2016

Copyright's Race, Gender and Age: A First Quantitative Look at Registrations

Robert Brauneis
The George Washington University Law School, rbraun@law.gwu.edu

Dotan Oliar
University of Virginia

COPYRIGHT'S RACE, GENDER AND AGE:
A FIRST QUANTITATIVE LOOK AT REGISTRATIONS

Robert Brauneis and Dotan Oliar*

ABSTRACT

On a per capita basis, do African-American authors produce more copyright registrations than non-Hispanic whites? Do men and women show a within-group bias in choosing co-authors? And what decade in the average musician's life is the most productive? This article provides answers to these questions – which happen to be yes, yes, and the 20s, respectively – and many more by statistically analyzing the 15 million entries that comprise the Copyright Office's full record of registered works from 1978 through 2012. It provides a variety of perspectives on individuals' creativity in modern-day America and on the beneficiaries of our copyright system along the axes of race, gender and age. Its findings suggest a need to promote greater diversity and equality in the processes of cultural production and the making of social meaning.

INTRODUCTION

Who creates the books, songs, movies, plays, art, and software that have formed education, culture and entertainment in the United States? What is the race, gender, and age of the authors of those works? Which authors are benefitting from the copyright system, and how do their demographic

* 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 access to the Copyright Office Catalog, we thank Register of Copyrights Maria A. Pallante and her staff at the U.S. Copyright Office. For valuable comments and discussions, we thank Michal Barzuza, Michael Birnhack, Alon Harel, Ariel Porat, Tomás Gómez-Arostegui, Ellen Goodman, Edward Lee, Lydia Loren, Zvi Rosen, and Matthew Sag. We presented earlier versions of the paper and benefitted from participants' comments at workshops at Tel Aviv University, the University of San Diego, St. John's University, Loyola Law School Los Angeles, the University of California, Berkeley, Lewis & Clark Law School, the Chicago IP Colloquium (co-sponsored by the Chicago-Kent College of Law and Loyola University Chicago School of Law), and the Christopher A. Meyer Memorial Lecture (co-sponsored by Meyer, Klipper & Mohr, PLLC, the United States Copyright Office, the Copyright Society of the USA, and the George Washington University Law School). The appended dataset, as well as the findings and conclusions in this Article, are our own, and have not been reviewed or endorsed by the U.S. Copyright Office.

characteristics compare with those of the population as a whole? This article pursues answers to those questions by examining a hitherto untapped data source: the United States Copyright Office Electronic Catalog, which is by far the world's largest registry of works under copyright.

Since 1978—the effective date of our current Copyright Act—the Copyright Office has kept its records digitally. Though this Catalog has been accessible to the public, it has only been available through an online search tool that is suitable for researching rights in particular titles but not for conducting statistical analyses of thousands or millions of records.[1] For the first time, the Copyright Office, through its Academic Partnership Program, has generously provided us a full copy of the Catalog as it stood in late 2014. We expended much work to reverse-engineer Office recordkeeping protocols that changed over time, clean the data, and, importantly, convert them from the Library of Congress's unique MARC archival format to a customary columns-and-rows dataset structure. All of these steps required judgment, to be sure, but conducting them was necessary for us to be able analyze these data statistically. As a service to the public, and to facilitate follow-up research by others, we are releasing the dataset we built and used and separate documentation explaining it.

Each registration record in the Catalog includes information about the dates of creation, publication, and registration of the works registered; the type of each work, whether they are literary, dramatic, musical or audiovisual works, works of visual art or computer programs; the names of individual authors and, in many cases, their birth and (if applicable) death years; the names of corporate authors; and the names of corporate and individual owners.[2] To these basic data that exist in the Office's records we added information about authors' ages and their probable gender and race, where no such information is solicited on the Office's registration forms.

We are able to calculate authors' ages by subtracting their birth year, where known, from the year in which they created their works. Establishing authors' gender is more difficult and we estimate it using their first names. While it is easy to determine 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 under the 1990 U.S. Census. Finally, determining the race or ethnicity of authors has posed the biggest challenge

[1] One of us 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, Nathaniel Pattison & K. Ross Powell, Copyright Registrations: Who, What, When, Where, and Why, 92 Tex. L. Rev. 2211 (2014).

[2] Beginning in 2008, the Catalog has included additional information, such as mailing addresses associated with claimants and authors, and their citizenship.

for us. To estimate race and ethnicity we have used data about the racial and ethnic distribution of last names from the 2000 U.S. Census. For a long list containing the last names borne by 90% of the U.S. population, we could determine the probability that a person bearing that last name has self-identified as being either of Hispanic ethnicity (regardless of race) or as a non-Hispanic who falls into one of the following mutually exclusive races as they are defined, named and used by the U.S. Census Bureau: white, black, Asian or pacific islander, native American or Alaskan, or two or more of these four races.

Part I below provides some basic information about the Catalog and about the subset of registration records that we analyze in this Article. Part II considers the race and ethnicity of individual authors identified in those records. Among other things, it reveals that Hispanic authors are very substantially underrepresented in the ranks of authors compared to their share of the population. It also shows that there are substantial differences in the race and ethnicity of authors of different types of works. Part III considers authors' gender. It begins by showing that two-thirds of all authors represented in copyright registrations in our 35-year window are male. Authorial gender disparity has generally decreased over that period, and in substantial ways. For example, women and men now produce registered, published literary works in roughly equal numbers. But this phenomenon of increased female authorship overtime has not been universal across all types of works. Women's share of registered musical and dramatic works, for example, has remained basically unchanged and in the minority, and women's share of registered visual art has actually decreased. Part IV focuses on the age of authors. It shows that the average age of authors overall has increased at about the same rate as the median age of the U.S. population as a whole. But it also shows that musical work authors are on average ten years younger than literary work authors, and that production of music is much more age-concentrated than production of literature. It may suggest that the human mind peaks in creativity at different ages for different subject-matters. Part V concludes. Four online appendices contain our dataset and additional statistics. Our findings reveal that copyright law has been oblivious to two questions: how do the incentives to create and access copyrighted works differ by the personal characteristics of individuals, and how should they? We argue that copyright policymakers should focus their attention on both questions and act to bring about a more diverse authorship scene that would enable all to participate meaningfully in shaping our cultural lives.

I. INTRODUCTION: THE DATASET OF ORIGINAL VALID MONOGRAPH REGISTRATIONS, 1978-2012

In this study, we focus on 14,598,621 original valid monograph registration records for the years 1978-2012 that were included in the Copyright Office Electronic Catalog as of September 30, 2014.[3]

A. What Are Original Valid Monograph Registrations, and why are we focusing on them?

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.[4]

The Catalog as received by us contained over 27 million records. We focus on a portion thereof—about 54%—that we call original valid monograph registration records. An original valid monograph registration is what most of us would imagine a typical copyright registration to be: the initial registration of a claim of copyright in a stand-alone work like a novel or a motion picture, which has not subsequently been cancelled. We selected those according to the following criteria, and for the following reasons:

- Monographs: A *monograph* registration is a registration for any work that is not a serial, serials being works published in a series such as magazines and newspapers that usually contain collections of articles, photographs and other materials created by a variety of authors. We decided to concentrate on monographs first and foremost because information about authorship and ownership of copyright in serials is relatively thin. The authorship and ownership reported in serial registrations are generally for the compilation – the selection and arrangement of the individual components – rather than for any of the individual contributions.[5] Serial registrations further contain no information about the type of work that is being registered, information that is recorded with regards to monographs. It is also the case that most economically significant works – successful motion pictures, video games, software, novels, songs, and so on – are registered as monographs.

