Reviewing Millions of Books:

Charting Cultural and Religious Trends with Google's Ngram Viewer

Roger Finke

Pennsylvania State University

and

Jennifer M. McClure

Samford University

Prepared for the annual meeting of the Association for the Study of Religion, Economics, and Culture, Boston, MA, March 21, 2015.

The authors would like to thank Roger Geiger, George Marsden and Grant Wacker for helping educate us on early American higher education and Nathaniel Porter for his assistance in assembling data from the Google Ngram Viewer. This project was made possible through the support of a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation.

Reviewing Millions of Books:
Charting Cultural and Religious Trends with Google's Ngram Viewer

Despite the importance of trend data for understanding key substantive and theoretical questions on American religion, almost no such data exist. This paper reviews how the Google Ngram Viewer and the Google book collection can generate detailed measures on religious trends. After demonstrating how the trends are generated, we evaluate the promise and potential limitations of these measures. Do the trends charted by the Ngram Viewer match known cultural and religious trends and is the pool of book authors and readers representative of the larger culture? We conclude with a discussion on how the tool can be used most effectively for future research.

Reviewing Millions of Books:
Charting Cultural and Religious Trends with Google's Ngram Viewer

Despite the importance of trend data for understanding key substantive and theoretical questions on American culture and religion, almost no such data exist. Unlike economics, education, employment, crime and other areas consistently covered by the census or other government agencies, the data collections on culture and religion have been fewer in number and less systematic. The Gallup Polls, General Social Survey and National Election Studies chart a few trends over the last few decades and church membership data have been estimated going back to 1776 (Gaustad 1976; Finke and Stark 1986; Stark and Finke 1988), but there is no source of data that offers multiple measures over the history of America.[1]

For this reason the millions of books scanned into Google Books offer scholars a unique opportunity to quickly review trends that have been ignored by other collections. When combined with Google's Ngram Viewer, over two billion words and phrases can be quickly searched and tracked overtime. Moreover, the search tool is highly sophisticated for both single and multiple word searches. The Ngram Viewer allows users to limit their searches by customizing them to parts of speech and inflection, or expand the searches by making them case insensitive or using wildcard characters. Including more than three million books and primarily covering the time period from 1800 to 2008, this resource offers us a unique window into the life of America for more than 200 years.

Yet, despite the great promise of this collection, many questions remain on how it can and should be used. To what extent do the words and word combinations used in books offer accurate measures of cultural and religious changes? Despite the remarkable precision, speed

[1] The General Social Surveys and National Election Studies are available for download free of charge from the Association of Religion Data Archives (www.theARDA.com).

and scope of the searches, do the results reflect the culture as whole? Finally, do the pool of authors writing books and the audiences reading them vary overtime as literacy and education levels increase?

The purpose of this paper is fourfold. First, we introduce the Google Ngram Viewer and review both the promise and the potential limitations. Second, we evaluate how well the trends charted by the Ngram Viewer match known cultural and religious trends. Third, we assess how the pool of authors and readers of books has changed over time. And, fourth, we discuss how the tool can be used most effectively for future research.

Google Books and the Ngram Viewer

In December 2004, Google announced an initiative to digitize more than 15 million books and to make the contents available for searching. They initially partnered with the university libraries of Harvard, Oxford, Stanford and Michigan, as well as the New York Public Library. Within a few years, however, their library partnerships increased to more than 40 and now included Princeton, Cornell, Texas, Virginia, Wisconsin and California.[2] The inclusion of Harvard and Princeton was especially crucial for the study of American religion and culture, because of their early founding dates and because they were each founded as seminaries and continue to support seminaries.

By 2010, the initiative had reached the goal of digitizing 15 million books and a research note in *Science* introduced the world to Google's Ngram Viewer (Michel et al. 2011a).[3] The initial Viewer, developed in 2009, relied on a collection of 5,195,769 digitized books,

[2] For information on many of the participating libraries go the Google Books info page: https://www.google.com/googlebooks/library/partners.html (viewed 11/4/14). For additional information, including a timeline of related events, go to: http://en.wikipedia.org/wiki/Google_Books (viewed 11/4/14).
[3] This article was published in *Science* on Jan, 14, 2011, but an online version of the article was available on Dec. 16, 2010.

representing approximately four percent of all books ever published. The authors explained that the subset of books was selected based on the quality of scans (i.e., the optical character recognition) and having complete information on the date and place of publication. Due to problems with dating, all periodicals were excluded.[4] Google later released a 2012 version of their Viewer that was based on "more books, improved OCR, improved library and publisher metadata" and included several improvements to the Viewer software.[5]

Michel and his co-authors coined the new data collection and analysis techniques as "culturomics" (2011a:176). They devoted much of their research note to demonstrating how their new tools could chart changes in grammar and vocabulary overtime. This included the transition from irregular to regular verbs (e.g., burnt to burned) and the formation of new words as well as the words that became obsolete overtime. But the authors clearly viewed the new measures as documenting more than changes in lexicon or grammar; they proposed that the Viewer allowed the social sciences and humanities a method "to investigate cultural trends quantitatively" (Michel et al. 2011a:176).

The authors accompanied these bold claims with a high level of transparency on where and why problems might occur. The brief seven-page *Science* article was initially published with an 85-page online supplement (Michel et al. 2011b) and more documentation soon followed. Some concerns could easily be addressed.[6] For example, when presenting the trends in charts they divide the number of matches in a given year by the total number of possible words

[4] Michel and his co-authors addressed many of the most frequently asked questions about data collection and the Ngram Viewer here: http://www.culturomics.org/Resources/faq

[5] For more information, see: https://books.google.com/ngrams/info (viewed 11/4/14). By 2012 Google Books included more than 20 million books (http://googleresearch.blogspot.com/2012/10/ngram-viewer-20.html, viewed 11/4/14).

[6] For the online supplement to Michel et al. (2011a), go to: www.sciencemag.org/cgi/content/full/science.1199644/DC1. Additional information can be found at: http://www.culturomics.org and https://books.google.com/ngrams/info.

for that year. And, to avoid sudden surges and dips from one year to the next, they do a smoothing of the data across three years or more. There were other challenges, however, that were more difficult to overcome.

Perhaps the most challenging is maintaining synonymy in the searches. Words often share the same spelling, but have very different meanings, refer to different people or are used in many different ways. Once again, the Viewer offers several options for reducing this problem. The simplest is to use multiple words rather than a single one. Rather than doing a single search for classical, the user can do separate searches for classical music, classical period and classical conditioning. The searches also can be refined by specifying the part of speech or if the word is capitalized. This allows the user to distinguish between worship as a verb or worship as a noun and between Catholic or catholic. These and other search options allow for a continual refinement of searches, yet the challenge of maintaining synonymy remains. For example, the names of Daniel Webster, John Smith, Horace Mann, Adam Smith and even Abraham Lincoln are shared by many in American history.

