Similar to authors belonging to other cross-sections of the population, the relative productivity of Jewish authors seems to vary across work types. Whereas the text column in Table 3 is almost entirely populated by last names that Jewish sources identify as borne by many with Jewish identity, there is a paucity of such last names in the art column.[67]

C. Methodology Revisited: Selection Bias in Assigning Probabilities

One might be concerned that our initial method above—which assigned authors probable races and ethnicities according to the racial and ethnic makeup of their last names in the general population—suffers from selection bias. To illustrate, assume that the last name "Williams" is shared equally by whites and blacks, but that blacks are registering copyrighted works at a rate double than whites. If so, we should be assigning two-thirds of the Williams registrations to blacks rather than only one-half. More generally, if people of different races and ethnicities have different per-capita propensities to register copyrighted works, our initial method above could mischaracterize racial and ethnic registration patterns.

How can one address such potential selection bias? Fortunately, we have registration counts not only for one last name, but for many. We further know the population racial makeup of each last name. These data make it possible to get a more accurate sense of racial and ethnic registration propensities. To illustrate, assume a stylized population with two races—black and white—and two last names—Smith and Williams. As we noted, about three-quarters of those bearing the surname Smith in the United States are white, and one-quarter are black; by contrast, about half of those bearing the surname Williams in the United States are white, and half are black.[68] As of the year 2000, there were about 2.3 million people in the United States bearing the surname Smith, and 1.5 million people bearing the surname Williams[69]—about one-and-a-half people named Williams for each person named Smith. Thus, all other things being equal, if white and black

---

67 Of course, because registrations for textual and musical works are by far the most numerous, *see supra* Table 1, one would expect that many last names that dominated in one of those categories would also appear in the list of top names for all works, and that is indeed the case: twenty-two of the twenty-five last names in the "all works" list also appear in either the textual works list or the musical works list, and five of them appear in both lists. The other three names in the "all works" list are the three top names in the list for art works, the third largest category of registrations.

68 *See* U.S. CENSUS BUREAU, *supra* note 33; *supra* text accompanying note 39.

69 *See* U.S. CENSUS BUREAU, *supra* note 33.

people were registering works at the same rate, we would expect to see about 1.5 copyright registrations by people named Smith for every one registration by a person named Williams. If, on the other hand, black people were registering works at a rate twice that of white people, we would expect to see only about 1.28 copyright registrations by people named Smith for every one registration by a person named Williams. It becomes apparent that just as any individual black-to-white registration ratio would entail a particular expected Smith-to-Williams registration ratio in the data, so can one estimate from any observed Smith-to-Williams registration ratio in the data the average black-to-white individual registration ratio that would fit the data best. If all people of the same race were exactly the same in terms of their production of registrations, working with just a few last names might very well be all that one would need in order to get correct estimates.

Reality, however, does not operate according to mathematical precision. As individuals vary in their propensity to register copyrighted works, the number of registrations under any family name has a random element to it. To find out which statistical registration tendencies of people of different races and ethnicities fit the registration data best, and to assess the strength of that fit, we used multiple regression analysis. Our data consist of registration counts for each of the 151,671 last names that appear in the 2000 U.S. Census data. For each last name, Census data contain the number of people bearing it in the population and its racial makeup. For some less popular last names, the Census data does not contain figures for one or more racial or ethnic categories. We have coded these cases as zero percent, and chose to use in our analysis only last names for which we have the racial makeup of more than 95% of the bearers. We dropped the 25,562 last names that did not meet this 95% threshold, leaving us with 126,109 last names on which we conducted our regression analysis. In this dataset, whites comprised 69.46% of the people for which we had data, blacks 12.27%, Hispanics 12.69%, and Asian & Pacific Islanders 3.38%. In the analysis, we weighted our observations (the various last names) according to the number of people they represent and used robust standard errors.

In all models below, we use as independent variables the number of people bearing a particular last name that are (1) black (Nblack), (2) white (Nwhite), (3) Hispanic (Nhispanic), and (4) Asian or Pacific Islander (Napi). In our first model, we use as our dependent variable the overall number of registrations under a particular last name. Our

remaining six models attempt to determine racial registration patterns of specific work types. In these models, numbered (2)–(7) below, the dependent variables are the number of registrations in each last name of music, text, art, drama, movies, and software. The results are:

TABLE 4. REGRESSION ANALYSIS OF COPYRIGHT REGISTRATIONS 1978–2012 BY RACE AND ETHNICITY

	(1) All	(2) Music	(3) Text	(4) Art	(5) Drama	(6) Movies	(7) Software
Nblack	0.0558***	0.0384***	0.00943***	0.00249**	0.00155***	0.000350*	0.000135***
	(0.00381)	(0.00266)	(0.00124)	(0.00105)	(0.000257)	(0.000200)	(4.84e-05)
Nwhite	0.0509***	0.0233***	0.0185***	0.00430***	0.00202***	0.000574***	0.000467***
	(0.00153)	(0.000599)	(0.000784)	(0.000588)	(0.000145)	(0.000112)	(1.73e-05)
Nhispanic	0.0124***	0.00849***	0.00175***	0.000732***	0.000398***	0.000173***	3.94e-05***
	(0.000663)	(0.000407)	(0.000134)	(6.92e-05)	(3.32e-05)	(3.46e-05)	(6.55e-06)
Napi	0.0356**	0.00836**	0.0158**	0.00645***	0.00142**	0.000789***	0.000449***
	(0.0139)	(0.00356)	(0.00653)	(0.00240)	(0.000648)	(8.47e-05)	(8.52e-05)
Constant	407.2***	157.5***	149.4***	44.58**	24.35***	8.528**	2.329***
	(65.72)	(32.60)	(28.14)	(18.11)	(5.056)	(3.422)	(0.663)
Obs.	126,109	126,109	126,109	126,109	126,109	126,109	126,109
R-squared	0.995	0.994	0.993	0.965	0.984	0.910	0.980

Robust standard errors in parentheses
***** p<0.01, ** p<0.05, * p<0.1**

The coefficients in the table designate the number of registrations one extra person of a particular race would contribute, on average, over our study's thirty-five-year span. As the first coefficient in the first model suggests, if the black population were greater by one person throughout the study's thirty-five-year span, we would expect the number of registrations to increase 0.0558 registrations on average (or, equivalently, if it were greater by eighteen additional individuals, we would expect to see one more registration). The coefficient on the number of blacks is greater than the one on the number of whites, which suggests that an additional black person would be expected to register more works than an additional white person. This difference, however, is not statistically significant. In other words, there is no compelling reason to think that blacks and whites markedly differ in their overall tendency to register copyrighted works. The following table notes whether the differences between the races are statistically significant.

TABLE 5. EXAMINING WHETHER THE DIFFERENCES IN AVERAGE REGISTRATION RATES BETWEEN PEOPLE OF DIFFERENT RACES AND ETHNICITIES ARE STATISTICALLY SIGNIFICANT

Model	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Registration Type	All	Music	Text	Art	Drama	Movies	Software
Differences							
Nblack - Nwhite	0.00490	0.0151***	-0.00908***	-0.00181	-0.000475	-0.000223	-0.000332***
	(0.976)	(4.814)	(-4.626)	(-1.137)	(-1.223)	(-0.737)	(-5.344)
Nblack - Nhispanic	0.0434***	0.0299***	0.00768***	0.00176*	0.00115***	0.000178	9.61e-05**
	(11.36)	(11.36)	(6.195)	(1.689)	(4.476)	(0.883)	(1.999)
Nblack - Napi	0.0201	0.0301***	-0.00640	-0.00396	0.000129	-0.000439**	-0.000314***
	(1.383)	(6.570)	(-0.959)	(-1.505)	(0.185)	(-2)	(-3.139)
Nwhite - Nhispanic	0.0385***	0.0148***	0.0168***	0.00357***	0.00163***	0.000401***	0.000428***
	(22.99)	(19.70)	(21.22)	(6.026)	(10.91)	(3.408)	(22.67)
Nwhite - Napi	0.0152	0.0150***	0.00268	-0.00214	0.000604	-0.000215	1.82e-05
	(1.082)	(4.151)	(0.406)	(-0.864)	(0.905)	(-1.492)	(0.208)
Nhispanic - Napi	-0.0233*	0.000131	-0.0141**	-0.00571**	-0.00102	-0.000616***	-0.000410***
	(-1.653)	(0.0363)	(-2.136)	(-2.362)	(-1.560)	(-6.591)	(-4.751)

t-statistics in parentheses
***** p<0.01, ** p<0.05, * p<0.1**

Table 5 notes the differences between the average tendencies of people of any two races or ethnicities to register copyrighted works in general and in specific areas of creativity, and the statistical significance of these differences. The first line does so for blacks and whites. For example, the first value in the first line, 0.00490, is simply the difference between the Nblack coefficient in model (1) in Table 4 above, 0.0558, and that of Nwhite there, 0.0509. That this difference, 0.00490, is positive reflects the regression result that on average, an additional black person in the population is expected to register more copyrighted works than an additional white person. That said, this result is not statistically significant, and therefore one should not have too much confidence that members of these two races truly differ in their overall registration rates.