---

[3] 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).

[4] The records in the Catalog are currently maintained in the Machine-Readable Cataloging (MARC) format for bibliographic records. For additional details on the history of the Catalog, see Online Appendix I, From the Copyright Office Catalog to the Original Valid Monograph Registration Datasets: Some History and Technical Details, available at [].

[5] *See* 17 U.S.C. §§ 103 (establishing copyright in compilations); 101 (defining "compilation").

- Original: We are focusing only on monograph registrations that are *original*. The excluded non-original monograph registrations are either supplementary registrations or renewal registrations. The former are registrations that are intended only to correct or amend earlier-filed registrations, and including them would have amounted to double-counting.[6] Renewal registrations concern works that originally obtained federal copyright before 1978, under the Copyright Act of 1909. Until 1992, renewals had to be filed to obtain copyright protection beyond the initial 28-year term; until the end of 2005, there remained some residual benefits to filing them.[7] Renewals can be useful for answering various policy questions,[8] but they would not be for purposes of investigating cultural production and registration since 1978. Registrations that are neither supplementary nor renewal registrations are original registrations. They are what most people think of what they think of copyright registrations – registrations filed in order to make an initial claim of copyright in a work or works and to gain the benefits of registration. Typically, only one original registration is filed for each work, and thus the number of original registration records has some correspondence to the number of works registered, although that correspondence turns out to be complicated, as we will explain further below.

- Valid: Finally, we are only considering original monograph registrations that were *valid* as of the date and time that the records

---

[6] 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 Online Table 2 of Online Appendix II, Additional Tables and Charts, available at [ ]. Under Copyright Office practice, 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. We have omitted these from consideration, since we would end up double-counting registrations if we included them. 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.

[7] 730,401 records in the 9/2014 Catalog are records of renewal registrations for works that originally gained federal copyright before 1978.

[8] *See, e.g.*, Landes & Posner, The Economic Structure of Intellectual Property Law, Ch. 8 (2003) (examining renewal registration rates in the context of proposing that copyrights be indefinitely renewable).

were extracted from the Catalog. A Catalog registration record is created when a registration application is granted. For a number of reasons, registrations can later be cancelled. When a registration is cancelled, the registration record is not removed from the Catalog, but is simply marked cancelled.[9] For most purposes, it is not useful to count cancelled registrations; counting them would in many cases amount to double-counting, since many works the registrations of which are cancelled end up being re-registered.[10]

For purposes of our analysis, we have further excluded registration records that had critical fields that were blank or contained invalid values.[11] Further, we have decided to consider only original valid monograph registration records that have registration dates from January 1, 1978 through December 31, 2012. As processing registration applications in the Copyright Office takes time, many registration applications filed in 2014 and even in 2013 had not entered the Electronic Catalog by September 30, 2014, the date of the Catalog version with which we are working. For that reason, statistics concerning registrations in 2013 and 2014 would not accurately reflect the number of valid registrations filed.[12] Applying all those additional criteria left us with 14,598,621 records. It is those records that comprise the dataset we are releasing, and those records that we will analyze below.

B. The Basic Information Available in Original Valid Monograph Registration Records: of Registrations, Works, Claimants and Authors

Registration records systematically include four different kinds of information:[13] information about the registration itself; about the work or

---

[9] 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 Table 2 of Online Appendix II, Additional Tables and Charts, available at [ ].

[10] 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.

[11] Those exclusions of an additional 590 records, detailed in Online Table 3, leave the dataset with 15,262,519 records.

[12] 663,884 original valid monograph registration records with registration dates in 2013 and 2014 were excluded.

[13] 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

works registered; about the claimant(s) of the work(s) registered; and about the author(s) of the work(s) registered.

1. *Information about the Registration.* For our purposes, the most important piece of information in this category is the effective date of the registration, which all registrations have.[14] That enables us to analyze registrations that were made in different years, and to show changes across time. The date of registration is determined by the Office as the date on which a valid application was received by it and is therefore objective and verifiable. For these reasons, and because the electronic catalog begins with registrations with effective dates on or after January 1st, 1978, we use this variable as our criterion for organizing our data along full years.

Most registrations also contain creation year data. Creation year is inferior as a running variable since not all registered works have one (though the vast majority do)[15] and since they are self-reported by registrants. We further do not have a 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 these can always be registered later. In any event, registered works' creation and registration dates are quite close: the average registered work in our dataset was created about one year and one month earlier,[16] and 56.87, 85.58, 93.93, 96.34, 98.29 percent of registered works were registered within 0, 1, 3, 5, and 10 years of creation, respectively.[17]

2. *Information about the Work or Works Registered.* Registration records contain a variety of information about the works to which they pertain. The three types of information that are most important to our analyses are

- the type of work;
- the year that the work was created; and
- the work's status as published or unpublished at the time of registration, and if published, the date of publication.

a. *Type of work.* Each registration record contains a two-letter code that identifies the work being registered as predominantly belonging to one of 11

---

this additional information was either irrelevant to our purposes or entered too inconsistently to be of use.

[14] 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).

[15] Some 104,091 registrations (about 0.72 percent) of the 14,472,367 registrations we focus on under the six major categories of work below do not have creation year data.

[16] The mean difference between the year of registration and the year of creation is 1.1 years.

[17] These numbers were calculated after omitting 9,129 registrations with a registration year earlier than their creation year, which are erroneous.

categories that are listed below in Table 1.[18] Some categories are quite broad and cover a very large number of registrations, while others are much narrower and cover a comparatively small number of registrations. Three are hybrid categories that cover two or more of the other categories, namely, "Sound Recording and Music," "Sound Recording and Text," and "Multimedia Kit."

For most of our analyses, we have decided to omit the three smallest categories, which are "Map," Sound Recording and Text," and "Multimedia Kit." These three categories together represent less than one percent of all registrations, and 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 categories that are related to the production of music, namely, "Musical Work," "Musical Work / Sound Recording," and "Sound Recording." It is undoubtedly interesting to separate these categories for certain purposes; indeed, one of us has written an article that uses shifts between these categories to reveal changing patterns in how music is created and registered.[19] Yet all three are closely related in the production of commercially distributed music, and the hybrid category already combines the other two, leading us to believe that combination of all three is appropriate for an initial analysis.

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." It is important to re-emphasize that some of these categories cover many more registrations than others: "Text" has over 18 times as many registrations as "Software." Naturally, trends in the largest categories will much more heavily influence totals than trends in the smallest categories. A more complicated Table in an online appendix shows the relationship between the categories we are using and other schemes for categorizing works

---

[18] 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.*, Compendium of U.S. Copyright Office Practices § 609(2) (3rd ed.) ("Works Containing Multiple Forms of Authorship") ("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 themselves recognize the hybrid character of many works in that category; for example, §102(2) defines one category as "musical works, including any accompanying words"; § 102(3) defines another category as "dramatic works, including any accompanying music." *See* 17 U.S.C. § 102.

