Replication Paper: 'Race, Writing, and Computation: Racial Difference and the US Novel, 1880-2000'

Amanda Su

5/8/2020

0.1 Abstract

So, Long, and Zhu (2019a) determine that novelists marked as "white" versus "black' produce different narratological effects with respect to the interaction of race and religious authority, finding that black writers who cite the Bible are more likely to cite it in a social context compared to white writers who cite the Bible in their novels. I was able to successfully replicate the results of the authors' paper. For my extension, I decided to reconstruct the paper's primary model using a Bayesian approach. I found that the results of the model were largely the same as that of the original. This corroborates and strengthens the paper's conclusions about how race and writing intersect across more than a century of U.S. fiction.

0.2 Introduction

In their project, So, Long, and Zhu (2019a) sought to test their hypothesis that novelists of different races produce different narratological effects in their works with respect to the Bible. They sampled a corpus of texts written by authors of marked black or white racial identities and published between 1880 and 2000, selecting only canonical works as a control. After performing a sequence alignment to determine whether or not certain texts cited the Bible, the authors used content analysis to observe the contexts from which they extracted the Bible alignments and looked for moments of sociality within the "scene" (the Bible citation and the surrounding context), defining sociality as the presence of two or more characters engaged in a dialogue or interaction. The authors ultimately constructed a mixed model that explains whether or not a scene is "social" as a function of the author's gender, race, whether or not they cited the Bible as a control variable, and the interaction of the race and Bible variables, also accounting for the random effect of a single novel. Their results conclude black writers who cite the Bible are more likely to cite it in a social context compared to white writers who cite the Bible in their novels.

I was able to replicate all results found by So, Long, and Zhu (2019a). The authors generously made their data available alongside their paper at Harvard Dataverse.[1] I used R to complete my replication.[2] My replication paper is publically accessible in my Github repository.[3]

For my extension, I decided to reconstruct the paper's primary model using a Bayesian approach. While

[1] So, Long, and Zhu (2019b)
[2] R Foundation (2020)
[3] Github repository.

the authors used the glmer function to fit a generalized mixed-effects model, I instead use stan_glmer to fit a Bayesian generalized linear mixed effects model with group-specific terms from the rstanarm package.[4] The Bayesian model combines prior information from the original population of texts with evidence from information contained in the sample to guide the statistical inference process. I found that the results of my model were largely the same as that of the original, proving that the original results are even more robust than So, Long, and Zhu (2019a) initially claimed and corroborating their conclusions about how an author's race influences how they contextualize the Bible in their work.

0.3 Literature Review and Paper Review

So, Long, and Zhu (2019a) bridges the scholarly fields of cultural analytics (also known as "computational criticism") and critical race studies. Cultural analytics is an emerging field wherein humanist scholars leverage the increasing availability of large digital materials and the affordances of new computational tools, allowing them to survey semantic and narratological patterns in the English-language novel at the scale of centuries and across tens-of-thousands of texts, according to So, Long, and Zhu (2019a). Cultural analytics scholars have explored a variety of topics, including genre and cultural prestige, but the topic of race and racial difference has remained relatively understudied. While recent scholarship on the relationship between computation and race has been critique-oriented, pointing to computation's role in intensifying racial stratification and reinforcing existing patterns of social inequality,[5] So, Long, and Zhu (2019a) seek to determine both computation's affordances and its inadequacies in the study of race and literature. They specifically research whether computational methods can reveal if and how racial difference is expressed in literature — language, style, and narrative.

To select their sample of novels, So, Long, and Zhu (2019a) drew from a larger corpus constructed from a list of the most frequently held novels by American authors published between 1880 and 2000 as catalogued by WorldCat. They narrowed down the original 6,000 authors represented in the corpus to only those novels written by authors with marked racial identities, labeling the authors only if they identified in one particular way or if their identity was documented in the scholarly record. So, Long, and Zhu (2019a) then selected novels written by authors who identified as "black" or "African-American" to represent their "corpus of novels by black authors" and created a parallel corpus of "white" writers, which far outnumber black writers in the larger corpus, by selecting works that similarly skewed canonical. So, Long, and Zhu (2019a) then used a sequence alignment method to identify quotations of repetitions of specific lines and phrases to determine textual commonality between texts. Throughout this process, the authors acknowledge several biases in their methods. In selecting the corpus, they omit African-American novels which are not traditionally marked as "novelistic" to maintain the corpus's canonical skew. Otheriwse, the comparison between distinguished black writers with a sea of high and low white writers of all genres would distort their results. So, Long, and Zhu

