The Dynamics of the Debate about Gay Rights: Evidence from U.S. newspapers

Alan Manning
London School of Economics

Paolo Masella
Universita' di Bologna *

October 5, 2021

Abstract

Changing attitudes are the result of a battle for hearts and minds in which agents for and against change try to persuade others. We know very little about this process. We develop a methodology for measuring the intensity and the contents of media coverage for and against an idea which we apply to attitudes to gay rights. We uncover several stylized facts: First, the diffusion process of both pro- and anti-gay rights language in U.S. newspapers follow an S-shaped pattern, characteristic of diffusion processes. Anti-gay rights coverage starts its diffusion process later but then catches up. Second, in the year gay marriages are introduced we observe a dramatic increase in coverage of both pro- and anti-gay rights language; the increase in the latter is larger. The rise in coverage is still present in the three years after the institutional change. Third, there is substantial spatial autocorrelation in media coverage.

JEL-Classification: J15, Z1
Keywords: Diffusion, Discrimination, Gay Rights

*We thank conference and seminar participants at the University of Sussex, University of Bologna, University of Roma Tor Vergata, UCL (SSEES Centre for Comparative Studies of Emerging Economies), MILLS Workshop 2017 (Fondazione Rodolfo DeBenedetti), ASREC 2017, University of Alicante, University of Padova, RES 2018. We thank University of Sussex (RDF competition) and University of Bologna (Almaidea competition) for fundings. Tuomas Ketola provided exceptional research assistance. A previous draft was circulated under the title "Diffusion of Social Values Through the Lens of US Newspapers". Contact Information: Alan Manning, Department of Economics, London School of Economics, Houghton Street, London, WC2A 2AE, UK (email: a.manning@lse.ac.uk); Paolo Masella, Department of Economics, University of Bologna, Piazza Scaravilli 2, 40126, Bologna, Italy (email: p.masella@unibo.it).

1 Introduction

Attitudes and values are shaped by family transmission (Bisin and Verdier (2001)), education (Cantoni et al. (2017)), political institutions (Alesina and Fuchs-Schündeln (2007)) and media (La Ferrara (2016)). Attitudes and values also change because individuals or groups argue the case, and, typically, there are those arguing against change as well as for it. Change in attitudes is often the outcome of a battle for hearts and minds between those proposing and opposing change. This process is less studied and often invisible in data which simply record whether someone supports or opposes a particular view.

To give a specific example (the application of this paper), consider attitudes to gay rights in the United States. Public opinion is shifting towards more liberal attitudes: according to the General Social Survey (GSS) in 1988 only 12% of adults agreed with the statements that "homosexual couples should have the right to marry one another" and 73% disagreed. By 2014 56% agreed and only 32% disagreed. Other measures of attitudes towards gay rights show similar, rapid, change. One might think that increasingly liberal attitudes to gay rights has been a process in one direction with those favouring increasing in number and those opposing decreasing. One of the contributions of this paper is to show that the process has not been a simple move towards more liberal views, that it has been actively contested. We show this by providing an analysis of the coverage of pro- and anti-gay rights language in US newspapers, an important (though not the only) arena where these debates play out. The actual beliefs people end up with can be thought of as the outputs of the battle for hearts and minds while our paper is focused on the inputs in that battle.

To define measures of intensity of pro- and anti-gay rights language in U.S. digitised newspapers we use a strategy similar to Laver et al. (2003) and Gentzkow and Shapiro (2010). We construct a training textual corpus from speeches in the U.S. Congressional Records from 1994 to 2012, selecting phrases that are diagnostic either of pro- or anti-gay rights views according to whether they are relatively more likely to be in a speech given by a pro-gay rights politician than one given by an anti-gay rights politician, where politicians are classified as pro-gay rights or anti-gay rights based on their voting records on gay-related issues. We then compute the frequency of these phrases in U.S. digitised newspapers. We complement these intensity measures with more detailed measures of the contents of the debate. We detect 20 different topics and label (most of) them based on the words they are most likely to include; as expected the most relevant topics are the ones connected to gay marriage, hate crimes, discrimination and HIV. For each of the newspaper articles we then also have an estimate of the share of text devoted to each of the 20 topics. As a result we end up with a unique and very rich dataset containing very high frequency and fairly geographically detailed information for a long period of time (starting from 1982 onwards).

Having constructed these measures of the intensity (and contents) of pro- and anti-gay rights

language, we first investigate how they have changed over time. They both follow an S-shaped pattern, characteristic of diffusion processes. The intensity of pro-gay rights language starts to rise earlier than the anti-gay rights one in the middle of the 1980s. This is not surprising as a challenge to the status quo is probably needed for attitudes to start to change. Inspection of the language used suggests that the start of pro-gay rights debate seems to be connected to the discussion of topics related to the HIV epidemic and the rights of people affected by the disease. The debate then also focused on protection against crimes targeting the gay communities and finally on legal unions and gay marriages. After 2000 the intensity of the pro-gay rights language rises at a lower rate.

This evolution of the frequency of pro-gay rights language is not that surprising given that public attitudes are becoming more supportive of the views expressed. What is surprising is the evolution of the intensity of anti-gay rights language. It starts increasing later than pro-gay rights language (in particular when the debate turned to gay marriage), also follows an S-shape eventually catching up with the number of pro-gay phrases. This might suggest that change has been actively opposed with the intensity of anti-gay rights language increasing even as public opinion moved in the opposite direction. This is something invisible in data that simply asks for opinions. A possible interpretation of this pattern in anti-gay rights language is that it is only required once the pro-gay rights views begin to spread (so starts after pro-gay rights language) but the fact that the tide is flowing against anti-gay rights views makes those opposed to this change increase their investment in opposing pro-gay rights views and invest almost as much as the proponents of pro-gay rights views do.

Consistent with this view is the behavior of the intensity of pro- and anti-gay rights language around periods of institutional change, specifically gay marriage. For the year gay marriages are introduced we observe a dramatic increase in the frequency of both pro- and anti-gay rights language in the press but with the increase in the latter being higher. Our hypothesis is that this represents the attempt by those opposed to change to prevent institutions moving in a more liberal direction. It is perhaps not surprising that the intensity of debate rises around the time of significant decisions but the rise in press coverage is still present in the three years after the change as those opposed to change might have been actively working to reverse gay marriage laws.

We also analyse county level measures of coverage. We present evidence of persistence over time in the degree newspapers cover pro- and anti-gay rights language: county level measures of coverage a decade ago seems to be a relatively strong predictor of the current county level measures. And we also document the existence of substantial spatial autocorrelation in coverage. Spatial autocorrelation might be driven by the spread of public opinion from a county to its neighbouring counties or also by newspapers in the same area discussing similar issues. We find that the spatial autocorrelation disappears when we control for state-time fixed effects; this suggests that it might be driven by news at state level rather than public opinion changes.

The paper contributes to a large literature studying the determinants of values, beliefs and cultural traits that are relevant for economic outcomes. Institutions such as property rights (Di Tella et al. (2007)) and the form of government (Alesina and Fuchs-Schündeln (2007)) have been found to influence trust and preferences for redistribution; language of education (Clots-Figueras and Masella (2013)) and contents of teaching (Cantoni et al. (2017)) shape individual ethnic identity and political attitudes, while macroeconomic shocks experienced at a young age affect preferences for redistribution during adulthood (Giuliano and Spilimbergo (2014)). The literature on understanding attitudinal change is, however, relatively small (though see Fernandez (2013) and Fogli and Veldkamp (2011)) and this paper aims to provide some insight into the process, in particular into the battle for hearts and minds at the root of attitudinal changes. We do not focus on public opinion, the output of that battle, but on the media debate, which we interpret as an input in the battle.

There are also connections with the literature that tries to measure discriminatory attitudes and in particular attitudes towards homosexuals. Beaman et al. (2009) and Burns et al. (2015) discussed possible drivers of attitudes towards discriminated groups: Beaman et al. (2009) find that prior exposure to female leaders reduces stereotypes about gender roles and negative biases against female leaders among male villagers, Burns et al. (2015) find that random exposure to roommates of a different ethnic group in the double rooms of the University of Cape Town weakens the prejudice associated with that group. More closely related to our work, a large social science literature discusses the effects of discriminatory attitudes towards homosexuals on labor market outcomes (see Tilcsik (2011), Plug et al. (2014) and Mize (2016)), but most importantly for our purposes this literature studies what are the drivers of changes in public opinion about the topic (see Andersen and Fetner (2008), Adamczyk and Pitt (2009), Lewis and Gossett (2008), Redman (2018), Harrison and Michelson (2015) and Chomsky and Barclay (2010)) and of changes in the related legislation (see Hansen and Treul (2015), Wald et al. (1996), Reynolds (2013), and Lax and Phillips (2009)). More recently Fernández et al. (2019) discuss the relationship between HIV epidemic, political mobilization and attitudes towards homosexuals. We look directly at the propagation of media coverage of pro- and anti-gay rights language using the lens of US newspapers and speeches in the US Congress.

We also relate to an emerging literature on the relationship between mass media and individual preferences, social and economic outcomes.[1] La Ferrara et al. (2012) and Kearney and Levine (2015) show how mass media and in particular TV programs are likely to have affected fertility decisions by introducing different role models, weakening old stereotypes and altering common perceptions about the role of women in society. Durante et al. (2019) investigate how media con-

[1] See La Ferrara (2016) and Della Vigna and La Ferrara (2015) for reviews of the literature and for a theoretical perspective on the topic.

tent affects political preferences, Lim et al. (2015) and Jetter (2017) discuss how the contents of newspapers may affect judicial decision and terrorist activities, respectively. Our paper does not investigate the causal impact of media content on social and economic outcomes, but instead focuses on media debate as the main outcome and identify patterns of diffusion of media content. To this end we directly measure the extent to which mass media (and newspapers in particular) report over time and across space key phrases that are related to a debate about a very controversial topic, such as discrimination against homosexuals.

Our study also contributes to the recent literature that adopts text analysis techniques to extract variables of interest for social scientists (see, for example, Stephens-Davidowitz (2014) who uses Google search data to measure racial animus, Schwarz et al. (2017) who detect legislators' preferences from their speeches in Parliament to better capture intra party differences, and Gentzkow and Shapiro (2010) who measure media bias using newspaper articles). Our work uses a mix of two different methods, a dictionary based and an unsupervised machine learning method,[2] to capture both the intensity of the debate in favour or against a given political and cultural change and what are the topics discussed by both sides of the debate.

Section 2 describes the procedures adopted to build our reference data sets. In Section 3 we discuss the patterns of temporal diffusion of press coverage we find, in Section 4 the coevolution between press coverage and institutional changes. Section 5 presents evidence of spatial diffusion while in Section 6 preliminary evidence of persistence over time of coverage of pro- and anti-gay sentiment is presented; Section 7 concludes.

2 Data

This section reports the strategies used to obtain both (i) measures of the intensity of coverage of pro- and anti-gay rights language by the press and also (ii) measures of the content of the articles using pro- and anti-gay rights language. The first set of measures help us isolate patterns of spatial and temporal diffusion of coverage of pro- and anti-gay rights language, our main proxy for the inputs in the battle for hearts and minds we study, the second set may shed some light on some of the mechanisms behind such patterns. In order to get coverage intensity measures we use a dictionary based method, while, as in recent work by Hansen et al. (2018), we use the Latent Dirichlet Allocation method, an unsupervised machine learning method, to have a better understanding of the contents of the relevant newspaper articles. The dataset generated has three important features: (i) it spans a fairly long timeframe so to make possible the study of the time diffusion of media coverage (ii) it contains fairly detailed geographical information that allows an

[2]Unsupervised machine learning methods have been also used to catalogue topics of discussion of speeches given in the US Congress (Quinn et al. (2010)) and in the FOMC meetings (Hansen et al. (2018)).

analysis of patterns of spatial diffusion (iii) it is high frequency and therefore suitable for event analysis studies of the coevolution between institutional change and coverage of pro- and anti-gay rights language.

2.1 Intensity of Coverage

This section describes the strategy used to obtain measures of the intensity of pro- and anti-gay rights language in the press. We follow a three step procedure following the approach used by Gentzkow and Shapiro (2010) to estimate the political bias of US newspapers. We start by isolating a reference corpus where we can identify documents (or parts of documents) that are expressions related to debates about gay rights. Then within this reference corpus we use an algorithm that identifies two word phrases (bigrams) that are an expression of pro- or anti-gay rights views. Finally, a script search for these bigrams within a corpus of U.S. digitized newspapers allows us to construct several measures of how strongly the press debate reflects pro- and anti-gay rights views at both national and local level.

2.1.1 Reference Corpus

As in Gentzkow and Shapiro (2010) we use text from the Congressional Record as our reference corpus, using all issues from 1994 to 2012, downloaded from thomas.loc.gov, and corresponding to the entire set of speeches given in the 103rd (only the part from 1st of January 1994) to 112th Congresses.[3] An automated script then identifies 1220282 speeches and the corresponding speakers.

We then separate the speeches along two dimensions, depending on the identity of the speaker and the content of the speech. We distinguish between speeches performed by pro-gay rights, anti-gay rights speakers and "indifferent" speakers according to the voting record of the speakers on issues related to gay rights. We use scorecards provided by a non-profit organisation, the Human Rights Campaign, that is the largest LGBT civil rights advocacy group and political lobbying organisation in the United States. It provides us with the set of votes held in each Congress that the organisation classified as relevant for LGBT civil rights and the corresponding voting record of each member of Congress. We classify a Senator/member of the House of Representatives as pro-gay rights in a given Congress if they voted in a pro-gay rights fashion in more than 75 per cent of the votes selected by HRC during that Congress, as anti-gay rights if they voted in a pro-gay rights fashion in less than 25 per cent of the votes, "indifferent" otherwise.

[3]Selecting text from the Congressional Record as our training corpus does not mean we suggest that political language drives media views on the topic of interest or vice versa that media shapes political debates; it is a method to identify the language used by both sides of the gay rights debate.

We also separate speeches into ”topical” and ”non topical” groups depending on whether they contain keywords signalling that the speech concerns gay rights. The keywords are: ”gay”, ”lesbian”, ”same sex”, ”transgender”, ”transsexual”, ”pro-gay”, ”anti-gay”, ” homo”, ”heterosexual” ”gender identity”, ”sexual identity”, ”LGBT”, ”GLBT”.[4] Because there is an arbitrary component in the keyword selection, we use several robustness checks in which we experiment with a larger and smaller set of keywords and with an algorithm that mitigates the human component by introducing a machine element (the algorithm is detailed in the Appendix).[5] Independently of the keyword set it is however worth noting that a large majority of the speeches are classified as ”topical” or ”non topical” based on three fundamental keywords: gay, homosexual and same-sex.

Then, as conventional in the literature, we eliminate extremely common words (stopwords) from the speeches and we stem each word left, that is we strip it to its linguistic root.[6] Finally we consider the entire set of pairs of two consecutive (stemmed) words which we will refer to as bigrams (the entire reference corpus contains 86717736 bigrams).

2.1.2 Reference Phrases

Within the training corpus we select the set of bigrams that best represent the language of the pro-gay rights and anti-gay rights politicians, identified by their voting record on gay-related issues as described earlier. To isolate such bigrams we use a two step strategy: first we identify ”topical” bigrams, that is bigrams related to LGBT civil rights topics, then within such ”topical” bigrams we select the ones that are diagnostic of the politician’s attitude to gay rights.

In order to isolate ”topical” bigrams, we calculate the frequencies of each bigram in both the ”topical” and ”non topical” set of speeches as defined in the previous subsection. To reduce the dimension of our datasets we then restrict our attention to a limited set of bigrams in ”topical” speeches (the 1500 most frequent bigrams) and in ”non topical” speeches (the 200000 most frequent bigrams).[7] Then within the restricted set of bigrams in the ”topical” speeches we identify as ”topical” bigrams only the bigrams that are either not present in the restricted set of bigrams in ”non topical” speeches or disproportionately more frequent in ”topical” speeches than in ”non topical” speeches (that is their frequency is 500 times higher in ”topical” speeches than in ”non topical” speeches). While we believe this strategy helps us disregard very generic phrases that may be only mildly related to the gay-rights debate, it may miss some substantial ideological language

[4] In the script this list is extended to include plurals, capital first letters, commas and dots, and spaces to surround the words.

[5] Both the smaller and larger list of keywords are reported by Table A5.

[6] For example, the words taxation, taxing, taxed would be all reduced to ”tax”. The outcome of this process, however, does not necessarily coincide with an English word. We use the stemmer introduced by Porter and his list of stopwords.

[7] The different thresholds reflect the different numerosity of the speeches in the ”topical” and ”non topical” subset of the reference corpus.

in topical speeches that isn't only present when discussing gay rights. In order to attenuate concerns related to the selection of the phrases we present exercises where we vary the following three relevant thresholds: (i) we focus on a larger (2000) and smaller (1000) set of most frequent bigrams contained in "topical" speeches (ii) we focus on a larger (250000) and smaller (150000) set of most frequent bigrams contained in "non topical" speeches (iii) we only focus on bigrams that are not present in the restricted set of bigrams in "non topical" speeches and neglect bigrams that instead are disproportionately more frequent in the restricted set of bigrams in "topical"speeches.

In the second step, within the set of "topical" bigrams, we identify the ones that are diagnostic of the views of the pro-gay rights and anti-gay rights speakers. For each of the "topical" bigrams we perform a Pearson Chi test whose null hypothesis is that the propensity to use the bigram is the same among pro-gay rights speakers and anti-gay rights speakers as defined in the previous sub-section. We then separate bigrams into pro- and anti-gay rights bigrams depending on whether they were relatively more frequent among pro- or anti-gay rights speakers. Then, within both the set of pro- and anti-gay rights bigrams we rank them according to their Pearson Chi value and finally select the top 30 bigrams diagnostic of the language of pro-gay rights speakers and the top 30 bigrams diagnostic of the language of anti-gay rights speakers. The complete list of the 60 bigrams is presented in Table 1. The language of pro-gay rights and anti-gay rights politicians are very different, with pro-gay rights speakers focussing on issues ranging from anti-discrimination policies to hate crimes, while bigrams by anti-gay rights speakers are more likely to be related to the marriage institution. Moreover, while pro-gay rights speakers seem to use the words gay and lesbian very frequently in their speeches, anti-gay rights speakers are more likely to use bigrams containing the words homosexual and same-sex. In the Appendix (Tables A1-A6) we show a list with larger set of bigrams (80) and several alternative lists of bigrams obtained using different criteria, that is by varying the initial set of keywords or the thresholds above mentioned.[8]

[Table 1 here]

For each of the final set of 60 bigrams, we then identify the unpreprocessed phrases that within the set of "topical" speeches are associated to the bigram. For instance, there are 11 unpreprocessed phrases associated with the bigram "base sexual" and they are: "based only on their sexual", "based on both sexual", "based on the sexual", "based on his or her sexual", "based on its sexual", "based on his or her sexuality", "based on their sexuality", "based on sexuality", "based on his sexuality", "based on sexual", "based on their sexual". In total we identify approximately 600 unpreprocessed phrases that are going to be searched in the newspaper corpus.[9]

[8] In Table A7 we also provide a detailed list of all the other phrases deleted throughout all the exercises performed while the reasons behind their elimination are given in the Appendix.

[9] Approximately 4000 if we consider the relevant bigrams for all the exercises performed throughout the paper.

2.1.3 Newspaper Corpus

Our main data source is the set of digitised newspapers provided by newslibrary.com which contains approximately 4500 newspapers, though more in recent years. We run an automated script that performs a search within this database for each of the previously identified unpreprocessed phrases and delivers the title, a short abstract and the day of publication of all articles containing this phrase.

Previous research (see Gentzkow and Shapiro (2010) and Baker et al. (2016)) has used this database and since then the number of newspapers and articles digitized provided by this source increased spectacularly.[10] In order to minimize potential selection issues related to the date of inclusion of each newspaper in the database and to the number of articles digitized within each newspaper we always use within newspaper variation and collect information on the total amount of articles digitised each year from each newspaper using an automated script that for each newspaper-year cell delivers the number of articles containing the empty space " ". This allows us to conduct robustness checks where we normalize the number of pro-gay rights and anti-gay rights phrases found by the total volume of news being produced.

Finally, we associate each newspaper to county (or State) geographical areas in the following manner: the classification in newslibrary.com identifies whether a paper is international (if so we drop it), national or local. We then assign counties to local newspapers in different ways depending on what information is available. For a large majority of the newspapers newslibrary.com reports a city of origin. In the simplest case we find the county in which the city is located and assign that as the county of the newspaper. However, sometimes the city in question is in multiple counties. In this case, we estimate in which of these counties the newspaper is read the most. This estimation is done in the following manner: we make a Google search for the name of the newspaper and the name of an individual county. We perform this search for all the counties in which the city is. Taking into account the population of the counties we use the equation below to estimate the prevalence of the newspaper in each county: prevalence=(number of Google hits)/(county population). In order to avoid to mechanically generate spatial autocorrelation, the county with the highest prevalence is then assigned to the newspaper. For some papers newslibrary.com does not report a city and therefore we had to find the newspaper's own website in order to understand where it is distributed. For a smaller set of newspapers we also have the number of copies sold in 2004 as reported by Gentzkow and Shapiro (2010). As a robustness check we often present results obtained weighting newspapers by copies sold in 2004.