[19] *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).

of authorship, including the eight categories in § 102(a) of the Copyright Act.[20] 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.

Table 1

Type-of-Work Categories in Original Monograph Registrations

Categories in Copyright Registrations	Our Abbreviations	Total Number of Registrations in the OVM 1978-2012 Dataset	Percentage of Total	Percentage of Total in our 6-Category Scheme with Combined Music
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%	

b. The Work's Year of Creation. As mentioned above, over 99 percent of original valid registration records contain information about the year that the work or works in question were created.[21] The year of creation is self-reported by registrants, and for that reason may not always be accurate.

[20] *See* Online Table 4, Online Appendix II, *available at* [ ].

[21] 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.

c. Publication Status and Date of Publication. Each registration record notes whether the concerned work or works were published at the time of registration. If the work was published at the time of registration, a date of publication is usually also included. A little over a half of all works were registered as published.[22]

d. The Number of Works Covered. It may be surprising to learn that there is often little or no way of telling from a registration record how many works are covered by the registration. In part, this difficulty stems from an inherent ambiguity in the term "work," and from a continuing tension between pre- and post-1976 Act conceptions. Before the 1976 Act, federal copyright protection attached upon publication, and therefore it was natural to think of a "work" as a unit of publication, even if portions of the content of that publication had been created by different people at different times. The 1976 Act states that a work is created upon fixation, which leads to a potentially different concept of work and count of works.[23]

Yet in some cases, it is uncontroversial that many works – sometimes thousands of works – are being registered in a single registration transaction. For example, litigation has revealed that some stock photography companies register thousands of photographs in a single transaction.[24] Because the photographs have been taken by many different photographers and are destined for completely separate use and sale, it is hard to see how each of them should not be treated as a separate work. Unfortunately, however, it is often difficult or impossible to tell from the registration record, the application, the registration certificate, or even the deposit, how many photographs are covered by a particular registration. In Online Appendix I, we have provided more detail about the information about number of works sometimes provided in registration records. For purposes of this Article, it is important to note that we are counting registrations, and not attempting to separately count works.

---

[22] Overall, 7,863,069 registrations, or about 54%, are for published works, while 6735551 registrations, or about 46%, are for unpublished works.

[23] *See* 17 U.S.C. § 101 ("A work is 'created' when it is fixed in a copy or phonorecord for the first time."). Note, however, that the second sentence of that definition suggests that a work could be created over time through multiple acts of fixation; in that case, the work is not defined by a continuous act of fixation, and another criterion must be found for distinguishing one work from another. *See id.* ("[W]here a work is prepared over a period of time, the portion of it that has been fixed at any particular time constitutes the work as of that time . . . .").

[24] *See* Alaska Stock, LLC v. Houghton Mifflin Harcourt Pub. Co., 2010 WL 3785720, *1 (D. Ak.) ("The compilations each contain between 500 and 6,000 photographs created by approximately 106 individual authors."), *reversed and remanded*, 747 F.3d 673 (9th Cir. 2013).

3. Information about Author(s) and Claimant(s). Each registration record contains information about the person(s) or organization(s) claiming to own copyright in the work or works registered, and the person(s) or organization(s) represented as having authored that work or works. In the case of claimants, the records reliably contain the name of every claimant, and a reasonably reliable indication of whether each claimant is an individual or a corporate entity. We parsed the individual claimant names into first and last names, and performed a variety of text searches and calculations where necessary to adjust for pseudonyms, "doing business as" names, and other alternate names. Technical details about these matters are covered in Online Appendix I.

Copyright registration records also contain a variety of information about authors. They contain names of authors, and a reasonably reliable indication of whether each author for copyright purposes is an individual or a corporate entity. As we did with claimants, we parsed the names of individual authors into first and last names, and also adjusted for pseudonyms and alternate business names. Somewhat confusingly, authors named in registration records may include both those who are authors for copyright purposes, and those who are not copyright authors but who may be authors for bibliographic purposes. For example, the author of a registered work for copyright purposes may be a corporate entity, as an employer for hire, but the registration record may also contain the name of the individual employee who is credited on the deposit copy with writing the work in question. Our analyses, and the datasets we are releasing, attempt to count as authors only those who are authors for copyright purposes. In part, this is because information about authors for other purposes is quite spotty and inconsistent.

To enable inquiry into the race and gender of authors, we have incorporated U.S. Census Bureau data in ways that we will detail below; we will also detail below how we calculated the age of authors at the time they created registered works.

Lastly, we should note that while information about claimants in registration records should be and is largely complete – the entire purpose of copyright registration is to make a claim of copyright, which cannot be done without identifying the claimant – the Copyright Office does not always require complete information about authorship on registration applications or deposits. For example, we know that the Copyright Office has allowed group registration of photographs by over 100 photographers with only three of the photographers identified.[25] We cannot solve this last problem, and therefore

[25] *See, e.g.*, Alaska Stock, LLC v. Houghton Mifflin Harcourt Pub. Co., 747 F.3d 673, 675-676 (9th Cir. 2014).

authorship information in the datasets, like authorship information in the Catalog, is unfortunately incomplete.[26]

II. RACE AND ETHNICITY

A. METHODOLOGY: INFERRING THE PROBABILITY OF AUTHORS' RACE AND ETHNICITY FROM THEIR LAST NAMES

Registration records do not specify individual authors' race or ethnicity so we use their last names as a proxy. Last names are often associated predominantly with particular racial or ethnic origins, and 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.[27] Under federal policy, the Census Bureau asked people to self-identify as members of one or more of six races—white, black, Native American or Alaskan, Asian, Hawaiian or other Pacific Islander, and "Some Other Race."[28] In addition, it asked them to separately note whether they are "Spanish, Hispanic, or Latino," which it regards as their ethnicity, rather than race.[29]

Based on answers to those questions in 2000, the Census Bureau provides the probability that holders of various last names are either of Hispanic ethnicity, regardless of their race, or are, alternatively, non-Hispanic and fall into one of five mutually-exclusive racial categories: white only, black only, native American or Alaskan only, Asian or Pacific Islander only, or is of two or more races.

Relying on this six-category taxonomy and terminology, we were able to assign probabilities of race or ethnicity to the vast majority of individual

---

[26] The lack of complete authorship information in the Catalog for works that are not made for hire is particularly unfortunate, because every such work in which copyright has been transferred is subject to the author's or authors' power to terminate the transfer, see 17 U.S.C. §§ 203, 304, and thus the registrations do not even identify all persons who have enforceable future interests in those registered works.

[27] *See* Frequently Occurring Surnames from the Census 2000, available at http://www.census.gov/topics/population/genealogy/data/2000_surnames.html (containing information on the probability that individuals with particular last names belong to one of six racial or ethnic categories).