[4] Gabry and Goodrich (2020)

[5] O'Neil (2016) and Noble (2018)

(2019a) also recognize that their crude, provisional identification of authors' racial identities is complicated by shifting social and historical circumstances and may have not have any implication on novels written under the sign of such identities. Generations of traditional literary and religious studies scholars have pointed to the Bible as the basis for the "Western cultural imaginary." Traditional scholars of literature and the Bible have long argued that even as the world of the novel has increasingly secularized, its commitment to religious ideas and language has persisted. Canonical literary scholar Northrop Frye conceived the idea of The Great Code to declare the Bible's universal commonality and significance.[6] A more modern inflection of this trope — "virality" — additionally explains the allure of the language of the Bible.[7] Indeed, the Bible possesses a distinct "resonance" that attracts both white and black writers. So, Long, and Zhu (2019a) seek to unsettle the narrative of the Great Code by excavating differences in how black and white novelists quote the Bible in their works. Their early attempts, however, were unfruitful. First, they assessed whether one group cited the Bible more frequently than the other by randomizing the race labels in their dataset and pulling from them a null distribution of quotation counts. They found that the actual amount of Bible quotation by each group was not significantly different from this null distribution. In other words, had they assigned the race labels randomly, they could have expected the same rates of quotation. They then tested whether or not white and black writers cited the Bible at different rates over time, finding that black writers did not explicitly cite the Bible more or less than white writers at any point in time. The authors then examined if novelists were citing different parts of the Bible, noting differences between the ideological orientation of the Old and New Testaments. However, after they analyzed whether chapters from either were being cited at different rates, the results were inconclusive. Then, they looked to the words surrounding the aligned Bible passages but could not conclude that white and black writers, as a whole, used a different vocabulary when invoking the Bible or discussed different topics.

So, Long, and Zhu (2019a) then readjusted their approach by examining whether writers differed in how they cited the bible by looking for moments of sociality, defined as the presence of two or more characters engaged in a dialogue or interaction. While not denying the centrality of the Bible in black communities and its frequency of citation among black writers, black studies scholars have argued black writers' invocation of the Bible usually occurs through a process of "critical modification and revision"[8]. This process takes several forms: irony, criticism, and dialogism. The latter refers to the Bible's mention as inciting dialogue among characters rather than occuring as a monologic polemic or sermon. Scholars argue that the Bible's appearance and quotation in novels by black authors tend to be very dialogic and interactive so as to question the Bible's normative or hegemonic "white" meaning. This understanding ultimately informed the creation of the authors' final model explaining the sociality of a text as a function of the novelist's race.

[6] Frye (1982)
[7] Prickett (1996)
[8] Valkeakari (2007)

0.4 Replication

To test their theory about novelists of different races producing different narratological effects in their works with respect to the Bible, So, Long, and Zhu (2019a) constructed a model that explains whether or not a text is "social" as a function of the author's gender, race, whether or not they cited the Bible, the interaction of the race and bible variables, and the random effect for each novel.

I was able to successfully replicate every aspect of the paper.

0.5 Extension

Table 1: Mixed Model Explaining the Fixed Effects of Author Gender, Race, Bible Citation, Race and Bible's Interaction on the Sociality of a Text and the Random Effect of Single Novels

