Gender Norms Do Not Persist But Converge Across Time

Shikhar Singla (ssingla@london.edu)
London Business School
Mayukh Mukhopadhyay
London Business School https://orcid.org/0000-0003-1412-0540

Article

Keywords:

Posted Date: June 9th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1720800/v1

Gender Norms Do Not Persist But Converge Across Time

Shikhar Singla* Mayukh Mukhopadhyay*

May 13, 2022

Abstract

Gender equality is a fundamental human right, considered essential for the creation of a just society and a key ingredient of the United Nations Sustainable Development Goals. A key determinant of gender inequality in terms of the gender wage gap and other dimensions are the attitudes and stereotypes associated with women (Blau and Kahn (2017), Fernández (2007), Klasen (2002)). Here, we investigate the evolution of gender norms for 160 years in the US. Socioeconomists have posited two fundamental and widely debated theories on the evolution of cultural norms across time. One school argues that cultural norms (including gender norms) should converge across time (Marx (1867), Bell (1976), Lerner (1958), Inglehart (1997)) as economies become more advanced and globalised and technological progress allows for easier information sharing. The other school states that cultural traits are highly persistent, passed down from generation to generation and will remain divergent across regions (Weber (1904), Huntington (1993), Huntington (1996), DiMaggio (1994)). To test these theories quantitatively, researchers need data on attitudes measured at appropriate levels of granularity, at a high frequency and over a long time series. However, no such data currently exists. Here we propose an unsupervised machine learning methodology to measure attitudes at the document level. We apply this methodology to 193 million pages of local newspaper text to produce localised attitudes towards women on four different dimensions: career vs family, attitudes towards abortion, attitudes towards feminism/suffrage, and violence against women. We establish five novel facts on the evolution of attitudes across time with these measures. (i) Attitudes are less persistent than the existing literature hypothesises. (ii) The persistence of attitudes varies considerably across different regions and dimensions. (iii) Attitudes exhibit cyclical patterns. (iv) Regional variation in attitudes decreases considerably over time and has fallen between 64% to 79%. (v) A decrease in transport costs that allows for easier information sharing is associated with a homogenisation of the norms. Our paper calls for more research on whether similar patterns exist across countries and the causal factors that make culture converge.

*London Business School; Correspondence: ssingla@london.edu

1 Introduction

The UN's Sustainable Development Goals (SDG) state: "Gender equality is not only a fundamental human right but a necessary foundation for a peaceful, prosperous and sustainable world". A critical determinant of gender equality in outcomes are the attitudes towards and stereotypes associated with women. Prior research has established that negative attitudes and stereotypes have material economic and social consequences. For example, gender discrimination towards women leads to a larger gender wage gap, lower labour force participation, and lower economic growth and development (Blau and Kahn (2017), Fernández (2007), Fernández and Fogli (2009), Klasen (2002), Fernández (2013), Fogli and Veldkamp (2011).

Here we investigate the evolution of gender attitudes for 160 years in the United States. Socioeconomists have posited two fundamental theories on the evolution of cultural norms across time. One school of thought argues that cultural norms (including gender norms) should converge across time as economies become more advanced and globalised, the cost of sharing information across regions decreases, and political and economic institutions become more homogenous (Marx (1867), Bell (1976), Bell (1973), Lerner (1958), Inglehart (1997), Inglehart (1977), Inglehart (1990)). The alternative view is that cultural traits are persistent and differences between regions should therefore remain as cultural values are passed (relatively unchanged) from generation to generation (Weber (1904), Huntington (1993), Huntington (1996), DiMaggio (1994)).[1] Both theories have received widespread attention among sociologists, historians, political scientists, economists, and cultural anthropologists.

Although there is hardly any empirical work on the convergence theory, the persistence of cultural and gender norms has received some attention using narrow sub-components of norms in limited geographical areas. However, there is a tension in the literature. One set of papers using high frequency data for a short period of time finds that cultural and gender

[1]Please refer to Inglehart and Baker (2000) for a discussion of both theories. They test the two theories empirically using data from 65 countries, but the power of their tests is limited as their time series is only 16 years long.

norms can change quickly because of changes in economic conditions, technology or institutions (Gruber and Hungerman (2008), Fernández (2007), Alesina and Fuchs-Schündeln (2007), Di Tella et al. (2007), Giuliano and Spilimbergo (2014), Bowles (1998), Bursztyn et al. (2020)). On the other hand, other work (using lower frequency data over longer time periods) has found that these norms are persistent across long periods of time (Putnam (1993), Michalopoulos and Xue (2021), Alesina et al. (2013), Guiso et al. (2006), Fukuyama (1996), Guiso et al. (2016), Tabellini (2010), Durante (2009), Voigtländer and Voth (2012), Bazzi et al. (2020)).

The primary challenge to testing these theories of cultural evolution and resolving this debate is to measure attitudes over a long time series, at a high frequency, and at a regional level. Prior research typically measures attitudes with nationally representative surveys (such as the General Social Survey (GSS) and the American National Election Studies (ANES) in the US) that mask geographic variability across regions. To partially address these shortcomings, empirical work in the social sciences has relied on multilevel regression and poststratification (MRP) to attempt to generate county level variation from nationally representative survey data (Gelman and Little (1997), Howe et al. (2015)). MRP requires a large survey (for example, more than 10,000 respondents) (Howe et al. (2015)) to work accurately. However, no such survey exists in the US. In addition, the time series component is restricted by the availability of GSS and ANES data, which only go back as far as the 1960s. Therefore, testing these theories is infeasible with surveys.

Alternative methods such as dictionary analysis (Henley (1989)) and linguistic expertise in the use of language (Bussmann and Hellinger (2003)) require time-consuming manual analysis that does not easily scale across groups/topics as well as time and areas. To address these concerns, a recent literature in the computational social sciences has used word vectors to measure attitudes and beliefs over time (Garg et al. (2018)). Word vectors are likely to capture human biases and stereotypes if they are present (however subtly) in the training text (Caliskan et al. (2017)). For example, these embeddings might reflect bias by associating women with family and men with career-related words (Caliskan et al. (2017),

Bolukbasi et al. (2016)) and can therefore be used to study the evolution of stereotypes and beliefs across time (Garg et al. (2018)). While this method is scalable and does not require expensive and time-consuming manual analysis by experts, it produces only one number per trained model. To produce more granular measures with this existing method, a separate word embedding model needs to be trained at the desired level of granularity. This is infeasible as training increasingly fine-grained models will be constrained by the availability of a sufficiently large corpus.

To solve this, we develop an unsupervised machine learning approach to measuring attitudes towards groups/topics. Our methodology combines the word vectors estimation literature (Hamilton et al. (2016b); Mikolov et al. (2013a)), dictionary generation literature (Rothe et al. (2016), Hamilton et al. (2016a)), and the sentence vector estimation literature (Arora et al. (2017)). The word vectors literature shows how to estimate word vectors efficiently and create dictionaries of positive and negative words from the trained vectors. Starting with well-trained word vectors, the sentence vector literature focuses on estimating vectors at the sentence and higher levels of aggregation. We combine these distinct but related strands of the literature by training a Word2vec model (Mikolov et al. (2013a)) and use it to create a dictionary of positive and negative words. We then use the word vectors from our trained models to estimate document level embeddings using the Smooth-Inverse-Frequency approach (Arora et al. (2017)) and calculate the cosine similarity between the dictionary and document vectors to create a measure of attitudes. This allows us to break the corpus constraint and move beyond model level numbers to produce document level measures of attitudes.

We apply our method to measure attitudes relevant to the gender equality targets of the UN SDG at the county (or county-equivalent), state and US levels from over 193 million pages (over 298 billion words) of newspaper data from over 13,000 local US newspapers from 1850 to 2009. Specifically, we measure attitudes towards women on four dimensions: career vs family for women (whether women are more associated with their careers or their families), moral vs immoral for abortion (whether abortion is considered moral or immoral), equal-

ity vs extremism for feminism/suffrage (whether feminists and suffragettes are perceived as fighting for equality or as extremists) and violence towards women. Newspapers are an ideal source to measure these highly localised attitudes as readers exhibit a preference for newspapers that reflect their beliefs (Gentzkow and Shapiro (2010), Gentzkow et al. (2014)). In equilibrium, the views of the readers and newspapers match.

We validate our measures by first testing the performance of our word vectors on standard evaluation metrics (Levy et al. (2015), Jastrzebski et al. (2017)). Our model outperforms the state-of-the-art Corpus of Historical American English (COHA) model on 16/17 tasks. We further strengthen our validation with regressions using external data in the spirit of Garg et al. (2018). Specifically, we run regressions with survey data, the Democratic party's vote share, labour force participation, educational attainment, per capita income, and crime data.

Using our measures, we establish five novel facts on the evolution of gender norms across time. To test whether gender norms remain persistent, we estimate autoregressive regressions using our measures. Using these regressions, we first show that attitudes are not persistent. The autoregressive coefficients in our regressions drop to 0.2 between 15 to 75 years for our four measures. The R-squared values of these regressions also decline substantially at horizons of 40 lags and remain below 20% thereafter. This suggests that past values do not explain a large share of the variation in current attitudes. Therefore, attitudes are not as persistent as the prior literature has argued.

Our paper also explains the tension in the literature between papers which find that cultural norms can change quickly and those which find culture moves very slowly. The first set of papers uses high frequency data on cultural norms for a short period of time. However, as historical data on cultural norms are unavailable over longer time periods, the second set relies on exogenous shocks to attitudes (whose magnitude cannot be precisely measured) and tests their relationship with current attitudes. Our data's high frequency and long time series allow us to explain this tension. We find that norms can change quickly (which supports the first set of papers) and that the findings of the second set can be explained by

a significant correlation in attitudes (even after a century), although accompanied by low coefficients and R-squareds. Second, the persistence of attitudes varies significantly across regions and dimensions. Third, attitudes exhibit cyclical variation. An improvement in attitudes is followed by a decrease/reversal in subsequent periods. The persistence coefficient oscillates from positive to negative (and vice versa) across time.

Fourth, we find strong evidence that attitudes towards women converge across regions rather than persist. Regional variation in attitudes as measured by the interquartile range, standard deviation, and mean absolute deviation has declined consistently across time. The interquartile range has declined between 64% to 79% from the first decade to the last decade of our sample for our four measures.

Fifth, we provide evidence that reductions in transport costs between regions, which allows information to flow easily across these regions, is correlated with a homogenisation in cultural attitudes. Specifically, we use data from Donaldson and Hornbeck (2016) who provide estimates of the lowest transport costs between county pairs during the 19th and 20th centuries. Building a railroad or waterway connecting two counties leads to a reduction in transport costs across these counties. Using this data, we show that a decrease in transport costs is associated with a convergence in gender norms between county pairs.

2 Results

Theories of the cultural evolution of attitudes towards women can be broadly classified into two schools of thought. One school, proposed by sociologists such as Karl Marx and Daniel Bell, has argued that culture should converge across regions and civilisations as improving economic conditions and globalisation lead to cultural homogeneity across regions over time. The other school of thought, advanced by Max Weber and Samuel Huntington (among others), argues that cultural values are an enduring and persistent influence on society as they are passed from generation to generation and are therefore unlikely to converge across time

(Inglehart and Baker (2000)).

To test this hypothesis, we construct measures of attitudes towards women on four dimensions (based on the gender equality component of the UN SDG) from 1850 to 2009: career vs family for women (whether women are more associated with their careers or their families), moral vs immoral for abortion (whether abortion is considered moral or immoral), equality vs extremism for feminism/suffrage (whether feminists and suffragettes are perceived as fighting for equality or as extremists) and violence towards women. We produce our measures of each newspaper page and aggregate them up to the county, state or country level for different parts of our analysis. Section 5 describes our methodology in detail.

2.1 Persistence, Variation in Persistence, and Cyclicality

The persistence theory posits that the culture or gender norms do not change much over time. Statistically, a time series variable is said to be persistent if past values of the variable strongly predict the current values. This suggests that running autoregressions and testing their coefficients are a natural test of this theory.

We run regressions of attitudes on their lagged values using the following regression specification:

$$Y_{it} = \alpha + \beta Y_{it-p} + \epsilon_{it} \tag{1}$$

where Y is the measure, i is the state, t is the time, and p is the lag. ϵ_{it} is the error term. We use robust standard errors.[2] We run 150 such regressions (for p = 1,2,...,150) for each of our four measures. Figures 1 and 2 show the results.