[28] *See* Elizabeth M. Grieco & Rachel C. Cassidy, Overview of Race and Hispanic Origin 2000: Census 2000 Brief, available at https://www.census.gov/prod/2001pubs/c2kbr01-1.pdf; Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity (Federal Register Notice, October 30, 1997), available at https://www.whitehouse.gov/omb/fedreg_1997standards.

[29] See *id.*

authors of registered works. The Census data provides probabilities for 151,671 last names. Our dataset contains 10,425,336 registrations of works of one of the six principal types that were created by individual authors. Of that group, 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.

It is important to have some general background understanding of how holders of the most popular last names self-identify racially and ethnically. There are some relatively popular last names that 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 25 most popular last names in the United States, and over 90% of people bearing those last names identified as Hispanic. Nyugen, 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. By contrast, however, those who self-identify as non-Hispanic white, or as non-Hispanic black, tend to share many 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% white, 47% black; Brown – 61% white, 35% black; and Jones – 58% white, 38% black. Thus, when, as shorthand, we make statements about the race or ethnicity of authors of a particular type of work, we are actually referring to the average of the aforementioned probable race or ethnicity of a certain cross-section of authors.[30]

The existence of many popular surnames that are shared in substantial percentages by people who self-identify as being of different races or ethnicities raises the possibility that authors who bear a particular surname may not be representative of holders of that surname in the general population. For example, non-Hispanic whites named "Williams" might become authors at a greater (or lesser) rate than non-Hispanic blacks named "Williams"; if that is the case—assume, for example, that all registrants carrying that last name are black—then the surname distribution in the general population will not provide accurate estimates of the racial or ethnic makeup of authors. We have devised two ways to measure and correct for such selection bias, which we detail below. In short, they affirm that the direction of our findings as to the average racial and ethnic registration

[30] 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 as 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.

tendencies of people of different racial and ethnic backgrounds is correct, and that the magnitude of the phenomena pointed at tends to be higher than population-based averages suggest.

B. MAIN FINDINGS

With regard our three principal demographic focuses – race or ethnicity, gender, and age – we will present a number of findings that were among the most interesting and striking to us as we worked through the dataset. In many cases, this was because the figures in question were somewhat counterintuitive, and seemed to call out for an explanation. Thus, this Article will likely raise more questions than it answers. Providing definitive explanations for surprising demographic findings is beyond the scope of this Article. It is our intention, however, to set out perplexing phenomena in the data and invite follow-on researchers to search for such answers.

1. Overrepresentation of non-Hispanic white authors in copyright registrations has grown between 1978 and 2012.

Many people believe that the United States has slowly been moving away from racial and ethnic discrimination, and as a corollary that the differences between races and ethnicities across various social and economic criteria are on the decline. Yet between 1978 and 2012, the proportion of non-Hispanic white authors reflected in copyright registrations, compared to their proportion in the population, has grown. In 1980, non-Hispanic whites accounted for 79.6% of the general population in the United States. [31] Our figures suggest that 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 non-Hispanic whites in the U.S. population has been decreasing. It dropped to 75.6 by 1990,[32] 69.1 percent by 2000,[33] and 63.7 percent by 2010.[34] While the percentage of non-Hispanic white authors represented in copyright registrations has also been dropping, it has not dropped nearly as much. It dropped to 77.41% in 1990; to 75.19% in 2000; and to 73.96% in 2010. Thus, as of 2010, non-Hispanic white authors

---

[31] Tbl. 43 p. 1-23, Single Years of Age by Race, Spanish Origin, and Sex: 1980.

[32] *See* U.S. Department of Commerce, Bureau of the Census, 1990 Census of Population: General Population Characteristics, p. 3, Tbl. 3, Race and Hispanic Origin: 1990 (noting that non-Hispanic whites were 188,128,296 and the whole U.S. population was 248,709,873), available at http://www2.census.gov/library/publications/decennial/1990/cp-1/cp-1-1.pdf.

[33] *See* U.S. 2000 Census Profiles of General Demographic Characteristics 1, Tbl. DP-1, available at http://www2.census.gov/census_2000/datasets/demographic_profile/0_United_States/2kh00.pdf .

[34] *See* U.S. Census Bureau, Tbl. 1, White Population 2000 and 2010.

were producing 116% of the registrations they would be if they were producing at a rate equal to their proportion of the general population – the rate at which they were producing registrations in 1980, three decades earlier.

Why are non-Hispanic 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 non-Hispanic white population is relatively older than the population of other racial and ethnic groups, and in particular has a smaller percentage of its population that is under 25, a segment of the population that produces very few copyright registrations. It is also possible that our methodology underestimates non-Hispanic white authors before 2000, because it allocates to last names the population distribution as of 2000, whereas non-Hispanic whites comprised a larger percentage of the population between 1978-1999 than in 2000 (although a smaller percentage between 2001-2012), which may suggest that non-Hispanic 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. Underrepresentation of Hispanic authors in copyright registrations, already prominent in 1980, has become extraordinary by 2010.

In 1980, Hispanics constituted 6.4% of the U.S. population, 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 producing at a rate equal to their proportion of the population. Since 1980, Hispanic population in the United States has grown tremendously: 9.0% in 1990, 12.5% in 2000, and 16.3% in 2010. 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 are 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 the largest underrepresentation of any racial or ethnic group. As mentioned above, in 2010 non-Hispanic whites were at 116% (73.96/63.7). To round out the figures, non-Hispanic blacks were at 120% (15.11/12.60); Asian or Pacific Islanders were at 83% (4.05/4.9); American Indian/Alaskan Natives were at 77% (0.7/0.9); and people of two or more races were at 62% (1.8/2.9).

What can explain the striking and growing underrepresentation of Hispanic authors? The relative age of the Hispanic population explains a small part of the difference. In 2000, Hispanics between 25 and 64 constituted 11.6%

of the total U.S. population between 25 and 64[35] – somewhat smaller than the 12.5% that Hispanics of all ages constituted of the total U.S. population. Yet in that year, Hispanics still only produced 6.8% of the copyright registrations. If we consider that as a percentage of the proportion of all Hispanics to the total U.S. population, we get a figure of 54.4%; if we consider it as a percentage of the proportion of Hispanics 25-65 to the U.S. population 25-65, we get a somewhat higher, but still dramatically low, figure of 58.6%. Similarly, we mentioned that in 2010, Hispanics were producing only 44.6% of the registrations they would be if they were producing at a rate equal to their proportion of the U.S. population; if we considered the proportion that Hispanics 25-64 constitute of the U.S. population 25-64 in 2010 – 14.6%[36] – that production rate would increase to 49.8%. However, that is still less than half the rate at which the U.S. population generally produces registrations.

A somewhat larger portion of the difference may possibly be explained by the fact that the general population of Hispanics in the United States includes a relatively large percentage who are unauthorized immigrants. Of the 50.5 million Hispanics in the United States in 2010,[37] approximately 8 million were unauthorized immigrants,[38] and that group accounted for a substantial majority of all unauthorized immigrants, estimated to be about 11 million in total.[39] 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

---

[35] *See* United States Census Bureau, Race and Hispanic or Latino Origin by Age and Sex for the United States: 2000, Table 8: Hispanic or Latino Origin Population; White Alone Not-Hispanic or Latino Origin Population; and Population Other than White Alone Not-Hispanic or Latino Origin, by Age and Sex for the United States: 2000, *available at* https://www.census.gov/population/www/cen2000/briefs/phc-t8/index.html (in 2000, Hispanic population 25-64 was 17,085,441; total U.S. population 25-64 was 146,992,887).