[10] In Gentzkow and Shapiro (2010) the total sample consists of 433 newspapers

2.1.4 Relevant Intensity Measures

We now define the coverage measures used in the empirical analysis. Within each newspaper n and in time interval t (usually month or year) for each of the unpreprocessed phrases p associated to one of the 60 selected stemmed phrases (30 pro-gay rights and 30 anti-gay rights) we calculate the number of articles reporting each unpreprocessed phrase at least one time,[11] and denote this variable as $Numb.ofArticles_{p,n,t}$. Then we calculate $Coverage_{g,n,t}$, that is the pro- (anti-) gay rights coverage measure (pro-gay rights if $g = pro$, anti-gay rights if $g = anti$) of a newspaper n within a given time interval t simply by taking the sum of $Numb.ofArticles_{p,n,t}$ over the entire set of the unpreprocessed phrases associated to the 30 pro- (anti-) gay rights stemmed phrases, that is $Coverage_{g,n,t} = \sum_{p=1}^{P_g} Numb.ofArticles_{p,n,t}$.[12]

We also consider measures that weight search results differently, that is $WCoverage_{g,n,t} = \sum_{p=1}^{P_g} W_p * Numb.ofArticles_{p,n,y}$. The weights W_p reflect the relevance of the bigram associated to the unpreprocessed phrase p and can be based either on the relative frequency of that bigram or on its relative Pearson Chi value within the set of pro- (anti-) gay rights bigrams. As discussed in the previous section for each bigram we calculated its total frequency within topical Congressional Speeches and a Pearson Chi value based on a test whose null hypothesis is that the propensity to use that bigram is the same among pro-gay rights speakers and anti-gay rights speakers.

Finally, we also build coverage measures at geographical level (at county, state or regional level) by taking the average of the coverage measure of the newspapers published within the same geographical area. For instance $Coverage_{pro,c,t}$ is the pro-gay rights coverage measure of a county c at time t and it is obtained simply by taking the average of $Coverage_{pro,n,t}$ over the set of newspapers published within the same county c at time t. In this case we disregard newspapers identified as national newspapers by the website. A newspaper n is assigned to a county c according to the procedure previously explained. Alternative county level measures are obtained weighting search results differently and weights again reflect relative frequency or relative Pearson Chi values; $WCoverage_{pro,c,t}$ is simply the average of $WCoverage_{pro,n,t}$ over the set of newspapers published within the same county c at time t.

[11] We denote as P_g the total number of pro-gay rights unpreprocessed phrases if g=Pro and the total number of anti-gay rights unpreprocessed phrases if g=Anti.

[12] Note that in the measure $Coverage_{g,n,t}$ the same article may be counted several times if the text of that article includes more than one unpreprocessed phrase and it may be counted among both the pro- and the anti-gay rights measure if the text contains unpreprocessed phrases associated with both the pro-gay and anti-gay rights stemmed bigrams. In the Appendix we provide robustness checks where we only consider articles that either contain only unpreprocessed phrases associated to pro-gay rights bigrams or only contain unpreprocessed phrases associated to anti-gay rights bigrams and we count each article only one time independently of the number of unpreprocessed phrases contained.

2.2 Contents of Coverage

As a second step we identify the subject matters of articles containing pro- or anti-gay rights language. In the previous exercise we downloaded the title, the day of publication and the abstract of all the articles containing at least once one of the phrases connected to the 60 bigrams selected. For this second exercise we use as our main corpus the entire set of abstracts downloaded during the first exercise.[13] As before we eliminate stopwords from the abstracts and we stem each word left, that is we strip it to its linguistic root. We further reduce the dimension of our dataset by eliminating stemmed words that are less informative. We do that by calculating for each stemmed word in the corpus the tf-idf value (term frequency-inverse document frequency), a measure that penalises stemmed words that are either very rare or appear in many abstracts. We then rank the stemmed words based on the tf-idf value and based on Figure A1 in the Appendix we decided to drop all the stemmed words with a value below 25, therefore remaining with a total of 137579 stems. At this stage each abstract therefore will be identified by a set of stemmed words. We then apply the Latent Dirichlet Allocation (LDA) method to the corpus as in Hansen et al. (2018).[14] LDA is an unsupervised machine learning algorithm that by analysing the words of a collection of documents is able to elicit the themes/topics running through each document/abstract. While the documents (and the words contained in them) are observed, the topic structure of the corpus is the hidden component of the model. By topic structure we mean the share of text of each document devoted to each topic and the distribution over the vocabulary (the set of 137579 terms) for each topic. The goal of the algorithm is to infer the topic structure of the corpus based on the observed set of words present in each document/abstract.

In order to estimate the LDA algorithm we first need to select the number of topics K. We choose K=20 because of interpretability after having experimented with different values. Then we apply the Markov chain Monte Carlo algorithm as described by Griffiths and Steyvers (2004).[15] The LDA algorithm then estimates how each abstract in the corpus is divided among the 20 topics, that is the fraction of the abstract devoted to each topic. As a result within a dataset where each observation represents an abstract of the corpus we build twenty additional variables (one for each of the 20 topics), each of them identifies for each abstract the fraction of its text devoted to a given topic. The LDA also estimates the the topic distribution over the vocabulary that, for simplicity, we represent graphically as a world cloud. Figure 1 shows the word clouds of the most relevant ones for our purposes. Topic 3 seems to be related to Supreme Court decisions about gay marriages,

[13] Since we are mainly interested in the diffusion over time of the contents discussed and since the number of abstracts is very different across years because of digitisation, we run the LDA on an artificially rebalanced sample/corpus, that is we generate a corpus where the number of abstracts is the same across years by randomly duplicating abstracts published in the years with lower initial number of articles.

[14] See Hansen et al. (2018) and Blei (2012) for a more accurate discussion of the method.

[15] We extract 10 samples from 8000 iterations.

Topic 4 to hate crimes, Topic 7 to HIV, Topic 9 to generic anti-discrimination policies, Topic 13 to religion, Topic 15 to gay in the militaries, Topic 17 to legal unions. In Figure A2 of the Appendix we report the word clouds of each of the 20 topics.

[Figure 1 here]

2.3 Survey Data

Finally we test the reliability of our measures of press coverage of pro- and anti-gay sentiment by relating them to survey data gathering opinions about gay-related issues such as gay marriages. We expect a larger share of respondents in favor of gay marriages the larger is the press coverage of pro-gay sentiment and the smaller the coverage of anti-gay sentiment. This is because either we expect press coverage to reflect the current opinion of the reader or to influence it.

We use data from the General Social Survey (GSS), and in particular the survey question on whether a respondent strongly agrees/neither agrees or disagrees/strongly disagree with the following statement "Homosexual couples should have the right to marry one another". As a dependent variable we have a variable going from 5 if the respondent strongly agrees with the statement to 1 if he strongly disagrees. In order to build the main explanatory variables (that is a regional yearly measure of pro- and anti-gay media coverage) for each newspaper-year[16] we consider the standard pro- (anti-) gay newspaper level coverage measure[17] and then take the average for all the newspapers within each of the 9 regions identified by the GSS.[18] We then simply regress the intensity of the opinion of an individual i resident in a region r on our regional measures of pro-gay and anti-gay press coverage controlling for survey year fixed effects, region of residence fixed effects, region specific time trends, age, race, and gender of the respondent.[19] Results reported in Table 2, column 1, show a significant positive (negative) correlation between the intensity of support towards gay marriage and the press coverage of pro- (anti-) gay sentiment.

Similar findings are documented when we consider different survey questions during the same survey period about similar topics, i.e. whether homosexuals should be allowed to make speeches in their community or teach in a college or university, whether libraries should allow books in favor of homosexuality, whether or not sexual relations between two adults of the same sex are wrong and whether same sex male/female couples can bring up a child as well as a male-female couple. Results are reported by Columns 2-7 of Table 2.

[Table 2 here]

[16] We only consider newspapers that have been digitised from 2004 as GSS collected opinions about this issue in 1988 and every 2 years starting from 2004. We drop 1988 as the number of digitised newspapers is extremely limited.

[17] See previous discussion on how this variable is generated.

[18] The regions are New England, Middle Atlantic, East North Central, West North Central, South Atlantic, Mountain, Pacific, East South Central and West South Central.

[19] Standard errors are clustered at region-year level.

3 Diffusion over time

3.1 Intensity of the debate

We now investigate how the frequency of pro-gay rights and anti-gay rights language in the press changes over time.

We start with a simple exercise: we consider the time period 1982-2014 and for each year we calculate our coverage measure at newspaper level. Then we run the following regression:

$$Coverage_{g,n,t} = \alpha_n + \beta_{g,t} + X_{n,t}\gamma + \varepsilon_{g,n,t}. \quad (1)$$

where the dependent variable $Coverage_{g,n,t}$, the coverage measure of type g (and we only have two types, pro-gay rights and anti-gay rights) of a newspaper n in the year t, is regressed on newspaper fixed effects α_n (to partially control for omitted variables related to the choice over time of what newspapers were digitised and included in the newspaper corpus) and on interactions of type and year fixed effects, $\beta_{g,t}$. $X_{n,t}$ is our proxy for the number of digitised articles, a variable that varies at newspaper-year level and that we introduce in some of our specifications.

[Figure 2 here]

Figure 2 (Panel A) plots the $\beta_{g,t}$ coefficients showing change over time. The level of the media coverage of both pro-gay rights and anti-gay rights language seems to follow an S-shape, characteristic of diffusion processes. Pro-gay rights media coverage starts to rise 5 to 10 years earlier than anti-gay rights coverage but then anti-gay rights coverage reacts, most likely to prevent attitudes to change in a more liberal direction, and catches up in the latest 10 years of the sample.[20] We also observe a spike in pro-gay rights coverage in 1993, most likely related to the legislation on gays in the military service being discussed around that year. After 2000 the intensity of the pro-gay rights language rises at a lower rate. Similar results apply if we use as a dependent variable the coverage measures obtained by weighting each article by the relative relevance of the bigrams contained. In Panels B and C of Figure 2 we use relevance measures based on relative frequency and on relative Pearson Chi Values, respectively.[21]

[Figure 3 here]

To further control for selection issues related to the choice over time of what fraction of newspapers articles were digitised and included in the newspaper corpus, Figure 3 Panel A includes as control variable X our proxy for the number of digitised articles for each newspaper-year cell (see the Data Section for details about how this variable is calculated); the patterns are consistent with previous discussion. As further robustness checks in Panel B we build our dependent variable

[20] The difference between the pro- and the anti-gay rights coefficients is significantly different from zero in the period between 1992 and 1997.

[21] See Section 2.1.4 for a discussion of how these measures are derived.

by dividing the main coverage measure in specification (1) by our proxy for the total number of articles. In Panel C we estimate regression (1) but using as weights for each newspaper the number of copies sold in 2004. Patterns in Figure 3 are quite similar to the ones discussed before; it is however less clear that the media diffusion of the anti-gay arguments, although increasing over time, presents an S-shaped pattern.

As robustness checks we use the set of phrases presented in Tables A1-A6 of the Appendix obtained by varying the set of keywords, the thresholds of inclusion of bigrams in the topical and non topical sets and the number of reference bigrams. Results are also robust when we consider articles that either only contain unpreprocessed phrases associated to pro-gay bigrams or only contain unpreprocessed phrases associated to anti-gay bigrams and we count each article only one time independently of the number of unpreprocessed phrases contained. See Figure A3 in the Appendix.

Finally, in Figure 4 we perform similar exercises but collapsed our information at monthly level, rather then yearly. Now, instead of 33 years on the X-axis we have 396 points for each month from January 1982 to December 2014. We estimate the coefficients for each month-type cell and we plot them in Figure 4. Patterns are again similar to those discussed previously and the shorter time interval does not seem to introduce too much variability in our estimates. Panel A reports the plot obtained using the standard measure $Coverage_{g,n,t}$ where t corresponds to monthly time intervals, Panel B and C the plots obtained using relevance measures based on relative frequency and on relative Pearson Chi Values where t again corresponds to monthly time intervals. Results obtained are consistent with previous discussion. In Figure A4 Panels A-C in the Appendix we report robustness checks analogous to the ones in Figure 3.

[Figure 4 here]

3.2 Content of the debate

We next present evidence that might help us understand how the contents of the debate has evolved over time, why the debate started in the second half of the 80s (in particular among articles containing pro-gay rights language) and why the coverage of anti-gay rights language started catching up with the coverage of pro-gay rights language with a delay of five to ten years. To this end for each stemmed phrase we consider the abstracts of the articles containing it at least one time. Then we generate two corpora, one consisting of the abstracts of the articles containing a pro-gay rights phrase the other consisting of the abstracts of the articles containing an anti-gay rights phrase.[22] Then for each year and for each topic we take the average of the share of an abstract devoted to a

[22]Please note that the same abstract can appear in both corpora if it refers to articles containing both pro- and anti-gay rights stemmed phrases.

topic within the sample of all the abstracts included in the pro/anti gay corpus. In Figure 5 we report these statistics for the main topics of interest, in Figure A5 of the Appendix we report statistics for all the twenty topics.

[Figure 5 here]

We start by noticing that there are several similarities in the topics structure of both the pro- and anti-gay corpus, similarities that may be driven by shifts in the political agenda. The start of the pro-gay debate seems to be related to the discussion of one topic in particular, the topic associated with stems such as HIV, AIDS, drug, health care, virus, etc. During the middle and the second part of the 80s around fifteen per cent of the debate was connected to issues such as HIV epidemics, anonymity of the tests, and care for people affected by the disease. The start of the pro-gay debate might therefore partly have its origin in more media attention to this disease and the welfare rights of the patients.

Interestingly, by looking at the evolution over time of the share of the articles devoted to each of the 20 topics within the pro-gay rights corpus we can also notice how the prevalence of the HIV topics ended at the end of the 80s and the beginning of the 90s when other topics began to be more important, such as the topic characterised by the protest against hate crimes and the demand for protection against them. In the middle of the 90s the salience of the topic about legal union among same sex partners started growing and peaked ten years later. Most recently, the topic associated with Supreme Court decisions (most likely about gay marriages) became by far the most important topic. The study of the evolution of the contents of the debate suggests an escalation of the rights demanded by the pro-gay rights group starting from welfare rights connected to HIV, moving to protection against crimes targeting the gay communities and finally to legal unions and gay marriages.

The anti-gay rights coverage started to catch up with the growth of pro gay rights coverage between 5 and 10 years later, most likely in connection with the increasing salience of topics such as legal unions between same sex partners (and later gay marriages). As we can see from Figure 5 Panel B in the middle of the 90s the legal union topic became crucial within the anti-gay rights debate while most recently the topic about Supreme Court and gay marriages covered one quarter of the entire anti-gay rights corpus. Preliminary evidence seems to suggest that when the rights demanded by the gay community (and covered by newspapers) turned to the right of marriage then the anti-gay rights groups reacted and became more vocal in the media.[23]

[23]Similar results are obtained if we adopt the strategy in equation (1) and we use as dependent variable $ShareTopic(i)_{g,n,t}$, that is the share of the corpus of type g within the newspaper n during the year t devoted to topic i.

3.3 Discussion

A careful empirical analysis of the many possible determinants of the time diffusion patterns in newspapers' language is beyond the scope of this paper. Newspapers might simply be reacting to some combination of outside events and political language. For example, pro- and anti-gay speech in Congress might be increasing over time, and local newspapers react by covering their local politicians' views. It may also be that newspapers have a fixed propensity to engage in ideological language and as gay rights became a more and more prominent issue in the last thirty years, newspapers shifted their ideological language towards that agenda. Finally, media bias in general may be due to demand factors and the time diffusion patterns in newspapers' language reflect underlying changes in population's attitudes towards gay rights.

As already mentioned we use speeches in Congress as a training dataset where we can isolate pro- and anti-gay language but we are definitely not able to establish a causal relationship between media and political debates. Fig. 6 shows the yearly diffusion of pro- and anti-gay language within the Congressional Record corpus that includes issues from 1994 to 2012, where pro- and anti-gay language are measured by the number of times the unprocessed pro-/anti-gay phrases selected in Section 2 (and searched within newspapers) are adopted by Congress people. When comparing pro- and anti-gay language patterns in newspapers and congressional speeches we find some similarities but also several differences. We observe anti-gay language spiking in 2004 in both datasets; this is probably because in 2004 both the Senate and the House of Representatives discussed a federal marriage amendment to legally define marriage as the union of a man and a woman. In both datasets we also observe that pro-gay language is often above anti-gay language; however, when we consider the corpus of congressional speeches we do not see the increasing pattern over time for both pro- and anti-gay language that we observe so clearly within the newspaper corpus. Certainly it is not possible to compare magnitudes in Figures 4 and 6 as the two measures are different and refer to different contexts; an analysis of the patterns emerging from the two figures may, however, help drawing some preliminary conclusions. While it is likely that newspaper texts partly react to political speeches (as the 2004 peak seems to suggest) we also believe it is unlikely the entire variation within the newspaper corpus can be explained by coverage of political speeches.

[Figure 6 here]

We then checked whether the diffusion pattern of pro-/anti-gay language differs depending on the ideology of the newspapers. We use two different variables to proxy for the ideology of a newspaper: what president they endorsed in 2004, as coded by Gentzkow et al. (2010) (we separate newspapers depending on whether they endorsed the Republican or the Democratic candidate), and their slant, as measured by Gentzkow and Shapiro (2010) (we rank the newspapers according to this measure and separate the sample in 4 groups of equal size). Panel A (B) of Fig. 7 displays the time diffusion graph for newspapers who endorsed the Republican (Democratic) candidate;

Panels C-F of Fig. 7 display the time diffusion of pro-/anti gay language for each of the 4 groups selected, with the fourth quartile being the group of newspapers with the highest Republican slant. Both exercises show that in more left wing newspapers we observe a stronger difference between pro-gay language and anti-gay language, in particular in the 90s. This seems to suggest that the ideology of a newspaper might have played a fairly important role.

[Figure 7 here]

As a further exercise we consider more closely the GSS question "What about sexual relations between two adults of the same sex–do you think it is always wrong, almost always wrong, wrong only sometimes, or not wrong at all". This is the question that has been asked more frequently and for a longer time in the GSS (from 1973). In Fig. 8 we plot this variable against time.[24] It seems clear that the change in attitudes seems to be starting after the 1990s, in particular from the 1993 survey onwards. This change in attitude seems therefore to occur later than the start of the diffusion of pro-gay coverage, which in Figures 2 and 4 seems to take place a little bit earlier, between 1985 and 1990. Obviously this does not imply by any means that media has a causal impact on attitudes, only that pro-gay coverage of media seems to predate the start of pro-gay attitudes diffusion.

[Figure 8 here]

4 Institutional changes

We will now try to understand how institutions and the intensity of coverage of pro-/anti-gay rights language in the press coevolve. We consider 116 institutional changes that took place in the United States until 2014. We study 41 pro-gay institutional changes introducing same-sex marriages that might have happened because of court decisions at the local or federal level, legislative decision or, more rarely, through referenda. Then we focus on 75 formal bans of gay marriages (i.e. changes that aim to restrict gay rights) such as executive orders by governors, statutory bans and constitutional amendments. For each institutional change we collected information about the month and the year of its implementation. Every State has been affected by either a pro-gay institutional changes or an anti-gay change or both of them.[25]

We take advantage of the high frequency feature of our dataset and use monthly coverage measures of pro-/anti-gay rights language at newspaper level. We first consider pro-gay institutional changes that introduced gay marriages and we focus on the month of change, 3 years before and after. We adopt an event analysis strategy and provide graphical analysis showing the evolution of a given outcome during each of the 36 months before and after the pro-(anti-) gay institutional change based on the following regression:

[24] A similar graph is also discussed in Fernández et al. (2019).

[25] Details of the institutional changes considered are in Tables A8 and A9 of the Appendix

$$Coverage_{g,n,t} = \alpha_n + \gamma_t + \sum_{m=-36}^{+36} d_{n,t}^{m} \beta_m + X_{n,t}\delta + \varepsilon_{n,t}. \quad (2)$$

t corresponds now to a given month. Standard errors are clustered at State level. We include newspaper fixed effects α_n and month fixed effects γ_t. The variables $d_{n,t}^{m}$ are defined for all integers t from -36 to 36 (so to only include observations from 3 years before the reform to 3 years after it). $d_{n,t}^{m}$ takes the value 1 if the closest pro-gay institutional change took place the month t-m, 0 otherwise. Therefore, if the state of circulation of newspaper n had a policy change in January 2004, $d_{n,t}^{12}$ takes the value 1 if t = January, 2005, $d_{m,t}^{24}$ takes the value 1 if t = January 2006, an so on. Since $d_{m,t}^{-36}$ is normalised to zero, the coefficients β_m indicate how the outcome variable changes with respect to three years prior to the institutional change.

We then present the graphical analysis of results obtained using specification (2), that is we plot β_m for our two dependent variables, the pro-gay coverage measure ($Coverage_{pro,n,t}$) and the anti-gay coverage measure ($Coverage_{anti,n,t}$). In Figure 9, Panel A, we show results using the standard pro-gay coverage measure (hollow circles) and using the standard anti-gay coverage measure (red circles) as dependent variable (the confidence intervals of each coefficient are reported by Figure A6 of the Appendix). We can see that more or less one year before the month gay marriages were introduced we observe a growing diffusion of press coverage of pro- and anti-gay arguments with a dramatic peak in the month of the institutional change; press coverage of anti-gay movements, however, rises much more. This might represent the attempt by those against gay rights to prevent institutions moving in a more liberal direction. The rise in press coverage (when compared with press coverage 3 years before the reform) is still present in the three years subsequent to the change, pointing at some degree of persistence in the effects of the introduction of gay marriages. In light of these findings we might also interpret the S-shaped pattern of the diffusion process of pro- and anti-gay language partly as the integral over a set of events that induced local spikes.

Figure 9, Panel B, shows results obtained performing the same empirical strategy but focussing on changes that restricted gay rights (the confidence intervals of each coefficient are reported by Figure A7 of the Appendix).[26] We observe patterns similar to the one presented in Panel A. A few months before the change we see a rise in both pro- and anti-gay coverage, the latter being substantially stronger. The rise in press coverage, however, appears to be less persistent then the one present after the introduction of gay marriage and after one year it seems to have disappeared.