As column (2) suggests, black individuals register music at significantly higher rates than that of members of any other race. Column (3) suggests that whites register textual works at a rate significantly higher than that of either blacks or Hispanics. They also register text

at a rate that is insignificantly higher than that of Asians and Pacific Islanders. Column (4) suggests that Asians and Pacific Islanders tend to register art at a higher rate than that of other races, although their advantage over whites and blacks is not statistically significant. Column (5) suggests that whites tend to register drama at a higher rate than that of members of other races, although their advantage over blacks and Asians and Pacific Islanders is not statistically significant. Column (6) suggests that Asians and Pacific Islanders tend to register movies at significantly higher rates than those of blacks and Hispanics, and at a substantially higher, though not statistically significant, rate compared to whites. Lastly, column (7) suggests that whites and Asians and Pacific Islanders tend to register software at significantly higher rates than that of blacks and Hispanics.

The regression analysis suggests that, as compared to whites, blacks tend to register more per capita (though the difference is not statistically significant, as suggested above) and Hispanics register at significantly lower rates across the board (as compared to blacks, whites, and Asians and Pacific Islanders). These results are qualitatively in line with the initial assignment above of race and ethnicity to authors according to their last name's distribution of race and ethnicity in the general population. Using that simpler method, we had earlier calculated that whites' percentage of registration compared to their portion of the U.S. population rose gradually from about 100% in 1980 to 116% in 2010,[70] and that blacks' percentages have hovered around 120% throughout the period.[71] We also saw that Hispanics' percentages have gradually decreased from 69.5% in 1980 to 44.6% by 2010.[72] Our regression analysis suggests that blacks' average registration rate advantage over whites is not statistically significant, but that their advantage relative to Hispanics is.

The percentages we reported at the outset, using population averages of last names, seem reconcilable with the regression result as they each point in the same direction. This is plausibly the case for a simple reason: if blacks are more productive at registering copyrighted works, then last names that are predominantly black should appear in our dataset of registered works more frequently than they are found in the general population. Conversely, if Hispanics infrequently register copyrighted works, then we should encounter last names that are predominantly Hispanic less frequently in our dataset of registered

70 *See supra* notes 42–45 and accompanying text.

71 *See supra* Section II.B.3.

72 *See supra* Section II.B.2.

works than we do in the general population. Assigning to each last name in our dataset its population distribution of racial and ethnic origin should therefore tend to rank in order correctly different races and ethnicities' average propensities to register.

Though our initial method tends to point us in the right direction, it still gives an inaccurate measure of the magnitude of the differences in average registration rates among races and ethnicities. Indeed, for example, the initial finding that blacks tend to register more works per capita than whites is in tension with the initial assumption that the racial makeup of authors' last names is the same as the one of those last names in the general population. The regression analysis gives a sense of the magnitude of the actual difference. Qualitatively, however, the two methods of analysis portray a similar picture of different racial and ethnic groups' average relative propensities to register copyrighted works.[73]

III. GENDER

A. Methodology: Inferring Gender from First Names

Registration records do not specify authors' gender.[74] They do, however, contain the authors' first names. In conducting gender statistics we rely on information elicited from the 1990 U.S. Census regarding the gender distribution of first names.[75] Accordingly, for each individual author in our dataset, we have calculated the probability that a person with that first name is male. When, for expositional clarity, we make statements below as to the gender makeup of a certain

---

73 We have limited our examination to white, black, and Hispanic authors as these are the three largest races and ethnicities in the United States, accounting for over 90% of the population. *See* U.S. CENSUS BUREAU, ACS DEMOGRAPHIC AND HOUSING ESTIMATES: 2011–2015 AMERICAN COMMUNITY SURVEY 5-YEAR ESTIMATES, https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_15_5YR_DP05&src=PT [https://perma.cc/Y4A7-V4RU] (last visited Jan. 2, 2018). Further, adding the other racial categories into our regression and mean-square difference analyses would introduce collinearity problems and involve basing statistical inference on what are often small sample sizes. As for our multiple regression analysis, adding them would not substantially alter the R-squared statistic.

74 *See, e.g.*, U.S. COPYRIGHT OFFICE, FORM TX (2012), http://www.copyright.gov/forms/formtx.pdf [https://perma.cc/5L8M-T2QH] (last visited Jan. 2, 2018) (not requiring authors registering textual works to note their gender).

75 We used first-name gender distribution and frequency data drawn from the 1990 U.S. Census. The data, containing gender distributions for 5164 first names, is available in part on a U.S. Census webpage. *See* U.S. CENSUS BUREAU, FREQUENTLY OCCURRING SURNAMES FROM CENSUS 1990 – NAMES FILES, https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html (last updated Sept. 2, 2014) (follow hyperlinks to "dist.female.first" and "dist.male.first") (containing files of male and female first names and their distributions).

cross-section of authors, we simply refer to the average of the probabilistic gender variable in that category.[76] Our dataset contained 10,465,488 registrations that reported at least one individual author. Of those, 982,234 registrations contained a first name that did not match any entry in the U.S. Census list of first names. Those were excluded, leaving 9,483,254 registrations on which we base our gender statistics.

First names are generally much more closely correlated with particular genders than last names are with particular races and ethnicities.[77] Eighty-two percent of the registrations in our dataset that have gender probabilities associated with them have probabilities of either 99% or higher male, or 99% or higher female. We will be using that 99%-minimum identified gender subset for a number of purposes below, where we make the assumption that creativity patterns of male and female authors in these categories are representative of authors as a whole.

B. Main Findings

1. Authors Are Two-Thirds Male

The most striking statistic about authors' gender is that two-thirds of the authors in our study are male.[78] At the same time, the data show a statistically significant time trend of increased female representation among authors of registered works.[79] While the rate of male authorship was about 70% in 1978, it steadily dropped to about 64% in 2012. Conversely, while the rate of female authorship was about 30% in 1978, it rose to about 36% in 2012. The female rate of participation in authorship has been lower than women's share in the labor force, which stood at 41.7% in 1978, and rose to 46.9% in 2012.[80]

What could explain persistent, though decreasing, overrepresentation of males? Any simplistic biological explanation would be,

---

76 For the purposes of conducting gender statistics we have excluded registrations that have no individual authors, such as registrations of works created by corporations, as these have no gender. A registration's gender has been calculated as the average gender of its individual authors. Gender statistics, such as for a category of works or for a year, have been calculated as the averages of the relevant registrations' probabilistic genders.

77 Shervin Malmasi & Mark Dras, *A Data-Driven Approach to Studying Given Names and Their Gender and Ethnicity Associations*, *in* PROCEEDINGS OF AUSTRALASIAN LANGUAGE TECHNOLOGY ASSOCIATION WORKSHOP 145, 146 (2014).

78 Authors are 66.63% male (averaging out the average gender profile per registration).

79 Regressing the general male authorship rate on time yields a -0.002 coefficient that is significant even at the 0.1% level.

80 *See* U.S. DEP'T OF LABOR, WOMEN IN THE LABOR FORCE, https://www.dol.gov/wb/stats/NEWSTATS/facts/women_lf.htm#one [https://perma.cc/LB5K-R9E3] (last visited Jan. 2, 2018).

among other things, difficult to square with the change over the thirty-five-year period, because biology could presumably not have changed that quickly. Sociological explanations may fit better with the fact that different types of works exhibit strikingly different gender-of-author splits and trends: different industries may be more or less male-dominated, and that domination may have changed more or less over time.[81] There may be other social and individual barriers to create in or enter different industries, which may have morphed differently over time.

It should be noted that about 28% of registrations have only corporate authors for copyright purposes, and we do not know the gender of the people who actually created those works. If women are more likely than men to be employees of, or work-for-hire contractors for, companies that register works under corporate authorship, it could explain at least some of our male-dominant findings. If creative women have become less inclined over time to create works as employees, it could explain at least some of the decline in male domination of individual author registrations. The findings would also be consistent with a conjecture that for some reason, women register the works they have created less often than men (but have become better at registering over time comparatively). As it is beyond the scope of this Article to find the causes for its empirical findings, we leave such exploration to future work.

2. Authors Prefer Same-Gendered Co-Authors

We looked at the gender of co-authors in registrations that included two or more individual co-authors that had first names that each appeared in the 1990 Census table (and thus had gender probabilities).[82] These criteria brought our data to 2,035,683 registrations. For expositional purposes, we present our data as those of Author 1 and Author 2. Author 1 is the first listed author on the registrations that meet the aforementioned criteria, and Author 2 is the second listed. Their gender probabilities are known according to their first names. For registrations with more than two authors we calculated the average gender probabilities of all authors except the first and treated that as the probable gender of Author 2. The probability

81 *See infra* Section III.B.3 (noting that the percentage of male authorship is substantially higher for some work types than for others).