Depending on the measure, the autoregressive coefficient drops to 0.2 in just 15 to 75 years. The R-squareds for these regressions remain below 20% at horizon of 40 lags or higher, which suggests that past attitudes do not explain a large share of the variation in our mea-

[2]The results are very similar if we use Newey-West standard errors (Newey and West (1987)) and are presented as a robustness check in Section 3.1.

sures (Figure 2). These results imply that gender norms are not as persistent as previously thought. However, there are clearly visible differences in the persistence of attitudes across our four measures.

Each of the measures exhibits cyclical variation across time. This is visible in Figure 1 as the regression coefficients oscillate from being positive and statistically significant to statistically insignificant, and finally to negative and statistically significant.

We then explore the variation in our measure across states. We rerun the analysis in figures 1 and 2, separately for each state using the following regression specification:

$$Y_t = \alpha + \beta Y_{t-p} + \epsilon_t \tag{2}$$

where Y is the measure, t is the time, and p is the lag. ϵ_t is the error term. We use robust standard errors. Figures 3 and 4 are box plots of the autoregressive coefficients and the R-squared values of these regressions using the distribution across all states. These results show that there is substantial variation in persistence. Patterns of cyclicality are visible in these plots as well.

Prior work has found that gender norms (Alesina et al. (2013), Michalopoulos and Xue (2021)) and culture more broadly (Putnam (1993), Guiso et al. (2006), Fukuyama (1996), Guiso et al. (2016), Tabellini (2010), Voigtländer and Voth (2012), Bazzi et al. (2020)) are persistent across time. These papers have at most two data points (but typically one): exogenous shocks to attitudes (which cannot be measured precisely) or historical culture (usually from 500 or more years ago) and culture today which they show is still affected by the past.

Our findings show that although the autoregressive coefficient can be statistically greater than zero for more than a century, their magnitudes are very small. This implies that historical culture only has weak predictability for culture today. The cyclicality that we document

is particularly important. As Figures 1 and 3 show, there is significant variation in the autoregressive coefficients across time. This means that a short time series or a limited number of data points could severely distort any conclusions on the persistence of norms across time. For example, if career vs family norms were measured 90 years apart, they would not appear to be persistent as the autoregressive coefficient is statistically insignificant. If they were measured 110 years apart, there would be a negative association which would change to positive if the horizon extends to 130 years. This highlights the importance of having a long and high frequency time series on cultural values to be able to analyse the persistence of culture.

Our paper also explains the inconsistency between papers which find that cultural norms can change quickly because of changes in economic conditions, technology or institutions (Gruber and Hungerman (2008), Fernández (2007), Alesina and Fuchs-Schündeln (2007), Di Tella et al. (2007), Giuliano and Spilimbergo (2014), Bowles (1998), Bursztyn et al. (2020)) and papers which find culture moves very slowly (Putnam (1993), Michalopoulos and Xue (2021), Alesina et al. (2013), Guiso et al. (2006), Fukuyama (1996), Guiso et al. (2016), Tabellini (2010), Voigtländer and Voth (2012), Bazzi et al. (2020)). The first set of papers uses high frequency data on cultural norms for a short period of time. However, as historical data on cultural norms are unavailable, the second set relies on exogenous shocks to attitudes (whose magnitude cannot be precisely measured) to test their relationship with current attitudes. The long time series and high frequency of our data allow us to explain the inconsistency between these findings. We find that norms can change quickly (which supports the first set of papers) and that the findings of the second set can be explained by a significant correlation even after a century, although accompanied by low coefficients and R-squareds.

2.2 Convergence Among States

A lack of persistence in gender norms does not imply that these norms converge across regions. This just means that there is significant variation in the measures across time but does not tell us whether this variation leads to a convergence or divergence in gender norms.

To test whether norms converge across states, we plot the interquartile range (a measure of dispersion) for each of our measures across states (Figure 5). On this metric, the dispersion in gender norms has consistently declined across time and across all measures. Specifically, the average dispersion declines (from the first decade to the last decade of the sample) by 72.5, 71.1, 63.9 and 78.8 percent for career vs family, feminism/suffrage, abortion and violence, respectively.

To formally test this, we calculate the Pearson correlation between the interquartile range and a year variable and run a regression between the two to test for any declining time trends in dispersion using the following regression specification:

$$dispersion_t = \alpha + \beta t + \epsilon_t \tag{3}$$

where $dispersion$ is the dispersion of each measure calculated using the interquartile range, t is year, ϵ_t is the error term. We use robust standard errors.

Figure 5 plots the line of best fit for this regression and the correlation with time for each of our four measures. For each of these measures, the coefficient of the regression and the correlation between dispersion and time are highly negative and statistically significant. These results imply that attitudes are less dispersed across time, which supports the cultural convergence hypothesis. These results also hold using two alternative measures of dispersion: the mean absolute deviation and standard deviation (Figures 6 and 7).

An implication of this convergence result is that gender norms should be more correlated across states over time. To investigate this, we divide our sample into three equal time periods (1850-1902, 1903-1955, 1956-2009). We plot the heatmap of pairwise Pearson correlations (Figure 8) and find that they are higher in the last period compared to the first period for all four measures.

To more formally test this, we use a Kolmogorov–Smirnov test, which tests for the equal-

ity of two distributions. We apply this test to every state pair in our sample across each of our three sub-periods. Figure 9 plots the proportion of state pairs where we fail to reject the null hypothesis of identical distributions. For each of our four measures, this proportion is steadily increasing across time.

2.3 Transportation Costs

What mechanisms could drive the convergence in culture across time? This could be due to many potential reasons which can be categorised into there major groups. First, economic globalisation could lead to economic cycles becoming more similar across regions as they are increasingly interlinked through goods and services trade as well as financial flows. Second, an exchange of views, either through immigration across regions facilitated by more efficient methods of transport or through modern forms of communication technology, could lead to a greater mixing of views and more homogeneity. Third, economic globalisation could drive economic and political institutions to become more similar across regions as common standards are helpful to facilitate economic integration.

Each of these potential mechanisms depends on the cost of sharing information across regions. A reduction in these costs should therefore be correlated with a convergence in norms across regions. To test this hypothesis, we use data from Donaldson and Hornbeck (2016) who calculate the lowest transport costs from 1850 to 1920 using railroads and waterways between each county pair in the US for each decade. This period was a time of major expansion of railroads and waterways across the US. As more railroads and waterways were constructed, transportation costs decreased.

Specifically, we run the following regression:

$$ln(absolute\ difference)_{it} = \alpha_t + \alpha_i + \beta ln(costs)_{it} + \epsilon_{it} \tag{4}$$

where $ln(absolute\ difference)_{it}$ represents the log of the absolute difference between the

measures in each county pair, $ln(costs)_{it}$ represents log of costs from Donaldson and Hornbeck (2016), i denotes the county pair, t denotes the year, α_t is the time fixed effect, and α_i is the county pair fixed effect. ϵ_{it} is the error term. We use robust standard errors. Since costs are only available decade-wise, we assume there are no changes in the costs between decades.

We find that a decrease in transport costs between county pairs is correlated with a smaller absolute difference in gender norms between these county pairs (positive coefficient between costs and absolute deviation). The results, using different regression specifications, are presented in Figure 10. This suggests that a substantial decline in transportation costs due to the enormous expansion in rail and waterway networks provides one of the explanations for the convergence in gender norms. The convergence in views in our broader sample might therefore be explained by the steady decrease in the costs of communicating and commuting over the last two centuries.

3 Robustness

3.1 Presence of Unit Roots

Statistical inference in time series models with unit roots is problematic. Although it is visible from our results on persistence that our measures do not possess unit roots, we also formally test whether this is the case. To do this, we use the Augmented Dicky-Fuller (ADF) test to find states that exhibit unit roots (we use a 10% significance level) and use Newey-West standard errors (Newey and West (1987)) to account for heteroskedasticity and autocorrelation. We then rerun our analysis excluding these states and find very similar results (Figures 11 and 12).

3.2 Migration

The US has experienced a large amount of external and internal (across states) migration in our sample period. This means that the convergence in cultural norms could simply be

driven by a migration of people with opposite views rather than by a generalised convergence in the views and attitudes of the majority of people in the country. This could also lead to the low persistence result as immigration leads to fluctuation in the average attitudes across years.

To rule this out, we use IPUMS census data and obtain the share of the local-born population in each state and use only those state-year pairs where at least 70% of the population is born in the same state. Our results are shown in Figures 13 and 14 and are similar to the main results.

3.3 Errors in the Measures

Measurement error could lead to a rejection of the null hypothesis of persistence due to the attenuation bias generated in an OLS regression. To ensure that our results are not driven by measurement error, we rerun our analysis by aggregating our measure across the whole of the US, where measurement error is likely to be significantly lower. This analysis also results in very similar patterns (Figure 15).

3.4 Analysis at the County Level

We test the theories of convergence and persistence using state level measures because they offer greater coverage across the US than county level measures. However, the lowest transportation costs data is available at the county level (Donaldson and Hornbeck (2016)). Therefore, we rerun our analysis of convergence and persistence at the county level to rule out the possibility that our results are driven by the choice of geographic aggregation and find very similar results (Figures 16 and 17).

4 Discussion

We are the first to provide evidence that gender norms, which is a component of cultural norms more broadly, can become similar over time across regions and is less persistent than

previously thought. This resolves a long-standing debate among scholars over the evolution of cultural norms across time. We also provide an answer to the inconsistency in the persistence literature where some papers find that cultural norms change fairly rapidly whilst others argue that they remain persistent over centuries.

Our methodology is also an inexpensive way to track granular attitudes associated with the UN SDG, which are a vital ingredient in the success of the UN's initiatives. By tracking attitudes across time and highly localised regions, our results can guide intervention on the geographic level and time frame of appropriate UN SDG policy.

Our paper also opens up additional questions on the persistence and convergence theories, such as testing whether these patterns hold across countries, across other components of culture, and the causal factors behind each of the facts that we document.

References

Alesina, A. and N. Fuchs-Schündeln (2007): "Good-bye Lenin (or not?): The effect of communism on people's preferences," *American Economic Review*, 97, 1507–1528.

Alesina, A., P. Giuliano, and N. Nunn (2013): "On the origins of gender roles: Women and the plough," *The quarterly journal of economics*, 128, 469–530.

Arora, S., Y. Liang, and T. Ma (2017): "A Simple but Tough-to-Beat Baseline for Sentence Embeddings," *International Conference on Learning Representations*.

Bazzi, S., M. Fiszbein, and M. Gebresilasse (2020): "Frontier culture: The roots and persistence of "rugged individualism" in the United States," *Econometrica*, 88, 2329–2368.

Bell, D. (1973): *The coming of the post-industrial society*, Basic Books.

——— (1976): *The cultural contradictions of capitalism*, Basic Books.

Blau, F. D. and L. M. Kahn (2017): "The gender wage gap: Extent, trends, and explanations," *Journal of economic literature*, 55, 789–865.

Bolukbasi, T., K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai (2016): "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," *Advances in neural information processing systems*, 29, 4349–4357.

Bowles, S. (1998): "Endogenous preferences: The cultural consequences of markets and other economic institutions," *Journal of economic literature*, 36, 75–111.

Bursztyn, L., G. Egorov, and S. Fiorin (2020): "From extreme to mainstream: The erosion of social norms," *American economic review*, 110, 3522–48.

Bussmann, H. and M. Hellinger (2003): *Gender across languages: The linguistic representation of women and men*, J. Benjamins.

Caliskan, A., J. J. Bryson, and A. Narayanan (2017): "Semantics derived automatically from language corpora contain human-like biases," *Science*, 356, 183–186.

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

DiMaggio, P. (1994): "Culture and economy," in *Handbook of economic sociology*, Princeton University Press and Russell Sage, 27–57.

Donaldson, D. and R. Hornbeck (2016): "Railroads and American economic growth: A "market access" approach," *The Quarterly Journal of Economics*, 131, 799–858.

Durante, R. (2009): "Risk, cooperation and the economic origins of social trust: an empirical investigation," *Available at SSRN 1576774*.

Eckert, F., A. Gvirtz, J. Liang, and M. Peters (2020): "A Method to Construct Geographical Crosswalks with an Application to US Counties since 1790," Tech. rep., National Bureau of Economic Research.