Accurately charting cultural trends over time also poses challenges. Prior to 1800, Michel and his colleagues (2011a) note that the number of books is too small to reliably quantify and after 2000 the method for collecting books moves beyond libraries, with publishers now submitting books for inclusion. As a result, they limited their analysis to the time period from 1800-2000. The earlier years of the 1800s are still more prone to error due to the reduced number publications per year, the reduced quality of the scans and the more limited metadata on the publications (Michel et al. 2011b). Since Ngram trends are time specific, Google has paid careful attention to accuracy of the date-of-publication. They omitted from the corpus of literature books that contained works from multiple years, as well as journals and periodicals which were often

dated with the first year of the publication, not the actual date of the specific piece. They also omitted books with poor Optical Character Recognition (OCR) quality and with inaccurate language metadata. These filters greatly increased the accuracy of the Google Ngram data (Michel et al. 2011b:6-8).

The most serious concerns, however, are the possible biases of the corpus that are related to the cultural topics being measured or to the variable of time. The authors acknowledge that the corpus of books is limited by the books acquired and preserved by the libraries, raising concerns that the books might not represent the larger body of books or the culture more generally. Yet another concern about the corpus, though not addressed by the authors, is about the change in authors and readers overtime. As the rate of literacy increases, does this change both the pool of readers and authors of the books being published over time? Were the authors and readers of books in the early 19th century an educated elite that failed to reflect the larger culture? As we will review below this bias poses one of the most serious threats to interpreting cultural and religious trends in the 19th century.

Despite these potential limitations, especially for the earlier years of the collection, the promises of this massive new source of information remain many. Below we attempt to assess the potential of the Google Ngram Viewer for accurately charting historical trends and for future research.

Assessing the New Measurement Tool

The greatest challenge for assessing the validity of new cultural and religious trend measures is that we have few criterion measures for comparison. Hence, we first assess if the reported trends are in agreement with well-known historical events. Do the measures accurately reflect what we

know occurred? Next, we evaluate if the new data are sensitive to subtle changes in language. In particular, we will look at how the words "Pentecostal" and "Fundamentalist" (and related variations) have changed over time. Does the new tool allow us to capture more subtle shifts? In short, we want to assess the Ngram tool's ability to measure prominent as well as subtle changes in American religious history.

Documenting the Well-known

When trends are well marked by distinctive words or a series of words, the Ngram tools nicely reflect some meaningful and often important historical time periods or events. Not surprisingly, when Americans are at war with another nation, the opposing nation is mentioned more frequently in books. As shown in Figure 1, Germany always receives substantial attention in American publications, but the interest surges during World War I and II. The mention of Japan slowly rises as Japanese immigration begins in the late 19th century, falls after the passage of the Immigration Act of 1924 and then peaks near the end of World War II. Vietnam is seldom mentioned prior to the 1950s, surges during the 1960s and early 1970s and has remained relatively high since the end of the war – though never reaching the levels of Germany or Japan.

But the Ngram trends often reflect important historical changes and events even when the events aren't as significant as war. The trend line for Mormons and other variants of the group's name (e.g., Latter-day Saints, Mormon, etc.) offers an important example. As shown in Figure 2 the Ngram results display four clear surges in activity: the late 1830s and early 1840s, much of the 1850s, the late 1860s and early 1870s and the late 1880s. Each of these surges reflected important controversies between the Mormon Church and the larger culture.

Founded in the early 1830s, the Mormon Church receives little attention until after 1838, when the governor of Missouri, Lilburn Boggs, issued an executive order referring to the Mormons as enemies and stating that they "must be exterminated or driven from the state if necessary, for the public peace" (Arrington and Bitton 1992:44). Shortly thereafter 17 Mormons were killed by an angry mob. The Mormons then moved to Nauvoo, IL, where their prophet, Joseph Smith, quickly becomes the mayor and military leader of the rapidly growing community. But after Joseph Smith and his brother Hiram were imprisoned and killed in 1844, the Mormons began their march west in 1846. Following this departure, the trend line drops sharply. In 1852, however, when their second prophet, Brigham Young, publicly acknowledges the practice of polygamy within the church, the start of the second surge begins. By 1856, the Republican presidential candidate, John C. Freemont campaigned against the "twin relics of barbarism, slavery and polygamy," and the eventual winner, James Buchanan, would remove Brigham Young as the territorial governor and send troops to Utah resulting in what was known as the Utah War (as quoted in Hansen 1981:144).

The trend line dropped during the years of the Civil War, but the campaign against polygamy continued once the war was over. There was a short drop during the late 1870s, perhaps due to the 1879 Supreme Court ruling upholding the Anti-Bigamy Act of 1862, but the drop was short-lived. Once the Edmunds Act of 1882 and the Edmunds-Tucker Act of 1887 were in place, polygamy was prosecuted as a felony and a pledged allegiance to anti-polygamy laws was required for voting and holding office. In the late 1880s, the trend line quickly moves to the highest level in history.[7] But when the fourth Mormon prophet, Wilford Woodruff, issued a manifesto in 1890 declaring that he and the church would submit to the laws of the nation, the

[7] The Edmunds-Tucker Act was upheld by the Supreme Court on May 19, 1890 and Woodruff's manifesto was announced on Sept. 25, 1890 (Arrington and Bitton 1992:183; Hanson 1981:145).

trend line plummets (Hansen 1981; Arrington and Bitton 1992). Since 1890, the Mormon membership has skyrocketed from 188,263 worldwide to more than six million in the United States alone, but the attention received in books remains far below the 1890 peak.

The examples just offered demonstrate that Ngram measures can reflect the cultural attention given to well-documented historical events and trends. As expected, the measures for Germany, Japan and Vietnam all surged when America was at war with each country, demonstrating the tools ability to capture major historical events. More impressive, however, is that Ngram measures could detect the cultural attention given to less prominent historical events involving the Mormon Church. Each of the four 19th century peaks on the Mormon graph reflected prominent historical struggles between the Mormon Church and the larger culture. These examples demonstrate some of the promises, but many challenges remain.

Documenting Subtle Shifts in Meaning

Whereas the words Mormons, Latter-day Saints and other variants have consistently referred to a specific religious group, other religious terms change overtime. This change in meaning reduces the ability of Ngrams to provide a standardized measure for charting trends of a specific group over time; but it does help to uncover important changes in how words are used. For example, Ngram trends help to illustrate how the commonly used words Pentecostalism and Fundamentalist have changed in meaning and use over time.

Over the past few decades, Pentecostalism has been the most rapidly growing segment of Christianity. The growth has been most stunning in the global South (Jenkins 2002), but Pentecostal groups in the U.S. have also shown remarkable growth. The Church of God in Christ has become one of the largest, if not the largest, of the historically African-American

denominations and the Assemblies of God has gone from 6,703 American members in 1916 to more than three million in 2012.[8] Most of the contemporary Pentecostals trace their heritage back to the African-American holiness pastor William J. Seymour and his revivals at 312 Azusa St. in Los Angeles (Wacker 2001).[9] Beginning in 1906 the young movement placed a strong emphasis on the gifts of the Holy Spirit, including speaking in tongues and healing.