[Figure 9 here]

We then check the robustness of the results in Figure 9 and in particular of the findings related to institutional changes that enlarged gay rights. In Panel A and B of Figure 10 we use weights based

[26] We consider as relevant thresholds when executive orders and statutory bans are signed by governors or constitutional amendments are approved by voters.

on the relative relevance of the unpreprocessed phrase. In Panel A we use a relevance measure based on relative frequency and Panel B based on relative Pearson Chi Values. Patterns are similar when we include among our control variables our proxy for the total number of digitised articles (Panel C), when we divide our coverage measure by our proxy for the total number of digitised articles (Panel D) or when we perform specification (3) but weighting newspapers by the number of copies sold in 2004 (Panel E).[27]

[Figure 10 here]

5 Spatial Diffusion

To analyze spatial variation and diffusion, we now construct our relevant measures at county level, that is $Coverage_{g,c,t}$ as explained by Section 3.1.4, by taking the average of the values of the coverage variable for all the newspapers within each county. We consider our coverage measures of pro-/anti-gay rights language at newspaper level and also report results using the ratio between the pro-coverage measure and the sum between the pro- and the anti-gay rights coverage measures.[28]

[Figure 11 here]

Figure 11 displays the ratio measure, that is $Coverage_{pro,c,t}/(Coverage_{pro,c,t}+Coverage_{anti,c,t})$ in U.S. counties in 2014 (this is the year with the highest number of digitised newspapers, indeed in 2014 3680 newspapers were digitised allowing us to have our relevant coverage measures for 1265 counties). In Figure 11 counties are divided in four groups of equal size according to their place in the distribution of the ratio measure. Counties coloured dark and light black are counties in the groups with the highest and second highest ratio measure, respectively; counties coloured light and dark grey are counties in the groups with the lowest and second lowest ratio measure, respectively (i.e. darker coloured counties are counties with higher ratio measure). The map shows the existence of substantial spatial correlation and, in particular, the concentration of pro-gay rights language coverage (high ratio measure) in the areas within the states of California and New York. This is formally tested using LISA ("Local Indicators of Spatial Association") maps that visualise counties that have statistically significant Local Moran values (LISA maps are introduced by Anselin (1995) and also adopted in Felkner and Townsend (2011)). Other pro-gay rights clusters

[27] Figure A8 in the Appendix reports robustness checks using the set of phrases presented in Tables A1-A6 of the Appendix obtained by varying the set of keywords, the thresholds of inclusion of bigrams in the topical and non topical sets and the number of reference bigrams. Figure A8 also shows that results are robust when we consider articles that either only contain unpreprocessed phrases associated to pro-gay bigrams or only contain unpreprocessed phrases associated to anti-gay bigrams and we count each article only one time independently of the number of unpreprocessed phrases contained. Figure A9 presents the same battery of checks but focusses on institutional changes that restricted gay rights.

[28] When both pro-gay rights and the anti-gay rights coverage measures are equal to zero we set the ratio measure equal to 1/2.

detected are in Michigan and Massachusetts, while anti-gay rights clusters (low ratio measure) in 2014 were detected mostly in Indiana, Kentucky, North Carolina, Oklahoma and Tennessee. The map displaying the pro-gay rights (black) and anti-gay rights (grey) clusters is shown by Figure 12.

[Figure 12 here]

To formally test for the existence of spatial autocorrelation between counties we calculate the Moran I statistic for each of our three main variables, that is $Coverage_{pro,c,t}$, $Coverage_{anti,c,t}$ and $Coverage_{pro,c,t}/(Coverage_{pro,c,t} + Coverage_{anti,c,t})$. We perform this exercise for each year from 2010 to 2014 (these are the years with a high enough number of observations at county level) and for the entire 5 years period 2010-2014 (in this case we consider only newspapers digitised for the all 5 years). Results are displayed in Table 3 and point to the existence of spatial autocorrelation. In the Appendix we then show results obtained weighting each article by the relative relevance of the unpreprocessed phrase contained and discuss the standard robustness checks related to the data construction explained in the data section.

[Table 3 here]

To further explore this issue we build a balanced panel dataset with observations at county-year level for the period from 2010 to 2014. To construct this dataset we rely only on newspapers digitised for the entire time frame. We end up with a sample of 1102 counties for 5 years. We then run a spatial lag model (with year fixed effects) using as dependent variables our 3 main measures: $Coverage_{pro,c,t}$ (columns 1, 4 and 7 of Table 4), $Coverage_{anti,c,t}$ (columns 2, 5 and 8 of Table 4) and the ratio measure (columns 3, 6 and 9 of Table 4). Table 4 displays results with (columns 4-6) and without county fixed effects (columns 1-3). Standard errors are always clustered at county level. Again results confirm the presence of county level spatial autocorrelation. However, when in columns 7-9 we add state-time fixed effects among the control variables, we observe a dramatic decrease in the coefficient of the spatial lag, that becomes very small in size and not always significantly different from zero. This seems to suggest the existence of important omitted variables at state-time level; events that took place at State level are likely to have generated the spatial autocorrelation between counties we detected. Spatial autocorrelation might therefore be driven by news or politics at state level rather then the spread of public opinion from a county to its neighbouring counties. In Tables A12 and A13 of the Appendix we have results from specifications with weighted dependent variables.

[Table 4 here]

6 Persistence

In this section we investigate whether there is persistence over time in the degree newspapers cover pro- and anti-gay rights language. For example, we assess whether the county level pro-gay rights coverage measure of a decade ago is a strong predictor of the current county level pro-gay rights coverage measure.

We focus on two points in time: the decade from 2005 to 2014 and the decade from 1995 to 2004. We only consider newspapers that have been digitised in both time periods; we do not require newspapers to be digitised for the entire period from 1995 to 2014, but in order to maximise data availability we require they have been digitised for at least one year in both the decades. For each newspaper we take a yearly average of the relevant coverage variable for both the time periods considered (1995-2004 and 2005-2014) and then we build a county measure of the coverage variable by taking the average of all the newspapers digitised within each county. Once we construct our measures at county level as previously described we end up with a sample of 585 counties

[Figure 13 here]

As usual we consider three measures, that is the pro-gay rights coverage measure at county level, the anti-gay rights coverage measure at county level and the ratio measure. For each of these variables we plot the log of the variable in 2005-2014 against the log of the same variable in 1995-2004 so to identify the elasticity between coverage measures today and coverage measures one decade ago. Results are displayed by Figure 13 and show that our measures seem to exhibit a certain degree of persistence over time. The elasticity is 0.63 when we consider pro-gay rights coverage and 0.58 when we consider the anti-gay rights coverage (both coefficients are significantly lower than one and greater than zero). Results are very similar if we restrict our sample only to newspapers digitised during the entire time frame considered, 1995-2014 (Figure A10 in the Appendix).

7 Conclusions

In this paper we study the battle for hearts and minds that was behind the change in the attitudes towards homosexuality. We do not focus on public opinion, the output of the battle, but on one of the inputs, the intensity of coverage of pro- and anti-gay rights language in US newspapers, an important arena where this debate plays out. We estimate the diffusion of pro- and anti-gay rights language in the media and how it is related to institutional changes surrounding gay marriage. Using a broad set of speeches given by Congresspeople and Senators in the last 20 years, we identify a set of phrases that are diagnostic of pro- and anti-gay rights views. We then build measures of coverage of pro-gay rights and anti-gay rights language based on the frequencies of

such phrases in a very large set of US newspapers digitised at different times in the last 20 years. We end up with a unique dataset containing high frequency and fairly geographically detailed information for a long period of time.

We document the existence of several important regularities in the data. The propagation of both pro- and anti-gay rights media coverage follows a S shaped pattern over time, characteristic of diffusion processes. We find that the diffusion in the media coverage of pro-gay rights language starts earlier, but that the diffusion of anti-gay rights language in the media catches up in the last 10 years. Moreover, we document the existence of substantial spatial autocorrelation across counties in media coverage of pro- and anti-gay rights language; interestingly, such spatial autocorrelation seems mostly to be driven by shocks taking place at state level. There is a very pronounced coevolution between coverage of pro- and anti-gay rights language and important institutional changes experienced by U.S. States, such as the introduction of gay marriages. When gay marriages are introduced press coverage of the rights language of both pro- and anti-gay rights language dramatically increase; the rise in coverage of anti-gay rights language is, however, overwhelmingly higher and is still present in the three years subsequent to the institutional change. Finally, we find that between counties differences in coverage of such language are fairly persistent over time.

Although we focus on gay rights, our approach and methodology should be of use in analysing how other attitudes also change. How people form their attitudes and how and why they change is a very important yet not well-understood process. We hope that this paper contributes to that understanding.

References

Adamczyk, A. and C. Pitt (2009). Shaping attitudes about homosexuality: The role of religion and cultural context. *Social science research 38*, 338–51.

Adena, M., R. Enikolopov, M. Petrova, V. Santarosa, and E. Zhuravskaya (2015, November). Radio and the rise of the nazis in prewar germany. *Quarterly Journal of Economics 130*(4), 1885–1939.

Alesina, A. and N. Fuchs-Schündeln (2007, September). Goodbye lenin (or not?): The effect of communism on people. *American Economic Review 97*(4), 1507–1528.

Andersen, R. and T. Fetner (2008, October). Economic inequality and intolerance: Attitudes toward homosexuality in 35 democracies. *American Journal of Political Science 52*(4), 942–958.

Anselin, L. (1995). Local indicators of spatial association-lisa. *Geographical analysis 27*(2), 93–115.

Baker, S. R., N. Bloom, and S. J. Davis (2016, November). Measuring economic policy uncertainty. *Quarterly Journal of Economics 131*(4), 1593–1636.

Beaman, L., R. Chattopadhyay, E. Duflo, R. Pande, and P. . Topalova (2009). Powerful women: Does exposure reduce bias? *The Quarterly Journal of Economics 124*(4), 1497–1540.

Bisin, A. and T. Verdier (2001). The economics of cultural transmission and the dynamics of preferences. *Journal of Economic Theory 97*(2), 298–319.

Blei, D. (2012). Probabilitic topic models. *Communications of the ACM 55*, 77–84.

Burns, J., L. Corno, and E. La Ferrara (2015). Interaction, prejudice and performance. evidence from south africa. *Working Paper*.

Cantoni, D., Y. Chen, D. Y. Yang, N. Yuchtman, and Y. J. Zhang (2017, April). Curriculum and ideology. *Journal of Political Economy 125*(2), 338–392.

Chomsky, D. and S. Barclay (2010). The mass media, public opinion, and lesbian and gay rights. *Annual Review of Law and Social Science 6*, 387–403.

Clots-Figueras, I. and P. Masella (2013, August). Education, language and identity. *The Economic Journal 123*(570), F332–F357.

Della Vigna, S. and E. La Ferrara (2015). *Social and Economic Impacts of the Media*, Volume Handbook of Media Economics. Elsevier.

Di Tella, R., S. Galiani, and E. Schargrodsky (2007). The formation of beliefs: Evidence from the allocation of land titles to squatters. *The Quarterly Journal of Economics 122*(1), 209–241.

Durante, R., P. Pinotti, and T. Andrea (2019, July). The political legacy of entertainment tv. *American Economic Review 109*(7), 2497–2530.

Enikolopov, R., M. Petrova, and E. Zhuravskaya (2011, December). Media and political persuasion: Evidence from russia. *American Economic Review 101*(7), 3253–3285.

Felkner, J. S. and R. M. Townsend (2011, November). The geographic concentration of enterprise in developing countries. *The Quarterly Journal of Economics 126*(3), 2005–2061.

Fernandez, R. (2013). Cultural change as learning: The evolution of female labor force participation over a century. *American Economic Review* (1), 472–500.

Fernández, R., S. Parsa, and M. Viarengo (2019). Coming out in america: Aids, politics, and cultural change. *NBER Working Paper* (25697).

Fogli, A. and L. Veldkamp (2011). Nature or nurture? learning and the geography of female labor force participation. *Econometrica 79*(4), 1103–1138.

Gentzkow, M. and J. M. Shapiro (2010, January). What drives media slant? evidence from u.s. daily newspapers. *Econometrica 78*(1), 35–71.

Gentzkow, M., J. M. Shapiro, and M. Sinkinson (2010, December). The effect of newspaper entry and exit on electoral politics. *American Economic Review 101*(7), 2980–3018.

Giuliano, P. and A. Spilimbergo (2014). Growing up in a recession. *The Review of Economic Studies 81*(2), 787–817.

Griffiths, T. L. and M. Steyvers (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences 101*, 5228–5235.

Hansen, E. R. and S. A. Treul (2015). The symbolic and substantive representation of lgb americans in the us house. *The Journal of Politics 77*.

Hansen, S., M. McMahon, and A. Prat (2018). Transparency and deliberation within the fomc: a computational linguistics approach. *The Quarterly Journal of Economics*.

Harrison, B. F. and M. R. Michelson (2015). God and marriage: The impact of religious identity priming on attitudes toward same-sex marriage. *Social Science Quarterly 96*(5), 1411–1423.

Jetter, M. (2017). The effect of media attention on terrorism. *Journal of Public Economics 153*(C), 32–48.

Kearney, M. S. and P. B. Levine (2015, December). Media influences on social outcomes: The impact of mtv's 16 and pregnant on teen childbearing. *American Economic Review 105*(12), 3597–3632.

La Ferrara, E. (2016, August). Mass media and social change: can we use television to fight poverty? *Journal of the European Economic Association 14*(4), 791–827.

La Ferrara, E., A. Chong, and S. Duryea (2012, October). Soap operas and fertility: Evidence from brazil. *American Economic Journal: Applied Economics 4*(4), 1–31.

Laver, M., K. Benoit, and J. Garry (2003, May). Extracting policy positions from political texts using words as data. *The American Political Science Review 97*(2), 311–331.

Lax, J. R. and J. H. Phillips (2009). Gay rights in the states: Public opinion and policy responsiveness. *American Political Science Review 103*(3), 367–386.

Lewis, G. and C. Gossett (2008). Changing public opinion on samesex marriage: The case of california. *Politics Policy 36*, 4–30.

Lim, C. S. H., J. M. J. Snyder, and D. Strömberg (2015, October). The judge, the politician, and the press: Newspaper coverage and criminal sentencing across electoral systems. *American Economic Journal: Applied Economics 7*(4), 103–35.

Mize, T. D. (2016). Sexual orientation in the labor market. *American Sociological Review 81*(6), 1132–1160.

Plug, E., D. Webbink, and N. Martin (2014, January). Sexual orientation, prejudice, and segregation. *Journal of Labor Economics 32*(1), 123–159.

Quinn, K. M., B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev (2010, 228). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science 54*(1), 209.

Redman, S. (2018). Effects of same-sex legislation on attitudes toward homosexuality. *Political Research Quarterly 71*(3), 628–641.

Reynolds, A. (2013). Representation and rights: The impact of lgbt legislators in comparative perspective. *American Political Science Review 107*(2), 259–274.

Schwarz, D., D. Traber, and K. Benoit (2017). Estimating intra-party preferences: Comparing speeches to votes. *Political Science Research and Methods 5*(2), 379–396.

Stephens-Davidowitz, S. (2014). The cost of racial animus on a black candidate: Evidence using google search data. *Journal of Public Economics 118*(C), 26–40.

Tilcsik, A. (2011). Pride and prejudice: Employment discrimination against openly gay men in the united states. *American Journal of Sociology 117*(2), 586–626.

Wald, K., J. W. Button, and B. A. Rienzo (1996). The politics of gay rights in american communities: Explaining antidiscrimination ordinances and policies. *American Journal of Political Science 40*.

Table 1

Pro Reference Phrases	*Anti Reference Phrases*
gay lesbian	tradit marriag
sexual orient	union man
gay men	same-sex marriag
speak hate	definit marriag
gay man	redefin marriag
base sexual	marriag union
crime base	marriag man
crime motiv	marriag law
men lesbian	marriag licens
lesbian american	homosexu marriag
orient gender	defens marriag
non-discrimin act	defin marriag
employ nondiscrimin	same-sex union
discrimin gay	marriag act
pass hate	promot homosexu
employ non-discrimin	issu marriag
enforc hate	tradit definit
gender ident	legal same-sex
lesbian gay	opposit sex
serv open	homosexu lifestyl
victim hate	legal union
lgbt communiti	say marriag
gay american	homosexu militari
gay coupl	marriag institut
allow gay	homosexu conduct
legal incid	right same-sex
introduc hate	marriag legal
regardless sexual	fundament institut
bisexu transgend	marriag import
peopl transgend	protect marriag

NOTE: The Table shows the 30 bigrams that are most diagnostic of pro-gay language in Congress and the 30 bigrams that are most diagnostic of anti-gay language in Congress.

Figure 1: Relevant Word Clouds

Topic 3

Topic 4

Topic 7

Topic 8

Topic 9

Topic 13

Topic 15

Topic 17

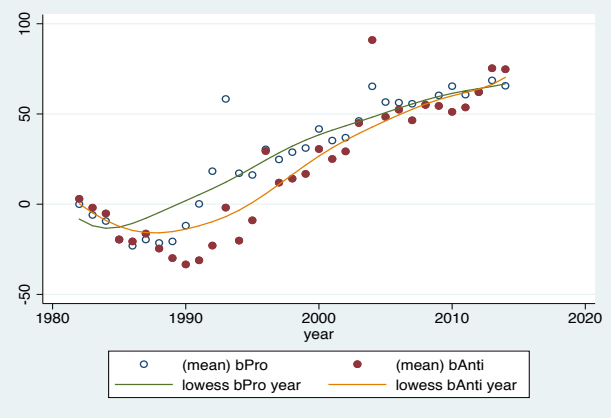
Table 2: Consistency of Coverage Measures

	(1) **marhomo**	(2) **spkhomo**	(3) **colhomo**	(4) **libhomo**	(5) **homosex**	(6) **ssfchild**	(7) **ssmchild**
Pro Coverage	.0115***	.0011	.0028***	-.0007	.0037	.0312***	.0311***
	(.0027)	(.0007)	(.0008)	(.0007)	(.0035)	(.0079)	(.0084)
Anti Coverage	-.0050***	-.0010***	-.0016***	.0001	-.0042**	-.236***	-.0259**
	(.0010)	(.0003)	(.0003)	(.0002)	(.0017)	(.0010)	(.0109)
Ind. Controls	X	X	X	X	X	X	X
Region FE	X	X	X	X	X		
Time FE	X	X	X	X	X		
Obs	8718	8375	8353	8378	8136	1228	1228

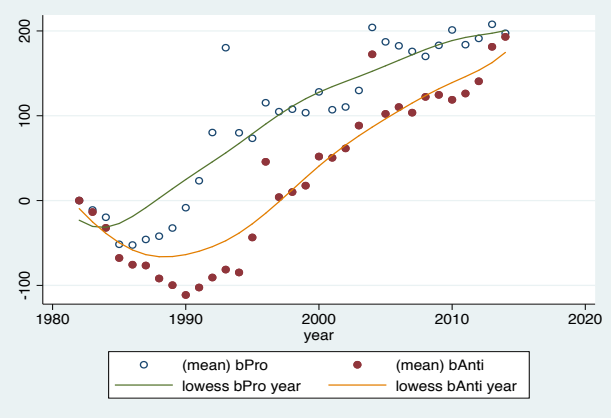
NOTE: In Column (1) we consider the question whether a respondent strongly agrees/agrees/neither agrees or disagrees/disagrees/strongly disagree with the following statement "Homosexual couples should have the right to marry one another". As a dependent variable we have the variable **marhomo** going from 5 if the respondent strongly agrees with the statement to 1 if he strongly disagrees. In Column (2) the question: what about a man who admits that he is a homosexual? Suppose this admitted homosexual wanted to make a speech in your community. Should he be allowed to speak, or not? We report results using as dependent variable **sphomo,** a Dummy variable equal to one if the respondent thinks he should be allowed. In Column (3) the question: Should such a person be allowed to teach in a college or university, or not? We report results using as dependent variable **colhomo,** a Dummy variable equal to one if the respondent thinks he should be allowed. In Column (4) the question: If some people in your community suggested that a book he wrote in favor of homosexuality should be taken out of your public library, would you favor removing this book, or not? We report results using as dependent variable **libhomo,** a Dummy variable equal to one if the respondent thinks the book should not be removed. In Column (5) the question: What about sexual relations between two adults of the same sex--do you think it is always wrong, almost always wrong, wrong only sometimes, or not wrong at all? We report results using as dependent variable the variable **homosex** with values going from 1 to 4, 1 if the respondent says it is always wrong, 4 if he says it is not wrong at all. In Column (6) the question: To what extent do you agree or disagree with the following statements? A same sex female couple can bring up a child as well as a male-female couple. The respondent is asked whether strongly agrees/agrees/neither agrees or disagrees/disagrees/strongly disagree with the statement. We report results using as dependent variable the variable **ssfchild** with values going from 1 to 5, 5 if the respondent says he strongly agrees, 1 if he says he strongly disagrees. In Column (7) the question: A same sex male couple can bring up a child as well as a male-female couple. The respondent is asked whether strongly agrees/agrees/neither agrees or disagrees/disagrees/strongly disagree with the statement. We report results using as dependent variable the variable **smchild** with values going from 1 to 5, 5 if the respondent says he strongly agrees, 1 if he says he strongly disagrees. The main explanatory variables are pro-gay coverage measure and anti-gay coverage measure at regional level. We consider newspapers digitised for the entire time period 2004-2014 and survey years from 2004. We control for age, race and gender of the respondent. In columns (1)-(5) we also control for survey year fixed effects, region of residence fixed effects and region specific linear time trends. Standard errors (clustered at region-year level) in parenthesis . * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$

Figure 2: Diffusion over time, Regression Approach, Yearly cells

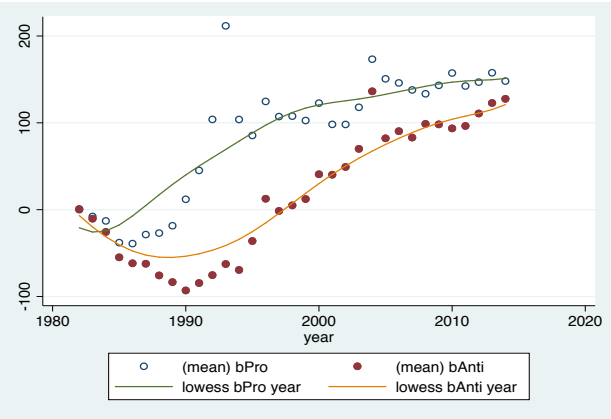
Panel A: Standard



Panel B: Frequency weights



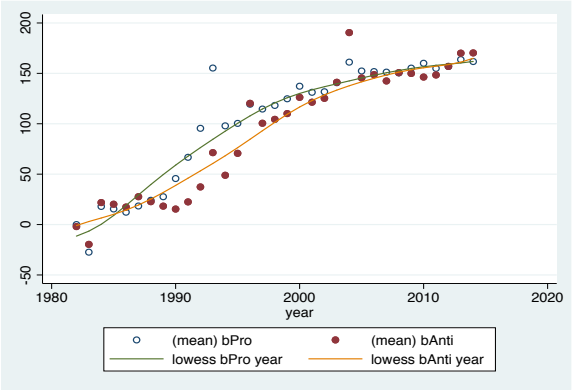
Panel C: Chi weights



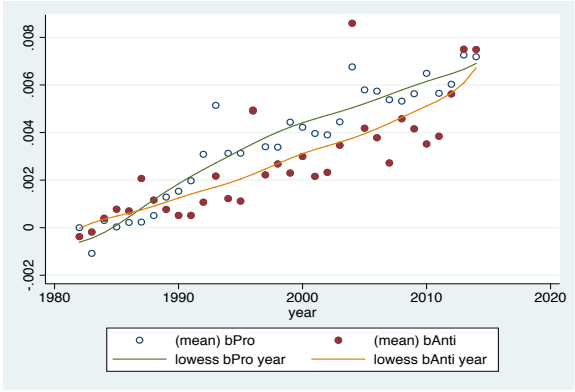
NOTE: We regress the coverage measure of type g (we only have two types, pro-gay and anti-gay) of a newspaper n in the year t on newspaper fixed effects and on interactions of type and year fixed effects. Panel A plots the coefficients **bPro (bAnti)** of the year fixed effects when type is pro- (anti-) gay against time; hollow (red) circles correspond to coefficients when type is pro-(anti-) gay. Panel B plots the same coefficients when the dependent variable, the coverage measure, reflects the relevance of the phrases covered by the newspaper and their relevance is based on their frequency whitin topical speeches in Congress. Panel C plots the same coefficients when the dependent variable, the coverage measure, reflects the relevance of the phrases covered by the newspaper and their relevance is based on their Chi values, that proxy how strongly we can reject the hypothesis they are as likely to be used by pro-gay Congressmen than anti-gay Congressmen.