Fernández, R. (2007): "Women, work, and culture," *Journal of the European Economic Association*, 5, 305–332.

——— (2013): "Cultural change as learning: The evolution of female labor force participation over a century," *American Economic Review*, 103, 472–500.

Fernández, R. and A. Fogli (2009): "Culture: An empirical investigation of beliefs, work, and fertility," *American economic journal: Macroeconomics*, 1, 146–77.

Firth, J. R. (1957): "A synopsis of linguistic theory, 1930-1955," *Studies in linguistic analysis*.

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

Fukuyama, F. (1996): *Trust: The social virtues and the creation of prosperity*, Simon and Schuster.

Garg, N., L. Schiebinger, D. Jurafsky, and J. Zou (2018): "Word embeddings quantify 100 years of gender and ethnic stereotypes," *Proceedings of the National Academy of Sciences*, 115, E3635–E3644.

Gelman, A. and T. C. Little (1997): "Poststratification Into Many Categories Using Hierarchical Logistic Regression," .

Gentzkow, M. and J. M. Shapiro (2010): "What drives media slant? Evidence from US daily newspapers," *Econometrica*, 78, 35–71.

Gentzkow, M., J. M. Shapiro, and M. Sinkinson (2014): "Competition and ideological diversity: Historical evidence from us newspapers," *American Economic Review*, 104, 3073–3114.

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

Gruber, J. and D. M. Hungerman (2008): "The church versus the mall: What happens when religion faces increased secular competition?" *The Quarterly journal of economics*, 123, 831–862.

Guiso, L., P. Sapienza, and L. Zingales (2006): "Does culture affect economic outcomes?" *Journal of Economic Perspectives*.

——— (2016): "Long-term persistence," *Journal of the European Economic Association*, 14, 1401–1436.

Hamilton, W. L., K. Clark, J. Leskovec, and D. Jurafsky (2016a): "Inducing domain-specific sentiment lexicons from unlabeled corpora," in *Proceedings of the conference on empirical methods in natural language processing. conference on empirical methods in natural language processing*, NIH Public Access, vol. 2016, 595.

Hamilton, W. L., J. Leskovec, and D. Jurafsky (2016b): "Diachronic word embeddings reveal statistical laws of semantic change," *arXiv preprint arXiv:1605.09096*.

Henley, N. M. (1989): "Molehill or mountain? What we know and don't know about sex bias in language," in *Gender and thought: Psychological perspectives*, Springer, 59–78.

Howe, P. D., M. Mildenberger, J. R. Marlon, and A. Leiserowitz (2015): "Geographic variation in opinions on climate change at state and local scales in the USA," *Nature climate change*, 5, 596–603.

Huntington, S. P. (1993): "The clash of civilizations?" *Foreign Affairs*, 72, 22–49.

——— (1996): *The Clash of Civilizations and the Remaking of World Order*, Simon & Schuster.

Inglehart, R. (1977): "The silent revolution," in *The Silent Revolution*, Princeton University Press.

——— (1990): *Culture shift in advanced industrial society*, Princeton University Press.

——— (1997): *Modernization and postmodernization: Cultural, economic, and political change in 43 societies*, Princeton university press.

Inglehart, R. and W. E. Baker (2000): "Modernization, cultural change, and the persistence of traditional values," *American sociological review*, 19–51.

Jastrzebski, S., D. Leśniak, and W. M. Czarnecki (2017): "How to evaluate word embeddings? on importance of data efficiency and simple supervised tasks," *arXiv preprint arXiv:1702.02170*.

Kaplan, J. (2021): "Uniform Crime Reporting (UCR) program data: A practitioner's guide," *CrimRxiv*.

Klasen, S. (2002): "Low schooling for girls, slower growth for all? Cross-country evidence on the effect of gender inequality in education on economic development," *The World Bank Economic Review*, 16, 345–373.

Lerner, D. (1958): *The passing of traditional society: Modernizing the Middle East*, Free Press.

Levy, O., Y. Goldberg, and I. Dagan (2015): "Improving distributional similarity with lessons learned from word embeddings," *Transactions of the association for computational linguistics*, 3, 211–225.

Marx, K. (1867): *Capital: A critique of political economy.*

Michalopoulos, S. and M. M. Xue (2021): "Folklore," *The Quarterly Journal of Economics*, 136, 1993–2046.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013a): "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013b): "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 3111–3119.

Newey, W. K. and K. D. West (1987): "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

Putnam, R. D. (1993): *Making democracy work*, Princeton university press.

Rothe, S., S. Ebert, and H. Schütze (2016): "Ultradense word embeddings by orthogonal transformation," *arXiv preprint arXiv:1602.07572*.

Tabellini, G. (2010): "Culture and institutions: economic development in the regions of Europe," *Journal of the European Economic association*, 8, 677–716.

Vella, F. (1994): "Gender Roles and Human Capital Investment: The Relationship between Traditional Attitudes and Female Labour Market Performance," *Economica*, 61, 191–211.

Voigtländer, N. and H.-J. Voth (2012): "Persecution perpetuated: the medieval origins of anti-Semitic violence in Nazi Germany," *The Quarterly Journal of Economics*, 127, 1339–1392.

Weber, M. (1904): *The Protestant ethic and the spirit of capitalism.*

Wilson, J. C. (2020): "Striving to Rollback or Protect Roe: State Legislation and the Trump-Era Politics of Abortion," *Publius: The Journal of Federalism*, 50, 370–397.

5 Methodology

One way to measure attitudes at the document level could involve a supervised machine learning approach, where training documents are labelled. This approach has two limitations. First, in the context of a historical study like ours, the training documents must be labelled by experts with considerable knowledge of the nature and circumstances of women for 160 years of US history, as well as semantic displacement in the use of language that refers to women. This makes obtaining these training documents or generating them through online platforms such as Amazon Mechanical Turk unfeasible. Even if training documents could be generated, the sheer number needed to effectively capture the nuances of the language across 160 years of US history would make this approach prohibitively expensive. In addition, for the classifier to work accurately, the training documents need to be very similar to the projection documents, which would make it harder to employ such a method for broader applications. Second, even the most fine-grained supervised machine learning is typically limited to a small number of classification categories and therefore cannot produce continuous measures of sentiment. To overcome these challenges, we propose a simple, multi-step unsupervised learning approach. Our machine learning pipeline involves three main steps.

5.1 Measures

The initial stage of our analysis (which precedes our machine learning pipeline) is to choose which attitudinal components to measure. We refer to the gender equality targets of the UN SDG which are relevant to the US to choose these components. In particular, SDG goal 5 aims to "Achieve gender equality and empower all women and girls."[3] We translate these targets into the following measurable components:

- Career vs Family for women - Targets 5.4 and 5.5. The association with career or family words is a popular measure of gender inequality in the literature (Caliskan et al. (2017)) and is related to these targets.

[3]Targets within this goal can be found at https://www.un.org/sustainabledevelopment/gender-equality/. Targets with numbers are "Outcome targets" (5.1-5.6), whereas lower case letters are "means of implementation targets" (5.a-5.c). We focus on outcome targets since means of implementation targets focus on the adoption of technology or increased financial resources for the outcome targets to be achieved.

- Moral vs Immoral for abortion - Target 5.6. Abortion has been an intensely debated topic in the US and has been used in prior work as well (Inglehart and Baker (2000)). The legal status of abortion is an important component of Target 5.6.[4]

- Equality vs Extremism for feminism and suffrage - Target 5.1. We choose feminism since it aims to end all forms of discrimination against women, which is the ultimate aim of Target 5.1. Universal suffrage is also an important component of Target 5.1 and was a prominent movement in the US.

- Violence towards women - Targets 5.2 and 5.3. Violence does not have any inherent opposing dimensions, so this simply measures violent language towards women. Although violence does not have sentiment, it does reflect attitudes towards violence through the news/stories in newspapers.

5.2 Word2vec

In the first step of our methodology, we train word embeddings using Word2vec (Mikolov et al. (2013a)). Word embeddings algorithms like Word2vec convert words to numbers by producing a vector representation of each word in the corpus in a way that captures their meaning effectively. Word2vec is based on the "distributional hypothesis" in linguistics (Firth (1957)), the idea that words that appear together in a sentence have a similar meaning. Mikolov et al. (2013a) show that the specific shallow neural network architecture used in Word2vec produces word vectors that capture meaning in an efficient way. We use the Skip-Gram implementation of Word2vec (Mikolov et al. (2013a) and Mikolov et al. (2013b)). This method uses context words in a fixed window to learn the 300 X 1 dimensional word embeddings of the centre words.

To train our word vectors, we collected and processed 193 million local newspaper pages published between 1850 and 2009 in the US. Our corpus contains approximately 298 billion words and is almost twice as large as the American version of the Google Books corpus,

[4]Detailed discussion on Target 5.6 can be found here: https://www.unfpa.org/sites/default/files/pub-pdf/UNFPA-SDG561562Combined-v4.15.pdf

which covers roughly 6% of all published books across history (Hamilton et al. (2016b)). Please refer to Sections 8.1 and 8.2 for more details on data collection and processing.

In the spirit of the COHA embeddings, we train 16 Word2vec models for each decade from 1850 to 2009 (Hamilton et al. (2016b)). This allows us to detect changes in word associations and, by extension, in the use of language across time. Due to computational constraints, we select a random subset of sentences until the training text reaches 1 billion words. We train our models on this 1 billion word corpus for each decade. Section 8.3 provides details on the Word2vec hyperparameters.

5.3 Dictionary Generation

We use the Word2vec models to create dictionaries from the top 100 most associated words (measured with cosine similarity) with the target words of our measures. The target words that our dictionaries are based on are a list of words related to each measure. For the career vs family and violence towards women measures, the set of target words is a list of synonyms for the word "woman" from the historical thesaurus of the Oxford English Dictionary (OED). For abortion and feminism/suffrage, our target words are a shorter list of synonyms that we select based on our judgement. Section 8.4 lists our target words.

One way to assign sentiment could involve calculating the cosine similarities between the target words and the sentiment related seed words, directly following Caliskan et al. (2017). This works effectively in their setting as they are interested in model level numbers and therefore require average associations between the target and seed words. However, this approach is inappropriate in our setting. To see this, let us consider an example where there is only one target word, "woman", with a higher association with family rather than career words. This means that any sentence referring to women will always be weighted by a score tilted towards family for "woman" irrespective of whether the other words in the sentence are related to family or career. This means that target words will not be enough, so we need a dictionary of associated words.

To measure overall attitudes, we need to classify the sentiment of words in our dictionary as positive or negative. In this context, positive and negative simply denote two opposite dimensions to classify words along. We, therefore, use the words positive or negative to represent the two dimensions of our measures. For the career vs family, abortion, and feminism/suffrage measures, the positive dimensions are career, moral, and equality, respectively. The other dimensions are negative.

We rely on the dictionary generation literature using word vectors to do this. The two main methods in this literature are DENSIFIER (Rothe et al. (2016)) and SentProp (Hamilton et al. (2016a)). We use DENSIFIER, which produces a sentiment dictionary of positive and negative words, starting from a few initial seed words. DENSIFIER (Rothe et al. (2016)) works by applying an orthogonal transformation to the embedding space to generate an ultradense embedding for each word. This ultradense embedding represents the polarity of each word on the chosen dimensions. The algorithm produces a continuous polarity score for each word in the corpus which signifies its association with the positive and negative dimensions. The higher (lower) this number is, the more positive (negative) the word is. This method achieves state-of-the-art performance in creating lexicons of words that capture sentiment well. We use DENSIFIER instead of SentProp as it is less sensitive to pre-processing and performs better with large vocabularies. As an input to the algorithm, we hand-curated a list of positive and negative seed words for each of our attitude dimensions which are listed in Section 8.5. We follow Caliskan et al. (2017) for seed words for our career vs family measure. We use a list of synonyms along the positive and negative dimensions for our other measures. Section 7 describes a robustness test for this modelling choice using SentProp.

Associated words change by decade, so for them to be comparable, we need a constant positive or negative definition for each word across decades. This is important as some words which were neutral in the past might have negative connotations today. This problem can be solved by using a constant valence for each word. This is generally done through lexicons. This is inappropriate in our setup as lexicons contain a fixed number of words and do not provide good coverage for all relevant words. To overcome this problem, we classify each

associated word as positive or negative and obtain their polarities using the DENSIFIER method with the seed words from the 1990-1999 model. We normalise the polarity scores to have a mean of zero and a standard deviation of one.

