Religousity and Public Insecurity in the US Senate

Emily Gade[1]
University of Washington
Jon Schaeffer
University of Washington
John Wilkerson
University of Washington
Anne Washington
George Mason University

Abstract

Nearly all members of Congress claim a religious affiliation, and 92 percent identify as Christians. This lack of variation calls for an alternative approach to studying the role of religion in legislative politics. We exploit a truly big data resource. The .GOV Internet Archive database contains 90 terabytes of .gov web page captures from 1998 to 2013 – more than nine times the amount of data as the print holdings of the Library of Congress. We use .GOV to investigate how members of Congress reference religion on their official websites. We hypothesize that legislators are more likely to employ religious rhetoric during times of public anxiety and insecurity. "Messy" distributed datasets such as .GOV are at the frontier of social science research. They pose new methodological challenges and require new research skills, but also offer valuable research opportunities not available elsewhere.[2]

[1]Corresponding Author: ekgade@uw.edu

[2]We thank Vinay Goel of the Internet Archive, Altiscale, the University of Washington eScience Institute Data Science Incubator for their assistance. This work was supported in part by the National Science Foundation under Grant No. 1243917 (Division of Social and Economic Sciences, Directorate for Social, Behavioral Economic Sciences). Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation.

"All politicians, Democrats and Republicans alike, love God. Or, more accurately, they love to use God to baptize their political agendas. In the Congressional Directory . . . no one is an atheist... You never know when it might help you to be religious," (Thomas 1999, 83).

1 Introduction

In 1996, a non-profit organization, the Internet Archive (IA) assumed the ambitious task of documenting the public web. The current collection contains over 450 billion webpage "captures" (downloads of URL linked pages and metadata) dating back to 1995. The best way to quickly appreciate the IA holdings is to visit the WayBack Machine website (`archive.org/web`), where individual historical website captures (e.g. the White House home page from Dec. 27, 1996) can be viewed.

The Internet Archive also curates a sub-collection: .GOV.[3] .GOV contains approximately 1.1 billion page captures of URLs with a .gov suffix (from 1996 through Sept 30, 2013). At the federal level, this archive includes the official websites of elected officials, departments, agencies, consulates, embassies, USAID missions and much more.[4] Whereas individual date-specific captures can be viewed using the Wayback Machine, the .GOV collection is a database that can be used to explore broader patterns across websites and over time. We also suspect that .GOV contains hidden treasures in the form of documents posted at one point in time by officials or agencies and later removed from official websites (the endless rows of shelves of dusty cardboard boxes in Indiana Jones and the Temple of Doom comes to mind!)

.GOV offers four types of data from each webpage capture: the link data (the page URL and every other url/hyperlink found on the page); the parsed text of the page; and the full content of the page (the text including html markup language; images; video files etc.). In this paper, we use parsed text of senators' congressional websites from 2006-2012 to explore how they use religion in their public communications. We also discuss the strengths and limitations of .GOV as a social science research database and encourage other to think about how questions they might be able to examine using historical government website content.

2 Religiosity in Congress

Article 6 of the United States Constitution states that *"no religious test shall ever be required as a qualification to any office or public trust under the United States."* While the framers sought to prevent Christianity from becoming the national religion, many of them believed (like de 'Tocqueville) that christian morality was essential to a successful democracy (see also Waldman 2009). The legal prohibition of state sponsored religion was also not intended to prevent religion beliefs from influencing the policy choices of elected officials or the voting decisions of the electorate. In "Religion in American Politics: A Short History," Lambert (2008) notes that Thomas Jefferson was accused of being unfit to lead "a Christian Nation." Even today, having a religious affiliation seems to be a virtual litmus test for elective office. Only one member of

[3]See the Internet Archive's description of their sub-collections here: `https://archive.org/details/additional_collections`

[4].GOV also includes state and local websites that use the .gov suffix.

the current 115th Congress claims no affiliation (compared to 20% of Americans) (Liu 2003).

Research also finds that the results have been controversial and partisan whenever Christians have organized in politics (Liu 2003). The rise of the Christian Right in the mid-twentieth century centered on a promise to restore "law and order" in the face of progressive political developments (Williams 2010; Henkin 1986). In the contemporary Congress, members' religious affiliations continue to predict their voting decisions on social policies and often divide them along party lines.

Polls indicate that the American public believes that the country is becoming much more secular. However Gill (2008) argues that Americans are as religious as ever. What has changed is that it is has become more socially acceptable to admit to having "no religious affiliation." As many European nations have actually become more secular, the absence of a state-sponsored religion in the U.S. has allowed religious organizations to maintain their memberships (see also Pfaff 2008).

Religious beliefs continues to shape many Americans' political behavior. According to a 2003 Pew Research Center Survey,[5] 38% of Americans say that religion sometimes plays a role in their voting decisions. Women are more likely to say that religion "frequently" influences their voting decisions than men (26% vs. 17%), Republicans more than Democrats or Independents (31% vs. 20% and 17%), and White Evangelicals (48%) and black Protestants (31%) more than Catholics (12%) or white Protestants (10%).

Studies of religion in the US Congress have largely focused on how members' reported affiliations impact their voting behavior, especially on cultural issues such as abortion or gay marriage (Guth 2014; Blackstone and Oldmixon 2015). A smaller number of studies explore member religiosity in more detail by documenting their associations and activities outside of government. Yet no standard, numeric measure of religiosity in Congress yet exists. We investigate how legislators use religion in their public communications. We hypothesize an electoral connection, but one that goes beyond the traditional culture wars.

The traditional way of measuring religiosity in other contexts has been through three central attributes. First, how literally an individual takes their religious text. Second, how often an individual prayers. Third, how often an individual attends religious services (Steensland et al 2000). This information is not available for most members of congress. We examine the frequency of religious references on congressional webpages as a measure of religiosity. To our knowledge, this is the first quantitative, temporally sensitive measure of religiosity in congress.

Why do members of the American public seem to prefer religious members of Congress? One explanation may be that religion helps many people to cope with stressful or traumatic events. One national survey found that 90% of respondents turned to religion to in the aftermath of 9/11 (Shuster et al 2001). We expect to find that elected officials are more likely to express religious solidarity with their constituents during times of public insecurity and anxiety.

Perceived cultural threats can certainly contribute to public anxiety. We therefore expect to find that legislators representing constituencies threatened by progressive politics should

[5]://www.people-press.org/2003/07/24/ii-religion-voting-and-the-campaign/

make more frequent references to religion in their webpage text. But we are also interested in the impact of anxiety-inducing events *other than* the "culture wars," such as terrorism and natural disasters (McTague and Pearson-Merkowitz 2013, 2015). It is the latter which we evaluate in this paper.