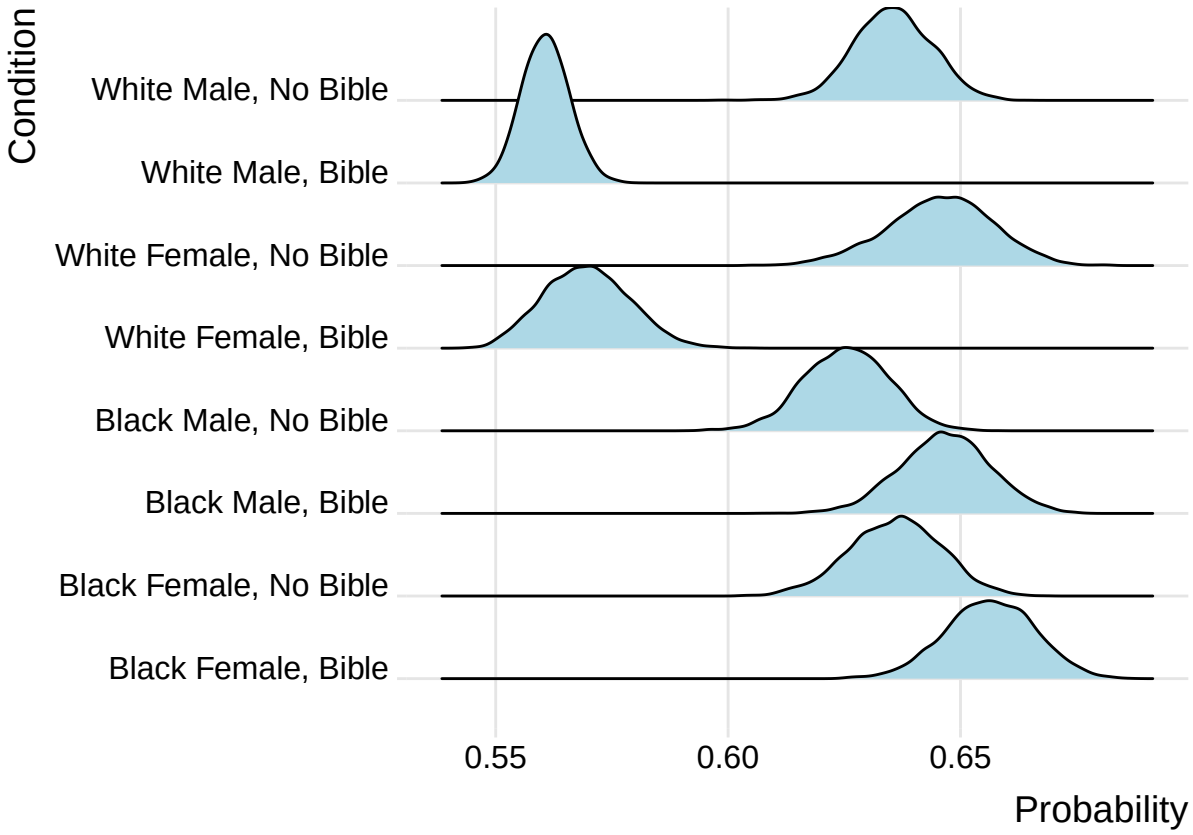
Statistic	Mean	St. Dev.
(Intercept)	0.6238750	0.3004405
gender	-0.3877192	0.2680429
race	-0.2363427	0.2853404
bible	-1.5162097	0.2541568
race:bible	1.9486787	0.3895490

I fit a mixed effects model explaining the fixed effects of an author's gender, race, their citation of the Bible, the interaction between race and Bible, and the random effects of single novels on the sociality of a text. I added "title" as a random effects variable as So, Long, and Zhu (2019a) did to ensure that no single novel might be contributing a disproportionate amount of Bible quotations and become the source of any specific effect. The "Bible" variable indicates whether the Bible is cited as a control to ensure that if there was an observed effect around gender or race in passages citing the Bible that this effect was tied to the fact that the Bible was being cited, and not that novels by white or black authors are inherently more "social." Whereas So, Long, and Zhu (2019a) decided to perform a maximum likelihood estimation of generalized linear models to determine the predicted values of model coefficients, I perform a full Bayesian estimation to find the average expected values for coefficients. Expected value averages are preferrable to predicted values because the latter contains both fundamental and estimation uncertainty, whereas the former only has to account for the estimation uncertainty caused by not having an infinite number of observations.ˆ[King, Tomz, and Wittenberg (2000)} As a result, predicted values have a larger variance than expected values. I ultimately found that the primary results of the original paper are largely unchanged even when using a Bayesian approach to create the model. Both models show that an author being male, the author being black, and the author citing the Bible negatively affect the sociality of a text while the main coefficient of interest — the interaction of race and bible or the author both being black and citing the bible while controlling for gender — has a positive effect on sociality.