However, the movements arising from the Azusa Street revivals were not the first American religious groups to describe themselves as Pentecostals. Prior to the Azusa Street movements, the word Pentecostal referred to holiness movements affiliated with the Methodist Church. These groups began feuding with the Methodist hierarchy in the mid-19th century and several split from the Methodists in the late 19th century to form the Association of Pentecostal Churches. This group later merged with the First Church of the Nazarene in 1907 to form the Pentecostal Church of the Nazarene. But the use of Pentecostal in their name was short lived. Twelve years later, they dropped the word "Pentecostal" to avoid any confusion over their ties with the new "Pentecostals" who were speaking in tongues (Melton 1991). From the 1920s forward the word was conceded to the new Pentecostals associated with the Azusa Street revivals.

Figure 3 displays how the use of the word Pentecostal increased in the 19th century as the holiness movements increased in number and voiced concerns about the modernizing trends of the Methodist Church. But the most significant increases came after 1960 with the rapid growth of the new Pentecostals. Figure 3 also shows the rise in the use of the word Pentecostalism since

[8] For clergy, congregations, and membership totals for the Assemblies of God from 1925 to 2009, as well recent survey findings and a geographical distribution, go to theARDA.com (http://www.thearda.com/denoms/D_1021.asp). For the most recent statistical overviews, go to: http://ag.org/top/About/statistics/index.cfm

[9] Charles Fox Parham is credited with introducing these "new" Pentecostal teachings to Seymour and others, but it was Seymour's Azusu Street Revivals that were most influential in launching the new movement (Wacker 2001).

the 1960s. Whereas, Pentecost is in the vocabulary of all Christian groups and Pentecostal sometimes goes beyond those traditions referring to themselves as Pentecostals, Pentecostalism seems to be distinctive to the movements arising from the Azusa Street revivals. In short, the graph captures the attention given to two very different Pentecostal movements, but understanding the graph requires knowledge of American religious history.

The word Fundamentalist also offers an example of how both the meaning and the use of a religious word can change over time. Initially, Fundamentalist was used to describe those who held tightly to the teachings outlined in a booklet entitled *The Fundamentals* that was published between 1910 and 1915. A Baptist editor would later coin the word fundamentalist in 1920 to describe those who are willing "to do battle royal for the Fundamentals" (Marsden 1980). Figure 4 shows that initial use of the term as a noun rises after 1920, falls in the early 1930s, and then shows a gradual increase. The use of fundamentalist as an adjective shows a similar start, first arising in the 1920s and having a similar level of use. But the trend lines begin to show a substantial separation in the 1950s. By 1970, the rate of using fundamentalist as an adjective was over six times higher than the use as a noun. By 1990, the rate was over 10 times higher. This suggests that the early usage of the word fundamentalist frequently referred to specific groups or people, but the word is now being used far more frequently to describe a type of group or person. Although not shown in Figure 4, it is also interesting to explore what words are paired with fundamentalist. When used as an adjective, fundamentalist Christians are the most frequent pairing since the 1980s. When used as a noun, however, Christian and Islamic have similar rates of describing a fundamentalist.

These two examples serve to illustrate both the challenge and opportunity of this research method. On the one hand, the challenge of synonymy is clearly evident. Words can change in

both their meaning and in how they are used over time. Thus, Ngrams can sometimes fail badly at offering a standardized measure over time. On the other hand, the Ngram searches can offer important insights into how the meanings and use of words have changed over time. Pentecostal (and related terms) referred to two different religious movements at two different points in time. The Ngram trend line captured both the growth of the holiness/Pentecostal movements associated with Methodism in the 19th century as well as the movements associated with the Azusa Street Revivals and promoting speaking in tongues and miraculous healing in the 20th century. The Ngram tool also helped to measure how the Fundamentalist (and related terms) has been used over time. Once frequently used as a noun referring to a specific group of religious believers in the 1920s and 1930s, by the close of the 20th century it was about ten times more likely to be used as an adjective to describe a group. Although failing to offer a standardized measure over time, each of these measures helps to identify important changes in American religious history.

Assessing the Corpus of Books

Given that even the most carefully researched books rely on "only" a few hundred other publications, it is easy to be impressed with the millions of books instantly accessed by Google Ngrams. Yet, even the most basic text on research methods warns that a large data collection does not ensure a representative collection. To what extent does the corpus of books reflect the interests of a broad cross section of the society or only a select stratum? One of the most frequently mentioned concerns is if the corpus represents the books actually being read? Since each book is given equal weight, the measure fails to account for the number of books actually

published and read for that year. Moreover, because the books are primarily drawn from major university libraries questions could be raised about the books being selected.

For religion and culture, however, one of the most significant biases might occur from the changes in the authors and readers of books over time. To what extent do the books represent the interests of the entire population or the interests of cultural and educated elites? As late as 1870, when the Department of Education offers their first statistical report, only two percent were high school graduates (Bureau of the Census 1975), and 20 percent of the population was estimated to be illiterate, compared to 0.6% in 1979 (Carter et al. 2015). The rate of literacy was obviously even lower as we move to the early years of the 19th century. How did this change in the potential pool of authors and readers change the ability of the books to serve as a measure of culture and religion? This potential bias raises at least two significant measurement concerns for the study of religion. First, to the extent that a bias is closely related to the variable of time, it will reduce the measures ability to accurately chart religious trends over time. Second, if the bias is related to a topic being studied, it will distort the perceived relationships religion holds with other topics of interest.

Below we review the distinctive problems this sampling concern poses for religion in the 19th century. We will first chart a couple of Ngram trends on religion and then we offer an assessment on how these trends are distorted by the changing pool of authors and readers.

Reviewing the Trends

When Michel and his colleagues reviewed trends on the mention of God in books, they noted that "God is not dead but needs a new publicist" (2011a: 182). Given the results shown in Figure 5, it is hard to challenge their conclusions. The trend line shows a continual drop from

1840 until about 1920 when it begins to plateau and eventually shows a modest rise at the end of the 20th century.[10]

Moreover, this trend is not limited to a single religious term. When we limit our search to terms more specific to Christianity and limit our attention to the latter half of the 19th century, a similar pattern is displayed. As shown in Figure 6, the mention of "Jesus" shows a similar decline throughout the 19th century and continues to decline until the 1930s. Although the trend line for "Jesus" does rebound after the 1970s, it never approaches the high rates shown earlier. Both God and Jesus seem to need a new publicist.

What is initially perplexing, however, is that atheism needs a new publicist as well. Like the trends shown for God and Jesus, the trend line for atheism shows a substantial drop in the latter half of the 19th century. As shown in Figure 7, atheism shows modest rebounds in the 1930s and 1960s, but the trend line ends the 20th century at one of the lowest levels for the entire 160 year span. So what is happening?

Rather than attributing these changes to a declining interest in God, Jesus or atheism, we suggest that the change can be attributed to the close relationship religion held with higher education in early America. Throughout much of the 19th century, early American colleges were largely staffed by clergy, financially supported by religious denominations and frequently established to help train clergy. As a result, the concerns and conversations of the highly educated clergy played a prominent role in determining both what was written and read in early America.