People use religion as a coping strategy in different ways. Some turn to religion as a source of positive support and meaning in difficult times. Others interpret events in a more negative religious light, for example as evidence of evil in the world, or as retribution for sinful behavior (Sohrabzadeh et al 2017). Do legislators' also vary in terms of the tone of their religious references?

3 Data

3.1 Dependent Variable: Religious references

Congressional websites are largely advertisements for the incumbent. They include autobiographical information as well as information about the lawmakers' policy priorities, institutional positions and available services. These websites also typically include archives of past speeches, press releases, and op-eds. As such, we contend that they provide a more complete and targeted portfolio than focusing on, for example, members' floor speeches or their press releases alone. The words themselves are from the Linguistic Inquiry and Word Count Dictionary (LIWC). The LIWC is a software that analyzes text for different themes and concepts in text, including religion (Pennebaker et al 2015).

3.1.1 Messy data

One important limitation of the .GOV data is that it is incomplete, and incomplete in ways that are not understood or well documented. Thus, it is simply impossible to download the entire content of the internet, or even a representative sample. The IA (as well as major search firms such as Google) capture web content by sequentially "crawling" from one page to another. Starting from "seed" URLs (web page addresses) a "bot" (software program) collects the content of all links found on the originating page, then all of the links on those pages (etc.). This sequential process inevitably offers an incomplete snapshot because the number of seed url is always limited and because the World Wide Web is changing as it is being crawled. In 2008, the official Google blog bragged that developers had collected 1 trillion unique URLs in a single concerted effort but also noted that "the number of pages out there is infinite."[6] Crawl results are also incomplete because many webpages are located behind firewalls (the "dark web"), or contain scripts to discourage bots from collecting content.[7]

The quality of the Internet Archive holdings has improved over time, due to advances in crawling technology and resources. Figure 1 illustrates this by displaying how frequently the White House website was captured during four time periods starting in 1997.[8] The White

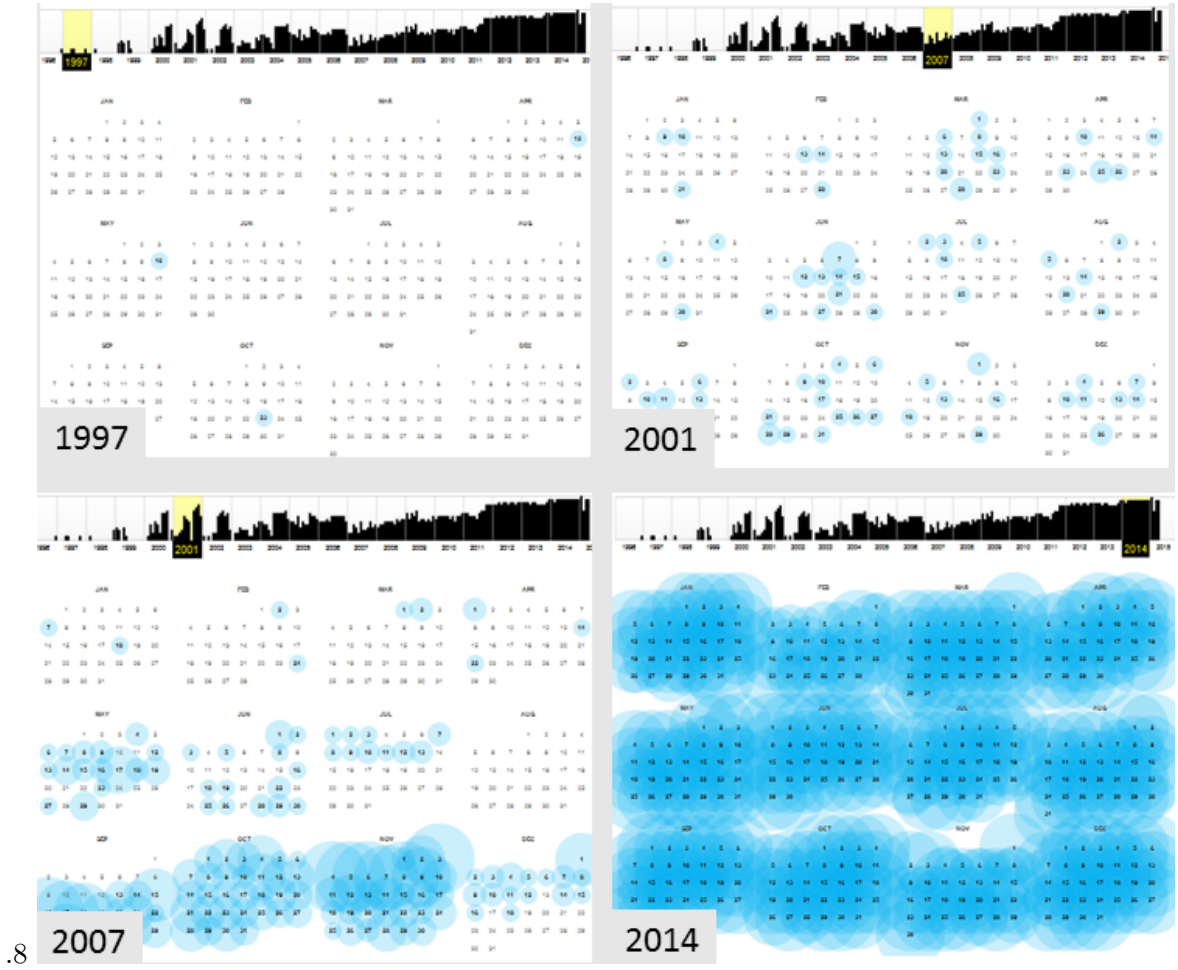
[6]See Google's Official Blog (July 25, 2008) for discussion at "We knew the web was big..." `http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html`

[7]The Internet Archive also deletes content at the owner's request.

[8]The graphs are copied from Wayback Machine search results for `whitehouse.gov`.