Table 2: Fixed Model Explaining the Effect of Author Gender, Race, Bible Citation, and Race and Bible's Interaction on the Sociality of a Text

Statistic	Mean	St. Dev.
(Intercept)	0.4095083	0.2059359
gender	-0.1859540	0.1676205
race	-0.1711943	0.2102669
bible	-1.3585511	0.1806839
race:bible	1.7370224	0.2818996

I fit an additional fixed effects model explaining the effects author's gender, race, their citation of the Bible, the interaction between race and Bible on the sociality of a text without the random effects of single novels. Using stan_glm instead of stan_glmer with random effects allows me to construct the resulting posterior distributions using the model and find the predicted likelihoods of a text being marked as "social" given the following conditions: a white female author not citing the Bible, a white female author citing the Bible, a white male author not citing the Bible, a white male author citing the Bible, a black female author not citing the Bible, a black female author citing the Bible, a black male author not citing the Bible, and a black male author citing the Bible.



Graph 1: Distribution of Predicted Likelihoods of a Text Being Social Given An Author's Race, Gender, and Citation of the Bible in Their Work

This ridges plot visualizes the resulting posterior distributions generated from my fixed effects model (Table 2). A comparison between the Black Female, Bible and Black Male, Bible ridges and the White Female, Bible, and White Male, Bible ridges reveals that the predicted likelihood of a text invocation's of the Bible occurring in a social context is greater when it is a black author citing the Bible than when a white author is citing the Bible, as concluded by both So, Long, and Zhu (2019a) and my own models. A comparison between the four "No Bible" ridges reveals that there is little difference between the predicted likelihoods of a black author's text being social and the predicted likelihoods of a white author's text being social when the authors do not cite the Bible. This suggests that the difference between the narratological effects produced by authors of difference races is especially impacted by the authors' reference of religious authority.

0.6 Conclusion

So, Long, and Zhu (2019a) sought to test their theory about novelists of different races producing different narratological effects in their works with respect to the Bible. They constructed a mixed model that explains

whether or not a text is "social" as a function of the author's gender, race, whether or not they cited the Bible as a control variable, and the interaction of the race and bible variables, also accounting for the random effect of a single novel. Their draw their corpus from a list of most frequently held novels by American authors published between 1880 and 2000, paring the sample down to just those novels by authors with known racial identities and selecting only white and black canonical authors as a control mechanism. Their results conclude black writers who cite the Bible are more likely to cite it in a social context compared to white writers who cite the Bible in their novels.

I successfully replicated all results found by So, Long, and Zhu (2019a). The authors generously made their data available alongside their paper at Harvard Dataverse.[9] I used R[10] to complete my replication, which is publically accessible in my Github repository.[11]

I extended on So, Long, and Zhu (2019a) by reconstructing the paper's primary model using a Bayesian approach. I use stan_glmer to fit a Bayesian generalized linear mixed effects model with group-specific terms, which adds priors on the regression coefficients and allows me to update these initial beliefs in the evidence of new data. My model yielded largely the same results as that of the original model, proving that the original results are even more robust than So, Long, and Zhu (2019a) initially claimed and thus strengthening their conclusions about how an author's race interacts with their writing across more than a century of U.S. fiction.

While analyzing the results of their model, the authors discourage presuming the reality of racial categories and their correlation with particular literary effects, as if the categories themselves were causing these effects. Rather, they seek to interpret the race variable as a construct of social conditions. With this understanding, further research could stratify the data by time period and situate these novels within their historial contexts. After all, though this replication paper confirmed the authors' conclusions that black authors were more likely to cite the Bible in a social context than white authors, the models still draw from the original sample of American novels published from 1880 to 2000, generalizing the average predicted results across a century of texts and significant historical events. Accounting for different historical and social circumstances, I would test if the observed relationship between an author's race and their produced narratological effects remains consistent throughout the decades represented in the corpus and study the heterogeneous effects across time periods. I would be especially interested in scrutinizing the effect of an author's race on how they write about the Bible during relevant historical events, such as the Civil Rights movement in the mid-20th century and the Fourth Great Awakening in the late 20th century.

[9] So, Long, and Zhu (2019b)
[10] R Foundation (2020)
[11] Github repository.