Figure 3: Diffusion over time, Yearly cells, Robustness

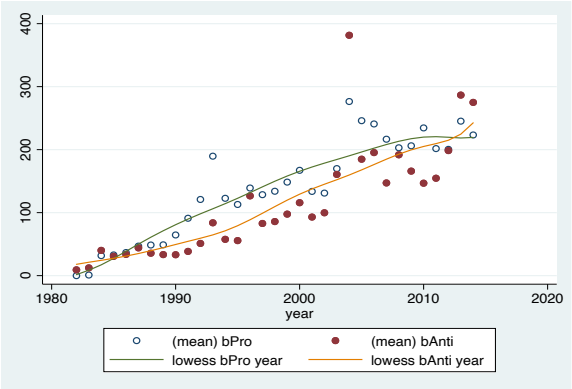
Panel A: control for n digitised articles



Panel B: share



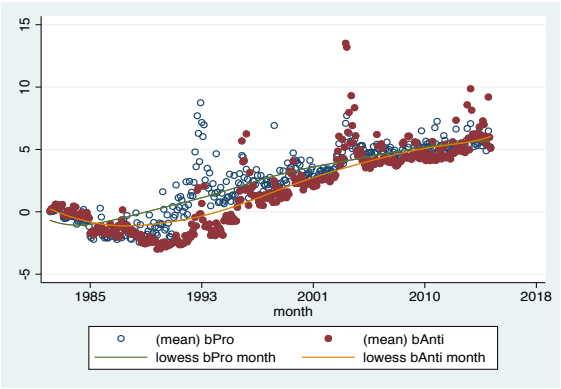
Panel C: Circulation weights



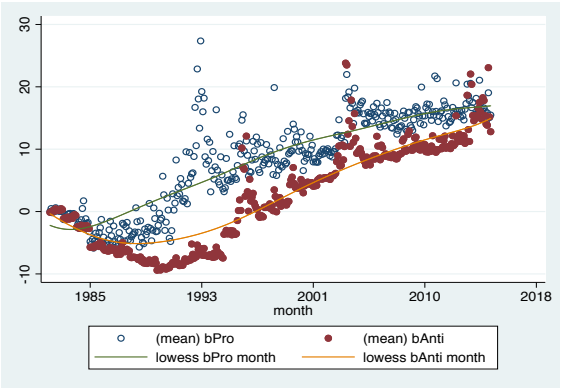
NOTE: We regress the coverage measure of type g (we only have two types, pro-gay and anti-gay) of a newspaper n in the year t on newspaper fixed effects and on interactions of type and year fixed effects. We plot **bPro (bAnti),** that is the coefficients of the year fixed effects when type is pro- (anti-) gay against time; hollow (red) circles correspond to coefficients when type is pro- (anti-) gay. Panel A includes in the main regression a proxy for the number of digitised articles per newspaper-year. In Panel B we use as a dependent variable the ratio between our main coverage measure and our proxy for the number of digitised articles per newspaper-year. In Panel C we perform the standard regression as in Panel A Figure 2 but using as weights for each newspaper n the number of copies sold in 2004.

Figure 4: Diffusion over time, Regression Approach, Monthly cells

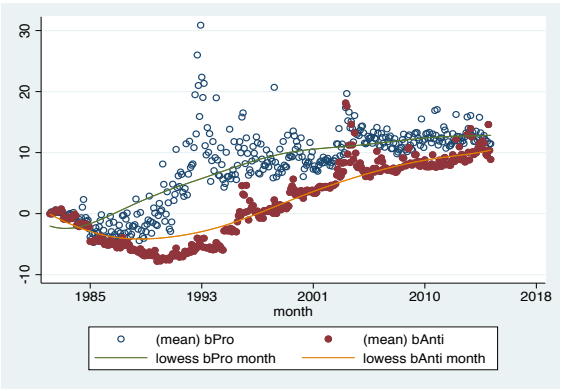
Panel A: Standard



Panel B: Frequency weights



Panel C: Chi weights

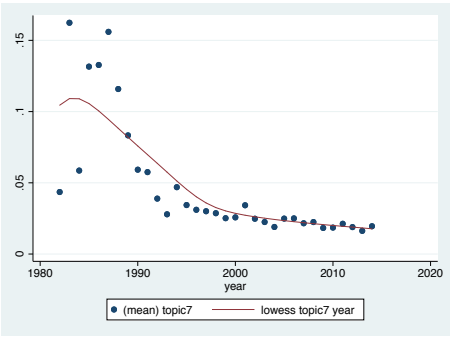


NOTE: We regress the coverage measure of type g (we only have two types, pro-gay and anti-gay) of a newspaper n in the month m on newspaper fixed effects and on interactions of type and month fixed effects. Figure 4 Panel A plots the coefficients **bPro (bAnti)** of the month fixed effects when type is pro- (anti-) gay against time; hollow (red) circles correspond to coefficients when type is pro- (anti-) gay. Panel B plots the same coefficients when the dependent variable, the coverage measure, reflects the relevance of the phrases covered by the newspaper and their relevance is based on their frequency whitin topical speeches in Congress. Panel C plots the same coefficients when the dependent variable, the coverage measure, reflects the relevance of the phrases covered by the newspaper and their relevance is based on their Chi values, that proxy how strongly we can reject the hypothesis they are as likely to be used by pro-gay Congressmen than anti-gay Congressmen.

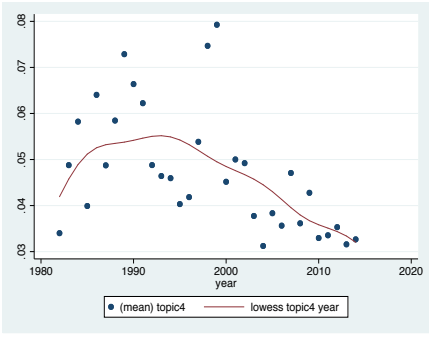
Figure 5: Evolution over time of the relevant topics

Panel A: Pro-gay corpus

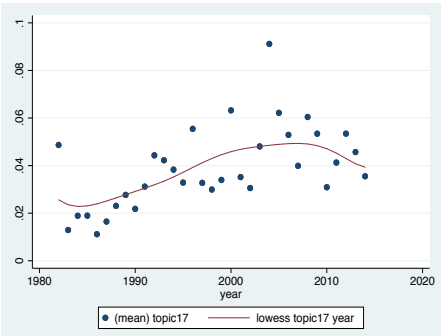
Topic 7 (HIV topic)



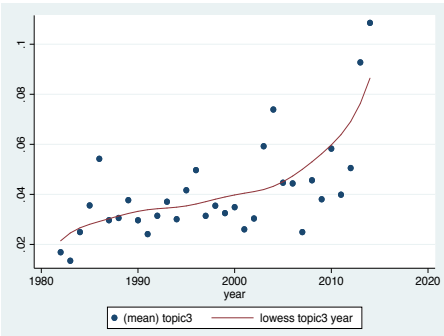
Topic 4 (Hate Crimes topic)



Topic 17 (Legal Unions)

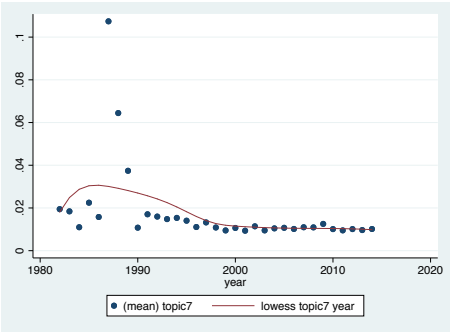


Topic 3 (Supreme Court - gay marriages)

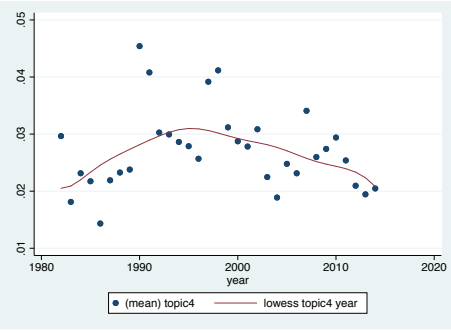


Panel B: Anti-gay corpus

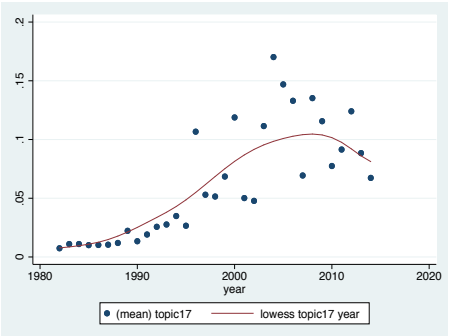
Topic 17 (Legal Unions)



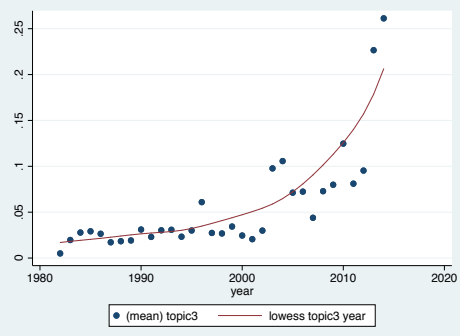
Topic 3 (Supreme Court - gay marriages)



Legal Unions

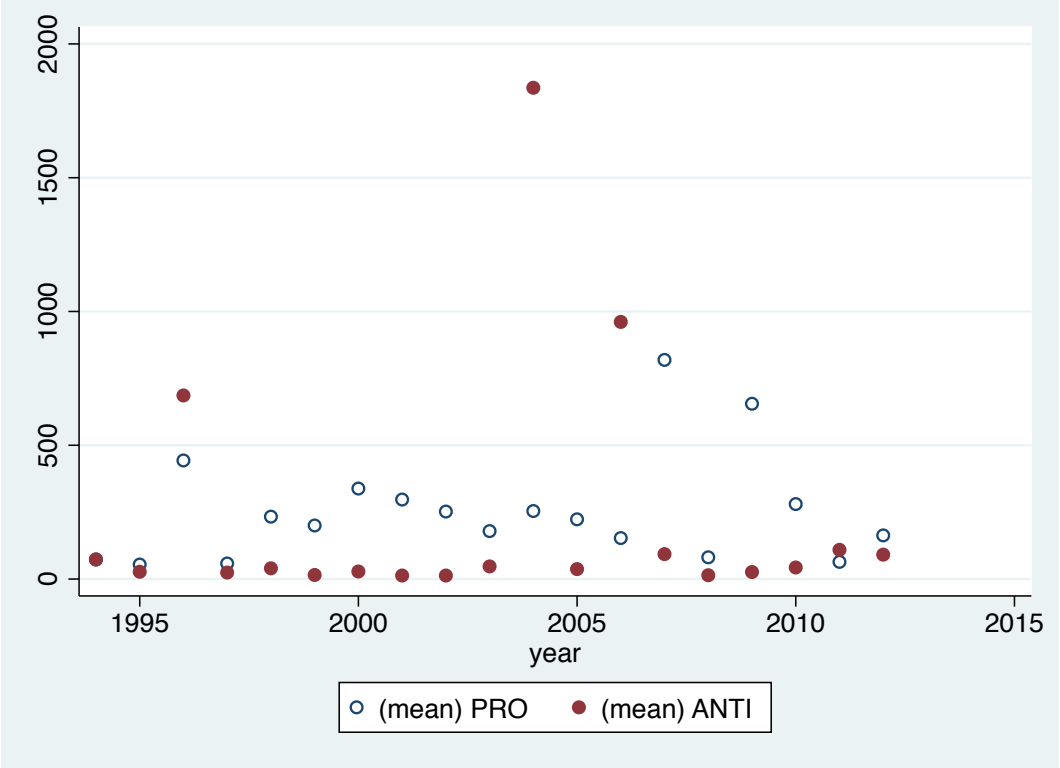


Supreme Court - gay marriages



NOTE: In Panel A (Panel B) we plot the share of text within the pro- (anti-) gay corpus devoted to each of the 4 topics considered against time.

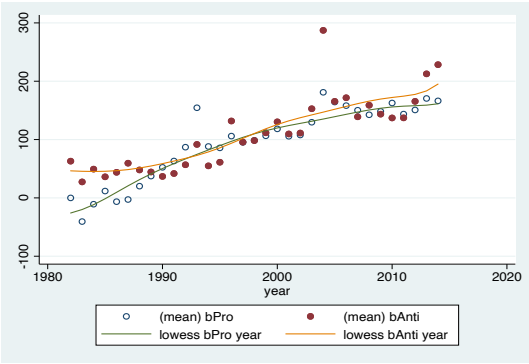
Figure 6: Frequency of Phrases in Congressional speeches



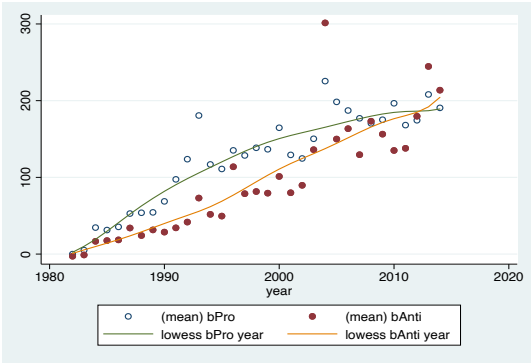
NOTE: Fig. 6 shows the diffusion over time of pro- (hollow circles) and anti-gay (red circles) language within the text of congressional speeches. Data are from 1994 to 2012.

Figure 7: Diffusion by Newspapers' Ideology

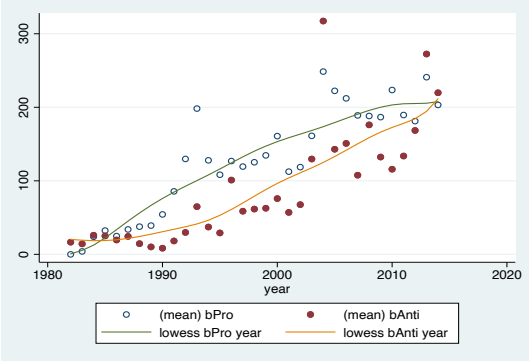
Panel A: Diffusion by Party Endorsement (Republican candidate)



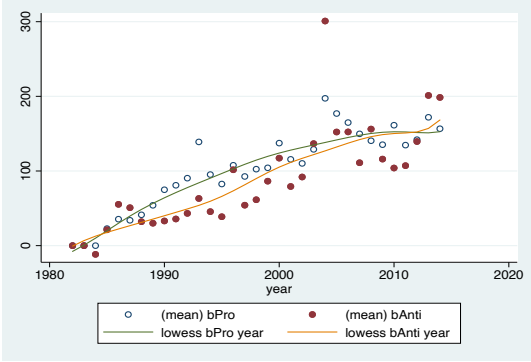
Panel B: Diffusion by Party Endorsement (Democratic candidate)



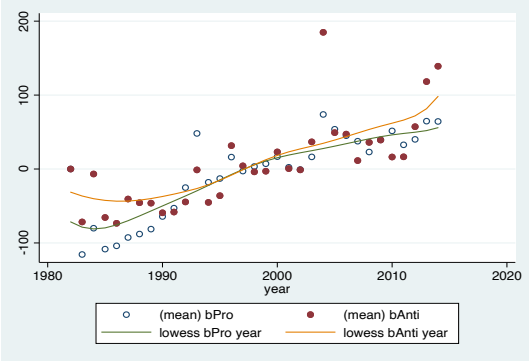
Panel C: Diffusion by Slant (First Quartile)



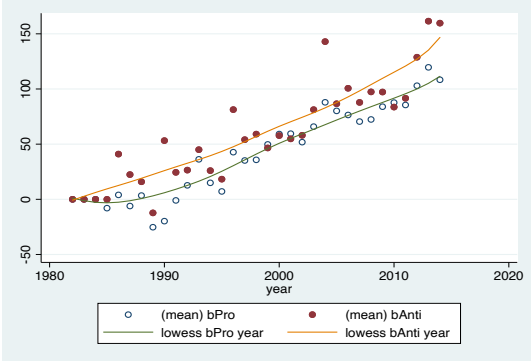
Panel D: Diffusion by Slant (Second Quartile)



Panel E: Diffusion by Slant (Third Quartile)

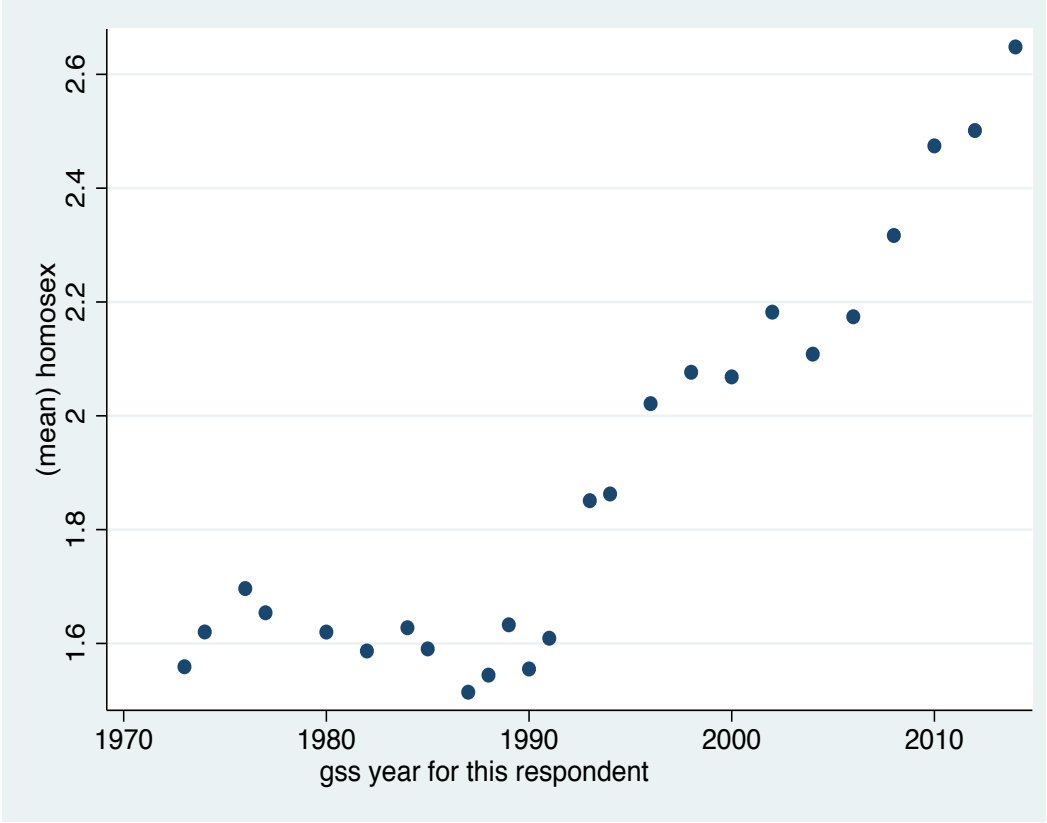


Panel F: Diffusion by Slant (Fourth Quartile)



NOTE: We regress the coverage measure of type g (we only have two types, pro-gay and anti-gay) of a newspaper n in the year t on newspaper fixed effects and on interactions of type and year fixed effects. Figure 7 plots the coefficients bPro (bAnti) of the year fixed effects when type is pro- (anti-) gay against time; hollow (red) circles correspond to coefficients when type is pro- (anti-) gay. Panel A (B) considers the sample of newspapers that in 2004 endorsed the Republican (Democratic) candidate. Panels C-F consider the sample of newspapers within each quartile of the (Republican) slant distribution.