To construct our dictionary, we only keep words that have a polarity of 2 standard deviations away from the mean ($polarity \notin (-2, 2)$). We intersect these words with the 100 most associated words for each of our target words and include only these in our dictionary. This gives us a dictionary of words with high polarity scores. Finally, we add all the positive word vectors and subtract the negative ones multiplied by their absolute polarity scores to create a vector for the dictionary.

To create a dictionary for the violence measure (which is an association measure and does not have a sentiment dimension), we cannot use DENSIFIER. Instead, we simply calculate the average cosine similarity between the most associated words and a larger set of seed words (provided in Section 8.5). We normalise these cosine similarities and keep the words with a polarity score of more than 2. Then, we add the word vectors multiplied by their polarity scores to generate the dictionary vector.

5.4 Page Level Vectors

To produce granular variation in sentiment at the document level, our method needs to go beyond Word-Embedding-Association-Test (WEAT) based metrics which produce model level numbers by averaging across the entire corpus (Caliskan et al. (2017), Garg et al. (2018)). We solve this problem by combining our dictionaries with document level embeddings.

To do this, we follow the approach of Arora et al. (2017), who propose the Smooth Inverse-Frequency (SIF) method, and show that it outperforms sophisticated supervised methods, including RNNs and LSTMs by around 10-30% in textual similarity tasks. The method involves simply removing the first principal component (PC) of the weighted average of the word vectors. The sentence vectors are composed of the weighted average of the word vectors. The weights depend on two parameters: a constant and a word frequency probability.

Each sentence vector is arranged in a matrix and the principal component is calculated. We select a random subset of sentences until the training text reaches 100 million words to train our SIF models. We use these randomly selected sentences to train a SIF model for each decade. We then subtract the first Principal Component calculated from the above models from the document vector (each newspaper page) to obtain a page embedding. We obtain the initial page embeddings by simply adding the vectors of each sentence in the page (Arora et al. (2017)).

5.5 Calculation of the Measures

We only keep newspaper pages that contain at least one of our target words for each measure. This ensures that our measure captures sentiment associated with our topics and not other unrelated topics. We calculate the angular distance between the dictionary vector and the SIF vector for each page to create our measures. We do not use cosine similarity as it is not a linear function and therefore does not have constant marginal effects. We compute the cosine similarity of the two vectors and use arccos to convert the cosine similarity into an angular distance.

Our newspaper data is at the city level, but the economic data for our validation regressions is only available at the county level. To account for this, we aggregate the measure to the county level (weighting each page by the number of words) for each year. County boundaries have changed substantially over the last two centuries, so we map each city to its 1990 county to keep a consistent county definition. Since our measures do not have comparison groups like the WEAT measures (Caliskan et al. (2017)) and it is not always possible to find such groups, we construct a synthetic comparison group. To do this, we create a dictionary of words with high polarity ($polarity \notin (-2, 2)$) and repeat the steps outlined above to create an angular distance measure for all pages. We aggregate this measure for all pages to the county level for each year. We subtract this from the measure obtained using only pages relevant to our topic to produce our final attitudes measure. Therefore, our measure is the difference between sentiment towards our topic and sentiment towards all topics.

Since it is an angular distance measure, an increase in the measures implies a more negative attitude, i.e. a lower association of women with career rather than family words, a lower association of abortion with moral rather than immoral terms, and a lower association of feminism/suffrage with equality rather than extremism. However, for the violence measure, an increase in the measure implies a lower association of women with violence related words which denotes an improvement in this particular measure. We also aggregate our measures at the state (used for primary analysis in the main text because of greater coverage) and national level using the above steps. Since we validate our measures at the county level, higher levels of aggregation can also be performed. Figure 18 shows the county coverage of our measures.

6 Methodology Validation

6.1 Model Validation

We validate the performance of our Word2vec model using the Jastrzebski et al. (2017) framework, which provides 17 tests across similarity, analogy, and categorisation tasks to evaluate performance. Table 1 shows the results. Our model performs better than the COHA model (used by Garg et al. (2018)) for the 1990 decade in 16 out of 17 tasks.

6.2 External Validation

Although our method relies only on accurately estimated word vectors, we follow Garg et al. (2018) and further validate all our measures by testing whether they track externally verifiable surveys, census data, political voting shares and crime data. All of our regressions are at the county level. We weight our validation variables using the geo-referencing procedure described in Section 8.6. In each of our regression specifications, we use time fixed effects and robust standard errors. Time fixed effects remove time-specific variation. This leaves us with variation at the county level for both our measures and the dependent variables.

Specifically, we estimate the following specification:

$$Y_{it} = \alpha_t + X_{it} + \epsilon_{it} \tag{5}$$

where Y represents the external variable, X represents our measure, i denotes the county, t denotes the year and α_t is the time fixed effect. ϵ_{it} is the error term. Statistically significant coefficients in our validation regressions will therefore imply that our measure captures variation at the county level.

6.2.1 Survey Data

We first use survey data from the ANES database to validate our attitude measures for women. The survey we use covers the period 1970-2000.

The survey question (VCF0834) asks (with small variations across years) respondents to rank on a 7 point scale whether they feel that women should have an equal role with men in running business, industry and government or whether they feel that a woman's place is in the home. 1 represents an equal role and 7 represents the view that a woman's place is in the home. We average the responses in a particular county and run a regression of the county level mean response on our attitude towards the career vs family measure. Column (1) of Table 2 shows the results. The coefficient is positive and statistically significant at the 5% level, which implies that a higher value of our measure (more association of women with family than career) predicts an increase in the average score on the survey question (survey respondents feel that a woman's place is in the home to a larger extent). The specification includes year fixed effects. That the coefficient is still statistically significant suggests that our measure is able to capture variation within counties.

We repeat the same exercise for our measures of attitudes towards abortion and feminism/suffrage using questions VCF0838 and VCF0225 in the ANES data, respectively. VCF0838 asks the respondents to rate on a 4 point scale whether they think if, by law, abortion should never be permitted, which is represented by 1 on this scale or if, by law, a

woman should always be able to obtain an abortion as a matter of personal choice which is represented by 4. An increase in our measure (higher association of abortion with immoral terms than moral terms) predicts (Table 2, column 7) a lower average score on the survey (respondents feel abortion should never be permitted, to a greater extent). The coefficient is statistically significant at the 5% level.

VCF0225 asks respondents to rate their feeling towards the women's liberation movement on a scale of 0 to 100. The higher the rating, the more favourable the respondent's attitude towards the movement is. Our measure on feminism/suffrage predicts the average rating score in a county (Table 2, column 9). The relationship is statistically significant at the 1% level.

6.2.2 Labour Force Participation Data

Our second set of validations uses economic data. We use historical data on labour force participation from the Integrated Public Use Microdata Series (IPUMS). County level data on labour force participation is available starting from 1940 for decennial census waves until 2000 (except for 1960). There was a change in the definition of the labour force in 1950, so we perform two regressions, one from 1940-1950 and one from 1970-2000. Table 2, columns 2 and 3 report the results for the 1970-2000 and 1940-50 data, respectively. The coefficients are positive and statistically significant at the 1% level. A higher association of women with career than family is correlated with an increase in the labour force participation rate. This evidence is consistent with Vella (1994), who finds that more traditional gender roles of women towards family are associated with lower human capital investment and female labour supply.

6.2.3 Educational Attainment Data

We correlate our measure with human capital investment using educational attainment data (Vella (1994)). Our measure predicts women's educational attainment at the 1% level (Table

2, columns 4 and 5) consistent with Vella (1994). As women are more associated with career than family, educational attainment (defined as the share of women above the age of 25 with high school education or higher) increases. The data covers the same years as the labour force participation regressions.

6.2.4 Income Data

Klasen (2002) shows that as gender inequality increases, economic growth decreases. We validate our measures using per capita income data from IPUMS. County level data is available for 1979, 1989, and 1999. A higher association of women with career than family predicts higher per capita income in that county (Table 2, column 6).

6.2.5 Political Data

Our fifth set of validations uses political data. We use the dataset from the Inter-University Consortium for Political and Social Research (ICPSR), which provides county level information on the share of Democratic votes in each US presidential election from 1950 to 1990. This regression is based on Wilson (2020) who show that after the Roe vs Wade (1973) ruling, the Democratic party has actively defended abortion access in contrast with the Republican party. We show that our measure on abortion is correlated with a higher democratic voting share after 1974. As abortion is associated more with immoral terms rather than moral terms, the democratic party's voting share decreases (Table 2, column 8). The coefficient is statistically significant at the 1% level.

6.2.6 Crime Data

We use county level crime reported by the FBI as a part of the annual Uniform Crime Reporting Program compiled by Kaplan (2021) from 1974 to 2009. We use the number of arrests of males in rape cases as a percentage of the population as our dependent variable. Since crimes do not report the gender of the victim, this is the only crime where the majority of victims are likely to be women. As the measure increases (less violence towards women),

the rate of arrests decreases (Table 2, column 10). The coefficient is statistically significant at the 1% level.

7 Methodology Robustness

To ensure that particular modelling choices do not drive our results, we run several different robustness tests and calculate the correlation between our measures and the measures produced with these alternative modelling choices at the page level. Table 3 reports these correlations. First, we recalculate our measures using the top 50 and 200 associated words to form our dictionaries. The correlation between our measure and these alternative ones varies between 0.93 and 0.99, so our results are robust to alternative choices for the numbers of associated words.

Second, we replace our target words for women (used for career vs family and violence measures) from the OED with target words from Garg et al. (2018). This change does not affect our measure substantially either, as the correlation is 0.90 and 0.99 for the violence and career vs family measures, respectively.

Third, we test whether our choice of sentiment assignment model drives our results. To do this, we use SentProp (Hamilton et al. (2016a)) instead of the DENSIFIER method. The correlations for career vs family, feminism/suffrage and abortion are 0.99, 0.69 and 0.92, respectively. These high correlations further validate our methodology.

8 Further Methodology Details

8.1 Data Collection

We obtained approximately 193 million pages of newspaper text from NewspaperArchive (www.newspaperarchive.com) and the U.S Library of Congress's Chronicling America database. To collect newspapers from NewspaperArchive, we used a web scraping pipeline built with the library 'Scrapy'. We obtained approximately 184 million newspaper pages across all

US states from 1850 to 2009. We complement this dataset with data from the Chronicling America Database which we downloaded through the Library of Congress's API. This gives us 16 million additional pages of newspaper text. We match these papers to the Newspaper-Archive sample on the basis of name, city and state and keep only papers that are unique, which shrinks the sample from 16 million to 9 million.

8.2 Data Processing

We use nltk's (https://www.nltk.org/) default sentence tokeniser and keep all words in the nltk English dictionary and convert them to lower case.

8.3 Word2vec Training

We trained 16 Word2vec models for each decade from 1850-2009. We used the Word2vec implementation in gensim (https://radimrehurek.com/gensim/) to train our models.

The rest of the hyperparameters are as follows: the embedding dimension is 300, the context window is 10 words, the minimum count is 100 and the number of iterations is 10.

8.4 Target Words

For the career vs family and violence measures, we curated the initial word list from the historical thesaurus of the OED. We removed polysemous (generic) words and words that should be the product of Word2vec (associated words) from this initial list to produce a list of final target words for women. The original word list for women can be accessed from the OED (https://ht.ac.uk/category/#id=39814). Our target words are as follows:

Career vs Family and Violence: quean, wife, woman, lady, bride, carline, carling, female, goddess, feminine, she, feme, maness, sister, madonna, moll, womankind, biddy, mother, bint, totty, dame, babe, virago

Feminism/Suffrage: feminism, feminist, suffrage, womanhood, emancipation, activism, motherhood

Abortion: abortion, aborted, unborn, infanticide, preborn

8.5 Seed Words

The seed words for our measures are as follows:

Career: professional, corporation, salary, office, business, career

Family: home, child, family, cousin, marriage, wedding

Moral: moral, morality, ethical, morals, conscience, decency

Immoral: immoral, reprehensible, despicable, immorality, irresponsible, disgraceful

Equality: activism, activist, reformer, revisionist

Extreme: extremist, radical, provocative, cynical

Violence: victim, assailant, attacker, perpetrator, suspect, stabbing, unconscious, accomplice, gunshot, assault, rape, rapist, raping, murder, molestation, abduction, slaying, mutilation, manslaughter, homicide, felony, harassment, molest, molester, abuser, abusive

8.6 Geo-referenced County Crosswalks

County boundaries have changed significantly over our sample period. To deal with this we first scrape data on the exact coordinates of the cities covered by NewspaperArchive. For the Chronicling America database we use a regular expression search to extract the city, state and place of publication. The library of congress database also provides information on the cities of circulation in the Chronicling America database. We delete all newspapers with coverage across multiple cities and match the remaining ones on paper name, city and state to the NewspaperArchive metadata.