[10] The authors began their graph in 1800 and warned that prior to this date "there aren't enough books to reliably quantify many of the queries" for English only sources. Since we restrict our search to American English books, our total collection for the early 19th century is greatly reduced. To avoid erratic fluctuations due to the small number of books in the early 1800s, we begin our graph in 1840. For additional information on the Ngram Viewer and to review the full quote given above, go to: http://www.culturomics.org/Resources/A-users-guide-to-culturomics.

A Shifting Pool of Authors and Readers

Prior to the Civil War the development of colleges in America was dominated by religious denominations. Colin B. Burke's (1982) detailed data collection on *American Collegiate Populations* found that from 1810 to 1860, 82 to 86 percent of college students were enrolled in colleges supported by religious denominations. One of the most obvious reasons for this dominance is that the religious groups were one of the few reliable sources of financial support for higher education (Ringenberg 1984; Marsden 1994; Burtchaell 1998; Thelin 2011). Burke explains that even the state colleges relied heavily on endowments "because no college before the Civil War was tied to a fixed and adequate tax base" (Burke 1982:44).

But the tie between religion and higher education in early America went far beyond financial arrangements. The faculty of these early institutions was dominated by clergy (Ringenberg 1984; Marsden 1994; Burtchaell 1998). In part, this is because many of these institutions were initially founded to train clergy for the ministry, but it also was the result of clergy being far more educated than the general population. The three dominant colonial religions, Congregationalists, Presbyterians and Episcopalians, all expected their clergy to have seminary educations at a time when literacy was low for the population as a whole (Finke and Stark 2005). The influence of the clergy and religion was even felt at the state schools. Historian William C. Ringenberg (1984:81) reports that "state universities almost invariably required students to attend chapel services and Sunday religious exercises. In many cases these requirements continued through the end of the century."

The Morrill Act of 1862, however, signaled a dramatic shift in higher education, both in the source of support and the training offered. The initial Act provided for the development and

support of land-grant colleges and this support continued to expand over the years that followed. But the Act also specified the type of education that should be provided. Although allowing for scientific and classical studies, the focus was on agriculture, the mechanic arts and training that would "promote the liberal and practical education of the industrial classes in the several pursuits and professions in life."[11]

Changes also were evident in higher education at the private schools during the final decades of the 19th century. One of the most significant changes for our measures was the waning influence of the clergy. This decline was clearly evident in the governing boards, faculty and administrators of the prominent Ivy League schools. The practice of college presidents being clergy ended at Harvard in 1869, at Yale in 1899 and at Princeton in 1902 (with the appointment of Woodrow Wilson). But this decline went far beyond the Ivy League. A study of governing boards at 15 private colleges found that clergy representation dropped from 39 percent in 1860 to 23 percent in 1900 and to seven percent by 1930 (Ringenberg 1984:127).

The training being provided at the private schools was beginning to change as well. Whereas many of these schools were founded for the training of clergy, the percentage of students training for the ministry was dropping. Using data from Burke's collection, Figure 8 charts the percentage of higher education enrollments that are in theological schools (1982:216). The combination of increasing numbers entering higher education (Burke 1982) and of the most rapidly growing American denominations not requiring seminary training for their clergy (Finke and Stark 2005) resulted in a sharp decline. As might be expected, the 1840-1900 trend line for percentage enrolled in theological schools looks remarkably similar to the Ngram Viewer rates

[11] For a transcript of the Morrill Act of 1862, go to: http://www.ourdocuments.gov/doc.php?doc=33&page=transcript.

for references to God, Jesus, and atheism, holding correlations of 0.82, 0.55 and 0.62 respectively.

Even this brief review and limited data helps to identify an important limitation for using the Ngram viewer to chart religious trends. Google's expansive collection offers an accurate picture of what was published, but the books fail to accurately represent the population as whole during most of the 19th century. Because of the clergy's high level of education and the prominent role they and their denominations played in the higher education of early America, the corpus of books held in these college libraries included many volumes touching on religion. Referring to the late 18th century, Ringenberg (1984:49) noted that because "clergymen donated most of the volumes, the libraries emphasized theology." The literary societies of the early 19th century would soon expand the collections, yet religion remained an area of focus.[12]

But the lack of representativeness also raises concerns about the books of early America accurately representing the religious interests and concerns of all Americans. Just as many secular interests of the culture might have been underrepresented by the books being published, many religious groups might also be ignored. To what extent are the Baptists, early Methodists and others not relying on a highly educated clergy (Finke and Stark 2005) underrepresented in the corpus of books? Even though religious topics are overrepresented in the early corpus of books, the religious authors and readers primarily came from only a few denominations. The clergy of select denominations remained highly influential in what was read and written in much of the 19th century.

Prospects for Future Research

[12] For an excellent overview of the dramatic transformation of educational institutions during the 19th and early 20th century, see David P. Baker's *The Schooled Society: The Educational Transformation of Global Culture,* 2014.

Many of the sampling concerns just raised, as well as the measurement issues discussed earlier, are not distinctive to the Google Ngram Viewer and the corpus of books it draws on. Like the Ngram Viewer, historians often rely on a subset of written sources. They have long debated how their accounts are limited by the sources available, with more recent accounts attempting to draw on new sources and methods for studying religion (Taves 2011). Likewise, the problem of synonymy is not limited to the measures used by the Ngram Viewer. Social scientists and historians struggle to understand how meanings might vary over time and across cultures. For survey research, and especially cross-cultural survey research, synonymy remains one of the greatest measurement challenges (see Smith 2017).

So, can the Ngram Viewer be used for research? We propose that the Ngrams Viewer can and should be used. Building on the examples just reviewed, we offer an assessment on the potential pitfalls and promises of this research tool. We suggest that the Ngram Viewer should be complemented with other research methods and propose that the Google book collection offers a readily available resource for understanding and interpreting the Ngram Viewer results.

Breadth of coverage

As just reviewed, scholars need to be aware of the overrepresentation of religious topics in the books of the 19th century. Because of the close relationship higher education and literacy held with religious groups and the highly educated clergy, both the authors and the readers of early American books were more likely to address religious topics. This produces misleading trend lines over time during the 19th century and underrepresents the interests of secular groups and the upstart religious sects and churches during this time period. Based on our preliminary assessment, however, this bias in the corpus is sharply reduced by the end of the 19th century.

Despite these concerns, however, the Ngram Viewer's impressive breadth of coverage offers great research promise for developing new measures and new insights. The size of the collection is the most impressive. Whereas, many historical accounts are limited by a small set of written sources that are often confined by region, religion or time period, the Ngram Viewer includes 3.4 million books covering diverse geographic areas, topics and time periods of America. But it is the diversity of the sources of books included in the corpus that is the most important. This diversity ensures that if a view was published it will most likely be included.

The benefit of drawing on a diverse set of sources, rather than a single collection or small group of sources, can be illustrated if we return back to the Mormon example discussed earlier. Using the recently developed Chronicle tool for searching *New York Times* news coverage throughout its history (http://chronicle.nytlabs.com/), Figure 9 charts the coverage given to Mormons.[13] Unlike the Ngram Viewer trend line, which accurately reflected multiple significant conflicts in Mormon history, the *New York Times* coverage shows a single sharp spike in the late 1850s that overshadows all other time periods. Thirty-five out of one thousand articles included the mention of Mormon, Mormons, Latter-day Saints, or LDS in 1858, but the rate was only 10 in 1,000 in 1870 and dropped to one in 1,000 in the late 1880s, when the federal government was seizing church assets and Mormon polygamists were being arrested following the passage of the Edmunds and the Edmunds-Tucker Acts (Hansen 1981:144-145). The *New York Times* coverage points to an important and newsworthy time in the 1850s, but other significant events are almost completely missing.