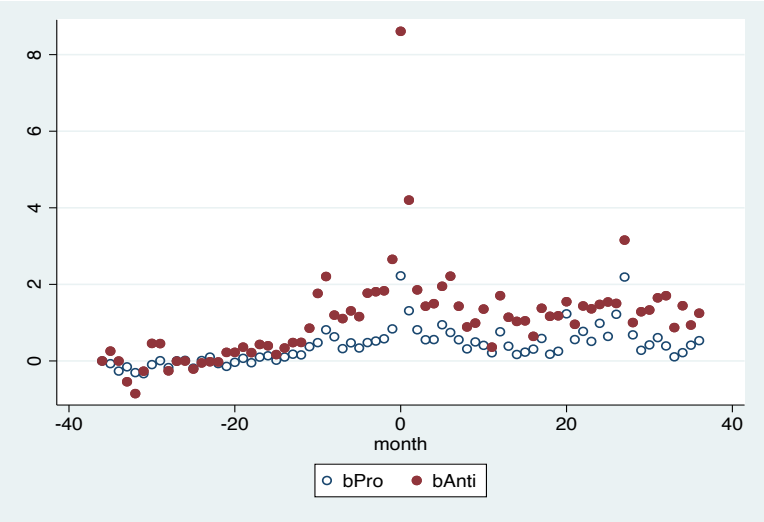
Fig 8: Pro Homosexual attitudes over time



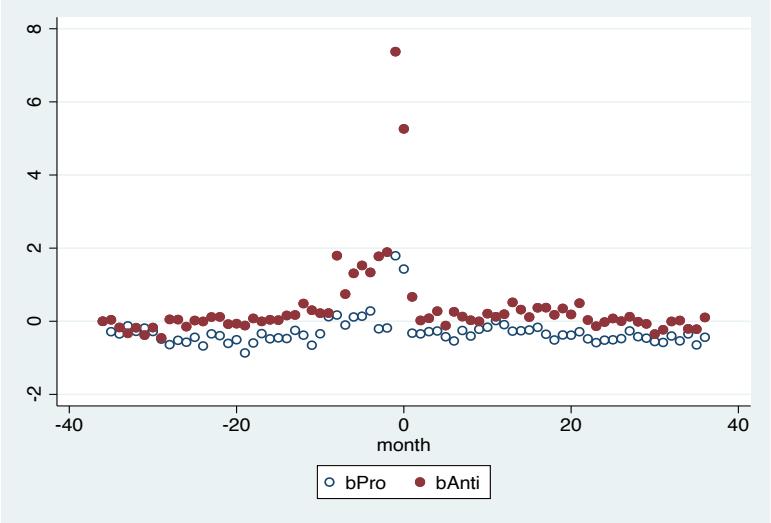
NOTE: We consider the survey question "What about sexual relations between two adults of the same sex--do you think it is always wrong, almost always wrong, wrong only sometimes, or not wrong at all?" We generate a variable with values going from 1 to 4, 1 if the respondent says it is always wrong, 4 if he says it is not wrong at all. We plot the average of this variable over time.

Figure 9: Institutional changes

Panel A: Introduction of gay marriages



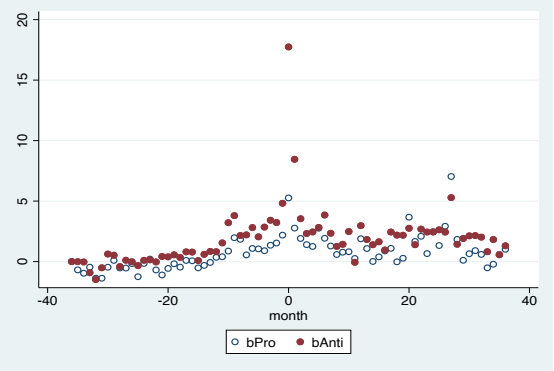
Panel B: Ban of gay marriages



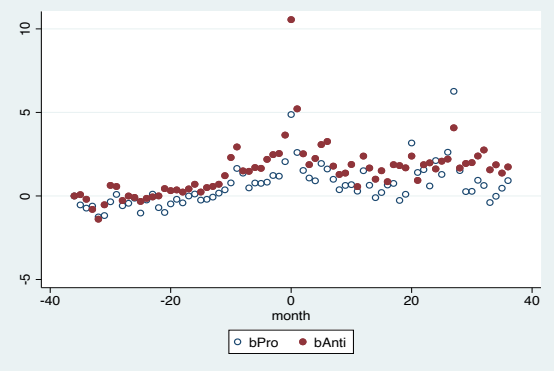
NOTE: In Panel A we consider the introduction of gay marriages, in Panel B we the consider ban of gay marriage. We include the month of implementation of the change, 36 months before and 36 months afterwards and we only consider newspapers that have been digitised throughout all the time considered. Hollow (red) circles correspond to the graphical representation of the event analysis exercise based on specification (2) in the paper when the dependent variable is pro- (anti-) gay coverage.

Figure 10: Introduction of gay marriage, Robustness

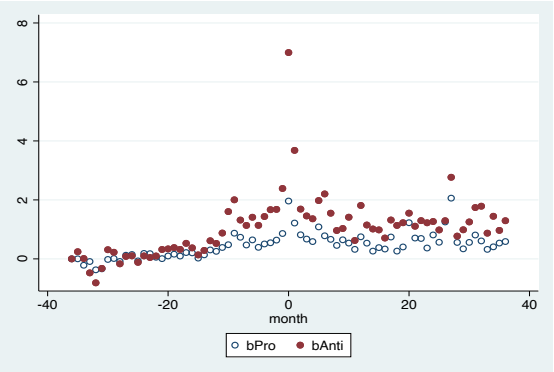
Panel A: Frequency weights



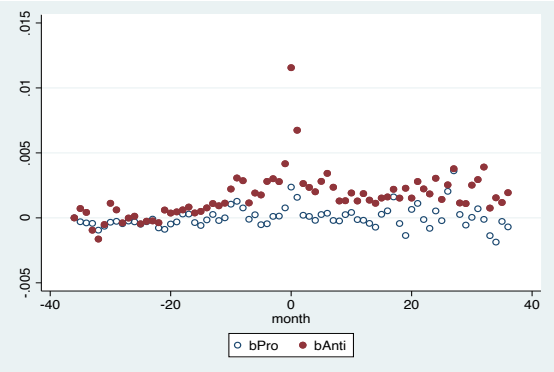
Panel B: Chi weights



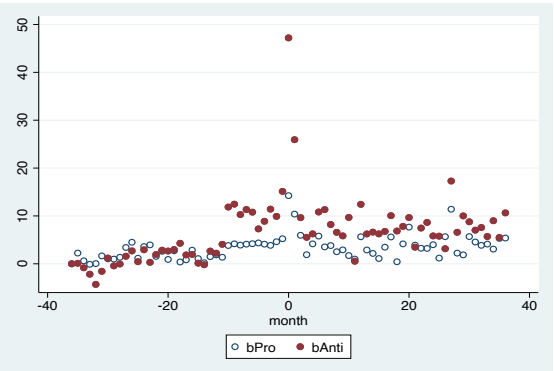
Panel C: Control for number articles



Panel D: Share



Panel E: Circulation Weights

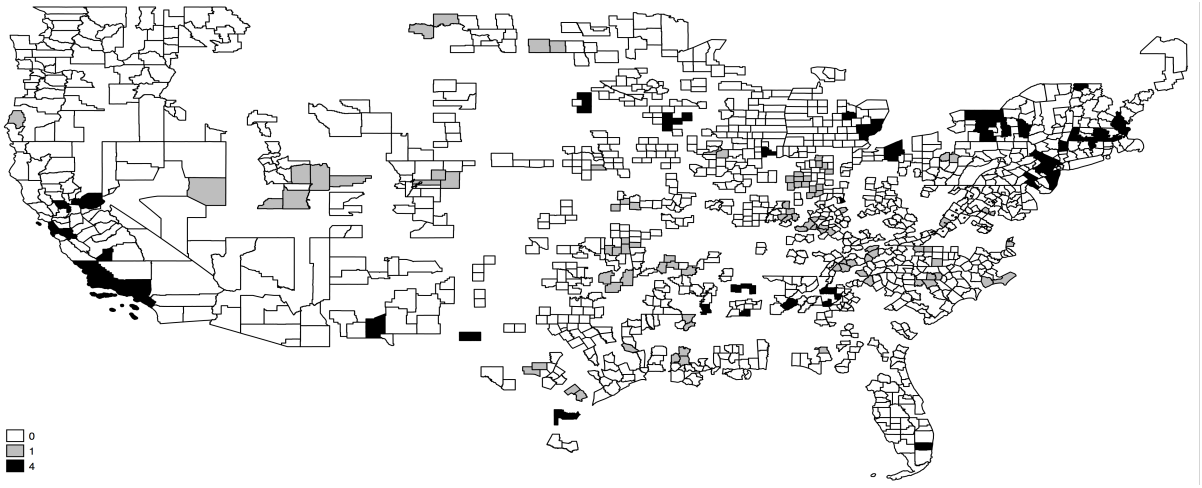


NOTE: We consider the introduction of gay marriages. We include the month of implementation of the change, 36 months before and 36 months afterwards and we only consider newspapers that have been digitised throughout all the time considered. Hollow (red) circles correspond to the graphical representation of the event analysis exercise based on specification (2) in the paper when the dependent variable is pro- (anti-) gay coverage. Panel A plots results when the dependent variable, the coverage measure, reflects the relevance of the phrases covered by the newspaper and their relevance is based on their frequency whitin topical speeches in Congress, Panel B when the dependent variable, the coverage measure, reflects the relevance of the phrases covered by the newspaper and their relevance is based on their Chi values, that proxy how strongly we can reject the hypothesis they are as likely to be used by pro-gay Congressmen than anti-gay Congressmen. In Panel C we include our proxy for the number of digitised articles as dependent variable. In Panel D we use as a dependent variable the ratio between our main coverage measure and our proxy for the number of digitised articles per newspaper-year. In Panel E we weight each newspaper n by the number of copies sold in 2004.

Figure 11: Map of the share of Pro-Gay Coverage, Pro/(Pro+Anti), by quartile groups

NOTE: Counties are divided in four groups of equal size according to their place in the distribution of the ratio measure, that is the share of pro-gay coverage, Pro/(Pro+Anti). Counties coloured dark black and light black are counties in the groups with the highest and second highest ratio measure, respectively; counties coloured light grey and dark grey are counties in the groups with the lowest and second lowest ratio measure, respectively. Statistics refer to the year 2014.

Figure 12:LISA Map of the share of Pro-Gay Coverage, Pro/(Pro+Anti), Clusters



NOTE: The Figure displays a map of the United States where black (grey) spots detects geographical clusters with high (low) share of pro-gay coverage, Pro/(Pro+Anti) (measured as the ratio between pro-gay coverage and the sum between pro- and anti-gay coverage) using the LISA methodology as in Anselin (1995) and Felkner and Townsend (2011). Statistics refer to the year 2014.

Table 3: Moran I

	2014	2013	2012	2011	2010	2010-2014
Pro Coverage	0.050*	0.025	0.032	0.055*	0.113**	0.098**
	(0.024)	(0.023)	(0.024)	(0.024)	(0.026)	(0.027)
Anti Coverage	0.110**	0.081**	0.091**	0.059**	0.082**	0.137**
	(0.024)	(0.023)	(0.024)	(0.025)	(0.025)	(0.027)
ratio	0.183**	0.110**	0.092**	0.091**	0.125**	0.219**
	(0.025)	(0.025)	(0.025)	(0.025)	(0.026)	(0.027)

NOTE: Table 3 displays the Moran I for the coverage of pro-gay language (row 1), anti-gay language (row 3), ratio of the coverage of pro-gay language over the sum of pro- and anti-gay language (row 5), for the years 2014 (column 1), 2013 (column 2), 2012 (column 3), 2011 (column 4), 2010 (column 5) and the entire time period 2010-2014 (column 6). In this last case we include only newspapers digitised between 2010 and 2014. Standard errors in parenthesis . * $p < 0.05$; ** $p < 0.01$

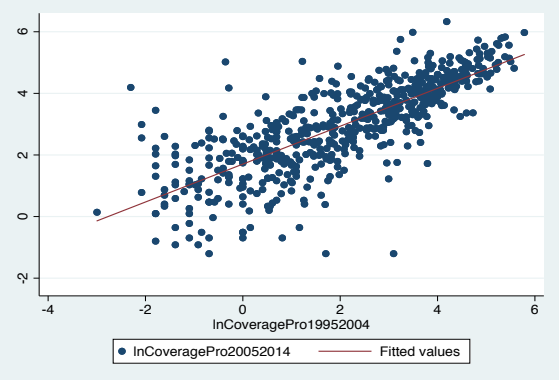
Table 4: Spatial Regression

	(1) Pro Coverage	(2) Anti Coverage	(3) ratio	(4) Pro Coverage	(5) Anti Coverage	(6) ratio	(7) Pro Coverage	(8) Anti Coverage	(9) ratio
Spatial Lag	0.074**	0.231**	0.117**	0.064**	0.246**	0.082**	-0.038	0.092*	-0.0003
	(0.015)	(0.03)	(0.011)	(0.017)	(0.032)	(0.013)	(0.023)	(0.040)	(0.012)
county F.E.	NO	NO	NO	YES	YES	YES	NO	NO	NO
year F.E.	YES	YES	YES	YES	YES	YES	YES	YES	YES
state*year F.E.	NO	NO	NO	NO	NO	NO	YES	YES	YES
obs	5445	5445	5445	5445	5445	5445	5445	5445	5445

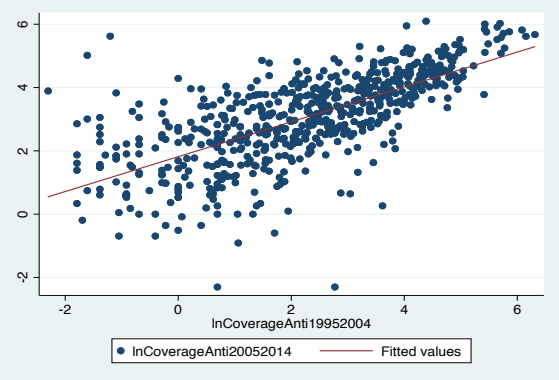
NOTE: Table 4 displays results for spatial models (command xsmle in Stata) using as dependent variable the coverage of pro-gay language (columns 1, 4 and 7), anti-gay language (columns 2, 5 and 8), ratio of the coverage of pro-gay language over the sum of pro- and anti-gay language (columns 3, 6 and 9). Standard errors (clustered at county level) in parenthesis . * $p < 0.05$; ** $p < 0.01$

Figure 13: Persistence

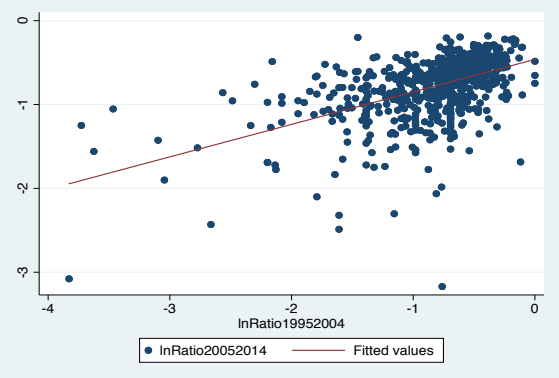
Panel A: Pro Coverage



Panel B: Anti Coverage



Panel C: Ratio



NOTE: We focus on two points in time: the decade from 2005 to 2014 and the decade from 1995 to 2004. We only consider newspapers that have been digitised for at least one year in both the decades. For each newspaper we take a yearly average of the relevant coverage variable for both the time periods considered (1995-2004 and 2005-2014) and then we build a county measure of the coverage variable by taking the average of all the newspapers digitised within each county. Panel A displays a scatter plot with the log of the pro-gay coverage measure in 2005-2014 on the vertical axis against the log of the pro-gay coverage measure in 1995-2004 on the horizontal axis. Panel B displays a scatter plot with the log of the anti-gay coverage measure in 2005-2014 on the vertical axis against the log of the anti-gay coverage measure in 1995-2004 on the horizontal axis. Panel C displays a scatter plot with the log of ratio measure in 2005-2014 on the vertical axis against the log of the ratio measure in 1995-2004 on the horizontal axis.

A Appendix (For Online Publication)

A.1 Algorithm for keywords

To attenuate the arbitrary choice of the keywords we also consider a dynamic approach. We implement an algorithm where the reference phrases obtained in the t-1 iteration are used as keywords in the subsequent iteration t. Ideally the algorithm should continue until the keywords in the final iteration (and therefore the reference phrases obtained in the previous iteration) perfectly coincide with the resulting reference phrases in that iteration. This way the choice of the final set of keywords would be partly ours (since we set the initial set of keywords in the iteration 0) partly coming from a learning process generated by the algorithm. We find that the set of anti-gay rights phrases perfectly converge after 4 rounds; the set of pro-gay rights phrases converges after 6 rounds, but to some sort of loop where the phrases "crime legisl" and "non-discrimin act" alternate with "report hate" and "introduc hate". This is most likely because both the set ["crime legisl","non-discrimin act"] and ["report hate", "introduc hate"] when used as keywords increase the total number of topical speeches containing such phrases, but also decrease their Pearson Chi values since these speeches are given by both pro-gay rights and anti-gay rights congressmen/senators. Table A6 reports the phrases selected at the end of this process. Results obtained using ["crime legisl","non-discrimin act"] instead of ["report hate", "introduc hate"] are very much equivalent.

A.2 Further Deleted Phrases

Phrases can be further deleted for the following reasons:

1) Phrases that contain a number are dropped. This applies for both written numbers and numeric ones. The numeric ones are dropped automatically by the script and the written ones by hand.

2) We also drop phrases containing names related to individuals, locations, organisations, court cases or legislative acts. Most names are dropped automatically in the script using named entity recognition and the ones not caught by it are dropped manually. Organizations such as Jones University (jone univer) are dropped manually. Court cases such as some v. Texas or Lawrence v. someone are dropped manually. Names of acts such as Family Abduction act are dropped manually.

3) A set of speeches that contain budget summaries enter the topical phrases and contain repeated phrases such construction plans. These phrases, such as "mi construct" are dropped manually.

4) Some speeches end up in the topical set due to a misunderstanding of the keywords. For example Homo Sapiens makes a speech topical even though it has nothing to do with the relevant topic. Phrases connected to such speeches are dropped.

5) There is an anomaly speech that repeats many times in the data. A single representative (Smith) uses the same speech to talk about hate crime 60 times without changing his wording. Phrases connected to such speeches are dropped.

6) There are also some typos that end up in the phrases. For example the extra space in same-sex. We drop these phrases.

Table A1: Threshold of Non Topical Bigrams

Non Topical: 150000		Non Topical: 250000	
Pro Reference Phrases	***Anti Reference Phrases***	***Pro Reference Phrases***	***Anti Reference Phrases***
gay lesbian	tradit marriag	gay lesbian	tradit marriag
sexual orient	union man	sexual orient	union man
gay men	same-sex marriag	gay men	same-sex marriag
speak hate	definit marriag	speak hate	definit marriag
gay man	redefin marriag	gay man	redefin marriag
base sexual	marriag union	base sexual	marriag union
crime base	protect tradit	crime base	marriag man
crime motiv	marriag man	crime motiv	institut marriag
men lesbian	marriag law	men lesbian	marriag law
lesbian american	marriag licens	lesbian american	marriag licens
orient gender	homosexu marriag	orient gender	homosexu marriag
non-discrimin act	defens marriag	non-discrimin act	defens marriag
employ nondiscrimin	defin marriag	employ nondiscrimir	defin marriag
discrimin gay	same-sex union	discrimin gay	same-sex union
pass hate	marriag act	employ non-discrimi	marriag act
employ non-discrimin	promot homosexu	gender ident	promot homosexu
enforc hate	issu marriag	lesbian gay	tradit definit
gender ident	tradit definit	serv open	legal same-sex
lesbian gay	support tradit	victim hate	opposit sex
serv open	legal same-sex	lgbt communiti	homosexu lifestyl
victim hate	opposit sex	gay american	legal union
lgbt communiti	homosexu lifestyl	gay coupl	say marriag
job discrimin	legal union	allow gay	homosexu militari
regardless sexual	say marriag	legal incid	marriag institut
gay american	homosexu militari	introduc hate	homosexu conduct
gay coupl	marriag institut	regardless sexual	right same-sex
allow gay	homosexu conduct	bisexu transgend	marriag legal
legal incid	right same-sex	ban gay	fundament institut
introduc hate	marriag legal	peopl transgend	marriag import
discriminatori polici	fundament institut	gay straight	protect marriag

NOTE: In columns (1) and (2) we show the 30 bigrams that are most diagnostic of pro-gay language in Congress and the 30 bigrams that are most diagnostic of anti-gay language in Congress when we consider only the 250000 most frequent bigrams in "non topical" speeches, in columns (3) and (4) when we consider only the 150000 most frequent bigrams in "non topical" speeches.