To create stable county mappings we use county geographical shape files from 1990 from the IPUMS-NHGIS database. We use these shapefiles and the geopandas function in python to map the latitudes and longitudes of each city to a 1990 county. This ensures that our measures are mapped to consistent 1990 county boundaries. Eckert et al. (2020) construct county crosswalks from 1790 and provide weights to map each county to a stable 1990 one. We use these weights to weight all the dependent variables in our validation regressions so that they are mapped to a consistent 1990 county. As the county crosswalks are only available by decade, we assume no changes in the files between decades.

8.7 Statistics and Github Repositories

All regressions were run in stata 16 using the reghdfe stata package. We used code from the Gensim (https://github.com/RaRe-Technologies/gensim), SocialSent (https://github.com/williamleif/socialsent), Fast Sentence Embeddings (https://github.com/oborchers/Fast_Sentence_Embeddings), Word Embeddings Benchmarks (https://github.com/kudkudak/word-embeddings-benchmarks) and Word Embeddings for Historical Text (https://github.com/williamleif/histwords) github repositories.

Data and Code Availability

Chronicling America data can be accessed via their API (https://chroniclingamerica.loc.gov/about/api) and we will provide computer code on how to access the NewspaperArchive data. All the other data sources that we use are publicly available and we will provide our trained models, measures, and code before publication.

Acknowledgments

We are grateful to the Wheeler Institute for Business and Development at London Business School for financial support. M.M. also acknowledges funding from the google cloud platform PhD student research credits program.

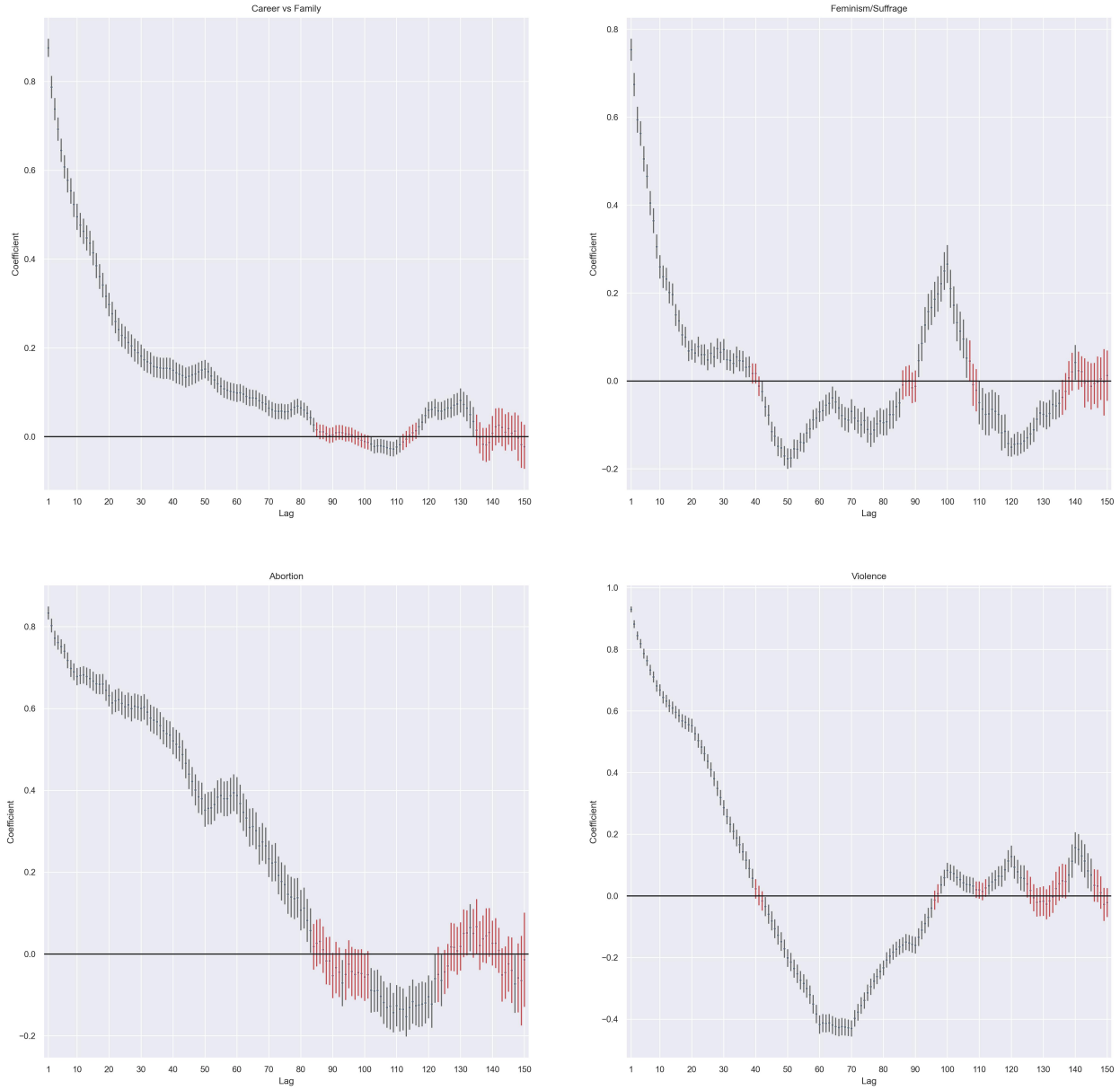
Author Contributions

Author Contributions: M.M. and S.S. conceptualised and designed the study, collected, and processed the data, and wrote and approved the final manuscript. S.S. designed, implemented, and validated the machine learning methodology.

Competing Interests

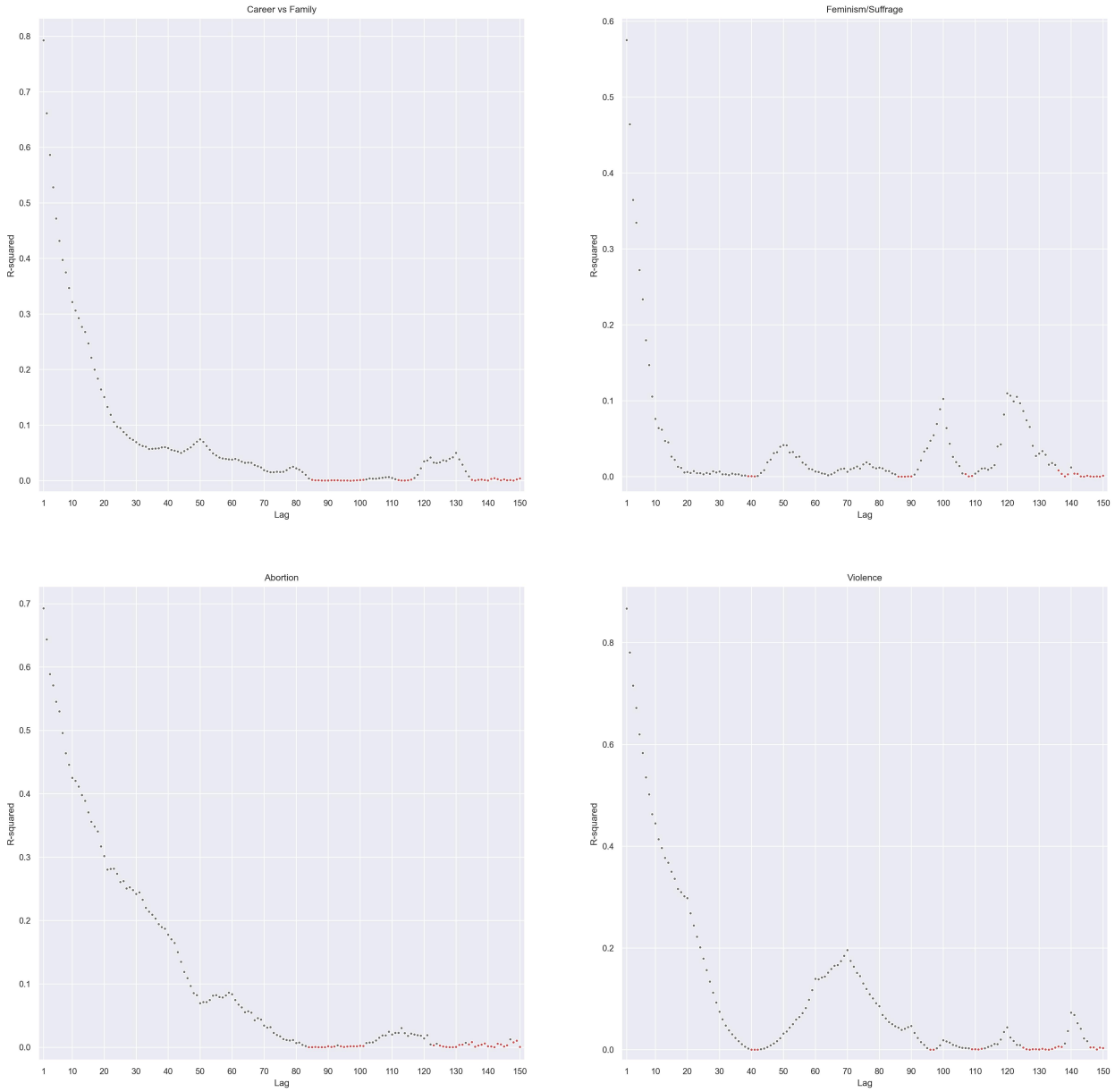
The authors declare no competing interests.

Figure 1: Autoregressive Coefficients



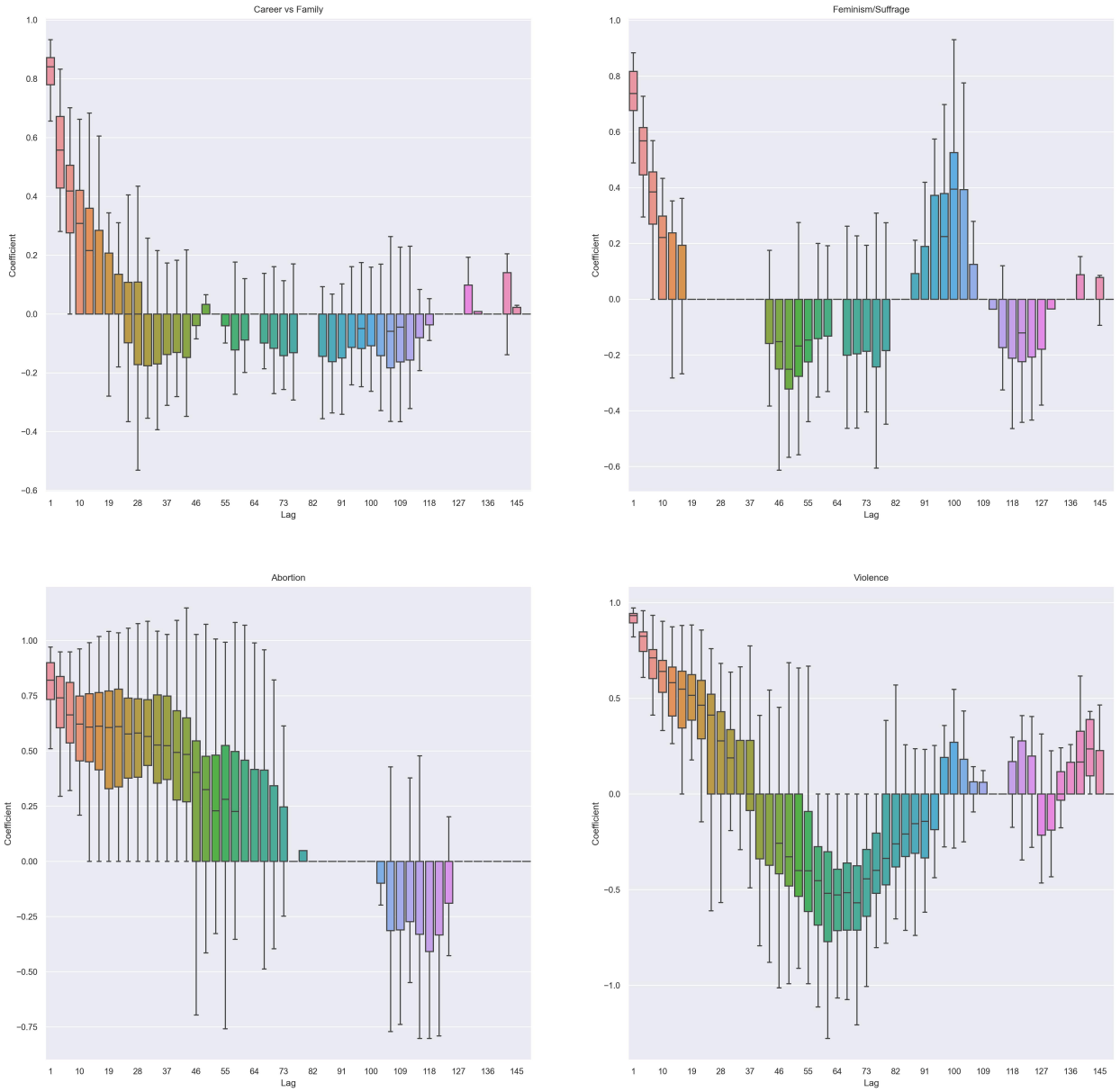
This figure plots the autoregressive coefficient of each lag using Equation 1. We use robust standard errors. The dot indicates the coefficient and the line indicates the 95% confidence intervals. Grey and red lines indicate significant and insignificant coefficients, respectively. The horizontal black line is plotted at 0.