[36] *See* United States Census Bureau, The Hispanic Population in the United States: 2010, Table 1: Table 1. Population By Sex, Age, Hispanic Origin, And Race: 2010, *available at* http://www.census.gov/population/hispanic/data/2010.html (in 2010, Hispanic population 25-64 was 23,560,000; total U.S. population 25-64 was 161,314,000).

[37] *See* Sharon R. Ennis, Merarys Rios-Vargas & Nora G. Albert, The Hispanic Population 2010 (2010 Census Briefs) 3 (Table 1), *available at* http://www.census.gov/prod/cen2010/briefs/c2010br-04.pdf

[38] *See* Michael Hoefer, Nancy Rytina, & Bryan C. Baker, Estimates of the Unauthorized Immigrant Population Residing in the United States: January 2009, *available at* https://www.dhs.gov/xlibrary/assets/statistics/publications/ois_ill_pe_2009.pdf (estimating that about 8,050,000 unauthorized immigrants in the United States originated from the countries of Mexico, El Salvador, Guatemala, Honduras, and Ecuador)

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

about 16% of Hispanics living in the United States are unauthorized immigrants, 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 44% or to 49%.

3. Non-Hispanic Black Authors are Slightly Overrepresented throughout the Study Period.

The non-Hispanic 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. Non-Hispanic black authors have also contributed a relatively stable and slightly rising percentage of copyright registrations, from 14.22% in 1980 to 15.11% in 2010. Thus, non-Hispanic 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 types of works

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 textual works. Lastly, Asians and Pacific Islanders have been strongest in art and software, and weakest in music and drama. The percentages of each type of work registered by authors of each of the races is as follows:

Table 2 – Percent of Registrations by Race and Work Types

	Text	Music	Drama	Art	Movies	Software	All
White	77.77	74.56**	77.82*	76.68*	76.96	78.52**	76.21
Black	13.57	16.07**	13.97*	12.57*	12.81	12.06**	14.61
Hispanic	4.65*	7.42**	5.76	5.65	6.55*	4.46**	6.09
Asian / Pacific Islander	4.27	1.86**	2.76*	5.63**	4.20	5.54*	3.25
Native American /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

Legend:
X** – Most prevalent type by race X* – Second most prevalent type by race
X** – Least prevalent type by race X** – Second least prevalent type by race

The strengths and weaknesses of white and Asian authors overlap somewhat: both are strong in software and are 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 reversed.

C. Methodology Revisited: Selection Issues in Estimating Race Probabilities

One might be worried that our method, which assigns authors probable races or ethnicities based on the population distribution of their last names, may suffer from a severe selection bias. To illustrate, assume that the last name "Williams" is shared equally by non-Hispanic whites and non-Hispanic blacks, but that non-Hispanic blacks are registering copyrighted works at a rate double than non-Hispanic whites. If so, we should be assigning two-thirds of the Williams registrations to non-Hispanic blacks rather than one-half. More generally, if people of different races had different propensities to register copyrighted works, our method could be substantially off the mark.

How can one detect and if necessary correct for such potential selection bias? We employed two methods to determine whether our method involves a severe selection bias, and we conclude that it does not.

First, we used multiple regression analysis using as our data the most popular 1000 last names in the U.S. under the 2000 Census. As our dependent variable we used the percentage of each last name among registered works. In the first model below, we used as the independent variable the percentage of each last name in the U.S. population. In the second model below, we added as an independent variable the non-Hispanic black to non-Hispanic white ratio of holders of each last name in the U.S. population. The third model uses, like the first two, the aforementioned population percentage variable and adds a Hispanic to non-Hispanic white ratio independent variable. Lastly, our fourth model uses all three aforementioned independent variables. We converted all our variables to a log form, with the interpretation that our coefficients are elasticities. Our results are as follows:

Table 3: Regression Analysis of Top 1000 Last Names: Population Frequency, Race and Ethnicity as Correlates of Copyright Registrations

VARIABLES	(1) logregistered	(2) logregistered	(3) logregistered	(4) logregistered
logUS	0.992***	0.951***	1.015***	0.996***
	(0.0248)	(0.0249)	(0.0163)	(0.0165)
logb2w		0.105***		0.0502***
		(0.0146)		(0.00984)
logh2w			-0.180***	-0.176***
			(0.00495)	(0.00496)
Constant	-0.190	-0.317	-0.495***	-0.546***
	(0.204)	(0.200)	(0.134)	(0.133)
Observations	1,000	999	1,000	999
R-squared	0.617	0.636	0.835	0.840

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

As Table 3 shows, the coefficient on the percentage of last names in the U.S. population came out across all four models in the neighborhood of 1 and is statistically significant. It means that if we increase the percentage that a last name captures in the U.S. population by one percent – such as from 2 percent to 2.02 percent – we would expect to see a one percent increase in the percentage of registrations under that last same in our sample as well.

Next, the coefficient on our black to white ratio came out positive and statistically significant across models 2 and 4 in which this ratio was used. It means that a one percent increase in a last name's ratio of black to white bearers – such as from a last name being 60% black and 15% white (a ratio of 4) to 60.6% black and 15% white (a ratio of 4.04) – correlates with a 0.05% (0.1%) increase in registrations under that last name under Model 4 (Model 2). It suggests that blacks register more works per capita compared to whites.

Next, the coefficient on our Hispanic to white ratio came out negative and statistically significant across models 3 and 4 in which this ratio was used. It means that a one percent increase in a last name's ratio of Hispanic to white bearers correlates with about 0.18% decrease in registrations under that last name. It suggests that whites register more works per capita compared to Hispanics.

We have devised an additional method to gauge the magnitude of any potential selection bias. We have normalized the per capita rate of registration of non-Hispanic whites to 1. We then assumed that blacks and Hispanics on average register copyrighted works per capita at certain multiples of the rate at which whites do, and have optimized over all ratios ranging from 1:100 to 100:1. For each assumed pair of such ratios, we calculated based on the last names' prevalence in the population and the racial and ethnic composition of that last name the expected proportions that we should encounter various last names in our sample of copyrighted works. We then calculated, per each pair of assumed racial and ethnic registration ratios, the mean square difference between the expected ratios of last names and the observed ones in our sample. The values that minimized this mean square difference were a 1.19 black to white ratio of per capita registration and a 0.28 Hispanic to white ratio.

These two estimation methods suggest to us that, as compared to whites, blacks register more than their population percentage and Hispanics substantially less than their population percentage. These results are qualitatively in line with assigning race and ethnicity to authors according to their last name distribution of race and ethnicity in the 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, and that blacks' percentages have hovered around 120% throughout the period. We saw that Hispanics' percentages were around 44%.

The percentages we reported initially pointed at the right direction for a simple reason: if blacks are highly 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 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 point us in the right direction as to different races and ethnicities average tendencies to register copyrighted works.