[13] The Chronicle (http://chronicle.nytlabs.com/) offers search features and graphing tools similar to the Google Ngram Viewer.

Even a highly respected news source cannot come close to matching the coverage or diversity of the Google book collection.[14] Because the breadth of the collection searched by Ngrams is so vast, it is less influenced by the distinctive interests of any single source. The views of a single editor or author or the interests of a select region or religion are less likely to fully shape the story told. Size alone does not ensure a representative description of the entire culture, but the vast size of the Google book collection does ensure that the publications held in libraries are represented.

Refining the Measures

When using the Ngram Viewer for research, the greatest challenge remains synonymy. Finding Ngram searches that identify a distinctive religious feature, person, event or group poses a major challenge when constructing religion measures. As a result, virtually all Ngram searches will require some refinement for improving the measures. As readily acknowledged by Michel and his colleagues (2011b:34), words often share the same spelling, but have very different meanings. Some words hold multiple meanings regardless of the time period, others change meaning or refer to a different person or group at different points in time. The Pentecostals and Fundamentalists examples illustrated how meaning changes can sometimes offer important insights into religious changes. When attempting to construct standardized measures over time, however, we want the meanings and the measures to remain consistent.

There is no easy or single solution for addressing the challenge of synonymy, but there are several ways the measures can and should be refined. First, like any research design, clearly

[14] An even broader online collection of news sources, named Chronicling America, is available through the Library of Congress (http://chroniclingamerica.loc.gov/) and is produced by the National Digital Newspaper Program (NDNP). Unlike Google Ngrams and the *New York Times*, however, this collection does not automatically break out results by year and depict them in a graph. Chronicling America currently provides information on American newspapers from 1836 to 1922.

defining the concept being measured is essential. What is included in the definition and what is excluded? This might seem obvious, but it is an essential first step for refining the searches that follow. Moreover, this step helps to assess if the concept can and should be measured with the Ngram Viewer.

A second method for refining the measures is effectively using the power of the Ngram Viewer. Although it is fun and somewhat addictive to do quick searches in the Ngram Viewer, generating precise measures typically requires a more careful use of the parameters available. The Viewer currently offers the following options for refining searches: wildcard search, inflection search, case insensitive search, part-of-speech tags and ngram compositions. Each of these options is explained in a lengthy footnote on the Ngram Viewer site (https://books.google.com/ngrams/info) and even more information is given on the Culturomics website (www.culturomics.org). As noted earlier, our use of part-of-speech tags helped us to better understand how the use of the word Fundamentalist changed over time and our use of a 2-gram (or bigram) allowed us to understand how the word was used differently when describing Christians and Muslims. In short, refining the measures requires a knowledge and understanding of the Ngram Viewer.

Refining the measures and improving their precision also requires knowledge of American religion. This knowledge improves the definitions offered, as well as the terms and parameters used for conducting searches. Like any research project, the Ngram searches are an iterative process. Some searches fail to offer a distinctive Ngram, or the meaning is unclear until additional searches are conducted. Holding knowledge of American religious history places both the search and the results of the search within the larger historical context. For example, the findings shown in Figure 3 on 19th century Pentecostals are meaningful if the researcher knows

that early Methodist holiness movements referred to themselves as Pentecostals. Without this knowledge, however, the results are confusing and meaningless.

Generating Quantitative Measures

Perhaps the most obvious use of the Ngram Viewer for future research is to develop quantitative measures of religion and culture. All of the data points shown in the Ngram Viewer graphs can be saved to a data file and merged with other measures over time for analysis. In an effort, to democratize access to these trend data, we created a dataset with more than 400 Ngram variables and 20 historical trend variables for dissemination from theARDA.com. The Ngram variables include both data for specific religious terms and composite data, where scales are created out of similar words (e.g., Atheist scale = atheist + Atheist + atheism + Atheism). The historical component of the dataset was drawn from many sources, the most common being the Historical Statistics of the United States (Carter et al. 2015). It contains a few of the measures collected by the census, including total population, GDP, median age, sex ratios and immigration. When available, it also includes measures on education and clergy training, such as the general levels of education, private institutions of higher education, and seminaries. See the Appendix A to this chapter for a list of all of our historical variables and their source.

By combining existing historical data sources with the Ngram measures, we are hoping to stimulate new lines of research. Yet, we offer the new measures with important precautions. First, the data file offers some simple Ngram and bigram searches, but for most research projects the measures will need to be refined. Our goal was to simply offer initial measures for some common religion concepts or terms. Second, like many trend data files, measures included in the file are often highly correlated, with numerous correlations above 0.9 or below -0.9. This high

level of multicollinearity should be taken into account when selecting the models and analysis used with data.

The most important precaution, however, is that researchers should seek additional information for understanding what the measures mean. For example, is increased attention due to fame or infamy? As noted earlier, an indepth knowledge of American religion will help to understand many of the measures, yet more information is often needed. Below we review how the Ngram Viewer and the Google book collection offer access to the primary sources used in the searches and how this can reveal more about the measures given.

Mixing Methods to Offer Meaning

Social scientists have long acknowledged that no single research design is the "right" choice for all research projects. Experimental designs offer advantages in isolating the independent variable and assigning causal ordering, surveys allow researchers to generalize findings to a larger population of interest, and field research allows for more in depth interviewing and observations in a natural setting. In most cases a mixed methods approach offers a far more complete understanding than any single method can provide (Axinn and Pearce 2006). The Ngram Viewer and the Google book collection offer unique resources and opportunities for mixing research methods.

Assigning meaning to Ngram measures is often difficult. Knowing that a religious person, group, belief or event is receiving more mentions in published books tells us little about what is being said. For example, Figure 2 shows the fluctuations in attention received by Mormons in 19th century books. But who was writing these books and what were they saying? Were they apologetic accounts written by Mormons, attacks on Mormons from the anti-

polygamy movements or something completely different? Answering these questions requires researchers to return to the books being searched.

Fortunately, the Ngram Viewer offers a complete bibliography of every book containing the Ngram(s) being searched. As shown in Figure 10, when the results are presented for the Ngram "Mormons," links at the bottom of the page allow the researcher to review a bibliography for each of the time periods listed. In the "Mormons" example shown, the researcher can link to the full citations of 300 books published between 1820 and 1857, 270 books published between 1858 and 1883, 221 books published between 1884 and 1890, and so on. When attempting to understand and interpret the findings for a select historical period, the researcher can quickly review citations for the actual sources.