Table A2: Threshold of Topical Bigrams

Topical: 1000		Topical: 2000	
Pro Reference Phrases	***Anti Reference Phrases***	***Pro Reference Phrases***	***Anti Reference Phrases***
gay lesbian	tradit marriag	gay lesbian	tradit marriag
sexual orient	union man	sexual orient	same-sex marriag
gay men	same-sex marriag	gay men	definit marriag
base sexual	definit marriag	speak hate	redefin marriag
speak hate	marriag licens	gay man	marriag union
gay man	redefin marriag	base sexual	union man
crime base	marriag union	crime base	marriag law
crime motiv	marriag man	men lesbian	marriag licens
men lesbian	protect marriag	lesbian american	homosexu marriag
lesbian american	institut marriag	orient gender	defens marriag
basi sexual	marriag law	non-discrimin act	marriag man
orient gender	defens marriag	employ nondiscrimin	same-sex union
non-discrimin act	homosexu marriag	discrimin gay	defin marriag
employ nondiscrimin	defin marriag	pass hate	marriag act
discrimin gay	same-sex union	employ non-discrimin	promot homosexu
open gay	marriag act	gender ident	tradit definit
pass hate	marriag state	enforc hate	legal same-sex
gay peopl	promot homosexu	lesbian gay	opposit sex
employ non-discrimin	issu marriag	serv open	homosexu lifestyl
enforc hate	tradit definit	victim hate	legal union
gender ident	legal same-sex	lgbt communiti	homosexu militari
lesbian gay	opposit sex	gay american	say marriag
serv open	homosexu lifestyl	gay coupl	marriag institut
victim hate	homosexu militari	allow gay	homosexu conduct
lgbt communiti	legal union	legal incid	right same-sex
introduc hate	say marriag	introduc hate	fundament institut
gay american	marriag institut	regardless sexual	marriag import
gay coupl	gay militari	bisexu transgend	protect marriag
allow gay	homosexu conduct	ban gay	believ marriag
legal incid	right same-sex	peopl transgend	marriag protect

NOTE: In columns (1) and (2) we show the 30 bigrams that are most diagnostic of pro-gay language in Congress and the 30 bigrams that are most diagnostic of anti-gay language in Congress when we consider only the 1000 most frequent bigrams in "topical" speeches, in columns (3) and (4) when we consider only the 2000 most frequent bigrams in "topical" speeches

Table A3: Only bigrams not present in Non-Topical Speeches

Pro Reference Phrases	*Anti Reference Phrases*
gay lesbian	same-sex marriag
gay men	redefin marriag
speak hate	marriag union
gay man	marriag man
men lesbian	marriag law
lesbian american	marriag licens
orient gender	homosexu marriag
non-discrimin act	same-sex union
employ nondiscrimin	promot homosexu
discrimin gay	issu marriag
pass hate	tradit definit
employ non-discrimin	legal same-sex
enforc hate	opposit sex
gender ident	homosexu lifestyl
lesbian gay	legal union
serv open	say marriag
victim hate	homosexu militari
lgbt communiti	marriag institut
gay american	homosexu conduct
gay coupl	right same-sex
allow gay	marriag legal
legal incid	fundament institut
introduc hate	marriag import
regardless sexual	protect marriag
bisexu transgend	believ marriag
ban gay	marriag defin
peopl transgend	recognit same-sex
gay straight	block societi
right gay	homosexu activist
crime statut	marriag tradit

NOTE: We show the 30 bigrams that are most diagnostic of pro-gay language in Congress and the 30 bigrams that are most diagnostic of anti-gay language in Congress when we consider as "topical bigrams" only bigrams that are not included among the 200000 most frequent bigrams in "non topical" speeches

Table A4: 40 Phrases

Pro Reference Phrases	*Anti Reference Phrases*
gay lesbian	tradit marriag
sexual orient	union man
gay men	same-sex marriag
speak hate	definit marriag
gay man	redefin marriag
base sexual	marriag union
crime base	marriag man
crime motiv	marriag law
men lesbian	marriag licens
lesbian american	homosexu marriag
orient gender	defens marriag
non-discrimin act	defin marriag
employ nondiscrimin	same-sex union
discrimin gay	marriag act
pass hate	promot homosexu
employ non-discrimin	issu marriag
enforc hate	tradit definit
gender ident	legal same-sex
lesbian gay	opposit sex
serv open	homosexu lifestyl
victim hate	legal union
lgbt communiti	say marriag
gay american	homosexu militari
gay coupl	marriag institut
allow gay	homosexu conduct
legal incid	right same-sex
introduc hate	marriag legal
regardless sexual	fundament institut
bisexu transgend	marriag import
ban gay	protect marriag
peopl transgend	believ marriag
gay straight	marriag protect
right gay	marriag defin
crime statut	recognit same-sex
gay communiti	block societi
number hate	homosexu activist
motiv hate	marriag tradit
immigr victim	believ tradit
ban same-sex	actual perceiv
lesbian communiti	thought crime

NOTE: The Table shows the 40 bigrams that are most diagnostic of pro-gay language in Congress and the 40 bigrams that are most diagnostic of anti-gay language in Congress.

Table A5: Set of keywords

Smaller set of keywords		Larger set of keywords	
Pro Reference Phrases	***Anti Reference Phrases***	***Pro Reference Phrases***	***Anti Reference Phrases***
gay lesbian	tradit marriag	hate crime	tradit marriag
sexual orient	definit marriag	sexual orient	union man
gay men	redefin marriag	gay lesbian	definit marriag
speak hate	same-sex marriag	gay men	institut marriag
gay man	marriag union	speak hate	marriag law
base sexual	marriag man	crime base	marriag union
men lesbian	union man	crime motiv	same-sex marriag
lesbian american	institut marriag	base sexual	redefin marriag
employ nondiscrimin	marriag law	gay man	marriag man
discrimin gay	marriag licens	famili valu	protect tradit
pass hate	homosexu marriag	basi sexual	marriag licens
non-discrimin act	same-sex union	pass hate	protect marriag
lesbian gay	defin marriag	men lesbian	defin marriag
serv open	defens marriag	enforc hate	defens marriag
orient gender	promot homosexu	lesbian american	marriag protect
employ non-discrimin	marriag act	employ nondiscrimin	marriag act
gay american	issu marriag	orient gender	tradit definit
gay coupl	tradit definit	non-discrimin act	issu marriag
victim hate	legal same-sex	victim hate	marriag state
allow gay	opposit sex	discrimin gay	homosexu marriag
legal incid	homosexu lifestyl	employ non-discrimin	fundament institut
introduc hate	legal union	open gay	same-sex union
bisexu transgend	say marriag	gay peopl	opposit sex
ban gay	homosexu militari	feder hate	promot homosexu
peopl transgend	marriag institut	lesbian gay	support tradit
gay straight	homosexu conduct	serv open	marriag institut
right gay	right same-sex	crime statut	marriag import
gender ident	marriag legal	motiv hate	say marriag
gay communiti	fundament institut	number hate	legal union
regardless sexual	marriag import	lgbt communiti	believ marriag

NOTE: In columns (1) and (2) we show the 30 bigrams that are most diagnostic of pro-gay language in Congress and the 30 bigrams that are most diagnostic of anti-gay language in Congress when we consider as keywords to define a speech as topical the set: "gay", "lesbian", "same sex", "transgender", "transsexual", "progay","antigay", " homo", "heterosexual" , in columns (3) and (4) when we consider as keywords to define a speech as topical the set: "gay", "lesbian", "same sex", "transgender", "transsexual", "progay", "antigay", " homo", "heterosexual" "gender identity", "sexual identity", "LGBT", "GLBT", "right of marriage", "marriage rights", "marriage equality", "respect for marriage", "defense of marriage", "family values", "don't ask don't tell", "between one man and one woman", "between men and women", "sanctity of marriage", "definition of marriage", "traditional marriage", "Institution of marriage", "protection of marriage", "heterosexual", "gender expression", "homophobia", "sex lives", "sexual conduct", "sexual preferences", "sexual disposition", "bisexual", "men who have sex with men", "MSM", "marriage in its traditional form", "sexual orientation".

Table A6: Set of keywords (Algorithm)

pro	anti
hate crime	tradit marriag
sexual orient	union man
gay lesbian	definit marriag
crime law	institut marriag
gay men	marriag union
crime motiv	outsid marriag
crime base	marriag man
base sexual	redefin marriag
pass hate	protect tradit
hate violenc	same-sex marriag
victim hate	togeth outsid
employ nondiscrimin	issu marriag
men lesbian	marriag law
child marriag	marriag protect
motiv hate	defin marriag
lesbian american	defens marriag
orient gender	coupl live
crime statut	tradit definit
employ non-discrimin	fundament institut
gay man	ident incom
enforc enhanc	marriag act
discrimin gay	marriag licens
number hate	protect marriag
gay peopl	marriag state
feder hate	homosexu marriag
current hate	believ marriag
open gay	support tradit
need hate	same-sex union
report hate	opposit sex
introduc hate	marriag institut

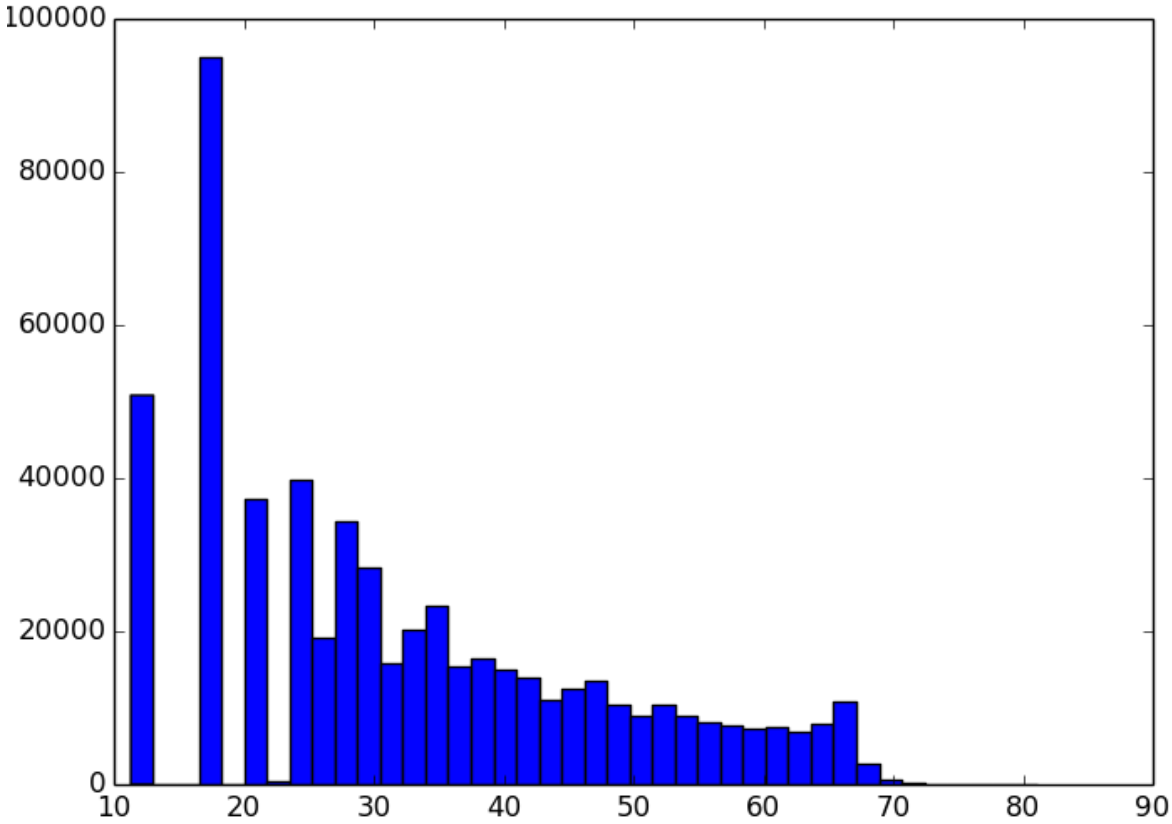
NOTE: We implement an algorithm where the reference phrases obtained in the t-1 iteration are used as keywords in the subsequent iteration t. Columns (1) and (2) show the 30 bigrams that are most diagnostic of pro-gay language in Congress and the 30 bigrams that are most diagnostic of anti-gay language in Congress after we implemented the dynamic algorithm described in Section A.1.

Table A7: Further deleted phrases

Number	Name/Orga	Budget	Wrong	Repeted speech	Typo
man one	Judici court	ms upgrad	First earth	connect incid	Same- sex
marriag	jame	ca widen	Clash civil	unit cohes	Repeal n't
union one	colonel	bus bus	Old-ag		
	joycelyn	pedestrian	Mad-mad		
	theatre	improv	Mellon		
	lsc grante	mi	Homo		
	captain	ca improv	gospel		
	suprem	apostl			
	jone univer	construct			
	Missouri	pursuant			
	Massachuse	construct			
	Jame byrd	ny town			
	Cherri	schedul			
	Lawrenc v.	ct construct			
	Enola gay	various			
	Fas citizen	pa			
	Shepard	md			
	Family	creek lake			
	Dr. mertz	tn			
	Vawa	condit			
	v. texa	request			
	day silenc	transit			
	kennedi	fl construct			
	mr.	ny			
	truth	grade separ			
	credit claus	project			
	mr. Hormel	row			
	jim Hormel	complianc			
		budget			
		nj construct			
		ga			
		intermod			
		investig			
		construct			
		inspect			
		pa design			
		ny improv			
		municip			
		construct			
		il construct			
		mn			
		upgrad			
		bus facil			
		int'l airport			
		state rout			
		bicycl			
		oh			
		ca			
		reservoir			
		tx construct			
		footwear			
		mixture			
		street bridg			
		free free			
		chang			
		free chang			
		construct			

NOTE: The Table reports the list of phrases manually deleted because of reasons described in Section A.2.

Figure A1: Number of stems for tf-idf values



NOTE: We plot the counts of stems on the vertical axis against tf-idf (term frequency–inverse document frequency) values on the horizontal axis. The formula for the stem v is $\text{tf-idf}_v = \log [1 + (N_v)] \times \log (D/D_v)$ where N_v is the number of times v appears in the dataset and D_v is the number of abstracts in which v appears.

Figure A2: Word Clouds

Topic 1

Topic 2

Topic 3

Topic 4

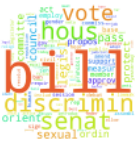
Topic 5

Topic 6

Topic 7

Topic 8

Topic 9



Topic 10

Topic 11

Topic 12

Topic 13

Topic 14

Topic 15

Topic 16

Topic 17

Topic 18

Topic 19

Topic 20

Table A8

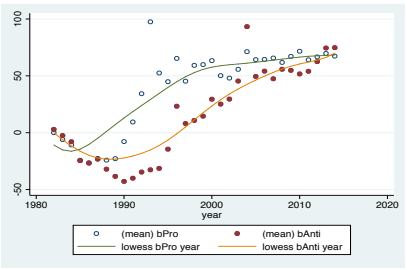
Topic 1		Topic 2		Topic 3		Topic 4		Topic 5		Topic 6		Topic 7		Topic 8		Topic 9		Topic 10	
film	0.012	meet	0.0168	rule	0.0371	polic	0.0213	school	0.058	wed	0.0113	aid	0.0698	republican	0.0251	bill	0.034	boy	0.1278
show	0.0117	center	0.0152	sex	0.0366	crime	0.0176	student	0.0413	look	0.0056	health	0.0226	democrat	0.0245	discrimin	0.0261	girl	0.1147
play	0.0098	inform	0.0116	suprem	0.0307	charg	0.0162	univers	0.0322	marri	0.0054	test	0.0192	elect	0.0196	hous	0.0257	son	0.0688
movi	0.0075	event	0.0115	judg	0.0282	man	0.0136	board	0.0159	night	0.004	diseas	0.0176	candid	0.0171	senat	0.0238	daughter	0.0615
star	0.0065	free	0.0112	decis	0.0195	hate	0.0121	educ	0.0148	man	0.0039	virus	0.0107	presid	0.0165	vote	0.0232	birth	0.0462
music	0.0065	today	0.0105	feder	0.0194	death	0.0093	colleg	0.0147	friend	0.0039	drug	0.0092	polit	0.0154	legisl	0.0178	ave	0.0305
theater	0.0064	club	0.0102	legal	0.0189	offic	0.009	campus	0.0101	hand	0.0038	blood	0.0091	campaign	0.0133	sexual	0.0165	baptist	0.0199
art	0.0056	communiti	0.01	appeal	0.015	murder	0.009	teacher	0.0099	white	0.0038	medic	0.009	parti	0.0129	council	0.016	juli	0.0157
perform	0.0055	com	0.0073	case	0.0148	arrest	0.0085	member	0.0091	ceremoni	0.0037	die	0.0083	senat	0.0094	orient	0.0143	deaco	0.0157
televis	0.005	support	0.0073	ban	0.0138	kill	0.008	program	0.0083	walk	0.0037	men	0.0077	vote	0.0081	pass	0.0126	hospit	0.0128
open	0.0048	servic	0.0064	marri	0.0123	sentenc	0.0076	public	0.0081	room	0.0035	infect	0.0074	presidenti	0.0076	committe	0.0116	sept	0.0126
stori	0.0046	park	0.0064	justic	0.0119	prison	0.0074	sexual	0.0077	befor	0.0034	public	0.0072	conserv	0.0073	protect	0.0115	east	0.012
night	0.0045	public	0.0061	attorney	0.0116	alleg	0.0068	communiti	0.0069	bride	0.0034	patient	0.0071	support	0.0072	ordin	0.0109	center	0.0111
book	0.0045	open	0.0059	constitut	0.0105	investig	0.0066	district	0.0067	could	0.0032	hiv	0.0067	hous	0.0069	propos	0.0107	drive	0.0094
festiv	0.0044	librari	0.0058	file	0.0084	victim	0.0066	bisexu	0.0058	hour	0.0032	offici	0.0066	district	0.0068	approv	0.0099	spring	0.0077
york	0.0041	noon	0.0055	district	0.0081	men	0.0064	parent	0.0055	around	0.0032	center	0.0065	voter	0.0065	measur	0.0082	buffalo	0.0074
best	0.004	famili	0.0052	act	0.0074	jail	0.0064	transgend	0.0054	love	0.0032	hospit	0.0063	sen	0.0063	act	0.0078	medic	0.0069
award	0.004	includ	0.0051	allow	0.0068	case	0.0063	polici	0.0052	home	0.0032	immun	0.0061	governor	0.0062	support	0.0072	merci	0.0066
seri	0.0038	illinoi	0.0048	lawsuit	0.0067	attack	0.0061	organ	0.0051	got	0.0032	case	0.0059	run	0.0062	employ	0.0069	born	0.0064
world	0.0036	happen	0.0048	order	0.0065	attorney	0.006	meet	0.0049	want	0.0032	spread	0.0054	race	0.0059	base	0.0067	provid	0.0061

Topic 11		Topic 12		Topic 13		Topic 14		Topic 15		Topic 16		Topic 17		Topic 18		Topic 19		Topic 20	
ave	0.0285	letter	0.0079	church	0.0739	famili	0.0193	militari	0.0343	organ	0.0159	sex	0.0397	tax	0.0143	percent	0.0245	report	0.0179
file	0.0184	editor	0.0073	cathol	0.0174	children	0.015	polici	0.0215	pride	0.0153	union	0.0236	compani	0.0104	report	0.0234	polic	0.0174
dougla	0.0134	think	0.0061	rev	0.0141	live	0.0131	presid	0.0192	communiti	0.015	amend	0.0223	employe	0.0102	women	0.0201	busi	0.0103
drive	0.0128	american	0.0052	bishop	0.0139	marri	0.013	ask	0.0144	parad	0.0138	legal	0.0216	million	0.0096	men	0.0193	area	0.0094
plaza	0.0127	recent	0.0052	christian	0.0112	parent	0.0125	forc	0.0127	protest	0.0116	constitut	0.0196	health	0.0088	studi	0.0151	block	0.0092
follow	0.0113	doe	0.0051	unit	0.0109	want	0.0108	serv	0.0126	celebr	0.0111	vote	0.0137	pay	0.0085	american	0.0145	town	0.0086
spokan	0.0103	articl	0.005	pastor	0.0099	love	0.0101	servic	0.0119	event	0.0105	civil	0.0137	servic	0.0083	accord	0.0143	street	0.0082
road	0.01	becaus	0.0049	minist	0.0093	friend	0.0099	tell	0.0105	annual	0.0091	voter	0.0129	benefit	0.0082	sexual	0.0115	incid	0.008
clerk	0.0091	polit	0.0048	episcop	0.0085	life	0.0085	ban	0.0104	support	0.0076	support	0.0126	fund	0.0077	number	0.0114	park	0.0078
record	0.0089	believ	0.0047	member	0.0081	mother	0.0085	scout	0.01	activist	0.0074	ban	0.0121	cost	0.0076	survey	0.0105	start	0.0077
judgment	0.008	word	0.0047	leader	0.0081	man	0.0083	open	0.0091	ralli	0.0072	man	0.012	money	0.0074	research	0.0097	build	0.0063
money	0.0072	read	0.0047	religi	0.0073	know	0.008	offic	0.0086	bisexu	0.0072	woman	0.0118	insur	0.007	releas	0.0089	road	0.0062
applic	0.007	know	0.0046	congreg	0.0073	home	0.0077	defens	0.0085	demonstr	0.0072	measur	0.0102	busi	0.0067	violenc	0.0087	depart	0.0058
inc	0.007	whi	0.0046	denomin	0.0072	becaus	0.0073	armi	0.0077	transgend	0.007	marri	0.01	plan	0.0061	show	0.0081	befor	0.0057
address	0.0067	ani	0.0043	baptist	0.0068	child	0.0069	arm	0.0075	york	0.0069	ballot	0.0096	budget	0.0057	increas	0.0078	today	0.0053
offic	0.0061	person	0.0043	methodist	0.0067	woman	0.0061	war	0.0072	men	0.0064	propos	0.0085	work	0.0056	sex	0.0077	store	0.0053
neb	0.006	seem	0.0043	sex	0.0058	husband	0.0061	member	0.0068	gather	0.0064	defin	0.008	worker	0.0052	poll	0.0071	charg	0.0053
bank	0.0058	societi	0.0041	god	0.0058	wife	0.0061	allow	0.0066	street	0.006	approv	0.0079	program	0.0052	found	0.0071	resid	0.0052
bull	0.0057	public	0.004	roman	0.0057	come	0.006	pentagon	0.0061	thousand	0.0054	benefit	0.0078	provid	0.0051	rate	0.0064	car	0.0051
divorc	0.0056	use	0.0038	open	0.0054	ago	0.0059	end	0.006	member	0.0053	allow	0.0078	job	0.005	among	0.0055	drive	0.0051

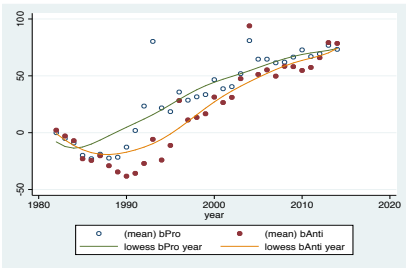
NOTE: Table A8 displays for each topic the probability distribution over the vocabulary. For each topic we report only the 20 terms with highest ptobabilties.