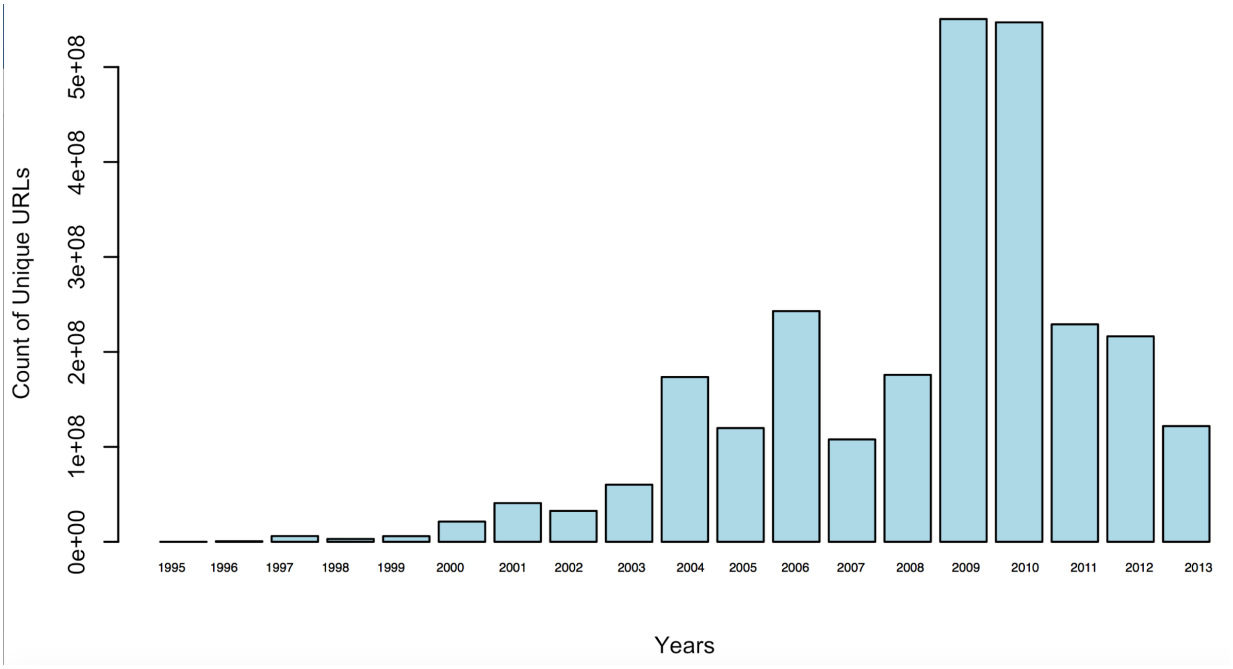
House website was crawled just 3 times in 1997. In 2001, it was not crawled at all in the month of August and then hundreds of times in the three months following the terrorist attacks on September 11. In 2007, it was crawled at least once a week. And in 2014, it was crawled at least once a day. The depth of these crawls also varies in ways that may have important implications for a research project.

Figure 1: Frequency of `whitehouse.gov` crawls (selected years)



We focus on a time period and domain where these holdings should be of exceptional quality. The Library of Congress (starting in 2004) now contracts with the IA to use congressional URLs as starting seeds during three months of each election cycle (November-January). This has dramatically increased .Gov captures (Figure 2). For example, the number unique (non-duplicate pages) .gov captures roughly triples from about 500 million to 1.5 million in the first election year (2004) of this new arrangement. This increase is not solely a product of the targeted crawl. It may also be due to an ever expanding web and improved technology, but the consistent spikes during election years does seem to indicate more comprehensive crawls.

Figure 2: Total .GOV Unique URLs



.GOV is not an ideal resource but it seems to be the best existing resource for studying government web presence historically. Congressional websites during election years also constitute some of the most complete results available. To investigate the content of senators' websites, we first created a root url regular expression to collect all page captures within a domain (e.g. hatch.senate.gov) for all US senators from 2006-2014. We examine only the text found on these pages and aggregate them by year (election years only). Finally, each year's collection only includes pages that contain new content compared to what was collected from the same domain during the previous election year.

Our dependent variable is the *proportion* of all words from all new pages of a website that are found in the "Religion" dictionary of the Linguistic Inquiry and Word Count (LIWC) project (Pennebaker et al 2015).[9] Notably, the LIWC dictionary does not have a political focus. It does not include terms associated with the culture wars, such as abortion, pro-life, homosexuality etc. Figure 3 displays a word cloud of the most common terms found on senators' websites (the size of the word corresponds to its relative frequency).

[9]We made a small number of additions because some obvious candidate terms (Christ, grace) were missing from the list.

Figure 3: Religion terms from Senators' websites

3.2 Independent Variables

We test three main predictors of differences in senators' religious references: their personal attributes; electoral considerations; and anxiety.

Personal attributes

Earlier, we noted that surveys indicate important differences in religiosity among Americans. Women, Republicans, white Evangelical Protestants and black Protestants are more likely to say that their religion frequently influences their voting decisions. The same should be true for senators, noting of course that we are examining their public rather than private behavior. Nevertheless, we hypothesize that there will be more religious references on the websites of female, Protestant, and Republican senators.

Electoral considerations

If congressional websites are important advertising tools, then the characteristics of legislators' constituencies should help to predict their substance. We expect that senators who represent more religious constituencies will make more frequent religious references. We test this hypothesis using a general measure of state religiosity - the percentage of citizens who report that they are "highly religious."[10] We found that this measure strongly correlates with a variety of others, such as the percent of citizens who report praying regularly, or the percent of citizens who identify as very conservative. We also test whether the religious denominations of

[10] http://www.pewresearch.org/fact-tank/2016/02/29/how-religious-is-your-state/?state=alabama

a state's citizens predict religious references by hypothesizing that states with more Evangelical Protestants (the denomination most likely to say that their religion "frequently" influences their voting decisions), and with more citizens who are Mormon. Finally, we ask whether senators make more religious references when they are up for reelection. Table 1 displays descriptive statistics of key variables.

Table 1: Descriptive Statistics

Var Name	Min	Max	Mean	SD
Freq Religious	0.000000001	0.002	0.0001	0.0001
Freq Religion No Islam	0.000000001	0.002	0.0001	0.0001
Freq DHS	0.000000001	0.004	0.001	0.0005
State Terror Attacks	0	8	0.28	0.90
FEMA Decs	0	470	30	56
% Very Religious (State)	0.33	0.77	0.54	0.10
Very Conservative (State)	2.83	3.87	3.40	0.18
% Evangelical (State)	2.21	43.97	15.759	11.83
Don't Pray	2.17	4.78	3.16	0.48
Conservatism (Senator)	-0.64	1	0.02	0.43
Jewish Faith (Senator)	0	1	0.12	0.33
Mormon Faith (Senator)	0	1	0.04	0.20
Female (Senator)	0	1	0.16	0.37
Up For Election	0	1	0.33	0.47
Republican (Senator)	0	1	0.47	0.50