Figure 2: R-squareds from Autoregressions



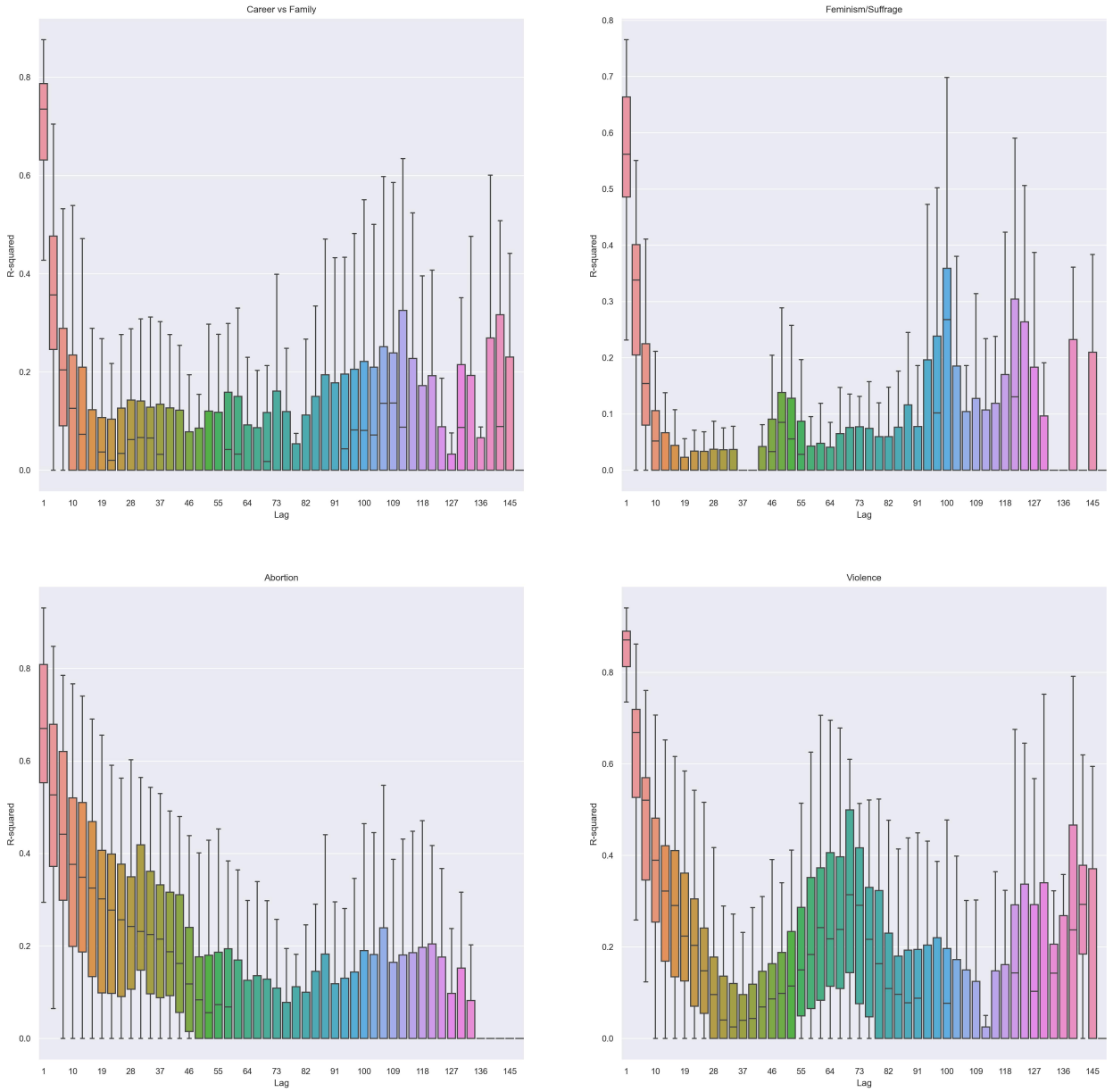
This figure plots the R-squareds from autoregressions for each lag using Equation 1. We use robust standard errors. Grey dots indicate significant R-squareds at the 5% level whereas red dots indicate insignificant R-squareds at the 5% level using an F-test.

Figure 3: Variation in Autoregressive Coefficients across States



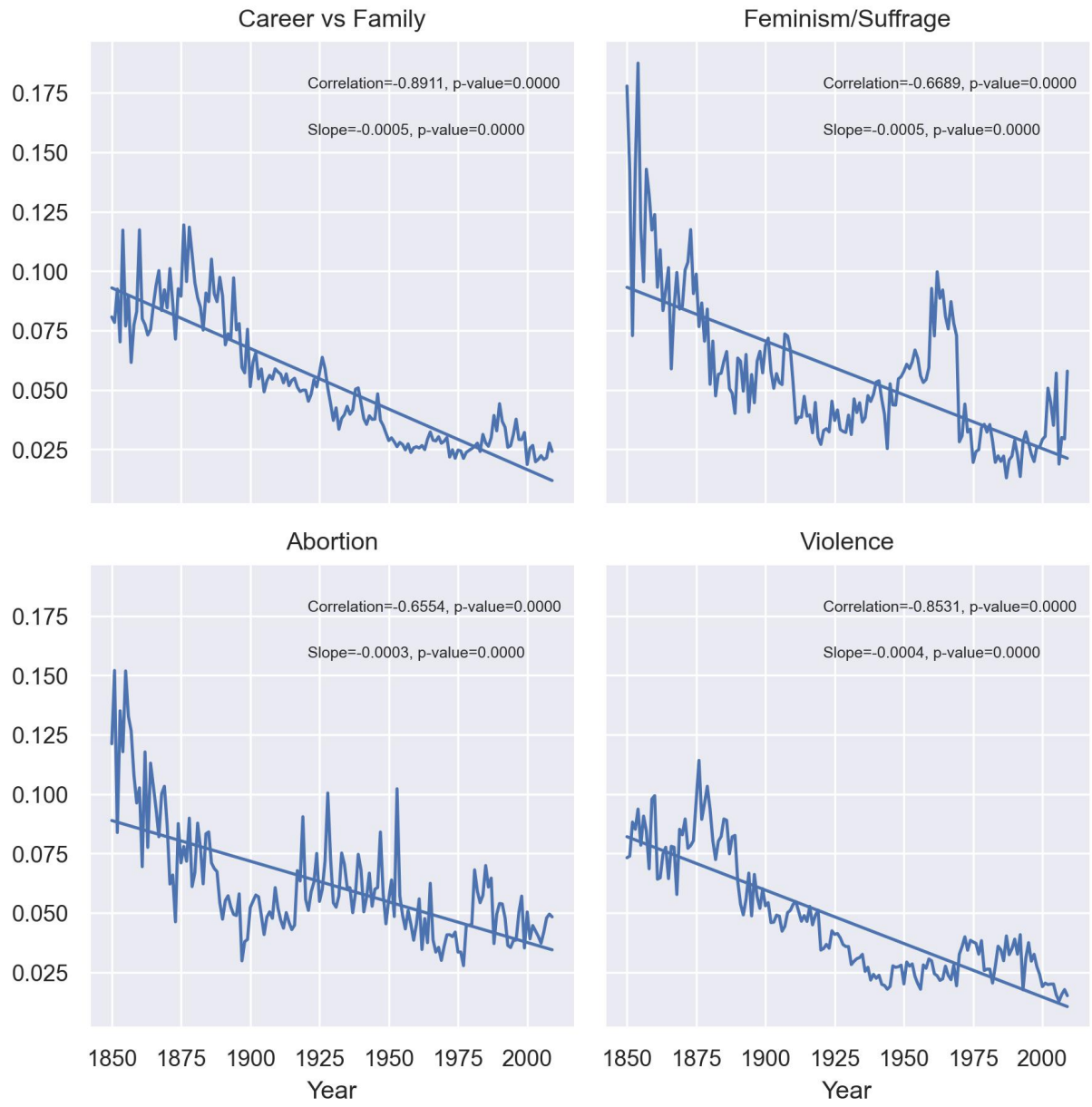
This figure plots the distribution of the autoregressive coefficient across states for $lags = 1, 4, 7, ..., 148$ using Equation 2. We only plot the distribution of coefficients for every third lag for ease of visualisation. We use robust standard errors. Insignificant coefficients at the 5% level are given a value of 0. The lower edge of the box, horizontal line inside the box and upper edge of the box represent Q1, median, and Q3, respectively, where Q1 is the first quartile and Q3 is the third quartile. The whiskers extend from Q1 - 1.5 $\times$ IQR to Q3 + 1.5 $\times$ IQR where IQR is the interquartile range. Please note for lags for which only a horizontal line is visible, the five points (Q1 - 1.5 $\times$ IQR, Q1, median, Q3, and Q3 + 1.5 $\times$ IQR) are all the same.