Figure A3: Time Diffusion, Robustness

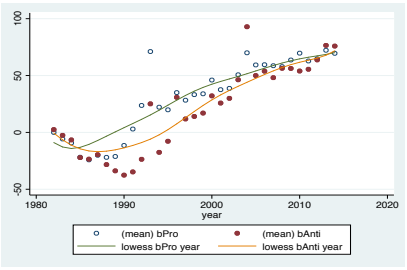
Panel A: Large Keys



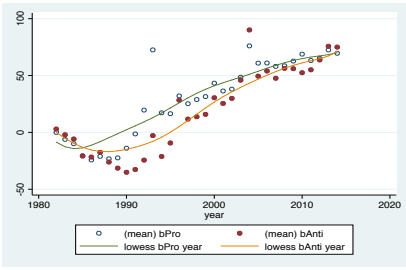
Panel B: Short Keys



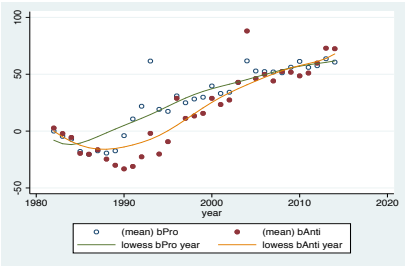
Panel C: 1000



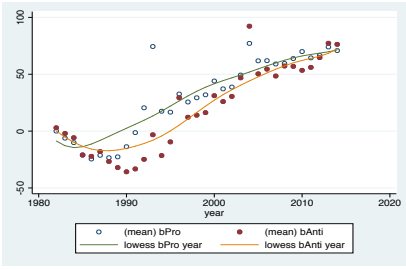
Panel D: 2000



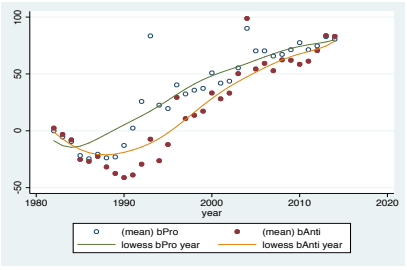
Panel E: 150000



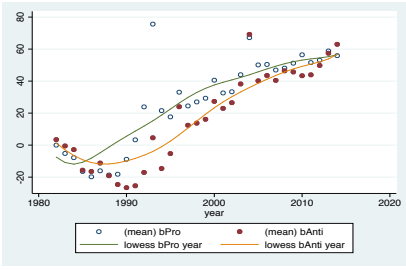
Panel F: 250000



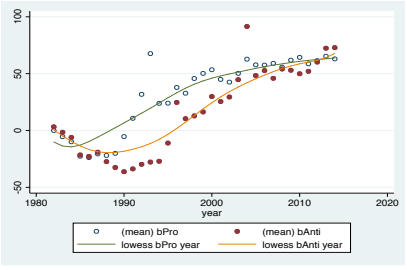
Panel G: 40



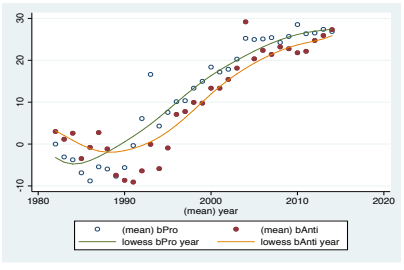
Panel H: Bigrams not in Non-Topical Speechs



Panel I: Algorithm for keywords selection



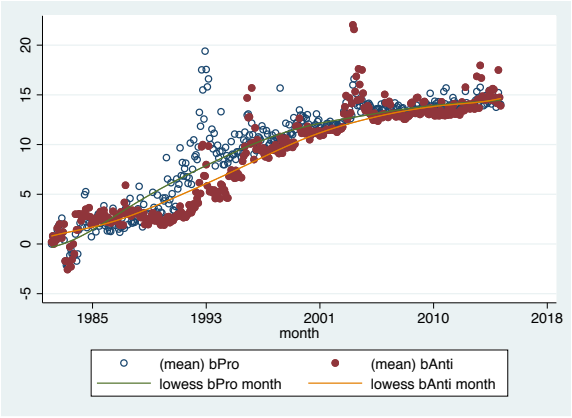
Panel L: No duplicates



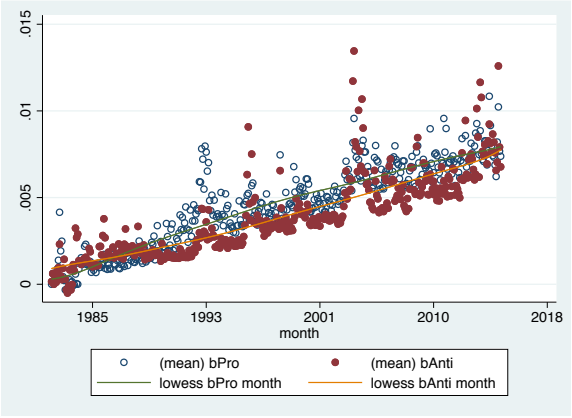
NOTE: We regress the coverage measure of type g (we only have two types, pro-gay and anti-gay) of a newspaper n in the year t on newspaper fixed effects and on interactions of type and year fixed effects. Panel A plots the coefficients **bPro (bAnti)** of the year fixed effects when we use the larger set of keywords to define speeches as "topical"; Panel B plots the coefficients when we consider the smaller set of keywords; Panel C plots the coefficients when we consider only 1000 most frequent bigrams within topical speeches; Panel D plots the coefficients when we consider the 2000 most frequent bigrams within topical speeches; Panel E plots the coefficients when we consider only 150000 most frequent bigrams within non topical speeches; Panel F plots the coefficients when we consider the 250000 most frequent bigrams within non topical speeches; Panel G plots the coefficients when we consider 40 pro-gay phrases and 40 anti-gay phrases; Panel H plots the coefficients when we consider as topical bigrams only bigrams that are not included in the 200000 most frequent bigrams in the non topical speeches; Panel I plots the coefficients when we consider reference bigrams chosen using the dynamic algorithm discussed in Section A1. In Panel L we only consider articles that either contain unpreprocessed phrases associated to pro-gay bigrams or contain unpreprocessed phrases associated to anti-gay bigrams and we count each article only one time independently of the number of unpreprocessed phrases contained.

Figure A4: Diffusion over time, Monthly cells, Robustness

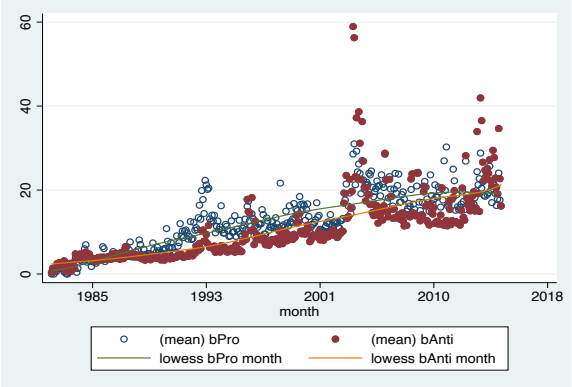
Panel A: control for n digitised articles



Panel B: share



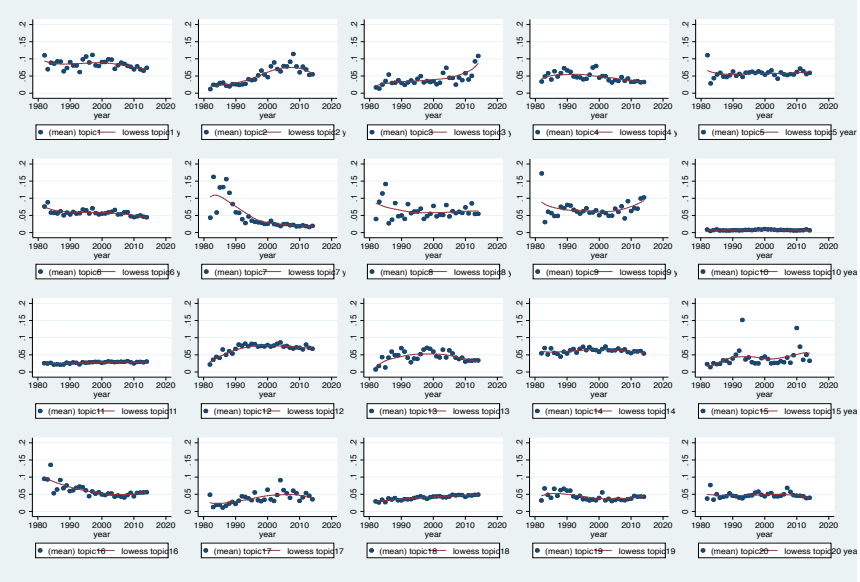
Panel C: Circulation weights



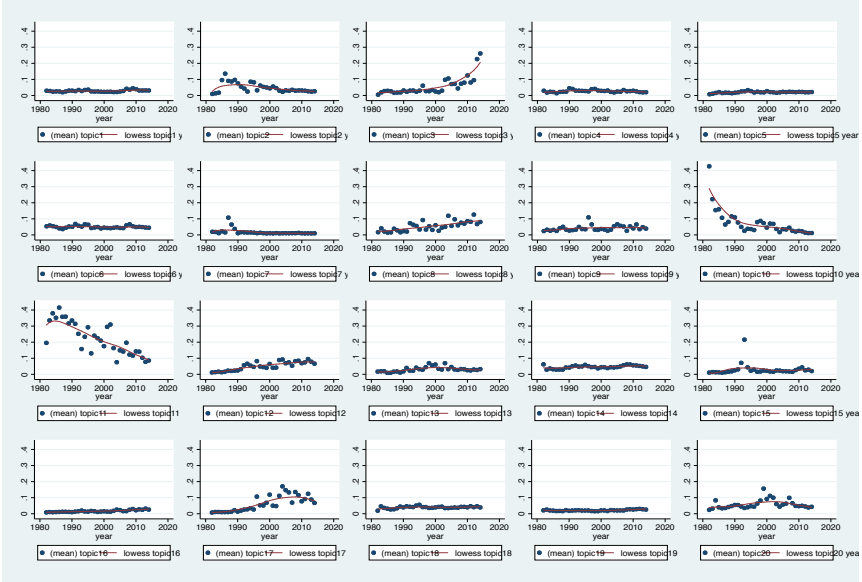
NOTE: We regress the coverage measure of type g (we only have two types, pro-gay and anti-gay) of a newspaper n in the month t on newspaper fixed effects and on interactions of type and month fixed effects. We plot the coefficients **bPro (bAnti)** of the month fixed effects when type is pro- (anti-) gay against time; hollow (red) circles correspond to coefficients when type is pro- (anti-) gay. Panel A includes in the main regression a proxy for the number of digitised articles per newspaper-year. In Panel B we use as a dependent variable the ratio between our main coverage measure and our proxy for the number of digitised articles per newspaper-year. In Panel C we perform the standard regression as in Panel A Figure 4 but using as weights for each newspaper n the number of copies sold in 2004.

Figure A5: Topics evolution over time

Panel A: Pro-gay corpus



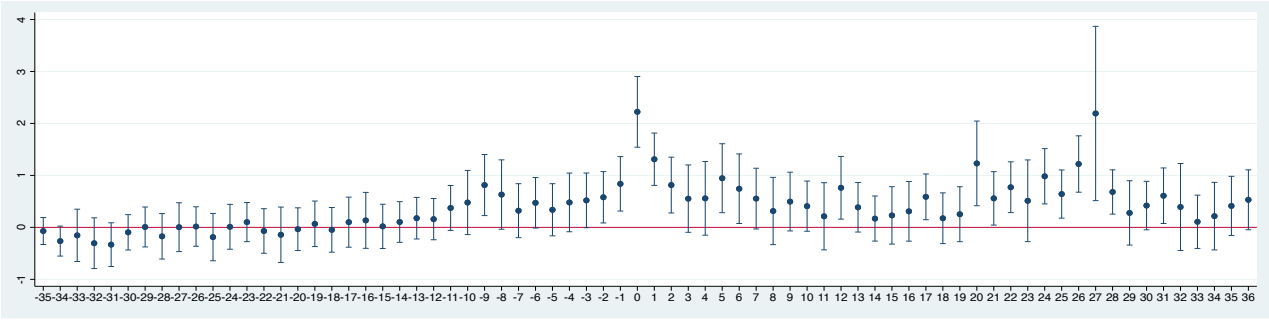
Panel B: Anti-gay corpus



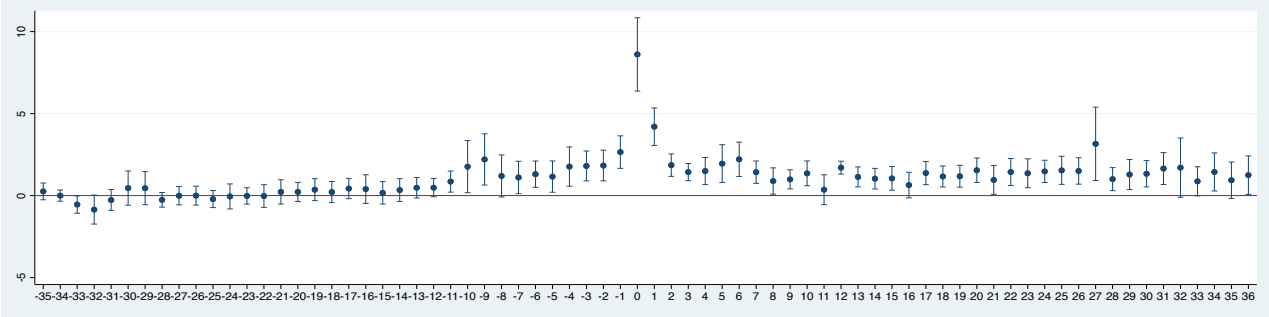
NOTE: In Panel A we consider the corpus of abstracts of articles containing pro-gay phrases and we plot the fraction of text devoted to each topic against year of publication; In Panel B we consider the corpus of abstracts of articles containing anti-gay phrases and we plot the fraction of text devoted to each topic against year of publication.

Figure A6: Introduction of gay marriages, Main Specification, Coefficients and Confidence Intervals

Panel A: Pro Coverage (Dependent Variable)



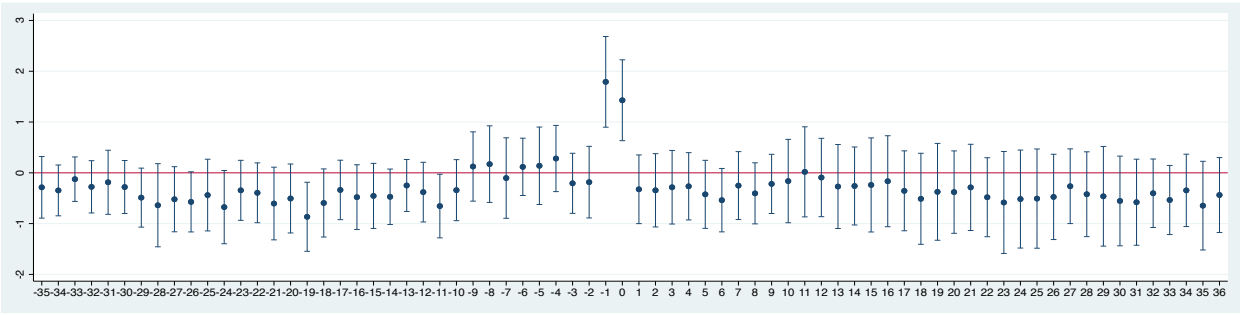
Panel B: Anti Coverage (Dependent Variable)



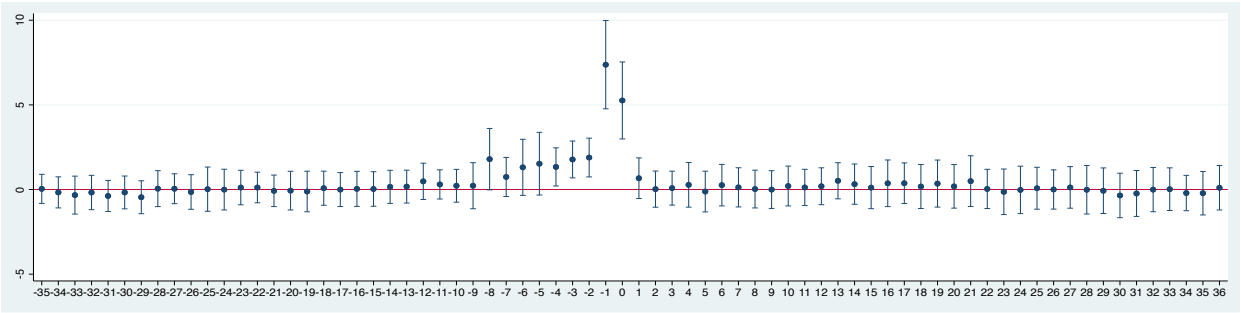
NOTE: We consider the introduction of gay marriages. Panel A (Panel B) presents the graphical representation of the event analysis exercise based on specification (2) in the paper when the dependent variable is pro- (anti-) gay coverage. We include the month of implementation of the change, 36 months before and 36 months afterwards (therefore we have 73 variables each corresponding to a given month) and we only consider newspapers that have been digitised throughout all the time considered. In Panel A we report the coefficients and their confidence intervals (at 5% level) of each of these variables obtained using Pro Coverage as dependent variable, in Panel B obtained using Anti Coverage as dependent variable. The omitted category is the variable equal to one if t was exactly 36 months before the reform.