82 *See* U.S. CENSUS BUREAU, *supra* note 42.

of Author 1 being male is slightly lower than that of Author 2—a difference in means of 0.00071 that is statistically significant.[83]

For the purposes of this subsection, we further classified as "male" any author who bore a name that had at least a 90% probability of use by a male, and as "female" any author who bore a name that had at least a 90% probability of use by a female. Dropping out names with intermediate probabilities, we were left with 1,708,442 observations with individual co-authors. As a result, 70.43% of our Authors 1 and 71.68% of our Authors 2 were male, and 29.57% of our Authors 1 and 28.32% of our Authors 2 were female.

Of the registrations where Author 1 was male, 80.79% of Authors 2 were male as well; where Author 2 was male, 79.39% of Authors 1 were male. Of the registrations where Author 1 was female, 50.02% of Authors 2 were female as well; where Author 2 was female, 52.23% of Authors 1 were female. In this more restricted sample of 1,708,442 observations, about 29% of authors are classified as females (29.57% of Author 1, 28.32% of Author 2) and 71% as males (70.43% of Author 1, 71.68% of Author 2). A random assignment of co-authors would result in about 71% of the males ending up with co-authors who were also male, while about 29% of the females would end up with co-authors who were also female. This suggests that both men and women likely have a significant preference for co-authors of their own gender. When we compared, on the one hand, the probable gender of Author 2 given that Author 1 is male to the probable gender of Author 2 given that Author 1 is female, the difference in means—negative 0.31—came out as statistically significant.[84] Males and females thus show a significant preference to co-author with similarly gendered co-authors.

3. Men and Women Register Different Types of Works

The summary figures concerning the gender of all authors lumped together mask gender variation across the different work types: some fields are more male-dominated than others. The work types sorted from the least to most male-dominated are art (54.34%

---

83 A t-test for the comparison of means came out with a t-statistic of -1.9574 that is associated with a two-tailed p-value of 0.05. The alternative hypothesis that Author 1 is more male than Author 2 has a p-value of 0.975 (so should be rejected) and the alternative that Author 1 is more female than Author 2 has a p-value of 0.025 (so should be preferred over the null).

84 The t-statistic came out as -430, with a p-value of (virtually) zero. The result was unchanged when we compared the probable gender of Author 1 given than Author 2 was male to the probable gender of Author 1 given that Author 2 was female.

male), text (57.45%), drama (69.99%), music (75.98%), movies (78.16%), and software (88.22%).

4. Gender Trends over Time Vary Across Types of Works

The degree to which the gender gap has or has not been bridged similarly varies by type of work. The upward trend of female authorship is driven mainly by the text category, in which the rate of female authorship has increased during our study period by 11.85%, from 33.98% to 45.83%. This category accounts for over a third of individual-author registrations.[85]

There has been an increase in the percentage of female authors with respect to movies (10.49%) and software (11.85%) as well, but these together account for only about 2.5% of all individual registrations.[86] At the same time, the music and drama categories, which account for about 44% and 5% of individual registrations, respectively, show a statistically flat time trend respecting female authorship.[87] Finally, the art category, which accounts for 11% of individual registrations, has a check-mark-shaped time trend with the percentage of male authorship generally decreasing from 1978 to 1984 and then generally increasing to 2012. While the 1978 (59.8%) and 2012 (59.1%) percentages of male art authorship are not markedly different, the trend is one of statistically significant increase in male authorship.

5. Age and Published Status by Gender: An Intricate Story

Men and women differ in the publication status of their registered works. Here we limit our inquiry to authors whose first name has a probability of 99%-minimum male or 99%-minimum female. For the study as a whole, 39% of works registered by men were published compared to 44% of works by women. If we considered those summary figures alone, we might speculate that women who register works tend to be, on average, more market savvy than the men who do as to the projects they invest in, or perhaps more risk averse.

However, the summary figures are influenced heavily by the differing types of works that men and women are likely to produce. Take, for example, the two largest categories of works: text and music. In both categories, male authors are more likely than female authors

---

85 The increased percentage of female authors of textual works over time is statistically significant at the 0.1% level.

86 These increases are significant at the 5% and 0.01% levels, respectively.

87 Music shows a positive and insignificant time trend of male authorship. Drama shows a negative and insignificant time trend of male authorship.

to register published works. Sixty-six percent of registrations for textual works by male authors were of published works, compared to 61% for female authors; the corresponding figures for musical works are 22% for male authors and 15% for female authors, and yet if we combined the categories of text and music, the percentage of registrations for published works by males would be 40%, and for females would be 45%. How is that possible? Fifty percent of all registrations by female authors are for text, whereas only 33% of all registrations by males are; conversely, 50% of all registrations by male authors are for music, whereas only 29% of registrations by female authors are. Because registrations by female authors are more likely to be for a type of work that is more often published at the time of registration, whereas registrations by male authors are more likely to be for a type of work that is less likely to be published, overall a smaller percentage of registrations by male authors are for published works. To round out the types of works, the percentage of published works by males (females) in movies is 73% (63%) and in drama is 7% (7%); in art and software greater percentages of registrations by female authors are for published works, those percentages being 36% (45%) in art, and 41% (44%) in software.

Further, keeping our inquiry to those with first names that are either 99%-minimum male or 99%-minimum female, we can also look at the different age profiles of registrants. Overall, the average male author is 39.39 years old, about two years younger than the average female author, who is 41.73. Male authors are on average younger than female authors in three of the six types of works: the average age of male (female) authors was 35.53 (37.89) in music, 42.39 (43.14) in movies and 39.63 (42.48) in software. Yet female authors are younger than males in the three remaining categories: the average age of male (female) authors was 46.84 (45.06) in text, 39.98 (38.11) in drama, and 42.66 (42.42) in art.

Once again, the overall figures are a little misleading, because they are influenced by the fact that the average ages of authors of textual works, whether male or female, are higher than the average ages of male or female authors of any other type of work, and registrations for textual works constitute a considerably larger proportion of all registrations by female authors than they do of all registrations by male authors. In other words, rather than coming to the conclusion that on average women have to be alive two years longer than men in order to create registered works, one could come to the conclusion that both women and men have to be alive longer to create textual

works than to create other works (men even longer than women), and that women specialize more in textual works than men do. However, note that there are also differences at the type-of-work level, and that female authors of music, in particular, are on average more than two years older than male authors of music.

IV. AGE

A. Methodology: Subtracting Birth Year from Year of Creation

Ascertaining the age of an author at the time a registered work is created is not as complicated as ascertaining the author's race or gender: just subtract the author's year of birth from the year of creation of the work. Although almost all registration records contain information about the year of creation of the registered work, many do not contain year of birth information. In addition, year of birth and year of creation are sometimes entered inaccurately, so that subtracting the first from the second may result in negative numbers or numbers that exceed 8000. We decided to filter the results and keep only values between zero and one hundred (not including these numbers). When a registration listed more than one author, we averaged the ages to obtain an average author age for that work. We ended up obtaining author age information for about 6.6 million registrations, or about 63% of the total.

In the set of registrations for which age-of-author information is available, the proportion of published works is substantially smaller than it is for all registrations in our study: 28% versus 54%. That may be because authors themselves are more likely to complete registration applications for unpublished works, and provide their year of birth because they know it, while many registration applications for published works are completed by employees of publishers, who do not immediately know the authors' years of birth, and simply leave the field blank. Whatever the reason for the difference in proportion of published works, it undoubtedly has some effect on the results. For example, because we know that authors of published works are on average older than authors of unpublished works,[88] the real average age of authors of all works in our study is almost certainly greater than the age we report below.

---

88 *See infra* Section IV.B.4.

B. Main Findings

1. Authors Are 40 on Average, Most Productive in Their Early 30s

The average author is just over 40—40.12 years old to be exact. Author productivity rises relatively quickly as authors advance into their 20s and 30s, and then declines more slowly, so the average age of 40 is above the years of peak production. For authors of all six types of works combined, the ten most productive years are those between the ages of 27 and 36. Production during those ten years accounts for 29.69% of all registrations; by comparison, if authors were equally populous and equally productive from 18 through 78, production over a ten-year period would account for 16.66% of registrations. On average, one-year age cohorts of authors each continue to produce at least one percent of all registrations through age 59; at age 60, authors drop below one percent, and at age 69, they drop below one half of one percent.