0.7 References

Frye, Northrop. 1982. *The Great Code: The Bible and Literature*. Harcourt Brace Jovanovich: New York.

Gabry, Jonah, and Ben Goodrich. 2020. *Rstanarm*. https://mc-stan.org/rstanarm/index.html.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation" 44 (2). American Journal of Political Science: 347–61.

Noble, Safiya. 2018. *Algorithms of Oppression*. New York: NYU Press.

O'Neil, Cathy. 2016. *Weapons of Math Destruction*. New York: Crown.

Prickett, Stephen. 1996. *Origins of Narrative: The Romantic Appropriation of the Bible*. Cambridge University Press: University of Florida Press.

R Foundation. 2020. *The R Project for Statistical Computing*. https://www.r-project.org/.

So, Richard Jean, Hoyt Long, and Yuancheng Zhu. 2019a. "Race, Writing, and Computation: Racial Difference and the Us Novel, 1880-2000." Journal of Cultural Analytics.

———. 2019b. "Replication Data for: Race, Writing, and Computation: Racial Difference and the US Novel, 1880-2000." Harvard Dataverse. https://doi.org/10.7910/DVN/6ANTB8.

Valkeakari, Tuire. 2007. *Religious Idiom and the African American Novel, 1952-1998*. Gainsville, FL: University of Florida Press.

Appendix

All results from So, Long, and Zhu (2019a) were successfully replicated. As an example, here is Figure 3 from page 22.