Figure A7: Ban of gay marriages, Main Specification, Coefficients and Confidence Intervals

Panel A: Pro Coverage (Dependent Variable)

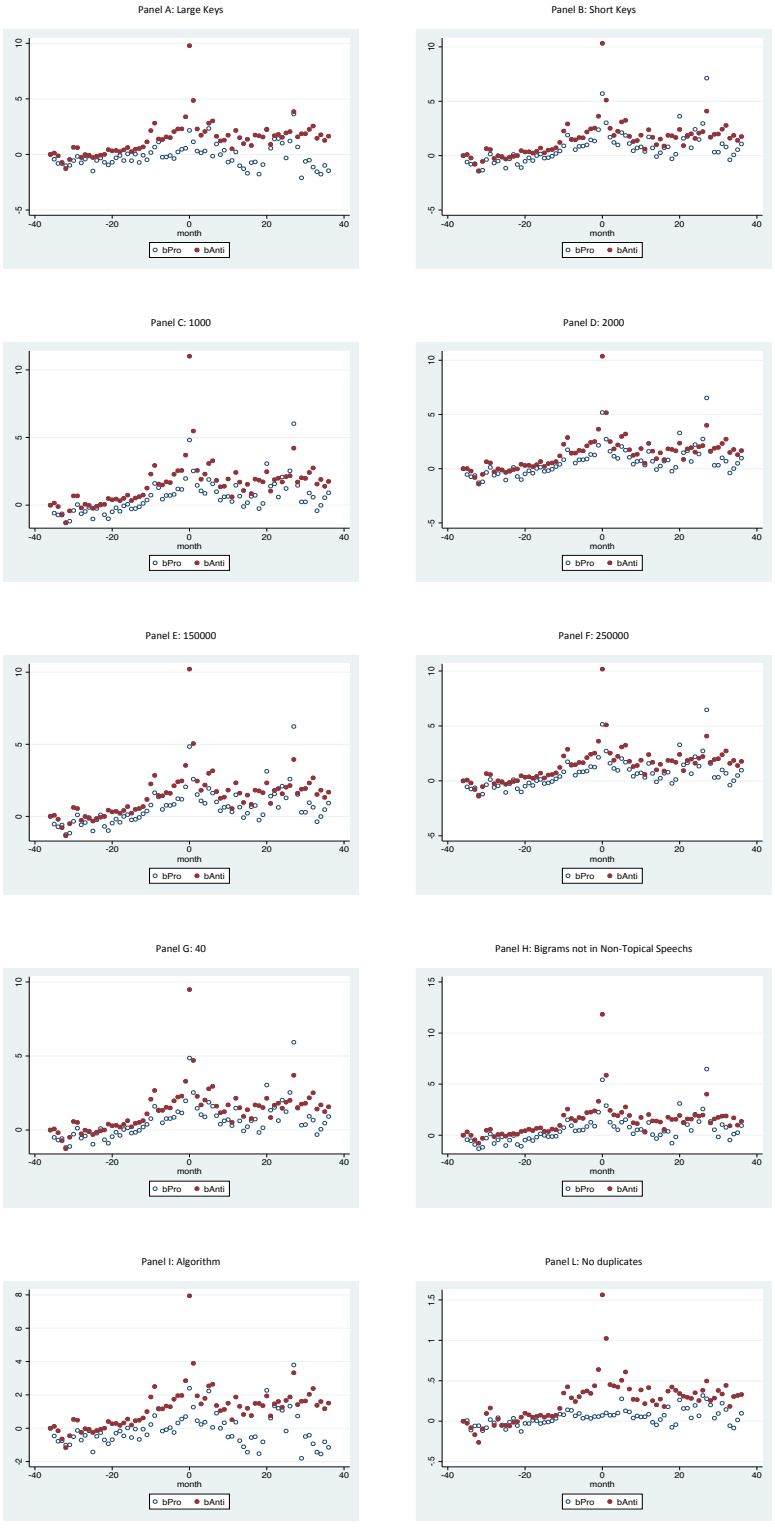


Panel B: Anti Coverage (Dependent Variable)



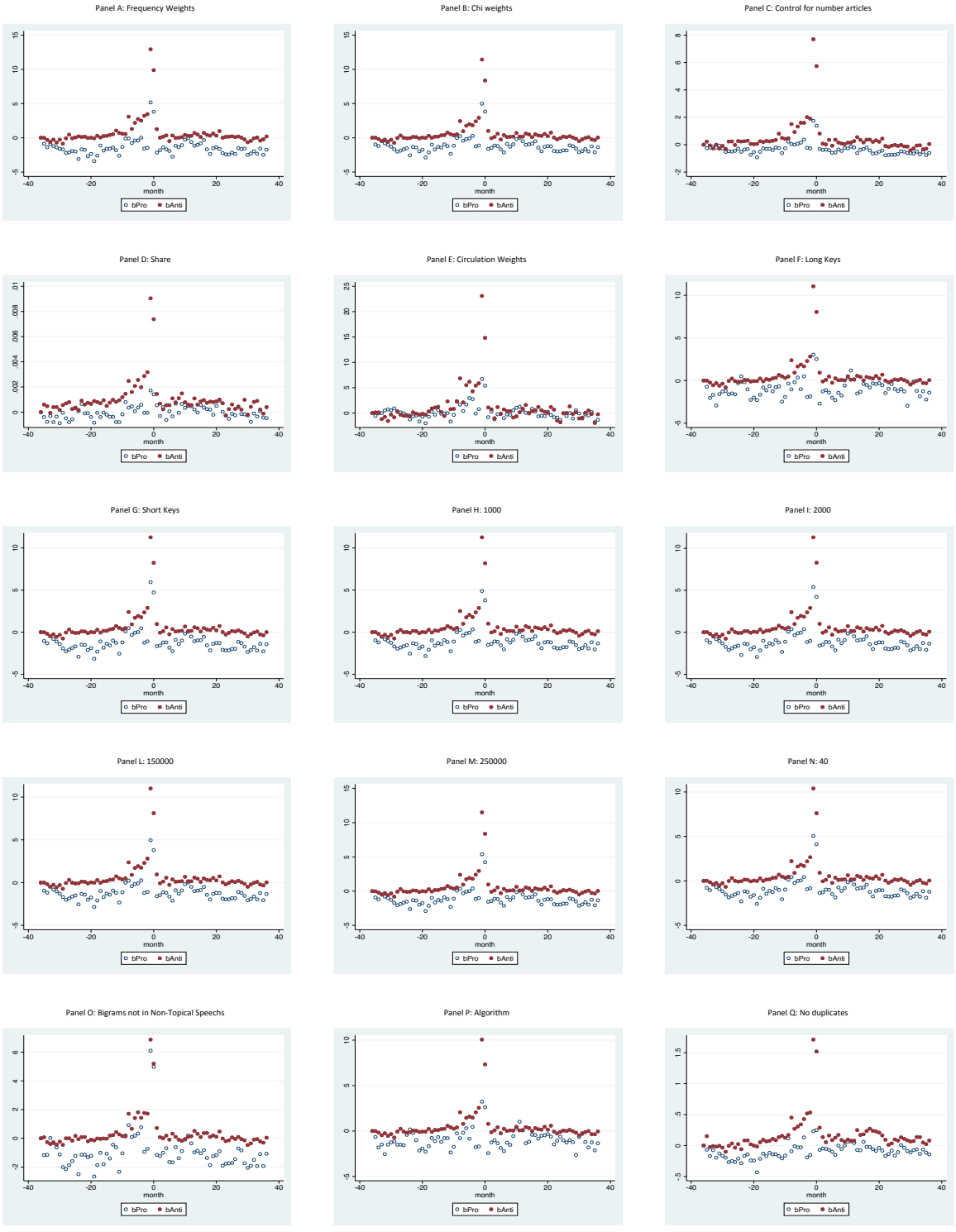
NOTE: We consider the ban of gay marriages. Panel A (Panel B) presents the graphical representation of the event analysis exercise based on specification (2) in the paper when the dependent variable is pro- (anti-) gay coverage. We include the month of implementation of the change, 36 months before and 36 months afterwards (therefore we have 73 variables each corresponding to a given month) and we only consider newspapers that have been digitised throughout all the time considered. In Panel A we report the coefficients and their confidence intervals (at 5% level) of each of these variables obtained using Pro Coverage as dependent variable, in Panel B obtained using Anti Coverage as dependent variable. The omitted category is the variable equal to one if t was exactly 36 months before the reform.

Figure A8: Introduction of gay marriages (Robustness)



NOTE: We consider the introduction of gay marriages. We include the month of implementation of the change, 36 months before and 36 months afterwards and we only consider newspapers that have been digitised throughout all the time considered. Hollow (red) circles correspond to the graphical representation of the event analysis exercise based on specification (2) in the paper when the dependent variable is pro- (anti-) gay coverage. Panel A plots the coefficients when we use the larger set of keywords to define speeches as "topical"; Panel B plots the coefficients when we consider the smaller set of keywords; Panel C plots the coefficients when we consider only 1000 most frequent bigrams within topical speeches; Panel D plots the coefficients when we consider the 2000 most frequent bigrams within topical speeches; Panel E plots the coefficients when we consider only 150000 most frequent bigrams within non topical speeches; Panel F plots the coefficients when we consider the 250000 most frequent bigrams within non topical speeches; Panel G plots the coefficients when we consider 40 pro-gay phrases and 40 anti-gay phrases; Panel H plots the coefficients when we consider as topical bigrams only bigrams that are not included in the 200000 most frequent bigrams in the non topical speeches; Panel I plots the coefficients when we consider reference bigrams chosen using the dynamic algorithm discussed in Section A1. In Panel L we only consider articles that either contain unpreprocessed phrases associated to pro-gay bigrams or contain unpreprocessed phrases associated to anti-gay bigrams and we count each article only one time independently of the number of unpreprocessed phrases contained.

Figure A9: Ban of gay marriages (Robustness)



NOTE: We consider the ban of gay marriages. We include the month of implementation of the change, 36 months before and 36 months afterwards and we only consider newspapers that have been digitised throughout all the time considered. Hollow (red) circles correspond to the graphical representation of the event analysis exercise based on specification (2) in the paper when the dependent variable is pro- (anti-) gay coverage. Panel A plots results when the dependent variable, the coverage measure, reflects the relevance of the phrases covered by the newspaper and their relevance is based on their frequency whitin topical speeches in Congress, Panel B when the dependent variable, the coverage measure, reflects the relevance of the phrases covered by the newspaper and their relevance is based on their Chi values, that proxy how strongly we can reject the hypothesis they are as likely to be used by pro-gay Congressmen than anti-gay Congressmen. In Panel C we include our proxy for the number of digitised articles as dependent variable. In Panel D we use as a dependent variable the ratio between our main coverage measure and our proxy for the number of digitised articles per newspaper-year. In Panel E we weight each newspaper n by the number of copies sold in 2004. Panel F plots the coefficients when we use the larger set of keywords to define speeches as "topical"; Panel G plots the coefficients when we consider the smaller set of keywords; Panel H plots the coefficients when we consider only 1000 most frequent bigrams within topical speeches; Panel I plots the coefficients when we consider the 2000 most frequent bigrams within topical speeches; Panel L plots the coefficients when we consider only 150000 most frequent bigrams within non topical speeches; Panel M plots the coefficients when we consider the 250000 most frequent bigrams within non topical speeches; Panel N plots the coefficients when we consider 40 pro-gay phrases and 40 anti-gay phrases; Panel O plots the coefficients when we consider as topical bigrams only bigrams that are not included in the 200000 most frequent bigrams in the non topical speeches; Panel P plots the coefficients when we consider reference bigrams chosen using the dynamic algorithm discussed in Section A1. In Panel Q we only consider articles that either contain unpreprocessed phrases associated to pro-gay bigrams or contain unpreprocessed phrases associated to anti-gay bigrams and we count each article only one time independently of the number of unpreprocessed phrases contained.

Table A9

State	year of first law	month of first law	year of second law	month of second law
Mississippi				
Iowa	2009	4		
Oklahoma	2014	10		
Wyoming	2014	10		
Minnesota	2013	7		
Illinois	2013	11		
Arkansas	2014	5		
New Mexico	2013	12		
Indiana	2014	6	2014	10
Maryland	2012	11		
Louisiana				
Idaho	2014	10		
Arizona	2014	10		
Wisconsin	2014	10		
Michigan	2014	3		
Kansas	2014	11		
Utah	2013	12		
Virginia	2014	10		
Oregon	2014	5		
Connecticut	2008	11		
Montana	2014	11		
California	2008	5	2013	6
Texas				
West Virginia	2014	10		
South Carolina	2014	11		
New Hampshire	2009	6		
Massachusetts	2004	5		
Vermont	2009	4		
Georgia				
North Dakota				
Pennsylvania	2014	5		
Florida				
Alaska	2014	10		
Kentucky				
Hawaii	1993	5	2013	11
Nebraska				
Missouri				
Ohio				
Alabama				
New York	2011	6		
South Dakota				
Colorado	2014	10		
New Jersey	2013	9		
Washington	2012	11		
North Carolina	2014	10		
Tennessee				
Nevada	2014	10		
Delaware	2013	5		
Maine	2012	11		
Rhode Island	2013	5		
Washington DC	2009	12		

NOTE: Table A9 shows the list of laws introducing gay marriage that are considered in the empirical specification (2) of the paper

Table A10

State	year of first law	month of first law	year of second law	month of second law	year of third law	month of third law
Mississippi	1996	8	1997	2	2004	11
Iowa						
Oklahoma	1996	4	2004	11		
Wyoming						
Minnesota	1997	6				
Illinois	1996	5				
Arkansas	1997	2	2004	11		
New Mexico						
Indiana	1986	3	1997	5		
Maryland						
Louisiana	1999	7	2004	9		
Idaho	1995	3	1996	3	2006	11
Arizona	1996	5	2008	11		
Wisconsin	2006	11				
Michigan	1996	6	2004	11		
Kansas	1996	4	2005	4		
Utah	1995	3	2004	3	2004	11
Virginia	1997	3	2004	4	2006	11
Oregon	2004	11				
Connecticut						
Montana	1997	4	2004	11		
California	2000	3	2008	11		
Texas	1997	4	2003	5	2005	11
West Virginia	2000	3				
South Carolina	1996	5	2007	2		
New Hampshire	2004	5				
Massachusetts						
Vermont						
Georgia	1996	4	2004	11		
North Dakota	1997	3	2004	11		
Pennsylvania	1996	10				
Florida	1997	6	2008	11		
Alaska	1996	5	1998	11		
Kentucky	1998	4	2004	11		
Hawaii	1994	6	1998	11		
Nebraska	2000	11				
Missouri	1996	7	2001	7	2004	8
Ohio	2004	2	2004	11		
Alabama	1996	8	1998	5	2006	6
New York						
South Dakota	1996	2	2006	11		
Colorado	2000	5	2006	11		
New Jersey						
Washington	1998	2				
North Carolina	1996	6	2012	5		
Tennessee	1996	5	2006	11		
Nevada	2002	11				
Delaware	1996	6				
Maine	1997	3				
Rhode Island						
Washington DC						

NOTE: Table A10 shows the list of laws banning gay marriage that are considered in the empirical specification (2) of the paper

Table A11: Moran I (weights)

	2010-2014 Frequency Weights	**2010-2014 Chi Weights**
Pro Coverage	0.106***	0.104***
	(0.027)	(0.027)
Anti Coverage	0.144***	0.147***
	(0.027)	(0.027)
ratio	0.207***	0.218***
	(0.027)	(0.027)

NOTE: Table A11 displays the Moran I for the coverage of pro-gay language (row 1), anti-gay language (row 3), ratio of the coverage of pro- gay language over the sum of pro- and anti-gay language (row 5). We consider the time period 2010-2014 and include only newspapers digitised between 2010 and 2014. In Column (1) the coverage measures reflect the relevance of the phrases covered by the newspaper and their relevance is based on their frequency whitin topical speeches in Congress, in Column (2) they reflect the relevance of the phrases covered by the newspaper and their relevance is based on their Chi values, that proxy how strongly we can reject the hypothesis they are as likely to be used by pro-gay Congressmen than anti-gay Congressmen. Standard errors in parenthesis . * p < 0.10; ** p < 0.05; *** p ‹ 0.01

Table A12: Moran I

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
ratio	0.201***	0.216***	0.217***	0.216***	0.221***	0.226***	0.205***	0.230***	0.229***
	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)
	40	1000	2000	150000	250000	ShortKeys	LargeKeys	Not in non top.	algorithm

NOTE: Table A12 display the Moran I for the ratio of the coverage of pro-gay language over the sum of pro- and anti-gay language for the time period 2010-2014 when we consider 40 pro-gay phrases and 40 anti-gay phrases (col. 1), when we consider only the 1000 most frequent bigrams within topical speeches (col. 2); when we consider the 2000 most frequent bigrams within topical speeches (col. 3); when we consider only 150000 most frequent bigrams within non topical speeches (col. 4); when we consider the 250000 most frequent bigrams within non topical speeches (col. 5), when we use the large set of keywords to define speeches as "topical" (col. 6); when we consider the smaller set of keywords (col. 7), when we consider as topical bigrams only bigrams that are not included in the 200000 most frequent bigrams in the non topical speeches (col. 8), when we consider reference bigrams chosen using the dynamic algorithm discussed in Section A1 (col. 9).Standard errors in parenthesis . * p < 0.10; ** p < 0.05; *** p ‹ 0.01

Table A13: Spatial Regression (Frequency Weights)

	(1) Pro Coverage	(2) Anti Coverage	(3) ratio	(4) Pro Coverage	(5) Anti Coverage	(6) ratio	(7) Pro Coverage	(8) Anti Coverage	(9) ratio
Spatial Lag	0.075***	0.229***	0.110***	0.063***	0.242***	0.076***	-0.041*	0.102**	0.004
	(0.014)	(0.039)	(0.011)	(0.0145)	(0.041)	(0.013)	(0.024)	(0.047)	(0.012)
county F.E.	NO	NO	NO	YES	YES	YES	NO	NO	NO
year F.E.	YES	YES	YES	YES	YES	YES	YES	YES	YES
state*year F.E.	NO	NO	NO	NO	NO	NO	YES	YES	YES
obs	5445	5445	5445	5445	5445	5445	5445	5445	5445

NOTE: Table A13 displays results for spatial models (command xsmle in Stata) using as dependent variable the coverage of pro-gay language (columns 1, 4 and 7), anti-gay language (columns 2, 5 and 8), ratio of the coverage of pro- gay language over the sum of pro- and anti-gay language (columns 3, 6 and 9). The coverage measures reflect the relevance of the phrases covered by the newspaper and their relevance is based on their frequency whitin topical speeches in Congress. Standard errors (clustered at county level) in parenthesis . * p < 0.10; ** p < 0.05; *** p < 0.01

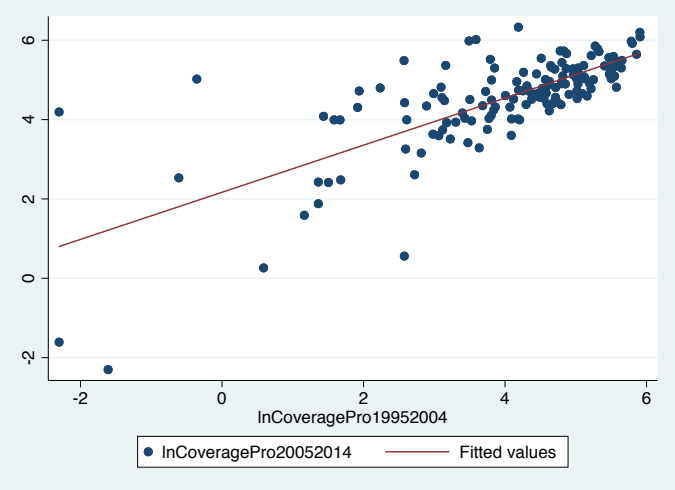
Table A14: Spatial Regression (Chi Weights)

	(1) Pro Coverage	(2) Anti Coverage	(3) ratio	(4) Pro Coverage	(5) Anti Coverage	(6) ratio	(7) Pro Coverage	(8) Anti Coverage	(9) ratio
Spatial Lag	0.074***	0.253***	0.115***	0.064***	0.271***	0.079***	-0.035	0.113**	0.014
	(0.013)	(0.038)	(0.011)	(0.015)	(0.040)	(0.013)	(0.021)	(0.048)	(0.012)
county F.E.	NO	NO	NO	YES	YES	YES	NO	NO	NO
year F.E.	YES	YES	YES	YES	YES	YES	YES	YES	YES
state*year F.E.	NO	NO	NO	NO	NO	NO	YES	YES	YES
obs	5445	5445	5445	5445	5445	5445	5445	5445	5445

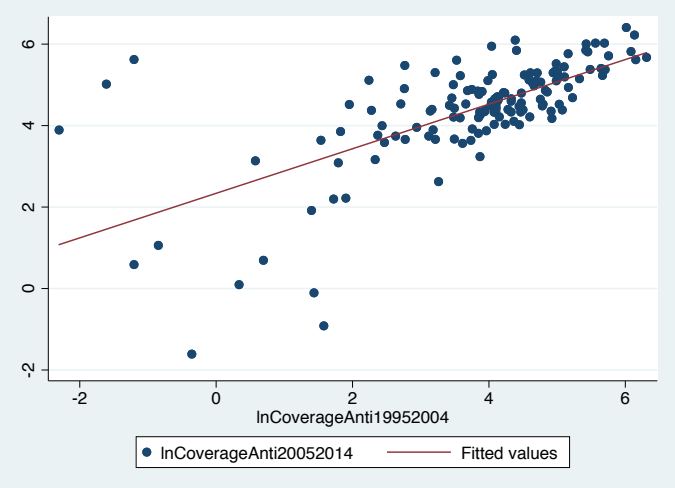
NOTE: Table A14 displays results for spatial models (command xsmle in Stata) using as dependent variable the coverage of pro-gay language (columns 1, 4 and 7), anti-gay language (columns 2, 5 and 8), ratio of the coverage of pro-gay language over the sum of pro- and anti-gay language (columns 3, 6 and 9). The coverage measures reflect the relevance of the phrases covered by the newspaper and their relevance is based on their Chi values, that proxy how strongly we can reject the hypothesis they are as likely to be used by pro-gay Congressmen than anti-gay Congressmen. Standard errors (clustered at county level) in parenthesis . * p < 0.10; ** p < 0.05; *** p < 0.01

Figure A10: Persistence (only newspaper digitized 1995-2014)

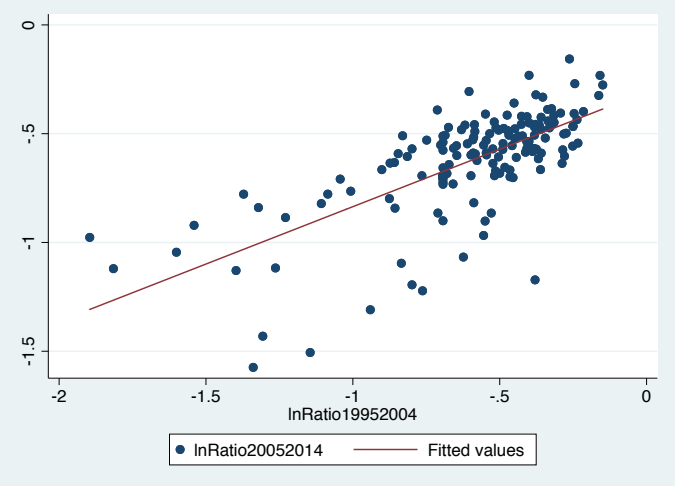
Panel A: Pro Coverage



Panel B: Anti Coverage



Panel C: Ratio



NOTE: We focus on two points in time: the decade from 2005 to 2014 and the decade from 1995 to 2004. We only consider newspapers that have been digitised for all the time period 1995-2014. For each newspaper we take a yearly average of the relevant coverage variable for both the time periods considered (1995-2004 and 2005-2014) and then we build a county measure of the coverage variable by taking the average of all the newspapers digitised within each county. Panel A displays a scatter plot with the log of the pro-gay coverage measure in 2005-2014 on the vertical axis against the log of the pro-gay coverage measure in 1995-2004 on the horizontal axis. Panel B displays a scatter plot with the log of the anti-gay coverage measure in 2005-2014 on the vertical axis against the log of the anti-gay coverage measure in 1995-2004 on the horizontal axis. Panel C displays a scatter plot with the log of ratio measure in 2005-2014 on the vertical axis against the log of the ratio measure in 1995-2004 on the horizontal axis.