TABLE 6. RATIO OF PERCENTAGE OF COPYRIGHT REGISTRATIONS TO PERCENTAGE OF U.S. POPULATION BY AGE GROUP, 1980–2012

Under 5 years	0.00
5 to 9 years	0.01
10 to 14 years	0.04
15 to 19 years	0.37
20 to 24 years	1.16
25 to 29 years	1.79
30 to 34 years	1.96
35 to 39 years	1.85
40 to 44 years	1.67
45 to 49 years	1.49
50 to 54 years	1.32
55 to 59 years	1.15
60 to 64 years	0.92
65 to 74 years	0.68
75 to 84 years	0.42
85 and over	0.30

Of course, those figures do not adjust for the age distribution of the U.S. population as a whole. If we divide the percentage of registrations produced by authors of various age groups by the percentage that those age groups represent of the U.S. population as a whole, we can generate a productivity ratio. If that ratio is more than one, then

that age group is producing more registrations than its population would suggest; if it is less than one, then it is producing less.

Table 6 shows the productivity ratios for sixteen age groups, using cumulative figures for both registrations and U.S. population from 1980 to 2012.[89] The highest ratio is for the age group of 30 to 34, which is producing copyright registrations at a rate of 1.96 times their percentage of the overall population, and there is a slow but steady decline in each succeeding age group after 35. All groups from 20 to 59 are producing at a ratio of greater than one. From birth to age 14, authors are producing almost no registrations at all, which of course makes sense, and also might be seen as slightly inflating the ratios from 15 upwards.

Authors of published works are, on average, five-and-a-half years older than authors of unpublished works: 44.10 versus 38.59. Although we do not know exactly what explains that age difference, it is not surprising that, by the time an author's work is being publicly distributed, he or she would usually be older. Although the size of the age gap differs somewhat across types of works and across time, it remains true for all types of works and for all years in this study that authors of published works are on average older than authors of unpublished works. Still, the five-and-a-half-year gap cannot reasonably be fully explained by the time it takes to publish. About 80% of published works are published in the same year that they are created, and over 98% are published within two years. Some of the explanation for the difference is thus likely found in the greater experience, market savvy, and access to publishers of older authors.

2. *Authors of Different Work Types Up to Ten Years Apart in Age*

The age of authors varies substantially according to work type. Overall, the average age of authors of registered music between 1978 and 2012 is 36.08. By contrast, the average age of authors of literary works across that same time period is 46.25, over ten years older. The average ages of authors in the other four categories fall between the extremes of music and literature. Authors of registered computer programs are on average 39.98 years old—the next youngest after music, but close to the overall average, and not much different from authors of dramatic works, who are on average 40.35 years old. Authors of

89 The years 1978 and 1979 are not included because the Census Bureau used different age groups before 1980. *See* U.S. CENSUS BUREAU, 1970 CENSUS OF POPULATION PART 1, at 1-263 (1973) (aggregating upper limit data into "75 years and over").

works of visual art and of motion pictures are virtually exactly the same average age, at 42.75 and 42.76 years old respectively.

3. *Different Age Concentration of Authors of Different Work Types*

Creators of music are not only youngest on average; production of music is also the most age-concentrated. As Table 7 below shows, music creators are on average most productive from 24 to 33. Production by authors of those ages accounts for over one third of all music registrations—35.77%. By contrast, the most productive decade for authors of literary works—from 33 to 42—not only takes place nine years later in life, but also accounts for the production of only 26.36% of all literary work registrations, a little over a quarter. Above the age of 53, creators of music begin to produce less than one percent of all registrations per year of life, and they drop to below one half of one percent above age 61. By contrast, authors of literary works continue to produce at least one percent of all registrations through age 66, and they do not drop below one half of one percent until after the age of 76.

TABLE 7. REGISTRATIONS BY AGE CONCENTRATION AND TYPE OF WORK, 1978–2012

Work Type	Average Author Age	Most Productive Decade	Percentage of Registrations Produced in That Decade	Last Year Producing at Least 1% of Registrations	Last Year Producing at Least One Half of 1% of Registrations
All	40.12	27–36	29.69	59	68
Literary	46.25	33–42	26.36	66	76
Music	36.08	24–33	35.77	53	61
Art	42.75	36–45	30.17	61	67
Movies	42.76	36–45	29.51	59	66
Drama	40.35	27–36	31.73	58	68
Software	39.98	32–41	34.00	57	64

Although, as we noted above, the authors of software and drama have similar average ages—39.98 and 40.35, respectively—their age profiles are somewhat different. Software peaks substantially later and stronger: its peak decade is 32–41, five years later than the peak decade of 27–36 for drama, and that peak decade accounts for 34.00% of all software registrations, versus 31.73% for drama. Yet at the same time, production of software tails off somewhat earlier, with production dropping below one percent at 57—versus 58 for drama—and be-

low one half of one percent at 64—versus 68 for drama. By contrast, the age profiles of art and movies are quite similar across the board. Both have peak decades of 36–45, accounting for 30.17% of registrations in the case of art, and 29.51% in the case of movies. Production of art drops below one percent at 61 and one half of one percent at 67, while the corresponding ages for movies are 59 and 66.

Thus far, we have considered average age data for the entire thirty-five-year period from 1978 through 2012. However, there are substantial changes in the average ages of authors over that period. We have noted that the average age of all authors was 40.12. Yet authors have on average been getting older throughout that thirty-five-year period. The average age of authors of works registered in 1978 was 37.63 years; by 2012, that figure was 44.64, seven years older. Authors actually rose in average age slightly less than the U.S. population overall. In 1978, the median age of the U.S. population as a whole was 29.5; by 2012, it had risen to 37.3, 7.8 years older.[90]

4. Diminishing Age Increase Associated with Published Status

Although the increase in average age of authors parallels the increase in average age of Americans generally, the increase in age is much greater for unpublished works than for published works. In 1978, authors of unpublished works were on average 34.64 years old; thirty-five years later, in 2012, that average age had increased to 43.75, an increase of over nine years. By contrast, authors of published

90 Data on median age was gathered from a variety of Census Bureau publications, including: U.S. CENSUS BUREAU, STATISTICAL ABSTRACT OF THE UNITED STATES: 1985, at 26 tbl.27 (1984), https://www2.census.gov/library/publications/1984/compendia/statab/105ed/1985-02.pdf; U.S. CENSUS BUREAU, STATISTICAL ABSTRACT OF THE UNITED STATES: 1995, at 15 tbl.14 (for the median age from 1980 through 1994), https://www2.census.gov/library/publications/1995/compendia/statab/115ed/tables/pop.pdf; U.S. CENSUS BUREAU, STATISTICAL ABSTRACT OF THE UNITED STATES: 2004–2005, at 12 tbl.11 (2003), https://www2.census.gov/library/publications/2004/compendia/statab/124ed/tables/pop.pdf (for the median age from 1995 through 2000); U.S. CENSUS BUREAU, STATISTICAL ABSTRACT OF THE UNITED STATES: 2012, at 11 tbl.8 (2012), https://www2.census.gov/library/publications/2011/compendia/statab/131ed/tables/pop.pdf (for the median age from 2001 through 2009); U.S. CENSUS BUREAU, AGE AND SEX COMPOSITION IN THE UNITED STATES: 2010, https://www.census.gov/data/tables/2010/demo/age-and-sex/2010-age-sex-composition.html (last updated July 6, 2016) (select "Table 1. Population") (for the median age in 2010); U.S. CENSUS BUREAU, AGE AND SEX COMPOSITION IN THE UNITED STATES: 2011, https://www.census.gov/data/tables/2011/demo/age-and-sex/2011-age-sex-composition.html (last updated July 6, 2016) (select "Table 1. Population: 2011") (for the median age in 2011); U.S. CENSUS BUREAU, AGE AND SEX COMPOSITION IN THE UNITED STATES: 2012, https://www.census.gov/data/tables/2012/demo/age-and-sex/2012-age-sex-composition.html (last updated Apr. 20, 2015) (select "Table 1. Population by Age and Sex: 2012") (for the median age in 2012).

works registered in 1978 were on average 42.29 years old; by 2012, they were 47.46 years old, an increase of only about five years. Thus, the difference in average age between authors of unpublished works and authors of published works in 2012—3.71 years—is less than half of what it was in 1978—7.65 years.

5. Varying Average Age Growth of Authors of Different Work Types

There is a wide disparity among age increases of authors of different types of works. Authors of software, who were on average 35.14 years old in 1978, were 45.31 years old in 2012, an increase of 10.16 years. Authors of literary works, an average of 42.97 years old in 1978, were on average 51.20 years old in 2012, an increase of 8.23 years. At the other end, authors of movies, an average of 40.93 years old in 1978, were only 3.73 years older in 2012, at 44.67 years old; authors of art, 40.68 years old in 1978, were on average only 4.27 years older in 2012, at 44.95 years old; and authors of dramatic works increased in age by only 5.57 years, from 38.03 years old to 43.60 years old. At the extremes, the spread between the average age of authors of music—the youngest—and the average age of authors of literary works—the oldest—increased. Those average ages were 9.44 years apart in 1978, and the gap increased to 11.19 years in 2012.