But the Ngram Viewer offers much more than a full bibliography. For many of the books cited, a single click takes the user to an ebook copy of the book or document. At the top of the document a small tool bar lists the number of times the Ngram is used in the book and allows the reader to easily jump from one mention of the word to another. Moreover, the researcher can quickly do searches for additional words of interest and once again can jump from one use of the word to another. As a result, detailed coding and reviews of the documents are greatly simplified for the ebooks. For example, a quick search can answer the question: did the books mentioning Mormons also mention polygamy? If so, how often was it mentioned and did the author support the practice?

Unfortunately, not every book cited is available as an ebook. For our example on the Ngram "Mormons," about 40 percent of the books from 1820-1890 were available from a link in Google Ngram. Yet, even this level of coverage greatly reduces the work needed to find and review the primary sources. Moreover, the complete bibliography allows researchers to request

additional books to secure a more complete and representative coverage. The Google book collection and Ngram Viewer offer new options for sampling sources, reviewing the text and coding the documents. But even if the review of the source materials is less systematic and complete, it can assist researchers in understanding the data generated.

Conclusions

The study of religion and culture more generally is handicapped by a lack of measures that can document changes over time. Google's Ngram Viewer, and the "culturomics" proposed by its developers, offers one option for filling this void. This paper has introduced and assessed the Ngram Viewer and the 3.4 million books it reviews for America.

We find that although the problem of synonymy in the searches remains a significant problem, the tool has great promise for future research. The Ngram Viewer was able to chart significant historical events in the history of the Mormon Church and captured subtle changes in meaning for the use of the words Fundamentalists and Pentecostals over time. The Mormon example illustrated how the Ngram Viewer could provide a standardized measure for some ngrams and the Fundamentalists and Pentecostals examples helped to demonstrate search parameters could help to capture changes in meaning over time. We propose that when knowledge of American religion is combined with a careful use of the Ngram Viewer search parameters, the synonymy of the searches are often improved and the subtle changes are more effectively detected.

We also conclude that the corpus of books has limitations not acknowledged by the developers. Whereas the corpus reflects the books placed in college libraries, the books fail to represent the entire culture, especially when literacy is low and books are not readily accessible.

This conclusion is especially important for religion measures. Because of the close tie between select religions and higher education in early America, the books secured in college libraries failed to fully represent the entire culture. Secular interests, as well as those of the rapidly growing upstart religions, are not fully represented.

Finally, we propose that the use of mixed methods will greatly improve our understanding of the measures generated. Whereas, most of our attention focused on the Ngram Viewer's quick production of quantitative measures over time, the tool provides ready access to the primary sources. Some are available with a single click as an ebook and all are listed in a bibliography for the time period of interest.

In summary, the Google Ngram Viewer is a powerful and promising research tool that has the capacity to generate large volumes of data for the 19th and 20th centuries. This tool is especially promising for religion, given the lack of measures consistently available over time. Yet, we found that the power of the tool must be balanced with an understanding of the corpus' limitations and the meaning of the measures. Like all research methods, the measures provided by the Ngram Viewer provide the greatest insights when complemented by the findings of other research methods.

Bibliography

Arrington, Leonard J. and Davis Bitton. 1992. *The Mormon Experience: A History of the Latter-day Saints, 2nd Edition.* Urbana and Chicago: University of Illinois Press.

Axinn, William G. and Lisa D. Pearce. 2006. *Mixed Method Data Collection Strategies.* New York: Cambridge University Press.

Baker, David P. 2014. *The Schooled Society: The Educational Transformation of Global Culture.* Stanford, CA: Stanford University Press.

Bureau of the Census.1975. *Historical Statistics of the United States, Colonial Times to 1970, Bicentennial Edition*. Washington, D.C.: U.S. Government Printing Office.

Burke, Colin B. 1982. *American Collegiate Populations: A Test of the Traditional View*. New York: New York University Press.

Burtchaell, James Tunstead. 1998. *The Dying of the Light: The Disengagement of Colleges and Universities from their Christian Churches.* Grand Rapids, MI: Eerdmans Publishing Company.

Carter, Susan B., Scott Sigmund Gartner, Michael R. Haines, Alan L. Olmstead, Richard Sutch, and Gavin Wright, editors. 2015. *Historical Statistics of the United States: Millennial Edition Online*. New York: Cambridge University Press. Retrieved March 6, 2014. (http://hsus.cambridge.org/HSUSWeb/HSUSEntryServlet).

Finke, Roger and Rodney Stark. 1986. "Turning Pews into People: Estimating 19th Century Church Membership." *Journal for the Scientific Study of Religion* 25(2): 180-192.

Finke, Roger and Rodney Stark. 2005. *The Churching of America, 1776-2005: Winners and Losers in our Religious Economy, Second Edition.* New Brunswick: Rutgers University Press.

Gaustad, Edwin. 1976. *Historical Atlas of Religion in America.* New York: Harper and Row Publishers.

Hansen, Klaus J. 1981. *Mormonism and the American Experience.* Chicago: The University of Chicago Press.

Jenkins, Philip. 2002. *The Next Christendom: The Coming of Global Christianity*. New York: Oxford University Press.

Marsden, George M. 1980. *Fundamentalism and American Culture: The Shaping of Twentieth-Century Evangelicalism, 1870-1925.* New York: Oxford University Press.

------. 1994. *The Soul of the American University: From Protestant Establishment to Established Nonbelief*. New York: Oxford University Press.

Melton, J. Gordon, (ed.). 1991. *American Religions: A Comprehensive Study of the Major Religious Groups in the United States and Canada, Vol. 1.* New York: Gale Research.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011a. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331:176-182. (Published online ahead of print: 12/16/2010.)

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011b. "Supporting Online Material for 'Quantitative Analysis of Culture Using Millions of Digitized Books."Retrieved Nov. 6, 2014 (http://www.sciencemag.org/content/suppl/2010/12/16/science.1199644.DC1/Michel.SOM.revision.2.pdf).

Ringenberg, William C. 1984. *The Christian College: A History of Protestant Higher Education in America.* Grand Rapids, MI: Eerdmans Publishing Company.

Smith, Tom W. 2017. "Improving Cross-National/Cultural Comparability Using the Total Survey Error Paradigm." In *Faithful Measures,* Roger Finke and Christopher D. Bader (eds.).

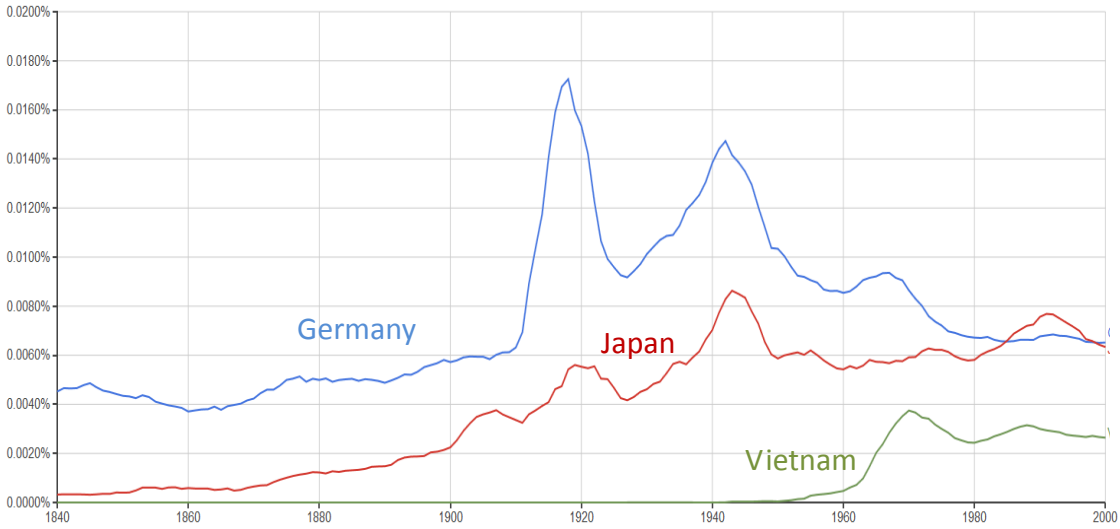
Stark, Rodney and Roger Finke. 1988. "American Religion in 1776: A Statistical Portrait." *Sociological Analysis* 49(1): 39-51.