Figure 4: Anxiety and DHS terms from Senators' websites

Public anxiety

We were not able to find any direct indicators of public anxiety at the state level. We ask whether the senators who make religious references also use terms associated with anxiety (also from LWIC), and whether they more likely to also mention terms on the Department of Homeland Security's social media monitoring wordlist (Figure 4).[11]. For example, Senator Charles Grassley (R-IA) made numerous religious references in a press release expressing sympathy for the victims of the Libya embassy attack (Figure 5). We also test several measures of events that should increase anxiety among constituents, including reported terrorist attacks[12] and FEMA declarations at the state level,[13] as well as counts of international terror attacks.[14]

[11] DHS wordlist can be found here: `https://gist.github.com/jm3/2815378`. Note that we only consider unigrams in this iteration and exclude some terms that seem likely to produce false positive results, e.g. state or country names

[12] `https://www.start.umd.edu/gtd/`

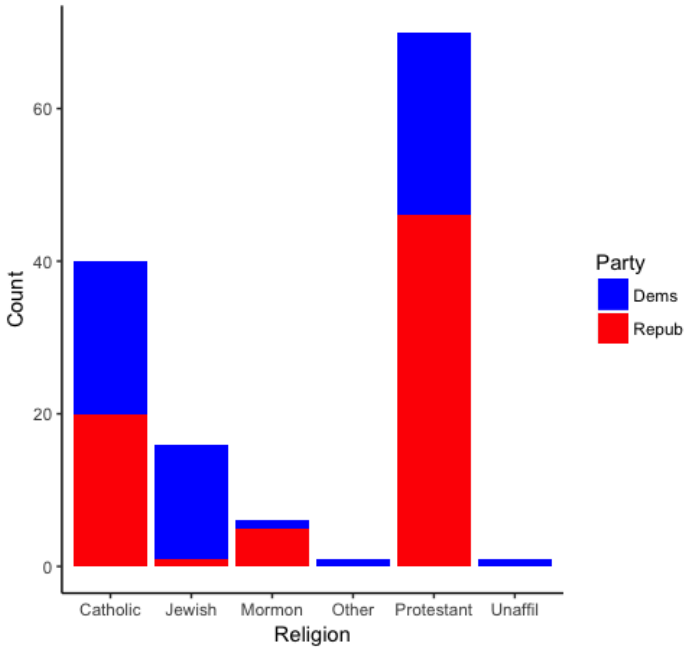
[13] `https://www.fema.gov/disasters/state-tribal-government`

[14] `https://www.start.umd.edu/gtd/`

Figure 5: Grassley press release

Figure 6 displays senators' reported religious affiliations by party.

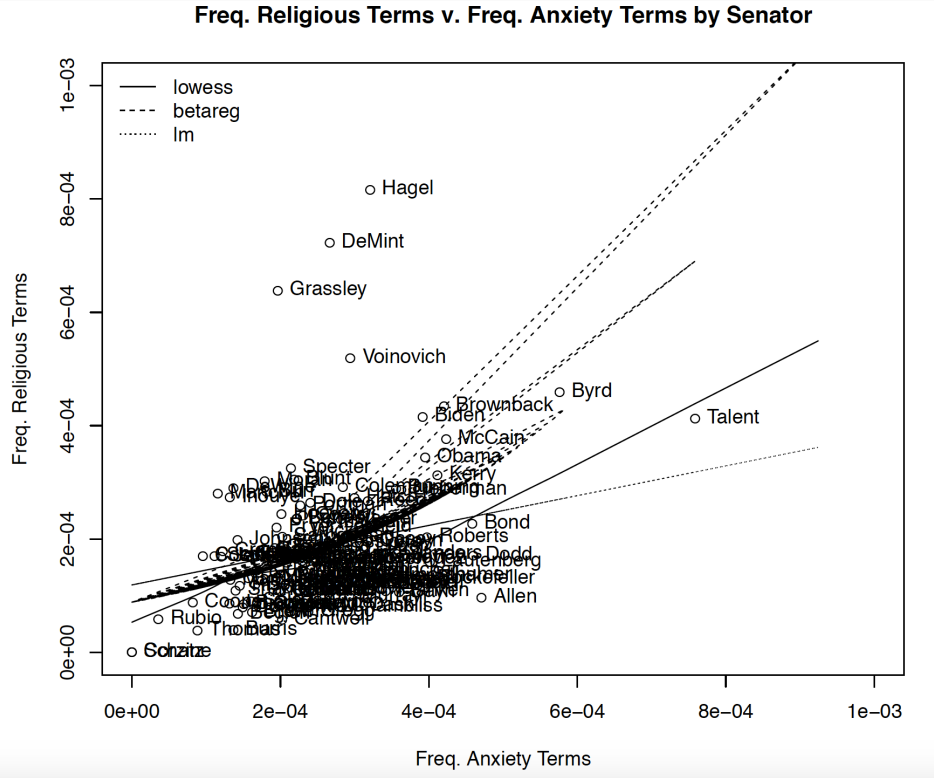
Figure 6: Senator Religious Affiliations by Party (2006-2012)



4 Findings

Figure 7 begins to explore whether senators are more likely to make religious references during times of anxiety. In this scatterplot the y axis is the dependent variable (frequency of religious references) while the x-axis is the frequency of anxiety references by the same senator in the same time period. The lines represent lowess, beta and linear bivariate regression models; all display a positive relationship.

Figure 7: Correlations between senator religiosity and anxiety mentions



4.1 Regression results

We employ a beta regression model in this analysis, given that the dependent variable is a proportion with values ranging between zero and one.[15] We test a beta regression predicting religious references on senators' websites. As discussed, our primary hypothesis is that senators will make more religious references during times of anxiety. Doing so is a way to express identify and solidarity with constituents who seek solace in religion. Senators who represent more religious states should also make more frequent religious references (for similar reasons). We also test whether senators' personal attributes predict their use of religious terms.

Several supportive findings emerge across the different model specifications in Table 2. First, senators make more religious references during years when they are also making more

[15]We bump any true zero values to .0000001 as beta regressions do not allow true zero values.