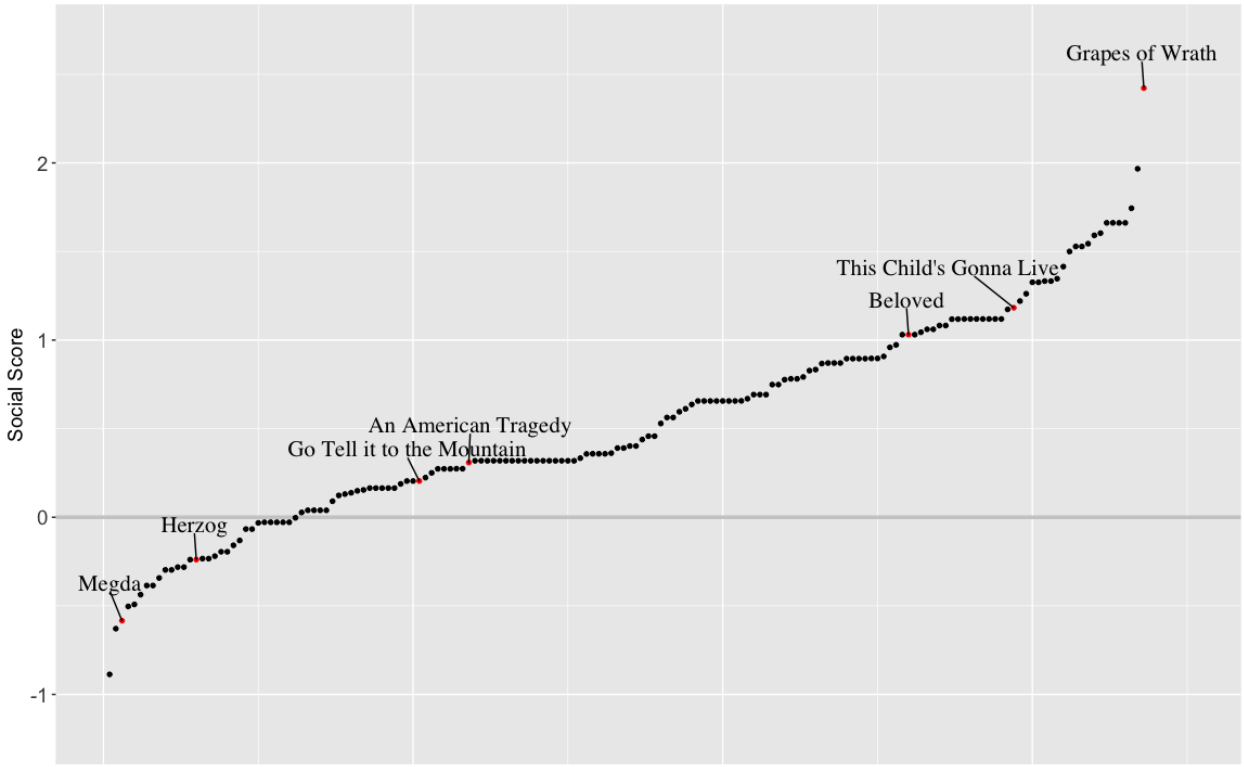
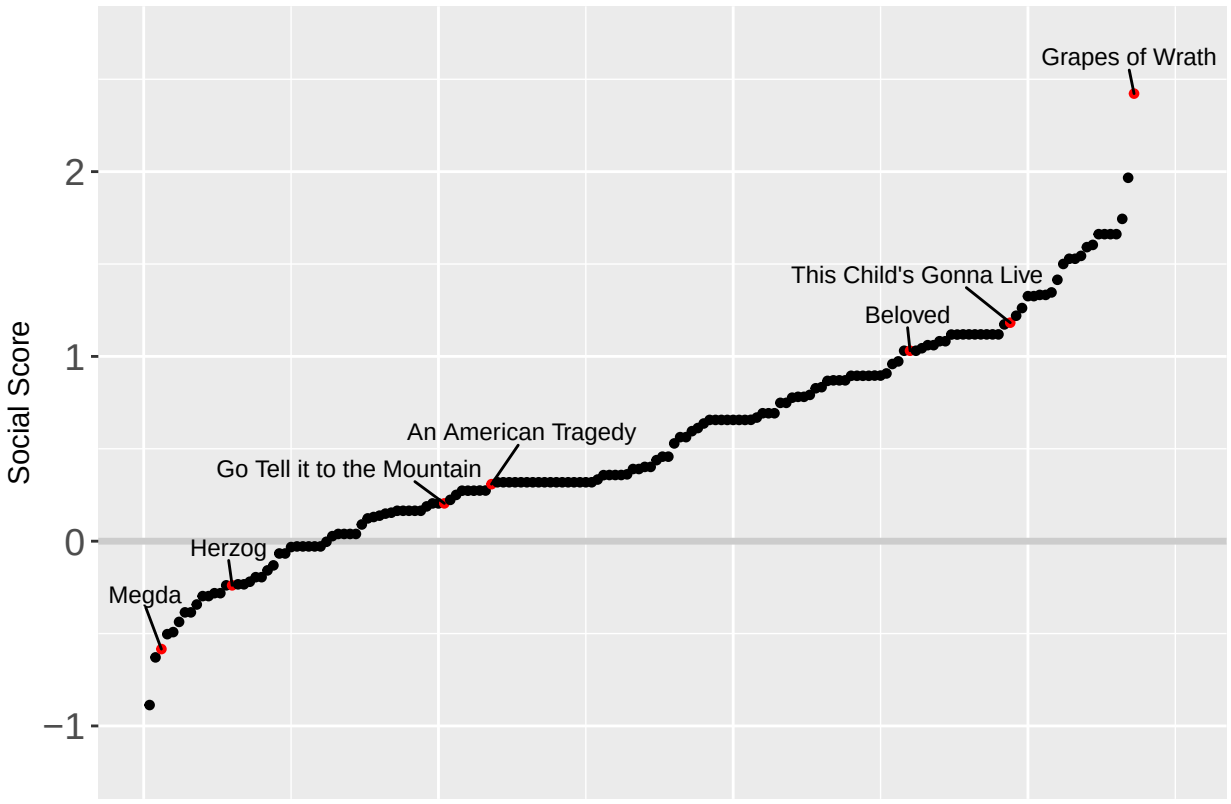


Figure 3: Plot showing the "social" score for all novels containing alignments with the Bible. Lower scores indicate novels where the Bible is less frequently cited in a "social" way, as we define the term. Scores closer to zero indicate novels where the "social" and "non-social" contexts are split evenly, as in James Baldwin's Go Tell it to the Mountain.

Here is my replication of the figure. All analysis for this paper is available at my Github repository.[12]

[12]Github repository.


Grapes of Wrath
This Child's Gonna Live
Beloved
An American Tragedy
Go Tell it to the Mountain
Herzog
Megda
Social Score
2
1
0
−1