6. *Authorship Has Become More Evenly Spread Across Age Groups*

TABLE 8. RATIO OF PERCENTAGE OF REGISTRATIONS TO PERCENTAGE OF U.S. POPULATION BY AGE, IN 1980, 1990, 2000, AND 2012

	1980	1990	2000	2012
5 to 9 years	0.00	0.00	0.01	0.01
10 to 14 years	0.03	0.03	0.04	0.10
15 to 19 years	0.31	0.40	0.38	0.52
20 to 24 years	1.24	1.27	1.14	1.04
25 to 29 years	2.47	1.80	1.63	1.41
30 to 34 years	2.35	1.94	1.67	1.49
35 to 39 years	1.89	2.09	1.61	1.45
40 to 44 years	1.53	1.72	1.61	1.43
45 to 49 years	1.33	1.42	1.76	1.47
50 to 54 years	1.13	1.15	1.61	1.45
55 to 59 years	0.92	0.93	1.30	1.49
60 to 64 years	0.73	0.73	0.98	1.50
65 to 74 years	0.57	0.55	0.75	1.06
75 to 84 years	0.45	0.34	0.48	0.59
85 years and over	0.39	0.29	0.28	0.37

We introduced above the productivity ratio for age groups of authors of registered works—the percentage of registrations produced by each age group divided by the percentage of the total U.S. population represented by that group. However, we only considered those ratios for the entire aggregated thirty-three-year period from 1980 through 2012. Those ratios have also changed over time, and in particular, they have substantially flattened out over adult age groups between 1980 and 2012.

Table 8 reveals that a few age cohorts seem to be extremely productive throughout their life. In 1980, authors of ages 25 to 29 were the most productive relative to their share of the population, and were producing registrations at a rate of 2.47 times that proportion. In 1990, these authors were ten years older, and the most productive age group was that of authors of ages 35 to 39, who were registering at 2.09 times their population share. A decade later, in 2000, 45- to 49-year-old authors were the most productive, registering at 1.76 times their population share. Finally, in 2012, twelve years later, the most productive group was 60- to 64-year-olds, but they were registering at only 1.50 times their proportion of the population, just barely edging out

younger age groups. Thus, on top of general age and time trends, there also seems to be a cohort creativity and registration effect.

Moreover, creativity has become less age-concentrated over time. Whereas in 1980, there is a creativity peak around the 25–29 cohort, by 2012, there is a high plateau of creativity: every age group between 25 and 64 was producing at a rate from 1.41 to 1.50 times their proportion of the population.

This flattening out of registration production over age groups is a major demographic shift, and deserves further study. Perhaps most optimistically, one might hypothesize that authors are now remaining more productive in their later years than they once were, and that creative production is spread out more evenly across the lifetime of authors. An alternative explanation, at least in part, might be that younger authors simply aren't using the registration system as much, so that a larger proportion of their creative production is not appearing in registration statistics. There is no question that part of the answer is that registrations of literary works, the authors of which have always been spread out more evenly by age, now account for a larger percentage of registrations than they once did, whereas registrations of music, the authors of which are on average younger and more concentrated by age, now account for a smaller percentage of registrations. However, even registrations of literary works, separated out from other registrations and adjusted for changes in age in the general U.S. population, have flattened out over age groups, with a later peak.

TABLE 9. RATIO OF PERCENTAGE OF LITERARY WORK REGISTRATIONS TO PERCENTAGE OF U.S. POPULATION BY AGE, IN 1980, 1990, 2000, AND 2012

	1980	1990	2000	2012
Under 5 years	0.00	0.00	0.00	0.00
5 to 9 years	0.00	0.00	0.01	0.01
10 to 14 years	0.01	0.03	0.03	0.05
15 to 19 years	0.09	0.13	0.14	0.16
20 to 24 years	0.37	0.38	0.38	0.38
25 to 29 years	1.38	1.05	0.75	0.75
30 to 34 years	2.33	1.66	1.06	1.14
35 to 39 years	2.46	2.02	1.30	1.33
40 to 44 years	2.14	2.17	1.47	1.45
45 to 49 years	2.33	2.51	2.30	1.77
50 to 54 years	1.62	1.72	2.28	1.56
55 to 59 years	1.34	1.46	2.04	1.74
60 to 64 years	1.16	1.29	1.93	2.03
65 to 74 years	0.93	0.96	1.59	1.84
75 to 84 years	0.85	0.69	1.12	1.27
85 years and over	0.80	0.48	0.70	0.87

Table 9 is similar to Table 8, but it breaks out the figures for literary works alone. In 1980, four 5-year age cohorts of authors—30–34, 35–39, 40–44, and 45–49—were producing registrations at over two times their proportion of the population. Only three such cohorts managed to do so in 1990, and the three that did so in 2000 were older—45–49, 50–54, and 55–59. Finally, in 2012, only one age group managed to produce registrations at two times their proportion of population. That age group was older still—60–64—and at a ratio of 2.03, barely broke two.

V. IMPLICATIONS

A. Implications for Copyright Theory

1. Implications for Utilitarianism

The major theory justifying copyright law in the United States is an instrumentalist or utilitarian theory.[91] According to this theory, the

91 *See* U.S. CONST. art. I, § 8, cl. 8 (vesting the intellectual property power in Congress as a means for the purpose of promoting progress in the arts and sciences); *see also* William Fisher, *Theories of Intellectual Property*, *in* NEW ESSAYS IN THE LEGAL AND POLITICAL THEORY OF PROPERTY 168, 169 (Stephen R. Munzer ed., 2001) (calling the utilitarian theory of intellectual property the "most popular" approach in the United States); Peter S. Menell, *Intellectual Property: General Theories*, *in* 2 ENCYCLOPEDIA OF LAW AND ECONOMICS 129, 130 (Boudewijn

law should balance two competing interests. On the one hand, to promote dynamic efficiency, society needs to give authors a strong right to exclude nonpayers as an incentive to create.[92] On the other hand, putting dynamic efficiency aside, wide distribution of works of authorship would maximize utility, which means that their price should not exceed the cost of their reproduction.[93] This would counsel reducing the author's power to exclude. Copyright law should configure a bundle of rights that would strike an optimal tradeoff between these two interests.[94] In a seminal article on the law and economics of copyright, Professor William Landes and Judge Richard Posner articulated this theory and devised a model that sets out the parameters of optimal copyright protection.[95]

How appropriate is Landes and Posner's framework and model for structuring lawmakers' thinking about copyright law? Under the method of scientific inquiry, a theory is judged by its ability to account for observed phenomena and to generate testable, falsifiable predictions about those that are yet to be discovered.[96] Falsification has an important constructive side to it: the discovery of phenomena that current theory did not predict and has a hard time explaining pushes researchers to develop new theoretical paradigms.[97]

Let us run a thought experiment, then: Had one asked copyright utilitarians, before we conducted this study, to spell out their predictions regarding copyright demographics, what would they say? We would be surprised if utilitarians predicted even a small subset of our findings, and we could not find anything in existing literature to anticipate those findings.

A major missing link in the utilitarian theory is a sense of the creative process, namely of the mechanism by which legal incentives to create result in original works of authorship. According to the current utilitarian mechanism—which is never stated explicitly, but is

---

Bouckaert & Gerrit De Geest eds., 2000) ("The utilitarian framework has been particularly central to the development of copyright law in the United States.").

92 *See* William M. Landes & Richard A. Posner, *An Economic Analysis of Copyright Law*, 18 J. LEGAL STUD. 325, 326 (1989).

93 *See id.*

94 *See id.*

95 *See id.* at 333–43.

96 *See* KARL R. POPPER, CONJECTURES AND REFUTATIONS: THE GROWTH OF SCIENTIFIC KNOWLEDGE 37 (5th ed. 1989) ("One can sum up all this by saying that *the criterion of the scientific status of a theory is its falsifiability, or refutability, or testability*.").

97 *See generally* 2 THOMAS S. KUHN, THE STRUCTURE OF SCIENTIFIC REVOLUTIONS 43–91 (2d ed. 1970) (arguing that falsification of previously held beliefs leads to "crises" which "are a necessary precondition for the emergence of novel theories").

rather implicit in the economic method—the author is an abstract agent, stripped of any individual characteristics, who merely responds to incentives.[98] Individuals become authors if the rewards to authorship outweigh those of alternative vocations. Further, authors are indifferent among types of works: they simply choose to create whichever work maximizes their net payoffs.