Table 2: Explaining Religiosity on Senators' Websites 2006-12

	Dependent variable:				
	Freq. Religious			Freq. Religious No Islam	
	(1)	(2)	(3)	(4)	(5)
Freq. Anxiety	2,302.626***		2,307.397***	2,162.948***	2,147.952***
	(261.803)		(261.912)	(265.816)	(267.262)
Freq. DHS	252.318***		257.318***	246.492***	251.262***
	(59.994)		(59.862)	(60.135)	(61.390)
Terrorist Attack (State)	−0.018	−0.079*		−0.021	−0.056
	(0.040)	(0.046)		(0.040)	(0.051)
FEMA Dec (State)	0.0003	0.001*		0.0002	0.0003
	(0.001)	(0.001)		(0.001)	(0.001)
Very Religious (State)	0.691*	1.549***	0.713**	0.830**	0.825**
	(0.360)	(0.399)	(0.353)	(0.361)	(0.364)
Mormon (State)					0.004
					(0.005)
Female (Senator)	−0.128	−0.057	−0.132	−0.098	−0.137
	(0.101)	(0.109)	(0.099)	(0.101)	(0.107)
Up for Election (Senator)	−0.015	−0.051	−0.015	−0.028	−0.025
	(0.070)	(0.079)	(0.070)	(0.070)	(0.071)
Conservatism (Senator)	0.257***	0.132	0.265***	0.245***	0.235***
	(0.087)	(0.097)	(0.087)	(0.087)	(0.089)
Mormon (Senator)	0.345**	0.336**	0.336**	0.361**	0.194
	(0.149)	(0.167)	(0.148)	(0.149)	(0.245)
Catholic (Senator)	0.087	0.022	0.088	0.084	0.102
	(0.079)	(0.089)	(0.078)	(0.079)	(0.080)
Constant	−9.972***	−9.475***	−9.987***	−10.053***	−10.057***
	(0.222)	(0.233)	(0.220)	(0.223)	(0.224)
Observations	391	391	391	391	383
R^2	0.224	0.034	0.223	0.214	0.216
Log Likelihood	3,066.789	2,992.338	3,066.598	3,083.884	3,020.428

Note: *p<0.1; **p<0.05; ***p<0.01

references to anxiety and to words on the DHS watch list. Senators representing more religious states ("percent highly religious") also make more religious references.[16] More conservative senators and Mormon senators make more religious references.[17]

The most surprising finding is the absence of any relationship between our measures of objective conditions (state terror attacks; state FEMA declarations) and senators' religious communications.

Earlier we noted that some people use religion as a positive coping strategy (as a source of support and inspiration), whereas others interpret events in a more negative religious light. To begin to explore differences in tone, we tagged 40 positive (e.g. faith, paradise, salvation etc) and negative religious terms (sin, doom, hell, etc) in the LIWC term list. We then created a ratio to compare relative positive and negative mentions among senators. We found that Protestant senators were significantly more likely to use positive religious terms, whereas Mormon senators were significantly more likely to use negative terms. Thus, not only do we find that Mormon lawmakers are more likely to make religious references, we also find that those references also differ in tone. Unfortunately we are not currently able to separate Evangelical Protestants from other Protestant Senators to know whether there are differences within that denominaton.

5 Discussion

In this paper, we explore an original and largely untapped data source to examine religiosity in Congress from a new perspective. Using the Internet Archive's .GOV database, we extract religious mentions from senators' official websites and test three sets of explanations for why some are more likely to make references to religion in their public communications. We find support for all three explanations. More conservative and Mormon senators are more likely to make religious references. Senators who represent more religious states make more religious references. And senators tend to use religious terms and terms associated with anxiety and public insecurity together. However, we are not able to connect their use of those terms to objectives conditions hypothesized to contribute to public anxiety and insecurity. This may be because our aggregation of these data to the annual level prevents sufficiently fine-grained analysis to detect any relationship that does exist.

These are intriguing findings represent a first cut. We hope to conduct a more focused analysis moving forward. Instead of examining content across entire websites annually, we would like to conduct a more focused analysis. Comparing religious attention during shorter time periods may not be practical due to the variability of .GOV crawls. We aim to next do a micro-analysis around key incidents, using day or month-level religious reference and relevant lags. We will probably end up with substantial numbers of missing values if we shift to a (e.g.) month/senator unit of analysis. A related concern is that crawls at time other than the three months of elections years supported by the Library of Congress may not crawl senators' websites to the same depth, making comparisons problematic.

[16] We only include the percent of citizens who are "highly religious" because it is strongly correlated with the percent who are evangelicals and state conservatism.

[17] There is no "evangelical protestant" option for senators' religious affiliations. We assume that, as is the case at the state level, senator conservatism is partly a proxy for evangelical.

At this point we think that the most useful next step would be to identify and examine more closely the specific web pages that contain religious content and, in particular, those that include both religious and anxiety terms. We could then better assess the validity of our current findings and develop methodologies for systematically identifying the contexts in which legislators make religious references in the their public communications. We can easily expand such an analysis to include members of the House of Representatives although, due the quirks of .GOV, the findings will inevitably violate conventional sampling assumptions. We appreciate your reactions and suggestions!

References

"Bash Shell Basic Commands." *GNU Software.* http://www.gnu.org/software/bash/manual/bash.pdf

Blackstone, Bethany, Elizabeth A Oldmixon. (July, 2015). Discourse and dissonance: religious agendas in the 104th Congress. *Research Politics.* Vol 2, Issue 3.

Boydstun, A. E., Bevan, S. and Thomas, H. F. (2014), "The Importance of Attention Diversity and How to Measure It." *Policy Studies Journal*, 42: 173–196. doi: 10.1111/psj.12055

Edwards, J., McCurley, K. S., and Tomlin, J. A. (2001). "An adaptive model for optimizing performance of an incremental web crawler". *Tenth Conference on World Wide Web* (Hong Kong: Elsevier Science): 106–113.

Gill, Anthony. (2009). *Political Origins of Religious Liberty.* Cambridge University Press.

Henkin, L. (1986). "Religion and the Constitution." *Jewish Social Studies*, 48(3/4), 325-328.

Hill, Peter, Ralph Hood. (1999). *Measures of Religousity.* Religious Education Press.

"The History of the Internet." *The Internet Society.* http://www.internetsociety.org/internet/what-internet/history-internet/brief-history-internet

"The Internet Archive." *Internet Archive.* https://archive.org/

Kahn, R. (1972). "Communications Principles for Operating Systems." *Internal BBN memorandum.*

Lambert, Frank. (2008). *Religion in American Politics: A Short History.* Princeton University Press.

Leiner et al. "Brief History of the Internet." http://www.internetsociety.org/sites/default/files/Brief_History_of_the_Internet.pdf

Licklider, J. C. (1963). "Memorandum for members and affiliates of the intergalactic computer network." M. a. A. ot IC Network (Ed.). Washington DC: KurzweilAI. ne.

Liu, J. (2003, July 24). "Religion and Politics: Contention and Consensus." http://www.pewforum.org/2003/07/24/religion-and-politics-contention-and-consensus/

Najork, M and J. L. Wiener. (2001). "Breadth-first crawling yields high-quality pages." *Tenth Conference on World Wide Web*, (Hong Kong: Elsevier Science): 114–118.