Though this method tends to point us at the right direction, it gives an inaccurate measure of the magnitude of the difference. This is so because, per each last name in our dataset of registered works, the method assigned the population distribution of races and ethnicities rather than one weighted by each race and ethnicity's average tendency to register. After all, our general initial finding that blacks tend to register more works per capita than whites

do is in tension with assigning per each last name in our dataset race and ethnic probabilities that are equal to those found in the population. Our last estimation method gives a sense of the magnitude of the actual difference, which is larger than the one initially estimated. Qualitatively, however, all three methods of analysis portray a similar picture of different racial and ethnic groups average relative propensities to register copyrighted works.[40]

II. GENDER

A. METHODOLOGY: INFERRING AUTHORS' PROBABILISTIC GENDER FROM FIRST NAMES

Registration records do not specify authors' gender.[41] They do, however, contain the authors' first names. It is not hard, for example, to tell the gender of a John or a Jane, but what about Pat or Terry? In conducting gender statistics we rely on information elicited from the 1990 U.S. census regarding the gender distribution of first names.[42] 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 category of creativity we simply refer to the average of the aforementioned probabilistic gender variable in that category.[43] 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

---

[40] We have limited our examination to white, black and Hispanic authors as these are the three largest races and ethnicities in the U.S., accounting for over 90 percent of the population. 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.

[41] *See, e.g.*, Form TX, at http://www.copyright.gov/forms/formtx.pdf (not requiring authors registering textual works to note their gender).

[42] 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, http://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html. *See* Frequency Occurring from Census 1990 – Names Files (containing files of male and female first names and their distributions).

[43] 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.

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. 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. Individual authorship of registered works is predominantly male, although it has decreased from 70 percent in 1978 to 64 percent in 2012.

The most striking statistic about authors' gender is that two-thirds of the authors in our study are male.[44] However, the data also shows a statistically significant time trend of increased female representation within the group of registering authors.[45] While the rate of male authorship was about 70 percent in 1978, it steadily dropped to about 64 percent in 2012.

What could explain the decreasing but persistent overrepresentation of males? Any simplistic biological explanation would be, among other things, difficult to square with the change over the 35-year period, since presumably biology cannot 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. It is important as well to recall that about 28% of the original valid registrations in our study period have only corporate authors for copyright purposes, and we don't know the gender of the people who actually created those works. Thus, it could be that women who create works are more likely than men to be employees of or work-for-hire contractors for companies that register works under corporate authorship. It may also be that for some reason, women register the works they have created less often than men (but have become better at registering overtime comparatively). As we have already

[44] N = 10,465,488 (number of works in our database in the six principal categories that have at least one individual author). Authors are 66.63 percent male (averaging out the average gender profile per registration).

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

suggested, our analysis may often raise more questions than it answers. We leave this question for future work by us and, hopefully, others.

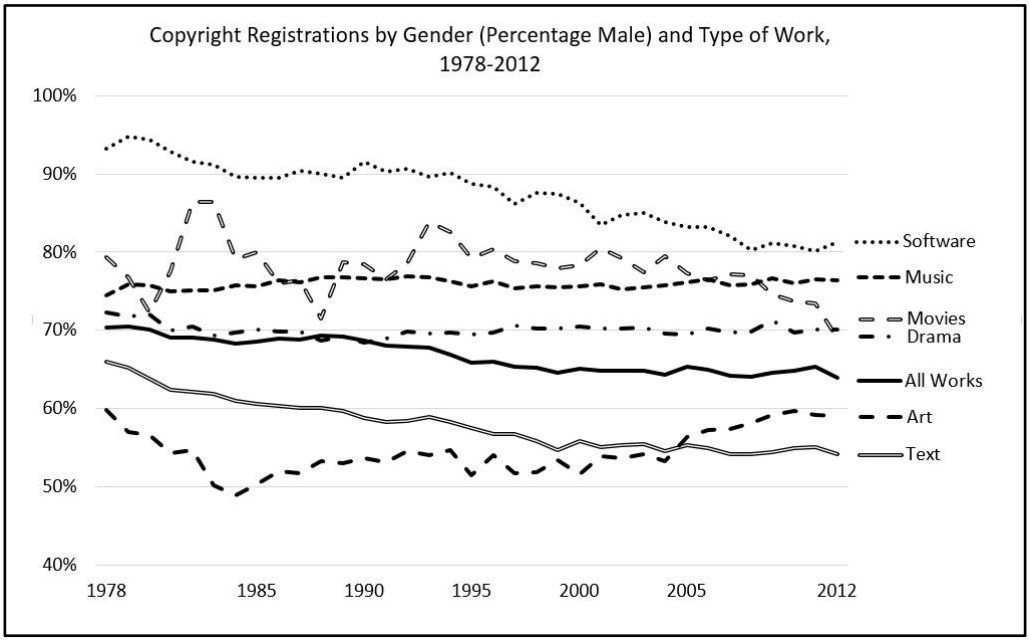
2. The gender of authors varies tremendously between types of works, from 54 percent male in art to 88 percent male in software.

The summary figures concerning the gender of all authors in the study mask a great variation in trends across the different work types. First, males and females differ in their patterns of authorship: some fields are more male-dominated than others. The work types sorted from the least to most male-dominated are art (54.34% male), text (57.45%), drama (69.99%), music (75.98%), movies (78.16%) and software (88.22%).

3. Gender trends across time also differ substantially among work types, with decreases in male share of authorship of over 10 percent in text, movies, and software, but very little change in music, art, and drama.

The degree to which the gender gap has or has not been bridged similarly varies by type of work. Figure 1 shows the trends in gender of author over the period of our study, as the percentage of authors who are male for the six major types of works as well as for all works combined. The increased female authorship trend is driven mainly by the text category, which experienced an increase in female authorship over the study period of 11.85%, from 33.98% to 45.83%, and which accounts for over a third of individual-author registrations.[46]

[46] The increased percentage of female authors of textual works overtime is statistically significant at the 0.1% level.



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 percent of all individual registrations.[47] At the same time, the music and drama categories, which account for about 44 percent and 5 percent of individual registrations, respectively, show a statistically flat time trend respecting female authorship.[48] Finally, the art category, which accounts for 11 percent of individual registrations, has a check-mark-shaped time trend with the percentage of male authorship generally decreasing from 1978-84 and then generally increasing to 2012. While the 1978 (59.8%) and 2012 (59.1%) percentages of male authorship are not markedly different, the time trend is one of statistically significant increase in male authorship.

4. Male authors register a somewhat higher proportion of unpublished works than female authors, and are on average a little younger, but these findings are heavily influenced by, and often reversed in light of, the differing types of works that male and female authors produce.

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 percent of works registered by men were published compared to 44 percent of works by women. If we considered those summary figures alone, we might

[47] These increases are significant at the 5 and 0.01 percent levels, respectively.
[48] Music shows a positive and insignificant time trend of male authorship, drama shows a negative and insignificant time trend of male authorship.