Figure 4: Variation in R-squareds of Autoregressions across States



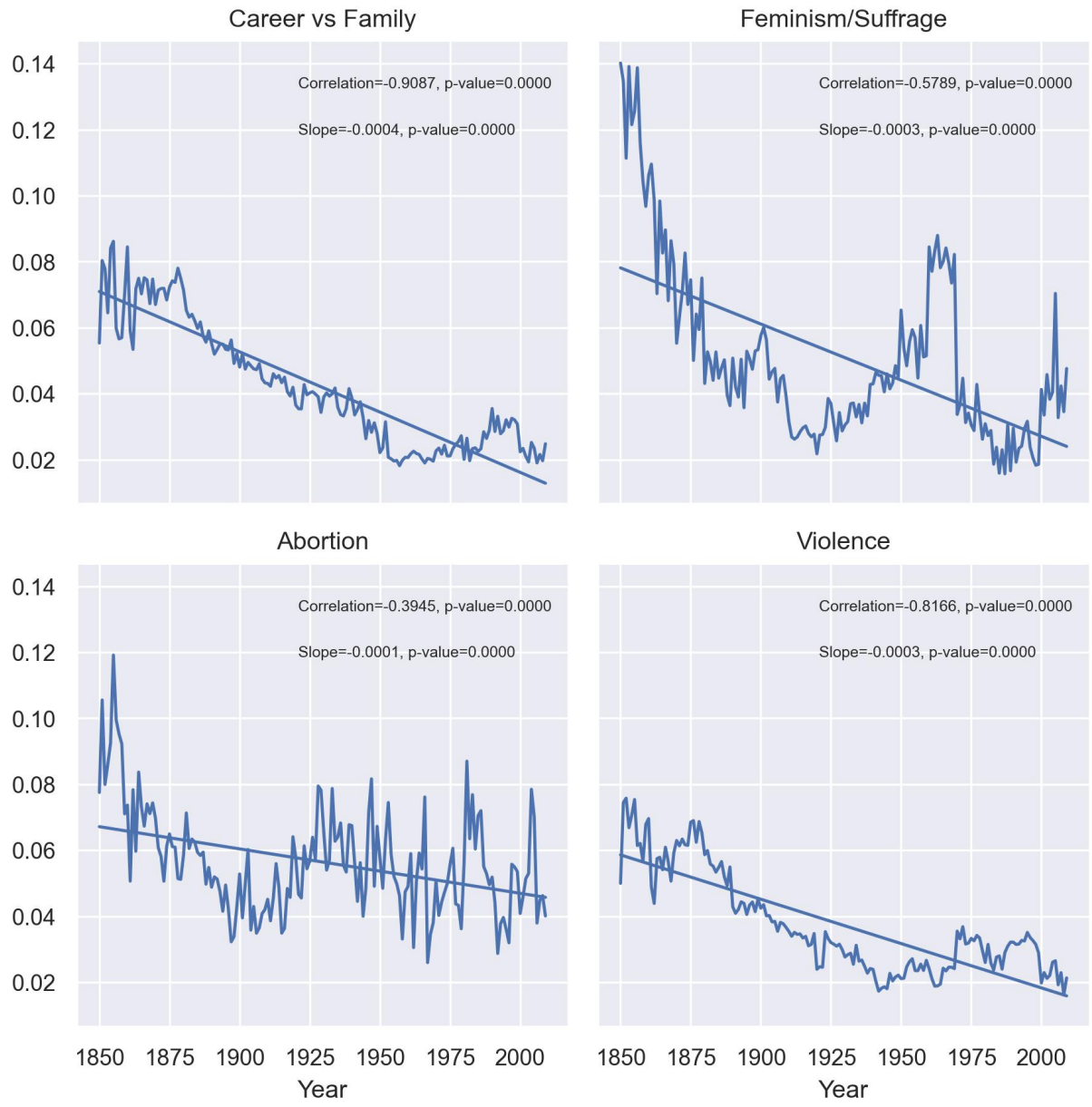
This figure plots the distribution of the R-squareds from autoregressions for $lags = 1, 4, 7, ..., 148$ using Equation 2. We only plot the distribution of R-squareds for every third lag for ease of visualisation. We use robust standard errors. Insignificant R-squareds at the 5% level using an F-test are given a value of 0. The lower edge of the box, horizontal line inside the box and upper edge of the box represent Q1, median, and Q3, respectively, where Q1 is the first quartile and Q3 is the third quartile. The whiskers extend from Q1 - 1.5 $\times$ IQR to Q3 + 1.5 $\times$ IQR where IQR is the interquartile range. Please note for lags for which only a horizontal line is visible, the five points (Q1 - 1.5 $\times$ IQR, Q1, median, Q3, and Q3 + 1.5 $\times$ IQR) are all the same.

Figure 5: Interquartile Range of the Measures across the States



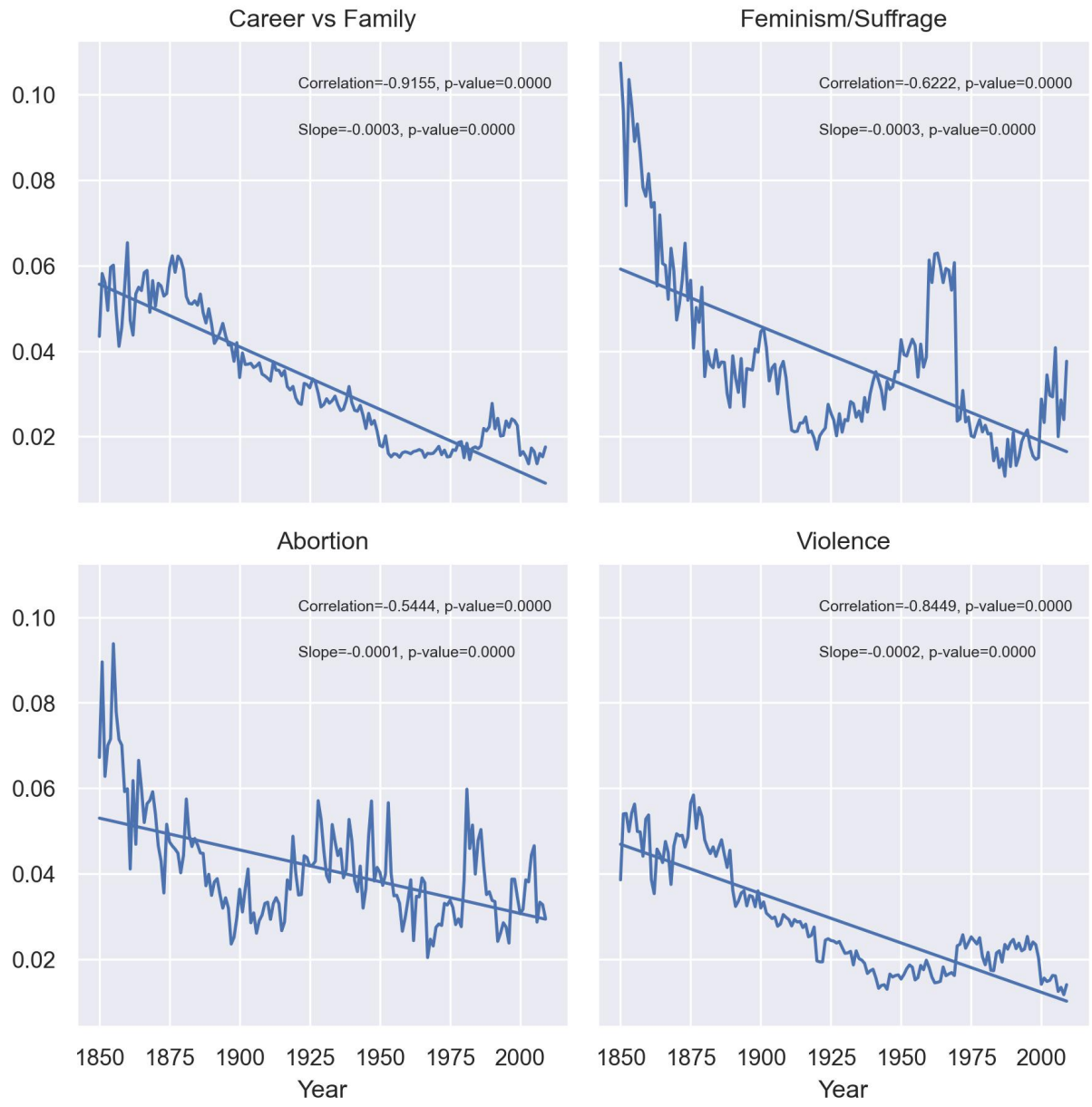
This figure plots the interquartile range of the state level measures over time. The correlation with the year variable and the slope of the regression from Equation 3 are in the top right corner. We use robust standard errors. The line of best fit is plotted.

Figure 6: Standard Deviation of the Measures across the States



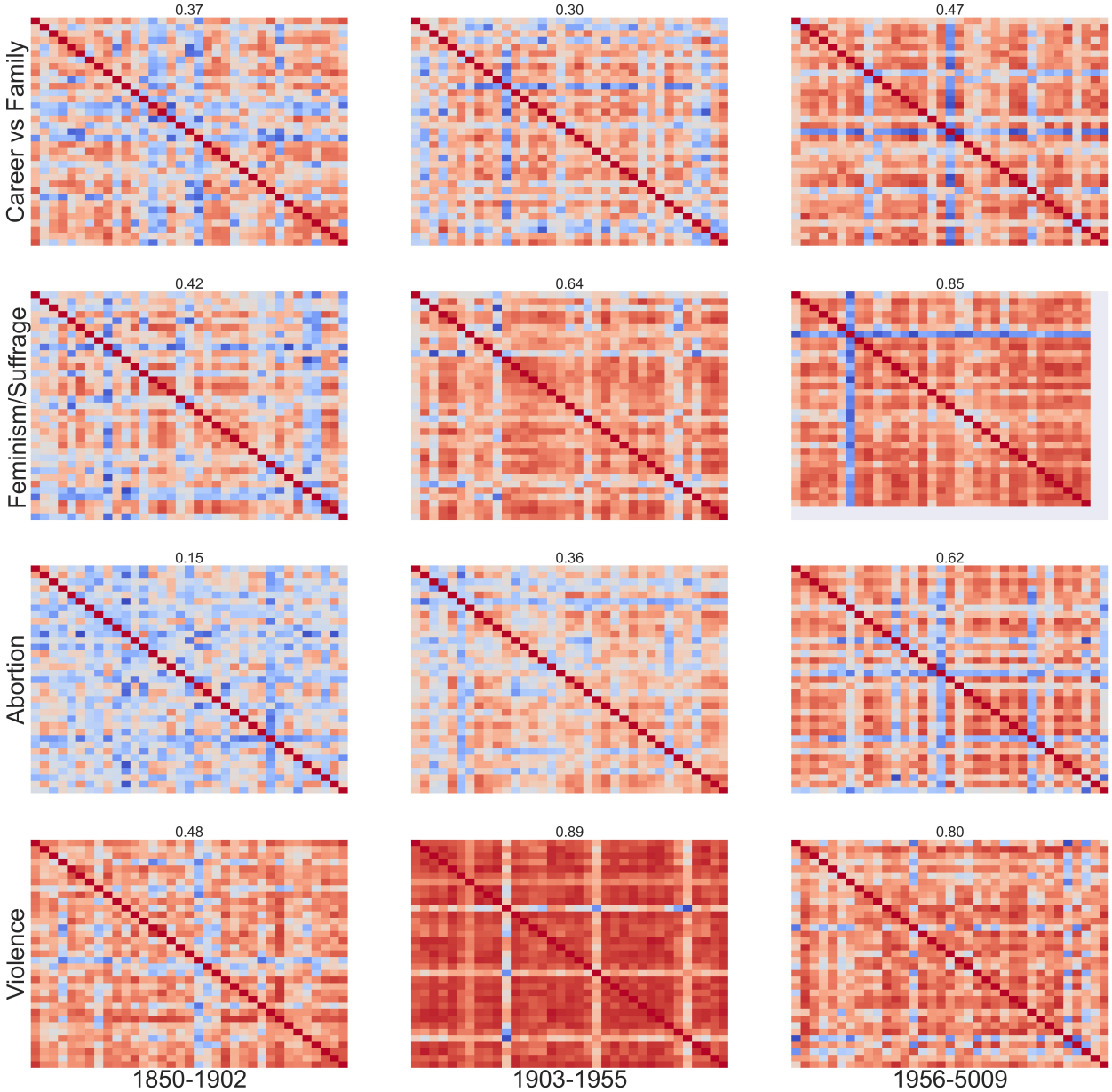
This figure plots the standard deviation of the state level measures over time. The correlation with the year variable and the slope of the regression from Equation 3 are in the top right corner. We use robust standard errors. The line of best fit is plotted.

Figure 7: Mean Absolute Deviation of the Measures across the States



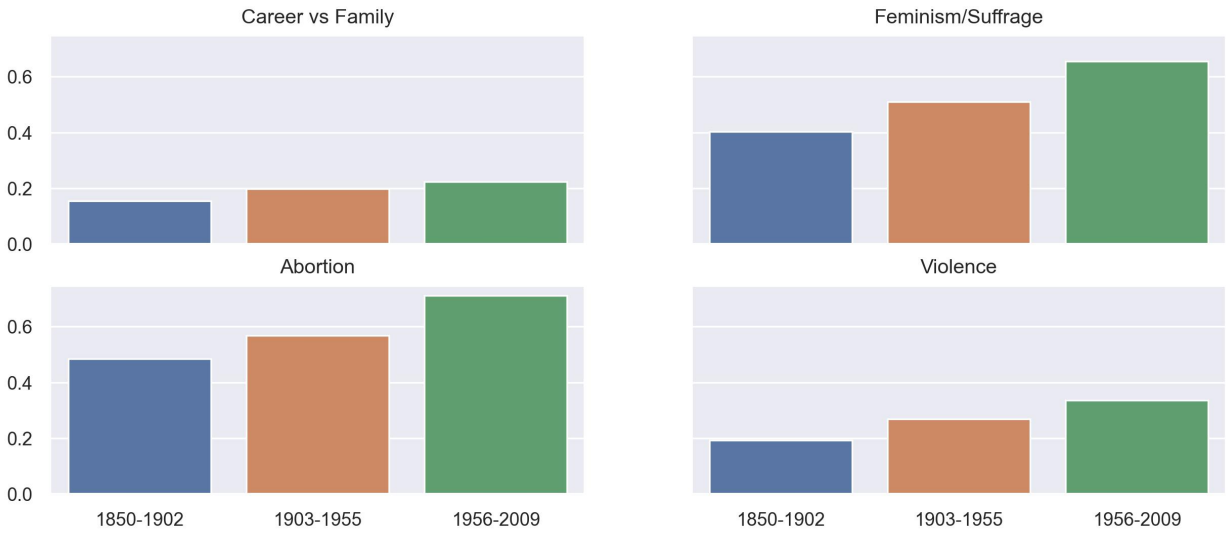
This figure plots the mean absolute deviation of the state level measures over time. The correlation with the year variable and the slope of the regression from Equation 3 are in the top right corner. We use robust standard errors. The line of best fit is plotted.