In light of our findings, we think that this mechanism fails to capture important aspects of the creative process, primarily because of the commonalities found among individuals who share similar demographics. In light of them, we do not think that utilitarian theorists can persist in applying armchair abstractions about the world. We cannot, for example, see how a utilitarian could seriously argue that policymakers wishing to advance efficient social production of creative works should not care—and indeed do not need to know (let alone know why)—that women, half of the population, currently comprise only 36% of registered authors, compared to a general female work force participation rate of 46.9%.[99] Why do women's returns on authorship fall below those of alternative vocations? Are there possible doctrinal changes that would greatly improve female authorship rates, even if they may, perhaps, slightly decrease male ones? Why do current incentives seem to motivate blacks and whites at relatively similar rates, but motivate Hispanics far, far less? Why do individuals of different races tend to create and register different types of works? Why, on average, is music created by people who are ten years younger than those who create novels? Although answering these questions is difficult, we cannot see how ignoring them completely is likely to result in optimal copyright law.

2. Implications for Lockean Labor-Desert Theory

The theoretical difficulties described above are not exclusive to the utilitarian theory of copyright, but apply more generally. Indeed, the utilitarian theory at least specifies a mechanism for the creative process, even if abstract. As we shall see, other theories suffer from similar, and at times worse, deficiencies.

Under the Lockean labor-desert theory, as applied—simplistically for present purposes—to copyright law, authors in the state of

---

98 *See, e.g.*, Ann Bartow, *Fair Use and the Fairer Sex: Gender, Feminism, and Copyright Law*, 14 AM. U. J. GENDER SOC. POL'Y & L. 551, 553 (2006) (faulting prominent and established economic analyses of intellectual property law in relying on "gender, sexual orientation, economic class, and race-neutral assumptions about human behavior").

99 *See supra* notes 79–80 and accompanying text.

nature have a natural right in their original works when they mix their intellectual labors with parts of the intellectual commons.[100] The acquisition of copyright is subject to several limitations, the major of which is that enough and as good is left for others in common. Under the social contract forming civil society, the state has a duty to protect people's natural right to property rightfully acquired.[101]

Just like utilitarianism, this labor theory of copyright is abstract, and the mechanism of acquisition is similarly uniform, individualistic, and ahistorical. There is therefore nothing in this theory that would predict or be capable of explaining the aforementioned patterns of copyright demographics. The findings further present numerous difficulties that are particular for labor theorists: If acquiring property is a natural right, is there a cause for concern that people of different races, ethnicities, genders, and ages get to enjoy and exercise their natural rights to different extents? Do some people have better access to the commons and an advantage in propertizing it? If so, are some demographics not leaving enough and as good for others? Should the state have an obligation to guarantee equal enjoyment of natural rights, rather than just equal opportunity to exercise them? The current state of discussion under the labor theory does not even begin to address these questions, and theorists writing under the labor tradition have much work to do to explain how the theory relates to, and can be reconciled with, observed patterns of creativity.

3. Implications for Personhood Theory

Under the personhood theory of copyright, authors have fundamental human needs and the state needs to allocate and enforce copyrights in order to best cater to them.[102] Scholars writing under this tradition similarly follow an abstract version of human nature.[103] Human character, and its fundamental human needs, are not only abstract, but they are also uniform across all people. Just like the prior two theories reviewed, there is nothing in the current explication of personhood theory that would predict or explain our findings above.

---

100 *See* JOHN LOCKE, TWO TREATISES OF GOVERNMENT § 27 (Peter Laslett ed., Cambridge Univ. Press 2d ed. 1967).

101 *See* Fisher, *supra* note 91, at 170; Justin Hughes, *The Philosophy of Intellectual Property*, 77 GEO. L.J. 287, 296–330 (1988).

102 *See* MARGARET JANE RADIN, REINTERPRETING PROPERTY 35 (1993); JEREMY WALDRON, THE RIGHT TO PRIVATE PROPERTY 3–5 (1988); Fisher, *supra* note 91, at 171; Hughes, *supra* note 101, at 330–50.

103 *See* RADIN, *supra* note 102, at 38–40.

The findings raise a series of difficulties that are particular to personhood theory. Is the state catering to the fundamental human needs of men better than to those of women? What are the fundamental needs of male authors, and how are they different from those of female authors? Do the fundamental needs of Hispanic authors differ from those of their white and black peers? Do the fundamental human needs of the young differ from those of the old, and if so, what is it about music copyrights that would appear to cater to the former particularly well? It would seem equally necessary that personhood scholars develop and add specificity to their theory in order to account for the aforementioned observed patterns of creativity.

4. Toward a Theory of Situated Authorship

It is time to update copyright theory. A more accurate description of copyright's creative process, suggested by the data, is that of incentives that operate together with social and psychological factors to motivate, as a statistical matter, different people (at least across race and gender) to create different types of works, at different ages.[104]

The findings above suggest that copyright theory needs to evolve from making only generalized, abstract, uniform, and individualistic assumptions about human incentives, nature, or personhood, and incorporate elements of social and cultural authorship.[105] Such an understanding is included in the situated understanding of creativity. As Fiorenza Belussi and Silvia Rita Sedita contend:

> The individualist approaches to creativity overestimate the role of the individual and of his/her abilities (the myth of the genius). On the contrary, the socio-cultural approach emphasizes the role played by contexts in the creation process: societies, cultures and historical periods. Accordingly, the individual is seen as a member of many overlapping social groups, each of them has its own network, with a specific structure and organization, which influences the creation of networks of—potentially creative—ideas. . . . Creativity is therefore "situated" in specific contexts.[106]

Such an approach should not be rejected outright because of a perceived misfit with a uniform incentive scheme embedded in copy-

104 *See supra* Sections II.B.3, II.B.4, IV.B.3.

105 *Cf.* RADIN, *supra* note 102, at 40 ("Communitarians see . . . [p]ersons [as] embedded in language, history, and culture, which are social creations; there can be no such thing as a person without society." (footnote omitted)).

106 Fiorenza Belussi & Silvia Rita Sedita, *Managing Situated Creativity in Cultural Industries*, 15 INDUSTRY & INNOVATION 457, 457 (2008).

right law. Copyright doctrine already recognizes, in substantial ways, that not all authors are alike, and that therefore one size does not always fit all. For example, copyright doctrine affords different bundles of rights and exemptions regarding different types of works, and thus provides different incentives to create and access them.[107] Copyright law also alters the bundles of rights that it recognizes in different types of legal entities: individuals and corporations have rights and limitations that differ in scope (individuals enjoy moral rights[108] and inalienable rights of termination[109]) and duration.[110]

Such an approach is moreover not foreign to copyright theory. There is a small but considerable group of scholars who have either moved away from author-uniformity assumptions[111] or who have otherwise emphasized the social, cultural, and historical side of authorship.[112] These theories have yet to explain, and did not predict, the findings in Parts II through IV above, but are at least more consistent with them.

B. Implications for Law and Policy

1. Implications for Copyright Law and Adjudication

Here we would like to make a modest normative claim: other things being equal, in cases of substantial disparities in authorship participation among various demographic groups, copyright law should adopt policies that promote authorial diversity and reduce minority groups' barriers to entry. This would seem to be justified under efficiency grounds of the utilitarian theory[113] as well as under the natural

107 *See, e.g.*, 17 U.S.C. § 114 (2012) (demarcating narrow copyrights in sound recordings).

108 *See id.* § 101 (excluding "any work made for hire" from the definition of a "work of visual art"); *id.* § 106A (defining the scope of moral rights of authors of works of visual art).

109 *See id.* § 203 (granting a right to terminate copyright transfers resecting "any work other than a work made for hire").

110 *See id.* § 302 (setting different copyright terms for works created by individual authors and for works made for hire).

111 *See, e.g.*, Yochai Benkler, *Free as the Air to Common Use: First Amendment Constraints on Enclosure of the Public Domain*, 74 N.Y.U. L. Rev. 354, 406–08 (1999) (differentiating between five different strategies for appropriation, and charting the disparate incentive impact associated with strengthening intellectual property rights).

112 *See, e.g.*, Julie E. Cohen, Configuring the Networked Self: Law, Code, and the Play of Everyday Practice 5–6 (2012) (exploring the ways in which cultural production is "mediated by context: by cultures, bodies, places, artifacts, discourses, and social networks," and arguing that "the production of the networked information society should proceed in ways that promote the well-being of the situated, embodied beings who inhabit it"); Burk, *supra* note 10, at 546; Jaszi, *supra* note 9, at 456 (deconstructing the romantic, individualistic concept of authorship and highlighting cultural, political, economic, and social influences).

113 *See* Max Nathan & Neil Lee, *Cultural Diversity, Innovation, and Entrepreneurship:*

rights and personhood theories. We believe that people bring something from themselves into their creativity, and that the authorship scene would integrate more insights, cater to more tastes, and generally be better and more interesting if a broader variety of people were involved in cultural production and had access to the means of making social meaning. Conversely, the artistic scene would be much duller if, by chance or by design, only one type of author—whether one race, one gender, or one age—participated. The more homogenous the creative class, the more normatively attractive the call for enhanced diversity.