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 to register published works. The percentage of registrations for published works by males (females) in text is 66 (61) percent, and in music is 22 (15) percent. 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 principal types of works, the percentage of published works by males (females) in movies is 73 (63) percent and in drama is 7 (7) percent; 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. And 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.

5. Both men and women exhibit significant preferences for co-authors of the same gender.

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). 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 but for the first and treated that as the probable gender of Author 2. The probability of Author 1 being male is slightly lower than that of Author 2—a difference in means of 0.00071 that is statistically significant.[49]

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. As a result, 70.43 (29.57) percent of our Author 1 and 71.68 (28.32) percent of our Author 2 were male (female).

Of the registrations where Author 1 (Author 2) was male, 80.79 (79.39) percent of Author 2 (Author 1) were male as well. Of the registrations where Author 1 (Author 2) was female, 50.02 (52.23) percent of Author 2 (Author 1) were female as well. In this last sample, 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 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

[49] A t-test for the comparison of means came out with a t-statistic of -1.9574 that associated with a two-tailed p-value of 0.05. The alternative hypothesis that Author1 is more male than Author2 has a p-value of 0.975 and the alternative that Author1 is more female than Author2 has a p-value of 0.025.

probable gender of Author 2 given that Author 1 is female, the difference in means—negative 0.31—came out as statistically significant.[50] Males and females thus show a significant preference to co-author with similarly gendered co-authors.

III. 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. However, although almost all registration records contain information about the year of creation of the registered work, many registration records contain year of birth information for authors. In addition, year of birth and year of creation are sometimes entered inaccurately, so that subtracting the first from the second results in a negative number, or a number that can reach as high as 8000. We decided to filter the results, so that only values of between 0 and 100 (not including these numbers) would be counted as the age of an author. When a registration listed more than one author, we averaged the ages to obtain an average age of author for the works registered. 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 completely 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, the real average age of authors of all works in our study is almost certainly greater than the age we report. That should be kept in mind when assessing the results in this section.

B. MAIN FINDINGS

[50] 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.

1. The average age of authors of the major types of works in our study is about forty, and the ten most productive years for authors are between 27 and 36.

Across the six categories of works that are the focus of this article, the average age of all authors of works registered from 1978 through 2012 is just above forty – 40.12 years old. Author productivity rises relatively quickly as authors advance into their twenties and thirties, and then declines more slowly, so the average age of forty 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 4:

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 years 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 4 shows the productivity ratios for sixteen age groups, using cumulative figures for both registrations and U.S. population over the period from 1980 to 2012. (1978 and 1979 are not included because the Census Bureau used different age groups before 1980.) 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 20 to 59. (In other words, if we simply didn't include the population below 20 in the calculation of the ratio, and calculated the percentage of the total adult population represented in each of the groups above age 19, the ratios would be lower.)

If we separate registrations of unpublished works from those of published works, we find that authors of published works are, on average, about four years older than the average for all works – 44.10 – while authors of unpublished works are on average a year-and-a-half younger – 38.59. While we don't 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 gap in age 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.

2. *Creators of literary works are on average ten years older than creators of music; creators of other types of works fall in between.*

We have already seen how the race and gender of authors differ substantially across work types; the age of authors also differs substantially. 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. That should at least raise the question of whether music and literature involve different kinds of creativity

that peak at different ages. 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 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. *Authorship of music is also the most age-concentrated, with the ten years from age 24 to age 33 accounting for almost 36 percent of registrations; authorship of literary works is the least age-concentrated, with the ten years from age 33 to age 42 accounting for about 26 percent of all registrations.*

Creators of music are not only on average the youngest; production of music is also the most age-concentrated. As Table 5 shows, music creators are on average most productive from 24 to 33. Production by authors of those ages accounts for over a third of all music registrations – 35.77%. By contrast, in the most productive decade for authors of literary works – from 33 to 42, which overlaps by only one year with the most productive decade of music creators – those authors produce 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 5: Registrations by Age Concentration and Type of Work, 1978-2012

Work Type	Average Age of Authors	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 Works	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 below 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% or 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 35-year period from 1978 through 2012. However, there are substantial changes in the average ages of authors over that period, and we now turn to examining trends in author ages across time. As we noted above, the average age of all authors of registered works from 1978 through 2012 is 40.12. Yet authors have on average been getting older throughout that 35-year period. The average age of authors of works registered in 1978 was 37.63 years; by 2012, by 2012, that figure was 44.64, seven years older. Authors actually rose in average age slightly less than the US 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. [51]

[51] Data on median age was gathered from a variety of Census Bureau publications, including U.S. Bureau of the Census, Statistical Abstract of the United States: 1985, Section 1: Population, Table No. 27 (Total Population, by Age and Sex: 1960 to 1983) (for the median age in 1978 and 1979); U.S. Bureau of the Census, Statistical Abstract of the United States: 1995, Section 1: Population, Table No. 14 (Resident Population, by Age and Sex: 1970 to 1984) (for the median age from 1980 through 1994); U.S. Census Bureau, Statistical Abstract of the United States: 2004-2005, Section 1: Population, Table No. 11 (Resident Population by Age and Sex: 1980 to 2003) (for the median age from 1995 through 2000); U.S. Census Bureau, Statistical Abstract of the United States: 2012, Section 1: Population, Table 8 (Intercensal Resident Population by Sex and Age: 2001 to 2009) (for the median age from 2001 through

4. *In 2012, authors of published works were on average only 3.71 years older than authors of unpublished works, less than half of the age difference in 1978, which was 7.65 years.*

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; 35 years later, in 2012, that average age had increased to 43.75, an increase of over nine years. By contrast, authors of published 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. We are not sure what explains this convergence of ages. One possibility would be that, with the development of low-cost distribution over the Internet, copies of a larger percentage of works are being publicly distributed, even obscure works of younger authors that would not have been published in an earlier era. However, the percentage of registrations that were for published works in 1978 – 58.70% -- is actually slightly higher than the percentage of registrations for published works in 2012 – 57.59% -- which seems to run against any simple explanation along those lines. Thus, the age convergence remains in need of explanation.

5. *While the average age of software authors has increased by more than 10 years between 1978 and 2012, and the average age of literary-work authors by more than 8 years, the average ages of authors of movies, art, and dramatic works has increased by only 3 ¾ years, 4 ¼ years, and 5 ½ years, respectively.*

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.

2009); U.S. Census Bureau, Community Population Survey Data on Age and Sex, available at http://www.census.gov/population/age/data/cps.html (for the median age from 2010 through 2012).

6. *Even adjusted for the aging of the general population, registration production is spread much more evenly across adult age groups in 2012 than in 1980, and younger authors do not account for as large a proportion of registrations as they once did.*

Table 6:

Ratio of Percentage of Copyright Registrations to Percentage of U.S. Population by Age Group, 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	0.03	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	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 23-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.