Taves, Ann. 2011. "Presidential Address: 'Religion' in the Humanities and the Humanities in the University." *Journal of the American Academy of Religion* 79(2): 287–314.

Thelin, John R. 2011. *A History of American Higher Education, Second Edition.* Baltimore, MD: The John Hopkins University Press.

Wacker, Grant. 2001. *Heaven Below: Early Pentecostals and American Culture.* Cambridge, MA: Harvard University Press.

Figure 1. Ngram Trends for Nations, 1840-2000



Note: This figure was made by searching "Germany, Japan, Vietnam" in the Google Ngram American English Corpus with three-year smoothing (https://books.google.com/ngrams).

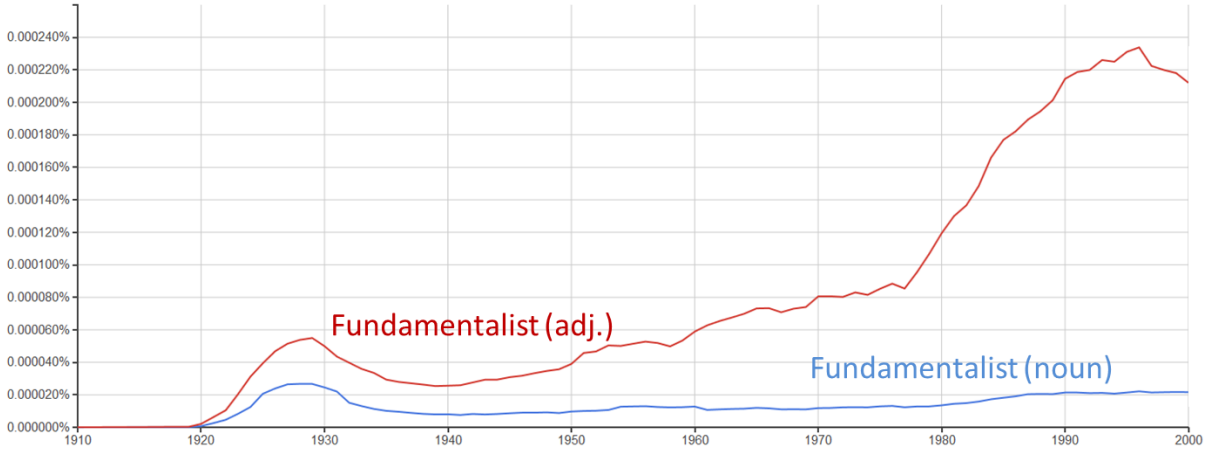
Figure 2: Ngram of Mormonism-related terms, 1820-2000

Note: This figure was made by searching "Mormon+mormon+Mormons+mormons+latter day saints+Latter day saints+Latter Day Saints+Latter Day saints+Latter day Saints+LDS+lds" in the Google Ngrams American English Corpus with three-year smoothing (https://books.google.com/ngrams).

Figure 3. Ngram Trends for Pentecostal vs. Pentecostalism

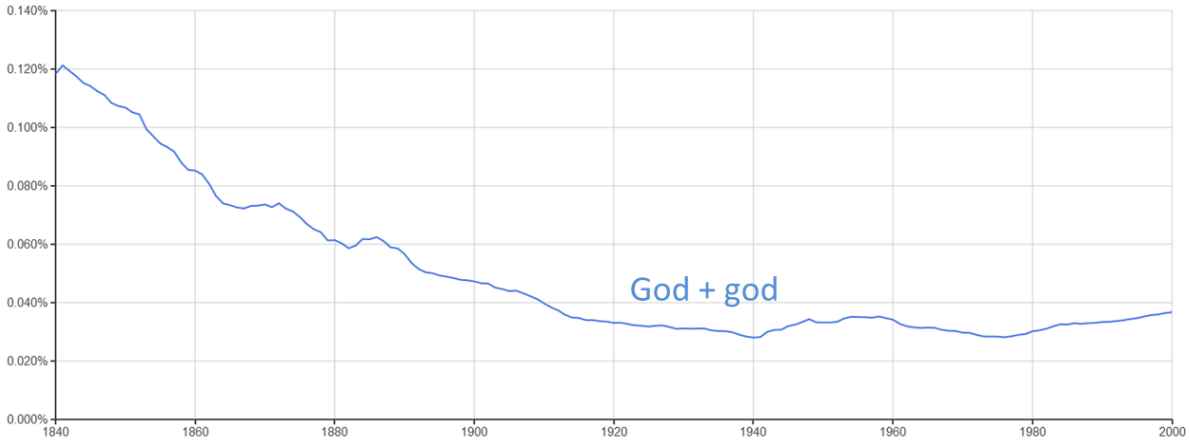
Note: This figure was made by doing a case-insensitive search for "Pentecostal, Pentecostalism" in the Google Ngrams American English Corpus with three-year smoothing (https://books.google.com/ngrams).

Figure 4. Ngram Trends for Fundamentalist, Comparing Its Use as a Noun and an Adjective, 1910-2000



Note: This figure was made by doing a case-insensitive search for "fundamentalist_ADJ, fundamentalist_NOUN" in the Google Ngrams American English Corpus with three-year smoothing (https://books.google.com/ngrams).

Figure 5. Ngram Trend for God, 1840-2000



Note: This figure was made by searching for “God+god” in the Google Ngrams American English Corpus with three-year smoothing (https://books.google.com/ngrams).