If this much is agreed, then our findings suggest that attention should be given to the fact that women's share of registered copyrights is only a little more than one third and that Hispanic authors are greatly underrepresented. Copyright law should consider policies that would tend to increase female and Hispanic participation, as well as other substantially underrepresented demographics.

At present time, we do not think we have the requisite data to suggest that the Copyright Act should be changed to literally provide for increased protection to Hispanics and women, for example. First, as we discuss below, we believe that more information needs to be gathered systematically in order to get a more complete and accurate sense of authors' demographics and the patterns of their creativity. We have not considered in this study class, wealth, or education level, for example, and these contexts may (or may not) require greater attention. Second, our data show that demographic patterns of authorship change over time, and having some express provision written into copyright law could make it difficult to change when it was no longer appropriate. Third, although we are concerned with the law's dispa-

---

Firm-Level Evidence from London, 89 ECON. GEOGRAPHY 367 (2013) (finding some support for claims that diversity is an economic asset, as well as a social benefit); Willemien Kets & Alvaro Sandroni, Challenging Conformity: A Case for Diversity (Nov. 15, 2015), https://mpra.ub.uni-muenchen.de/68166/1/MPRA_paper_68166.pdf (arguing that diverse groups outperform homogeneous ones when innovation is needed); Beth Comstock, *Want a Team to Be Creative? Make It Diverse*, HARV. BUS. REV. (May 11, 2012), https://hbr.org/2012/05/want-a-team-to-be-creative-mak [https://perma.cc/KUZ3-ATYV]; Steve Denning, *Why Is Diversity Vital for Innovation?*, FORBES (Jan. 16, 2012, 7:42 AM), http://www.forbes.com/sites/stevedenning/2012/01/16/why-is-diversity-vital-for-innovation/#efcd4814e7c9; Katherine W. Phillips, *How Diversity Makes Us Smarter*, SCI. AM. (Oct. 1, 2014), https://www.scientificamerican.com/article/how-diversity-makes-us-smarter [https://perma.cc/6ZCW-6WZ9]; *cf.* Nigel Bassett-Jones, *The Paradox of Diversity Management, Creativity and Innovation*, 14 CREATIVITY & INNOVATION MGMT. 169 (2005) (suggesting that diversity in the workplace has a creativity benefit but a misunderstanding and conflict concomitant cost).

rate impact, correcting it with disparate treatment may be counterproductive.

One possible way forward currently may be to authorize the Librarian of Congress to decide every three years whether certain classes of authors are significantly underrepresented, and then to allocate funds to increase outreach to members of those groups, seeking to promote authorship and registration. As a procedural matter, the Copyright Office currently exercises similar authority under the Digital Millennium Copyright Act[114] anticircumvention provisions, which direct the Librarian to make determinations in a rulemaking proceeding every three years, upon the recommendation of the Register of Copyrights, for evaluating and adopting exemptions from the prohibition against circumvention of access controls.[115] As a substantive matter, race-conscious marketing efforts have been implemented by the Department of Housing and Urban Development under the Fair Housing Act,[116] and have been upheld by a number of courts.[117]

Courts can take account of the findings in this Article without additional legislation. One way in which they can do so is by purging copyright caselaw from any vestige of disparate application of the law. The expectation of such disparate application in courts may discourage individuals in the discriminated-against group from creating and registering works in the first place.

Take the fair use doctrine, for example. Fair use is an equitable doctrine that allows judges to excuse an activity that would otherwise be infringing. The fair use inquiry is explicitly open-ended: the factors to be considered only "include" the four that are listed in § 107 of the Copyright Act.[118] In particular, as the Supreme Court stated in *Camp-*

---

114 Digital Millennium Copyright Act, Pub. L. No. 105-304, 112 Stat. 2860 (1998).

115 17 U.S.C. § 1201(a)(1)(C) (2012) (instructing the Librarian of Congress to engage in rulemaking every three years).

116 42 U.S.C. §§ 3601–3619 (2012); *see* 24 C.F.R. § 200.610 (2017) ("Each applicant for participation in FHA subsidized and unsubsidized housing programs shall pursue affirmative fair housing marketing policies in soliciting buyers and tenants . . . ."); *id.* § 200.620(a) (requiring applicants to FHA housing programs to "publiciz[e] to minority persons the availability of housing opportunities").

117 *See, e.g.*, S.-Suburban Hous. Ctr. v. Greater S. Suburban Bd. of Realtors, 935 F.2d 868 (7th Cir. 1991), *cert. denied*, 502 U.S. 1074 (1992); Steptoe v. Beverly Area Planning Ass'n, 674 F. Supp. 1313 (N.D. Ill. 1987). We recognize that classifications based on immutable characteristics, and on race or ethnicity in particular, will justifiably face high constitutional hurdles. *See, e.g.*, Fisher v. Univ. of Tex. at Austin, 136 S. Ct. 2198, 2210, 2214–15 (2016) (upholding affirmative action program at public university and explaining that the University must tailor its approach in light of changing circumstances, ensuring that race plays no greater role than is necessary to meet its compelling interest).

118 *See* 17 U.S.C. § 107.

bell v. Acuff-Rose Music, Inc.,[119] the doctrine rather "[requires] courts to avoid rigid application of the copyright statute when, on occasion, it would stifle the very creativity which that law is designed to foster."[120] As such, it provides judges with discretion in characterizing each factor's pull, and in balancing the factors against each other. Analyzing courts' fair use adjudication, Rebecca Tushnet has argued that courts' analysis of the factors suffers from an implicit gender bias.[121] For example, the analysis of market harm under factor four and commercial use under factor one tend to disfavor the not-for-profit authors and consumers of fan fiction, who are predominantly female.[122] Judges exercising their discretion in fair use cases (as well as other contexts) would be right to be self-conscious about the gendered disparate impact of their decisions.

2. *Implications for Para-Copyright Federal Authorship Policy*

Our findings have implications beyond copyright law. Increasing overrepresentation of white authors is a warning signal. It suggests that policies outside of copyright law—such as educational, labor, health, fiscal, housing, and tax policies—may have an effect on authorship skills and opportunities across races and ethnicities, genders, and ages, and that those policies may need to be reconsidered. Our research methods and findings also suggest that other areas of creativity, such as patent law, are ripe for demographic review of the situated inventor.[123]

Copyright law is not the only federal law that provides authors with incentives to create. The National Endowment for the Arts ("NEA") is a federal agency, created in 1965, which "funds, promotes, and strengthens the creative capacity of our communities by providing all Americans with diverse opportunities for arts participation."[124] In fiscal year 2015, it had a budget of about $146 million, and provided "more than 2,300 grants in every Congressional district in the country."[125] Half its grants were "intended to reach underserved popula-

---

119 510 U.S. 569 (1994).

120 *Id.* at 577 (alteration in original) (quoting Stewart v. Abend, 495 U.S. 207, 236 (1990)).

121 *See generally* Rebecca Tushnet, *My Fair Ladies: Sex, Gender, and Fair Use in Copyright*, 15 AM. U. J. GENDER SOC. POL'Y & L. 273 (2007).

122 *Id.* at 300–04.

123 In particular, we are beginning to work on similar analyses in the area of patent law, drawing on the availability of inventor names for all patents.

124 *See* NAT'L ENDOWMENT FOR THE ARTS, https://www.arts.gov [https://perma.cc/CTN7-G7DT] (last visited Jan. 3, 2018).

125 JANE CHU, NATIONAL ENDOWMENT FOR THE ARTS 2015 ANNUAL REPORT 4 (2016), https://www.arts.gov/sites/default/files/2015%20Annual%20Report.pdf.

tions."[126] Rather than operate at copyright law's level of uniformity and generality, the NEA sees that each and every artist is unique. Writing, for example, in the context of its poetry grants, the NEA sees that "poets come from all walks of life, each with a different story and unique perspective."[127] As the NEA makes individual choices about which authors to give grants to, and as diversity is one of its stated values, taking as one of its criteria whether the author comes from an unrepresented demographic seems appropriate.

Further in the context of rewards, each year the President awards the National Medal of Arts to "individuals or groups" who "are deserving of special recognition by reason of their outstanding contributions to the excellence, growth, support and availability of the arts in the United States."[128] In 2015, the President awarded these medals, among others, to "authors, a poet, . . . [a] historian, . . . and a higher education program."[129] As part of his or her discretion, the President can consider ways to encourage authorship in underrepresented demographics.

3. Implications for State and Local Law

Creativity-related initiatives are not limited to the federal government. Indeed, the NEA, as an organ of the federal government, works closely with, and awards forty percent of its budget to, state art agencies and regional arts organizations.[130] These state and regional actors further make decisions about which individuals, art groups, and projects to support. For example, many localities and nonprofit groups make special housing available for artists.[131] States have better information than the federal government about local communities and individuals that face particularly potent entry barriers into authorship. States and local governments can take author demographics into account in making artist support decisions.