As Table 6 shows, in 1980, authors of ages 25 to 29 were the most productive age group relative to their proportion of the population, and were producing registrations at a rate of 2.46 times that proportion. By 1990, the most productive age group was composed of authors of ages 35 to 39, who were producing registrations at 2.09 times their proportion of the population. A decade later, in 2000, the most productive group was the 45- to 49-year-olds, producing at a rate of 1.76 times their proportion of the population. Finally, in 2012, the most productive group was 60- to 64-year-olds, but they were producing registrations at a rate of only 1.50 times their proportion of the population, just barely edging out to younger age groups, and 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 7:

Ratio of Percentage of Literary Work Copyright Registrations to Percentage of U.S. Population by Age Group, 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	0.80	0.48	0.70	0.87

Table 7 is similar to Table 6, but it breaks out the figures for literary works alone. In 1980, four five-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. CONCLUSION: INCENTIVES AND ACCESS TO WHOM?

The Copyright Office Electronic Catalog is used on a daily basis to locate particular registered works and recorded documents, but it is also a remarkable yet heretofore untapped source of data about the copyright system and about American cultural production. In this article, we have focused in our analysis of the Catalog on the demographics of authorship. Establishing the race, ethnicity, gender, and age profiles of authors of text, music, art, movies, drama, and software helps us understand who produces American culture, and how American cultural production has changed over the past two generations. It also helps us understand the demographic profiles of the authors who benefit from the operation of the copyright system. And it raises important questions that we cannot hope to answer in this article. Why are Hispanic authors so dramatically underrepresented in copyright registrations? Why have gender disparities substantially decreased in the areas of text, movies, and software, but not in the areas of music, drama, and art? Why are the authors of literary works on average ten years older than the authors of music? We hope that future research will address these questions, and help us to gain further understanding of creativity, cultural production, and the copyright system.

Viewed against the background of the predominant theory of copyright law, our research reveals that copyright law has failed to address two important questions. The familiar predominant theory is the incentive theory, according to which copyright law involves a tradeoff between incentives and access. On the one hand, copyright law provides incentives to create as a way to overcome the public good market failure. Without copyright protection creators may not be able to appropriate a sufficient portion of the social value of their works, and therefore may be deterred from creating socially valuable works. At the same time, there is countervailing social interest in affording the public—be it users or follow-on creators—with adequate access to works created by others.

In charting this tradeoff, copyright doctrine recognizes, in substantial ways, that one size does not fit all. For example, copyright doctrine affords different bundles of rights and exemptions as it regards different types of works, and thus provides different incentives to create and access different types of works. Copyright law alters the bundle according to types of legal entities - individuals and corporations differ in the scope of their rights and

limitations (individuals enjoy moral rights[52] and inalienable rights of termination[53]) and their duration.[54]

Our empirical analysis suggests that copyright law has been oblivious to two questions: First, how do the incentives to create and access differ by the individual characteristics of individuals, and second, how should they? Take the first question first. We have seen that the average age of creators of music is about 36, compared to the average age of creators of text, which is about 46. This suggests that despite a facially uniform copyright term for individual authors, effectively music is protected for longer. Thus, society is holding out a greater carrot for those who create music. Society, at least at the margin, is signaling to authors that they should invest their efforts in some creative fields over others. Perhaps this disparate incentive scheme is good policy or not; our analysis brings this previously unseen aspect of our copyright system to light.

On to the second, normative question. Here we would like to make a modest normative claim: other things being equal, a diverse group of authors is superior to a homogenous group of authors, and copyright law should adopt policies that promote authorial diversity and minimize barriers to entry. 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.

If this much is agreed, then our findings suggest that we have work to do. Much 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, whether it is through reduced fees, stronger incentives to create, easier ability to access, or other measures.

It is also possible that courts could begin to address underrepresentation without additional legislation. Fair use, for example, is an existing equitable exception to infringement that is explicitly open-ended,

---

[52] 17 U.S.C. § 101 (excluding "any work made for hire" from the definition of a "work of visual art"); 17 U.S.C. § 106A (defining the scope of moral rights of authors of works of visual art).
[53] 17 U.S.C. § 203 (granting a right to terminate copyright transfers resecting "any work other than a work made for hire").
[54] 17 U.S.C. § 302 (setting different copyright terms for works created by individual authors and for works made for hire).

since the factors to be considered only "include" the four that are listed in § 107 of the Copyright Act.[55] Rather, as the Supreme Court stated in *Campbell v. Acuff-Rose Music, Inc.*,[56] it "permits and 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."[57] If the law should in particular be fostering creativity by underrepresented authors, then it could be appropriate to take into account that underrepresented status when determining whether use by an author is a fair use. In *Suntrust Bank v. Houghton Mifflin Co.*,[58] for example, the U.S. District Court for the Northern District of Georgia decided that a novel called "The Wind Done Gone" likely infringed Margaret Mitchell's novel "Gone with the Wind," and enjoined defendant Houghton Mifflin Co. from distributing it. The author of "The Wind Done Gone," Alice Randall, did not deny that she borrowed some elements from "Gone With the Wind." Indeed, the entire point of "The Wind Done Gone" was to retell the story of the classic novel of the Civil War American South from the point of view of a slave, rather than from that of the daughter of a white plantation owner. As Randall wrote, "[Gone With the Wind], more than any other work I know, has presented and helped perpetuate an image of the South that I, as an African-American woman living in the South, felt compelled to comment upon and criticize."[59] On appeal, the Eleventh Circuit vacated the District Court's injunction, and remanded for consideration of the fair use defense on the merits, signaling that it considered the case for fair use to be strong; the parties settled before the District Court issued any decision on fair use.

Neither of the Eleventh Circuit opinions mentioned that Alice Randall was an African-American woman, but the judges surely knew that she was, and may have informally weighed that fact in favor of the fair use defense. Perhaps that weight should become more formal. In otherwise close cases, in which market harm to the copyright owner is far from clear, the fact that a decision against fair use would blunt the expression of an author from an underrepresented group would seem a reasonable consideration that should tip the balance in favor of fair use. To be sure, more research is needed, but the problem can no longer be evaded. We must consider whether copyright law can afford to stick with facially neutral policies that turn out to have a disparate impact on (or mask) who gets to be an author.

Our findings also have implications beyond copyright law. Increasing overrepresentation of white authors is a warning signal, not just within

---

[55] 17 U.S.C. § 107.

[56] 510 U.S. 569 (1994).

[57] *Id.* at 577.

[58] 136 F. Supp. 2d 1357 (N.D. Ga. 2001) ("Suntrust I"), *vacated by* Suntrust Bank v. Houghton Mifflin Co., 268 F.3d 1257 (11th Cir. 2001) ("Suntrust II").

[59] Declaration of Alice Randall, at 1, Suntrust I, 136 F. Supp. 2d 1357 (No. 1:01 CV-701).

copyright, but within society at large. It suggests that policies outside of copyright law – educational policies, labor policies and others – may not be effective in generating authorship skills and opportunities across races and ethnicities, and that those policies may need to be reconsidered. Our research methods and findings also suggest that other areas of creativity are ripe for demographic review. 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.

We believe that our research is 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. This might include not only race, gender and age, but also income, education, residence, and other data. It also might include information about the natural persons who create the works of which business entities are technically the authors, since the legal fiction of corporate authorship should not obscure the fact that in all instances, it is actually human beings who are creating works protected by copyright. Better information will enable both analysis and action to achieve a more open and diverse authorship scene that would enable all to participate equally in shaping our common cultural lives.

.