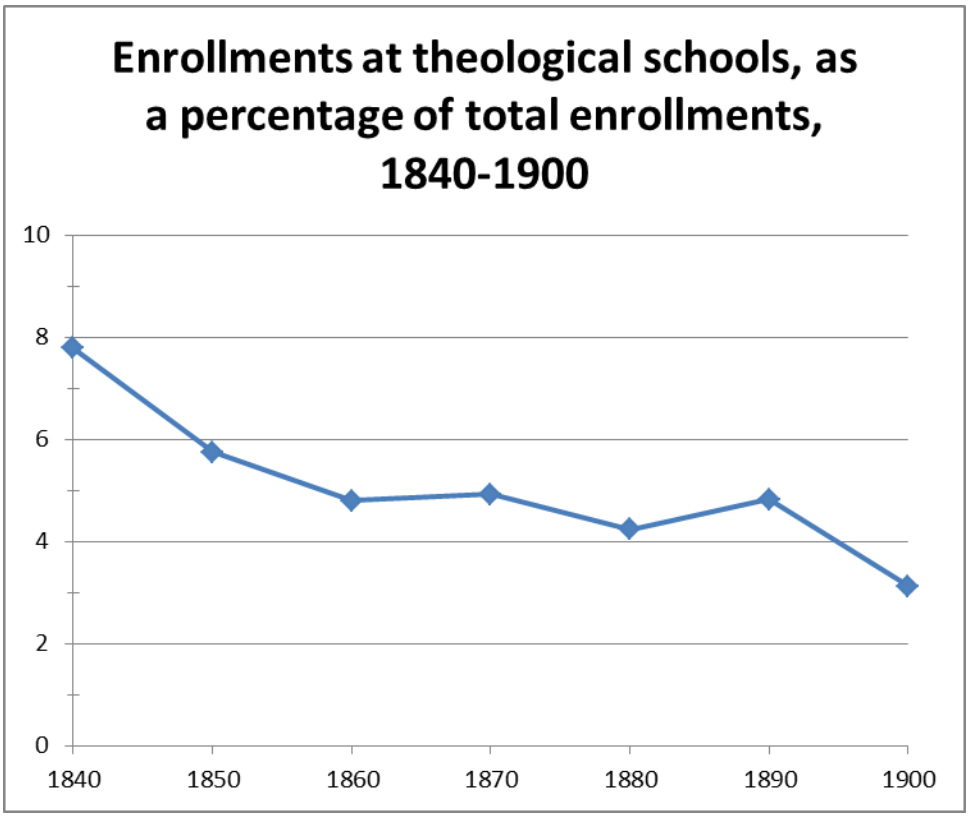
Figure 6. Ngram Trend for Jesus, 1840-2000

Note: This figure was made by doing a case-insensitive search for “Jesus” in the Google Ngrams American English Corpus with three-year smoothing (https://books.google.com/ngrams).

Figure 7. Ngram Trends for Atheism and Atheist, 1840-1900

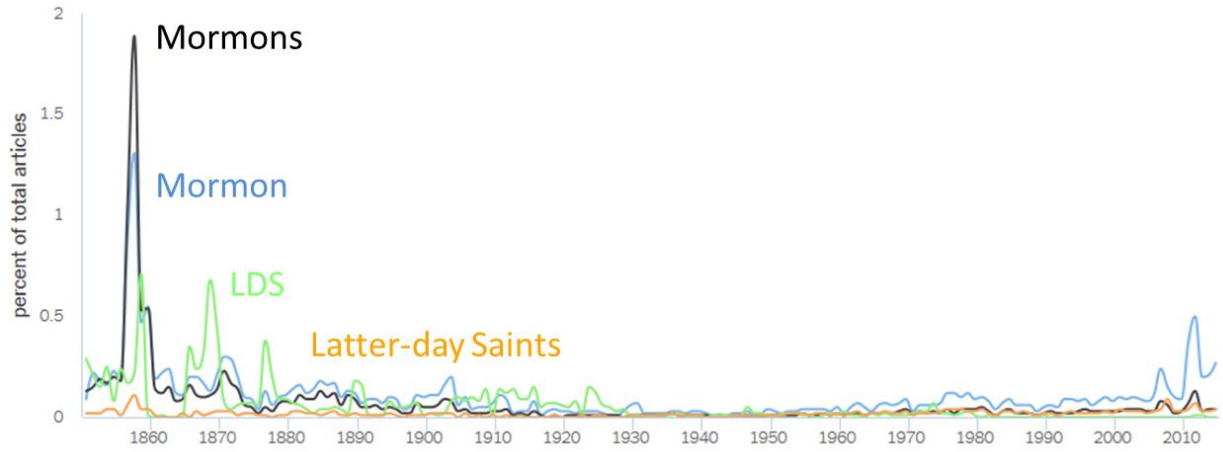
Note: This figure was made by doing a case-insensitive search for "Atheist, Atheism" in the Google Ngrams American English Corpus with three-year smoothing (https://books.google.com/ngrams).

Figure 8. Trends in Theological Enrollments, 1840-1900



Source: Burke 1982:216

Figure 9. Use of Mormonism-related Terms in the New York Times, 1851-Present



Note: This figure was made by searching for “Mormon,” “Mormons,” “LDS,” and “Latter-day Saints” on http://chronicle.nytlabs.com/.

Figure 10. Ngram for “Mormons” with Links to Google Books

Note: This figure was made by searching for “Mormons” in the Google Ngrams American English Corpus with three-year smoothing (https://books.google.com/ngrams).

Appendix A.

Historical Measures in the Current ARDA Ngrams Dataset

Measures from the Historical Statistics of the United States (Carter et al. 2015)

Variables	**Years**	**Linear Interpolation?**
Total population	1800-2000	Yes
Percent urban	1800-2000	Yes
Sex ratio-males per 100 females	1800-2000	Yes
Median age	1800-2000	Yes
Percent foreign born	1850-2000	Yes
GDP, real in 1996 dollars	1800-2000	No
GDP per capita, real in 1996 dollars	1800-2000	No
Percent of 18-24 year olds enrolled in higher education	1904-1995	Yes
Percent at least 14 years old who are illiterate	1870-1979	Yes
Percentage of workers who are farmers or farm laborers	1860-1990	Yes
Percent of 5-20 year olds enrolled in school--primary and secondary	1860-1994	Yes
Higher education enrollment, in thousands	1869-1995	Yes
Percent of immigrants who are from Western Europe--Northwestern Europe, Germany, Greece, Italy, Portugal, Spain or Other	1820-1997	No
Percent of immigrants who are from Eastern Europe--Poland, Other Central Europe, Eastern Europe	1820-1997	No
Percent of immigrants who are from Asia	1820-1997	No
Percent of immigrants who are from North America	1820-1997	No
Percent of immigrants who are from South America	1820-1997	No
Percent of immigrants who are from Africa	1820-1997	No
Percent of immigrants who are from Oceania	1820-1997	No
Percent of immigrants who are from Other	1820-1997	No
Total number of immigrants	1820-1997	No
Voter turnout in presidential elections	1824-2000, every four yrs	No

Data from the Yearbook of American and Canadian Churches

Variable	**Years**	**Linear Interpolation?**
Number of clergy	1925-2000	No

Data from the National Center for Education Statistics

Variables	**Years**	**Linear Interpolation?**
Number of higher education institutions	1870-2000	Yes
Percent of higher education enrollment that is in private institutions	1950-2000	Yes

Data calculated from the Association of Theological Schools Directory

Variable	**Years**	**Linear Interpolation?**
Number of seminaries	1800-2000	No

Calculated variables from multiple data sources

Variables	**Years**	**Linear Interpolation?**
Percent of the population that is clergy (HSUS and YACC)	1925-2000	No
Percent of higher education institutions that are seminaries (NCES and ATS)	1870-2010	No