---

126 *Id.*

127 *Id.*

128 *National Medal of Arts*, NAT'L ENDOWMENT FOR THE ARTS, https://www.arts.gov/honors/medals [https://perma.cc/CM2X-TJEM] (last visited Jan. 3, 2018).

129 *President Obama to Award 2015 National Humanities Medals*, NAT'L ENDOWMENT FOR THE HUMAN. (Sept. 13, 2016), https://www.neh.gov/news/press-release/2016-09-14 [https://perma.cc/D589-TVXK].

130 *See* CHU, *supra* note 125, at 6.

131 For example, Artspace Projects, Inc., has worked with over thirty communities to develop artists' housing. *See Our Places*, ARTSPACE, http://www.artspace.org/our-places [https://perma.cc/D3PX-FH7Z] (last visited Jan. 3, 2018).

4. Implications for Comparative Copyright Law

The United States is unique in having a widespread industry practice of registering copyrights, which is, at least in part, due to historical circumstance. In the past, registration was one of various formal prerequisites to obtaining[132] or maintaining[133] copyright protection in the United States. In 1989, the United States joined[134] the Berne Convention for the Protection of Literary and Artistic Works, which bars signatories from imposing any such formality.[135] Accordingly, in some member states—such as the United Kingdom[136]—there is not even the option to register copyrights, while in others the option still exists, though it is often limited to particular classes of works.[137] In the decades prior to joining the Berne Convention, as part of making its law Berne-compliant, the United States made registration permissive.[138] At the same time, seeing the public benefit of having a public registry of rights in intangibles, Congress provided several incentives to encourage voluntary registration.[139] Regardless of the particular reason,

---

132 *See* Copyright Act of 1790, ch. 15, § 3, 1 Stat. 124, 125 (repealed 1802) (providing for the sole right of publication); *id.* ("[N]o person shall be entitled to the benefit of this act . . . unless he shall before publication deposit a printed copy of the title of [the work] in the clerk's office of the district court where the author or proprietor shall reside: And the clerk of such court is hereby directed and required to record the same forthwith, in a book to be kept by him for that purpose . . . ."); Copyright Act of 1831, ch. 16, § 4, 4 Stat. 436, 437 (nearly identical language).

133 *See* Copyright Act of 1909, Pub. L. No. 60-349, § 23, 35 Stat. 1075, 1080 (repealed 1976) (allowing for the renewal of copyright beyond the then-initial twenty-eight-year term of protection, subject to an application for renewal to, and its registration by, the Copyright Office).

134 *See* Berne Convention Implementation Act of 1988, Pub. L. No. 100-568, § 9, 102 Stat. 2853, 2859. The Act provided that its effective date would be the date that the Berne Convention entered into force in the United States. *See id.* § 13, 102 Stat. at 2861.

135 Berne Convention for the Protection of Literary and Artistic Works art. 5(2), Sept. 9, 1886, 1161 U.N.T.S. 3 (amended Sept. 28, 1979) ("The enjoyment and the exercise of these rights shall not be subject to any formality . . . .").

136 *See How Copyright Protects Your Work*, U.K. GOV'T, https://www.gov.uk/copyright [https://perma.cc/4XDN-8ZAV] (last visited Jan. 3, 2018) ("There isn't a register of copyright works in the UK.").

137 For example, Russia allows for the registration of only computer programs and databases; Germany and Austria allow for the registration of "literary, scientific, and artistic works" that were published anonymously or pseudonymously. *See* WORLD INTELLECTUAL PROP. ORG., WIPO SUMMARY OF THE RESPONSES TO THE QUESTIONNAIRE FOR SURVEY ON COPYRIGHT REGISTRATION AND DEPOSIT SYSTEMS 2, http://www.wipo.int/export/sites/www/copyright/en/registration/pdf/registration_summary_responses.pdf [https://perma.cc/C4WQ-PRQ5] (last visited Jan. 3, 2018).

138 *See* 17 U.S.C. § 408(a) (2012). Refusal to deposit, in the face of an express request to deposit, can result in a fine. *Id.* § 407(d)–(e).

139 The current benefits of registration are the ability to file an infringement action regarding a U.S. work, *id.* § 411; the availability of statutory damages and attorney's fees as remedies, *id.* § 412; a prima facie presumption of validity of the certificate of registration for registrations made within five year of publication, *id.* § 410(c); and the ability to record the registration with

the annual number of registrations in the United States today is greater than that in all other countries with public registries combined.[140]

It would be informative for policymakers to examine how author demographics in the United States compare with those abroad. A comparative look may enable policymakers to assess the effects of factors internal as well as external to copyright law on the participation of authors from various demographics. For example, feminist theorists have criticized U.S. copyright law as embodying male values. Arguably the view of copyright law as centered around the right to exclude and the view of intellectual property as a commodified asset that is distinct from its creator and subject to perfect alienation is a male[141] or even white-male one.[142] One scholar who shares this view suggested that the European doctrine of inalienable moral rights (which has been incorporated only marginally and reluctantly into U.S. law[143]) preserves the bond between an artist and her work and is therefore more in line with feminist values.[144] Examining female participation rates in jurisdictions with a strong moral rights doctrine may shed light on such feminist critique of copyright law, and on the desirability of proposed legal reform. More generally, we may learn about the disparate incentives of copyright enactments on various demographics by taking a comparative look at registration patterns, and more broadly authorship patterns, in other countries.

U.S. Customs and Border Protection to prevent the importation of infringing copies, 19 C.F.R. §§ 133.31–.37 (2017).

140 *See* WORLD INTELLECTUAL PROP. ORG., STANDING COMM. ON COPYRIGHT AND RELATED RIGHTS: SURVEY OF NATIONAL LEGISLATION ON VOLUNTARY REGISTRATION SYSTEMS FOR COPYRIGHT AND RELATED RIGHTS, Annex 2, at 1 (2005) (showing that the United States had 2,844,127 copyright registrations between 1998 and 2002 while Argentina had the next highest number of registrations with only 282,488).

141 *See* Burk, *supra* note 10, at 547 (discussing the feminist critique of the "masculine separation" that the property concept involves, which, "[i]n the case of literary property, . . . necessitates clear separations between author and text, reader and text, and author and reader").

142 *See* Linda J. Lacey, *Of Bread and Roses and Copyrights*, 1989 DUKE L.J. 1532, 1536–37 (relying on "[t]he feminist insight that universal, 'objective' statements about human nature are really just illusions created by middle-class white males" to recommend the incorporation of the European moral rights doctrine into U.S. copyright law).

143 *See* 17 U.S.C. § 101 (defining a "work of visual art" narrowly, for example by limiting the concept to works that are not reproduced in more than 200 copies); *id.* § 106(a) (protecting only the moral rights of integrity and attribution, and only regarding works of visual art). Moral rights were historically foreign to the Copyright Act, and a narrow version thereof was added as part of bringing U.S. copyright law closer to the letter and spirit of the Berne Convention. *See* Visual Artists Rights Act of 1990, Pub. L. No. 101-650, §§ 601–610, 104 Stat. 5128, 5128–33.

144 *See* Lacey, *supra* note 142, at 1536–37, 1548–53, 1583–84, 1594–95.

5. *Implications for Evidence-Based Policymaking*

We believe that our research is at times only suggestive because our data are not perfect. This may be a call upon Congress and the Copyright Office to collect, either through the application form or by other means, more demographic information about authors.[145] This might include not only race, ethnicity, gender, and age, but also income, education, residence, and other data. This data collection also might include information about the natural persons who create copyrighted works for businesses, as the legal fiction of corporate authorship should not obscure the human identity of these authors. Better information will enable both analysis and action to achieve a better, more open and diverse authorship scene that would enable all to participate equally in shaping our common cultural lives.

Conclusion

The author is the major figure in copyright law. Lawmakers need to have a good understanding of the author and the process of authorship, but copyright theory has not shed much light on these questions to date. In this Article, we have used registration data from the Copyright Office in order to examine who the author is empirically. Our findings show that authors of different races and ethnicities, genders, and ages tend to create different types of works, and at different rates. They also show that these patterns of creativity have changed over time. These findings were not predicted by any of the current theories of copyright law, and are consistent with only a small number of them. We hope that our findings give scholars and lawmakers better insight into the process of cultural production, and that they will ultimately encourage better, empirically grounded, copyright theory, law, and policy.

145 Recent years have seen a growing recognition of the need to base intellectual property law on evidence rather than faith or speculation. *See* John M. Golden, Robert P. Merges & Pamela Samuelson, *The Path of IP Studies: Growth, Diversification, and Hope*, 92 Tex. L. Rev. 1757, 1758–59 (2014); *see also* Mark A. Lemley, *Faith-Based Intellectual Property*, 62 UCLA L. Rev. 1328 (2015) (pushing for basing intellectual property law on evidence, and criticizing natural law and faith-based approaches).