Pennebaker, J.W., Booth, R.J., Boyd, R.L., Francis, M.E. (2015). *Linguistic Inquiry and Word Count: LIWC2015.* Austin, TX: Pennebaker Conglomerates

Pew Research Center. (2015) *Faith on the Hill: The Religious Affiliations of the 114th.* Pew Research Center.

Pfaff, S. (2008). "The Religious Divide: Why Religion seems to be Thriving in the United States and Waning in Europe." In J. Kopstein S. Steinmo (Eds.), Growing Apart? America and Europe in the Twenty-First Century (pp. 24–52). New York: Cambridge University Press.

"Pig Manual." *Apache Systems* `https://pig.apache.org/docs/r0.7.0/piglatin_ref1.html`

"The Rise of 3G." *THE WORLD IN 2010.* International Telecommunication Union (ITU)). `<www.itu.int/ITU-D/ict/material/FactsFigures2010.pdf>`.

Sagiroglu, S., Sinanc, D. (2013, May). "Big data: A review." In *Collaboration Technologies and Systems (CTS), 2013 International Conference (pp. 42-47).* IEEE. `https://xa.yimg.com/kq/groups/72986399/1585974627/name/06567202.pdf`

Shuster, M. Stein B. Jaycox L. Collins R. Marshall G. Elliot M. et al (2001) "A national survey of stress reactions after the Septemer 11 2001 terrorist attacks." *New England Journal of Medicine* 345(20), 1507-1512.

"A "ssh" key (Secure Shell)" (2006). `http://tools.ietf.org/html/rfc4252`

Sohrabizadeh S. Jahangiri K. Janzani R. (2017) "Religiosity, Gender and Natural Disasters: A Qualitative Study of Disaster -Stricken Regions in Iran." *Journal of Religion and Health.* DOI 10.1007/s10943-017-0398-9.

Steensland, Brian, Lynn D. Robinson, W. Bradford Wilcox, Jerry Z. Park, Mark D. Regnerus, and Robert D. Woodberry. (2000) "The measure of American religion: Toward improving the state of the art." *Social Forces.* 79, no. 1.

Vance, A. (2009). "Hadoop, a Free Software Program, Finds Uses Beyond Search". *The New York Times.*

Waldman, S. (2009). *Founding Faith: How Our Founding Fathers Forged a Radical New Approach to Religious Liberty (Reprint edition).* New York: Random House Trade Paperbacks.

Williams, Daniel K. (2010). *God's Own Party: The Making of the Christian Right.* New York: Oxford University Press. pp. 1-2.

A Appendix I

The following script flags senator webpages that include one or more mentions of the listed terms and stores a count of those captures. We begin with an overview of the process of running jobs on the cluster, and then provide specific code. For questions, please contact the author (ekgadeuw.edu).

A.1 Overview

Running scripts on the cluster requires a basic understanding of bash (Unix) shell commands using the Command Line on a home computer (on a Mac, this is the program "Terminal"). For a basic run down of bash commands, see `http://cli.learncodethehardway.org/bash_cheat_sheet.pdf`.

Begin by opening a bash shell on a home desktop, and using an ssh key obtained from Altiscale to log in. Once logged in, you will be on your personal workbench and now have to use a script editor (such as Vi `http://www.catonmat.net/download/bash-vi-editing-mode-cheat-sheet.pdf`). Come up with a name for the script, open the editor, and then either paste or write the desired script in the editor, close and save the file (to your personal workbench on the cluster).

Scripts must be written in Hadoop-accessible languages, such as Apache Pig, Hive, Giraph or Oozie. Apache languages are SQL-like, which means if you have experience with SQL, MySQL, SQLlite or PostgreSQL (or R or Python), the jump should not be too big. For text processing, Apache Pig is most appropriate, whereas for link analysis, Hive is best. The script below is written in Apache Pig and a manual can be found at `https://pig.apache.org/`. For an example of some scripts written for this cluster, see `https://webarchive.jira.com/wiki/display/Iresearch/IA+-+GOV+dataset+-+Altiscale`. May be easiest to it "clone" the "archive analysis" file hosted on GitHub from Vinay Goel `https://github.com/vinaygoel` or three basic scripts from Emily Gade `https://github.com/ekgade/.govDataAnalysis` and use those as a launch point. If you don't know how to use GitHub, see here: `https://guides.github.com/activities/hello-world/`(it is actually quite straightforward).

Because Apache languages have limited functionality, users may want to write user defined functions in a program like Python. A tutorial about how to do this can be found at `https://help.mortardata.com/technologies/pig/writing_python_udfs`.

Once a script is written, you will want to run it on a segment of the cluster. This requires another set of Unix style Hadoop shell commands (see `http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/FileSystemShell.html`). Users must then specify the file path(s), the desired output directory, and where the script can be found.

A.2 Getting a Key

As discussed above, this script is run from your workbench on the cluster. To gain access, you will need to set up an SSH "key" with Altiscale (see `http://documentation.altiscale.`

com/configure-ssh-from-mac-linux). Once you have obtained and sent your SSH key to Altiscale, you can log in using any bash shell from your desktop with the command "ssh altiscale".

A.3 Locating the Data

The Altiscale cluster houses 9 "buckets" of .GOV data. Each bucket contains hundreds or thousands of Web Archive Files (older version are "ARC" files, newer version are "WARC" files, but they have all the same fields). Each WARC/ARC file contains captures from the same crawl, but it (a) won't contain all of the captures from a given crawl, and (b) since the crawl is doing a lot of things simultaneously, captures of a single site can be located in different WARC files.

With so much data, there is no simple "table" or directory that can be consulted to locate a specific web page. The best way to find specific pages is to use Hive to query the CDX database. See Vinay Goel's GitHub for details about how to query CDX: `https://github.com/vinaygoel/archive-analysis/tree/master/hive/cdx`. If a user know exactly what he or she wants (all the captures of the whitehouse.gov mainpage, or all the captures from September 11, 2001), the CDX can tell you where to find them. Otherwise, users will want to query all of the buckets because there is no easy way to learn where results are stored. (Though we advise first testing scripts on a single bucket or WARC file.)

First, use the command line with SSH interface to query the data directories and see which buckets or files to run a job over. This requires the Hadoop syntax to "talk" to the cluster where all the data is stored. The cluster has a user-specific directory where users can store the results of scrapes. A user's local work bench does not have enough space to save them.

Whenever users "talk" from a user's local workbench to the main cluster, users need to use 'hadoop fs -' and then the bash shell command of interest. For a list of Hadoop-friendly bash shell commands, see: `http://hadoop.apache.org/docs/current1/file_system_shell.html`. For example, the line of code

```
hadoop fs -ls
```

pulls a listing of the files in your personal saved portion of the cluster (in addition to the local workbench, each user has a file directory to save the results). As well,

```
hadoop fs -ls /dataset-derived/gov/parsed/arcs/bucket-2/
```

would draw up all the files in Bucket 2 of the parsed text ARCS directory.

A.4 Defining Search Terms

Scripts that deal with text are best written in Apache Pig. Hadoop also supports Apache Hive, Giraffe and Spark. To find and collect terms or URLs of interest, users will need to write a script. For example, users might write a script to flag any captures that have a mention of a global warming term, and return the date of the capture, URL, page title, checksum, and the parsed text. This script is saved on your local workbench and needs to have a .pig suffix. Users will need to use some sort of bash editor to write and store the script such as vi (details about how to use vi can be found above). Script is below. The first four lines are defaults and also

set the memory.

Script begins:

```
SET default_parallel 100;

SET mapreduce.map.memory.mb 8192;
SET mapred.max.map.failures.percent 10;
set mapreduce.reduce.memory.mb 16000;
set mapreduce.reduce.java.opts -Xmx8196m

REGISTER lib/ia-porky-jar-with-dependencies.jar;
```

The sequence file loader pulls the files out of the ARC/WARC format and makes them readable. Note, when they were put into the ARC/WARC format, the were run through a HTML parser to remove the HTML boilerplate. However, if the file was not in HTML to begin with, the parser will just produce symbols and this won't fix it. Users will have to deal with those issues separately.

This block allows you to load user defined functions from a Python file:

```
REGISTER 'extrawordsEKGJonJohn.py' USING jython AS myfuncs;
DEFINE FROMJSON org.archive.porky.FromJSON();
DEFINE SequenceFileLoader org.archive.porky.SequenceFileLoader();
DEFINE SURTURL org.archive.porky.SurtUrlKey();
```

When loading data on the command line (instructions below), give the data a name (here $I_Parsed_Data) and make sure to use the same "name" for the data in the command line command. This is a stand-in for the name of the directory or file over which you will run a script.

```
Archive = LOAD '$I_PARSED_DATA' USING SequenceFileLoader() AS
(key:chararray, value:chararray);
```

The below code block says: for each value and key pair, pull out the following fields. Chararray means character array - so a list of characters with no limits on what sort of content may be included in that field. The next line selects the date string. The full format is year, month, day, hour, second. Also note that because Pig is under-written in Java, users need two escape characters in these scripts (whereas only one is needed in Python). Note that "myfuncs.Threat_countWords" loads the Python UDF (below). If a user has function which selects certain URLs of interest and groups all other URLs as "other", they would run it only on the URL field. And, if a user has a function that collects words of interest and counts them as well as total words, the user should run that through the content field.

```
Archive = FOREACH Archive GENERATE FROMJSON(value) AS m:[];

Archive = FILTER Archive BY m#'errorMessage' is null;
```

```
ExtractedCounts = FOREACH Archive GENERATE myfuncs.pickURLs(m#'url'),
    m#'url' AS src:chararray,
    SURTURL(m#'url') as surt:chararray,
    REPLACE(m#'digest','sha1:','') AS checksum:chararray,
    m#'date' as date:chararray,
    myfuncs.Threat_countWords(m#'boiled');
```

In Pig, and the default delimiter is '' (new line) but many '' appear in text. So one must get rid of all the new lines in the text. This will affect our ability to do text parsing by paragraph, but sentences will still be possible. Code to get rid of the '' (new line delimiters) which are causing problems with reading in tables might look something like this:

```
UniqueCaptures = FOREACH UniqueCaptures GENERATE REPLACE(content, '\n', ' ');
```

To get TOTAL number of counts of webpages, rather than simply unique observations, merge with checksum data:

```
Checksum = LOAD '$I_CHECKSUM_DATA' USING PigStorage()
AS (surt:chararray, date:chararray, checksum:chararray);
```

Then join, flatten and output to the directory you listed in the command line:

```
FullCounts = FOREACH CountsJoinChecksum GENERATE
    ExtractedCounts::src as src,
    Checksum::date as date,
    ExtractedCounts::counts as counts,
    ExtractedCounts::URLs as URLs;

GroupedCounts = GROUP FullCounts BY URLs;

GroupedCounts = FOREACH GroupedCounts GENERATE
    group AS src,
   FLATTEN(FullCounts);

GroupedCounts = FOREACH GroupedCounts GENERATE
    src AS src,
    date AS date,
    SUBSTRING(date, 0,4),
    SUBSTRING(date, 4,6),
    URLs AS URLs,
    FLATTEN(counts);
STORE GroupedCounts INTO '$O_DATA_DIR';
```

The UDFs mention here are written in Python and can be seen in at the bottom of this Appendix.

A.5 Running the Script

To run this script, type the following code into the command line, after having logged in the Altiscale cluster with your ssh key. Users will select the file or bucket they want to run the script over, and type in an "output" directory (this will appear on your home/saved data on the cluster, not on your local workbench). Finally, users need to tell Hadoop which script they want to run. The $I_PARSED_DATA was defined as the location of the data to run the script over in the script above. Here we telling the computer that this bucket is the $I_PARSED_DATA. Next, one must load the $CHECKSUM data, and finally, give the output directory, and the location of your script.

The following should be run all as one line:

```
pig -p I_PARSED_DATA=/dataset-derived/gov/parsed/arcs/bucket-2/
-p I_CHECKSUM_DATA=/dataset/gov/url-ts-checksum/
-p O_DATA_DIR=place_where_you_want_the_file_to_end_up
location_of_your_script/scriptname.pig
```

A.6 Concatenate Results

Finally, we concatenate results across buckets on the cluster before outputing them.

```
SET default_parallel 20;

SET mapreduce.map.memory.mb 8192;
SET mapred.max.map.failures.percent 10;

REGISTER lib/ia-porky-jar-with-dependencies.jar;

DEFINE FROMJSON org.archive.porky.FromJSON();
DEFINE SequenceFileLoader org.archive.porky.SequenceFileLoader();

WordCounts = LOAD '$I_WORD_COUNTS' AS (url1:chararray, timestamp:int,
year:int, month:int, url:chararray, word:chararray, count:int);

-- note this is the location of the output from the previous script

GroupedCounts = GROUP WordCounts BY (year, month, url, word);

AggregatedCounts = FOREACH GroupedCounts GENERATE
    group.year AS year, group.month AS month, group.url AS url, group.word AS word,
    SUM(WordCounts.count) as count;

STORE AggregatedCounts INTO '$O_DATA_DIR';
```

A.7 Exporting Results

Lastly, to remove results from the cluster users need to open a new Unix shell on their local machine that is NOT logged in to the cluster with their ssh key. Then type the location of the

file they'd like to copy and give it a file path for where they'd like to put it on their desktop. For example:

The following should be run all as one line:

```
scp -r altiscale:~/results_location
/location_on_your_computer_you_want_to_move_results_to/
```

For additional scripts and for those with programing experience, see Vinay Goel's GitHub at `https://github.com/vinaygoel/archive-analysis`. For stepwise instruction of a wordcount script, see Emily Gade's GitHub at `https://github.com/ekgade/.govDataAnalysis`.

Python UDFs:

```
from collections import defaultdict
import sys
import re
from string import punctuation

@outputSchema("URLs:chararray")
def pickURLs(url):
    try:
        names =set([
            'tomudall',
            'wicker',
            'menendez',
            'warner',
            'moran',
            'voinovich',
            'webb',
            'whitehouse',
            'wyden'])

        regexp = re.compile ('([a-z]+)?(\.?\/?senate\.gov\/?\.?)([a-z]+)?')
        results = []

        result = regexp.search(url)
        if result is not None:
            if len(result.group(1)):
                if result.group(1) in names:
                    return(result.group(1))
            if len(result.group(3)):
                if result.group(3) in names:
                    return(result.group(3))

    except:
        pass
    return 'other'
```

```
# counting words

#define output schema as a "bag" with the word and then the count of the word
@outputSchema('counts:bag{tuple(word:chararray,count:int)}')
def Threat_countWords(content):
    try:
        Threat_Words = set([
            'afterlife',
            'agonstic',
            'alla',
            'allah',
            'altar',
            'amen',
            'amish',
            'angel',
            'upset',
            'vulnerable',
            'worry'])
    except:
        pass
    threat_counts = defaultdict(int)
    threat_counts['total'] = 0

    if not content or not isinstance(content, unicode):
        return [(('total'), 0)]
    splitcontent = content.lower().split()
    threat_counts['total'] = len(splitcontent)
    for word in splitcontent:
        if word in Threat_Words:
            threat_counts[word] += 1

    # Convert counts to bag
    countBag = []
    for word in threat_counts.keys():
        countBag.append( (word, threat_counts[word] ) )
    return countBag
```

B Appendix II: Lists of Terms

B.1 Religious Words

afterlife, agonstic, alla, allah, altar, amen, amish, angel, angelic, angels, baptist, baptize, belief, bible, biblic, bishop, bless, buddha, catholic, chapel, chaplain, christ, christen, christian, christmas, church, clergy, confess, convents, crucify, demon, demonic, demons, devil, divine, doom, episcopal, evangelic, faith, fundamentalist, gentile, god, goddess, gospel, hashanal, heaven, hell, hellish, hells, hindu, holier, holiest, holy, hymn, imam, immoral, immortal, islam, jesuit, jesus,

jew, jewish, juda, karma, kippur, koran, kosher, lord, lutheran, mecca, meditate, mennonite, mercifull, mercy, methodist, minister, ministry, missionary, mitzvah, mohammad, monastry, monk, moral, morality, morals, mormon, mosque, muhammed, mujahids, muslim, nun, orthodox, pagan, papal, paradise, passover, pastor, penance, pentecost, pew, piet, pilgrim, pious, pious, pope, prayer, preach, presbyterian, priest, prophet, protestant, puritan, quran, rabbi, rabbinica, ramadan, religion, rite, ritual, rosary, sabbath, sacred, sacrifice, saint, salvatior, satan, scripture, sect, sectarian, seminary, shia, shiite, shrine, sikh, sin, sinner, soul, spirit, sunni, temple, testament, theology, torah, vatican, veil, worship, yiddish, zen, zion, christian, christianity, hell, monastery, pagans, believer, believers, blessed, bless, wrath, almighty, christ, grace

B.2 DHS Words

anthrax, antiviral, assassination, attack, avalanche, avian, bacteria, biological, blizzard, bomb, botnet, breach, burn, calderon, cartel, closure, cocaine, collapse, conficker, contamination, crash, deaths, decapitated, disaster, earthquake, ebola, emergency, enriched, epidemic, evacuation, execution, exercise, explosion, explosive, exposure, extremism, farc, flood, flu, fundamentalism, gang, gangs, gunfight, guzman, h1n1, h5n1, hacker, hamas, hazardous, hazmat, heroin, hezbollah, hostage, hurricane, incident, infection, influenza, islamist, jihad, juarez, keylogger, kidnap, listeria, lockdown, looting, magnitude, malware, matamoros, methamphetamine, mexicles, michoacana, militia, mitigation, mudslide, mutation, narcos, narcotics, nogales, outbreak, pandemic, pirates, plague, plume, quarantine, radiation, radicals, radioactive, recovery, recruitment, relief, resistant, response, reynose, reyosa, ricin, riot, rootkit, salmonella, sarin, screening, security, shooting, shootout, sinaloa, smugglers, smuggling, sonora, spammer, spillover, standoff, storm, strain, symptoms, taliban, tamaulipas, tamiflu, temblor, terror, terrorism, threat, tijuana, tornado, torreon, toxic, trafficking, tremor, trojan, tsunami, twister, typhoon, vaccine, violence, virus, warning, wildfire, yuma, zetas

B.3 Anxiety Words

afraid, alarm, anguish, anxiety, apprehension, aversion, bewilderment, confusion, desperate, discomfort, distraught, distress, disturb, dread, emotional, fear, feared, fearing, fears, frantic, fright, hesitant, horrific, horrible, humiliating, impatient, inadequate, insecure, irritation, misery, numerous, obsession, obsess, overwhelm, panic, petrify, pressure, reluctant, restless, saw, scare, shake, shy, sicken, startle, strain, stress, stunned, stuns, tense, tension, terrified, terrifying, terror, tremble, turmoil, uncertain, uncomfortable, uneasy, unsure, upset, vulnerable, worry, fearful, worried, scared, suffer, suffering, need, help, miserable, apprehensive, bewildered, confused, disturbed, fearful, frightened, humiliated, miserable, obsessed, overwhelmed, panicked, petrified, scared, shaken, sickened, startled, strained, stressed, tragic, trembling, instability, upsetting, concerned