Figure 8: State Pair Correlations



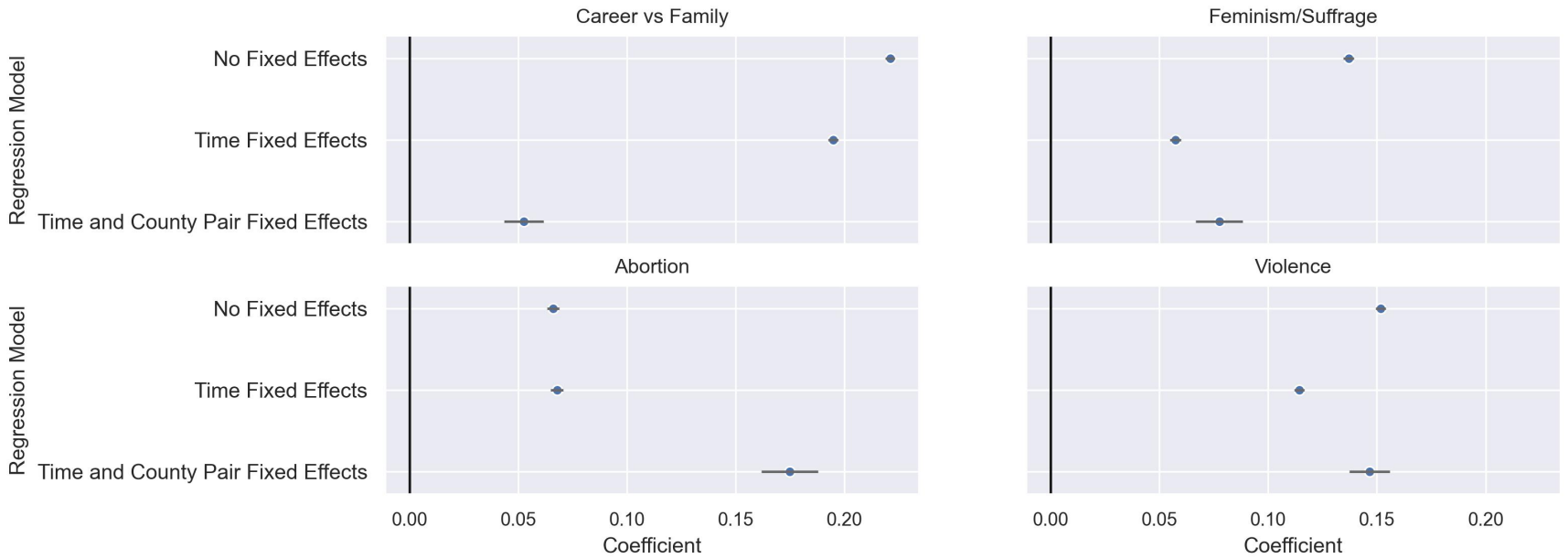
This figure plots the heatmap of the Pearson correlation between the measures for all possible state pairs. The sample is divided into three equal time periods. The median correlation for each period is on top of each heat map.

Figure 9: The Kolmogorov-Smirnov Test for the Equality of Two Distributions



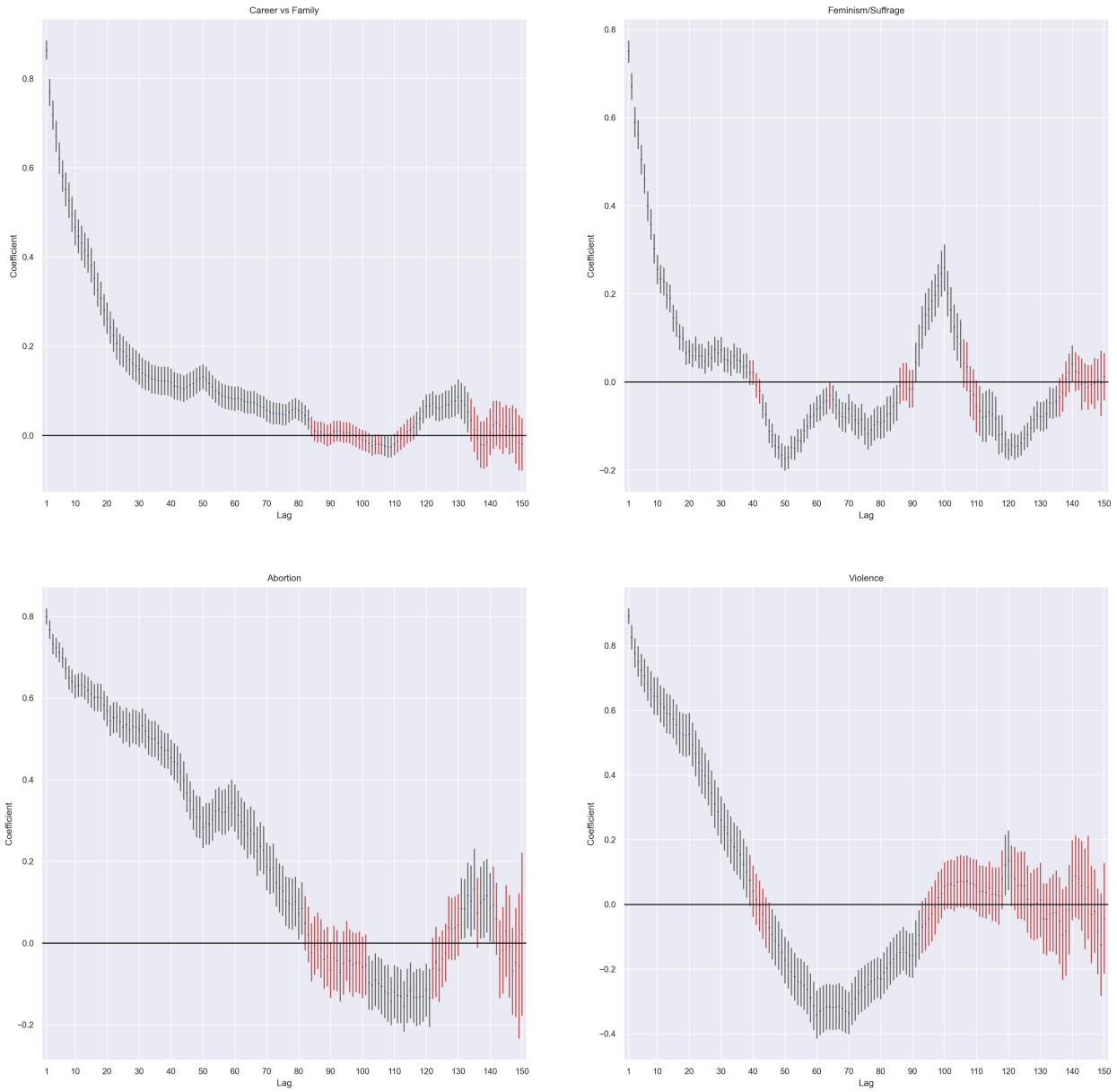
This figure denotes the proportion of state pairs where the Kolmogorov-Smirnov test fails to reject the null hypothesis of identical distribution. The sample is divided into three equal time periods.

Figure 10: The Role of Transportation Costs in Convergence of Gender Norms



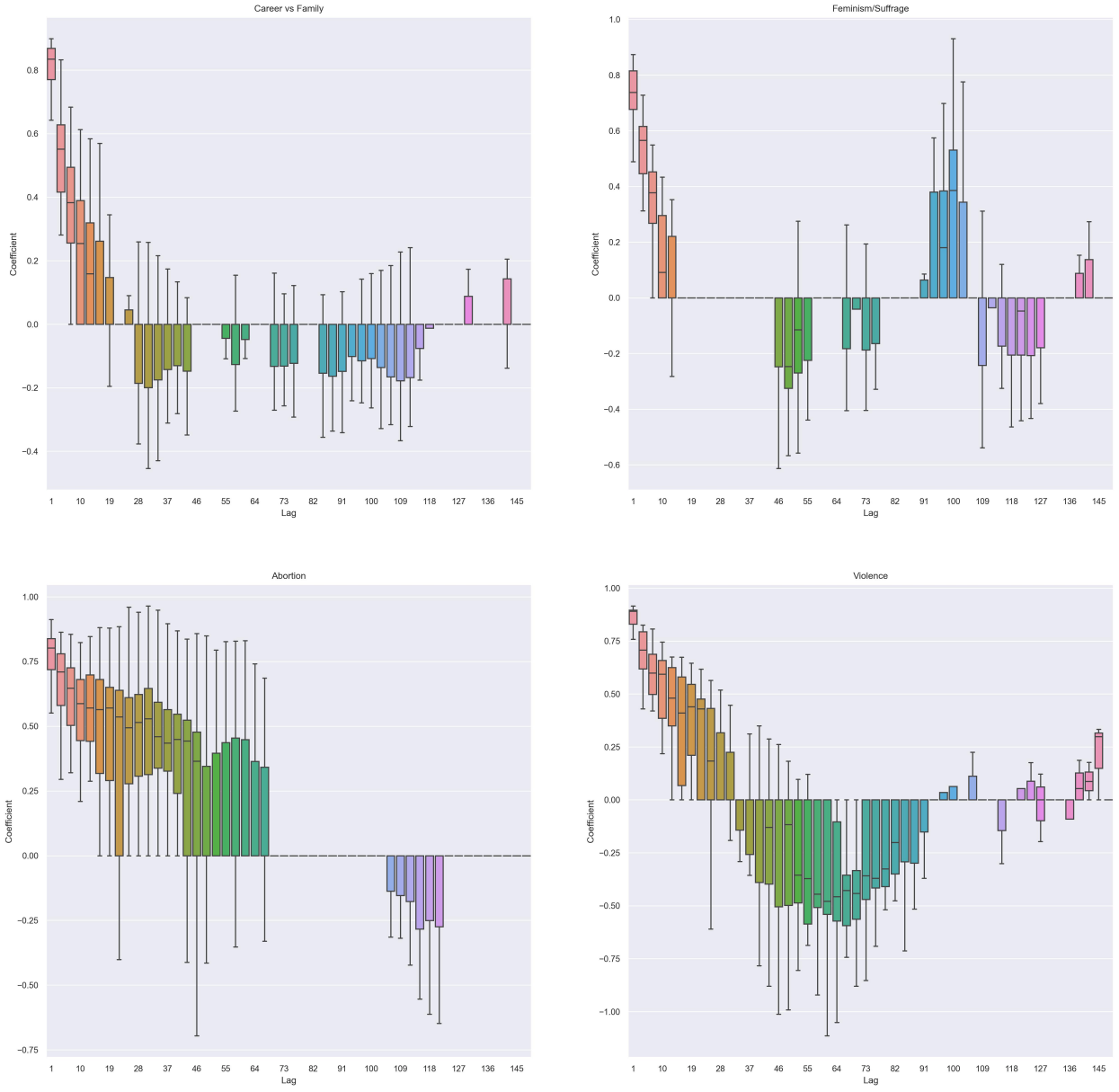
This figure plots the coefficients from the regression of the log of the absolute difference in the measures and log of transport cost for each each county pair using Equation 4. We use robust standard errors. The figure plots the coefficients using no fixed effects, time fixed effects alone, and both time and county pair fixed effects. The dot indicates the coefficient and the line indicates the 95% confidence intervals.

Figure 11: Robustness Test for Presence of Unit Roots: Persistence



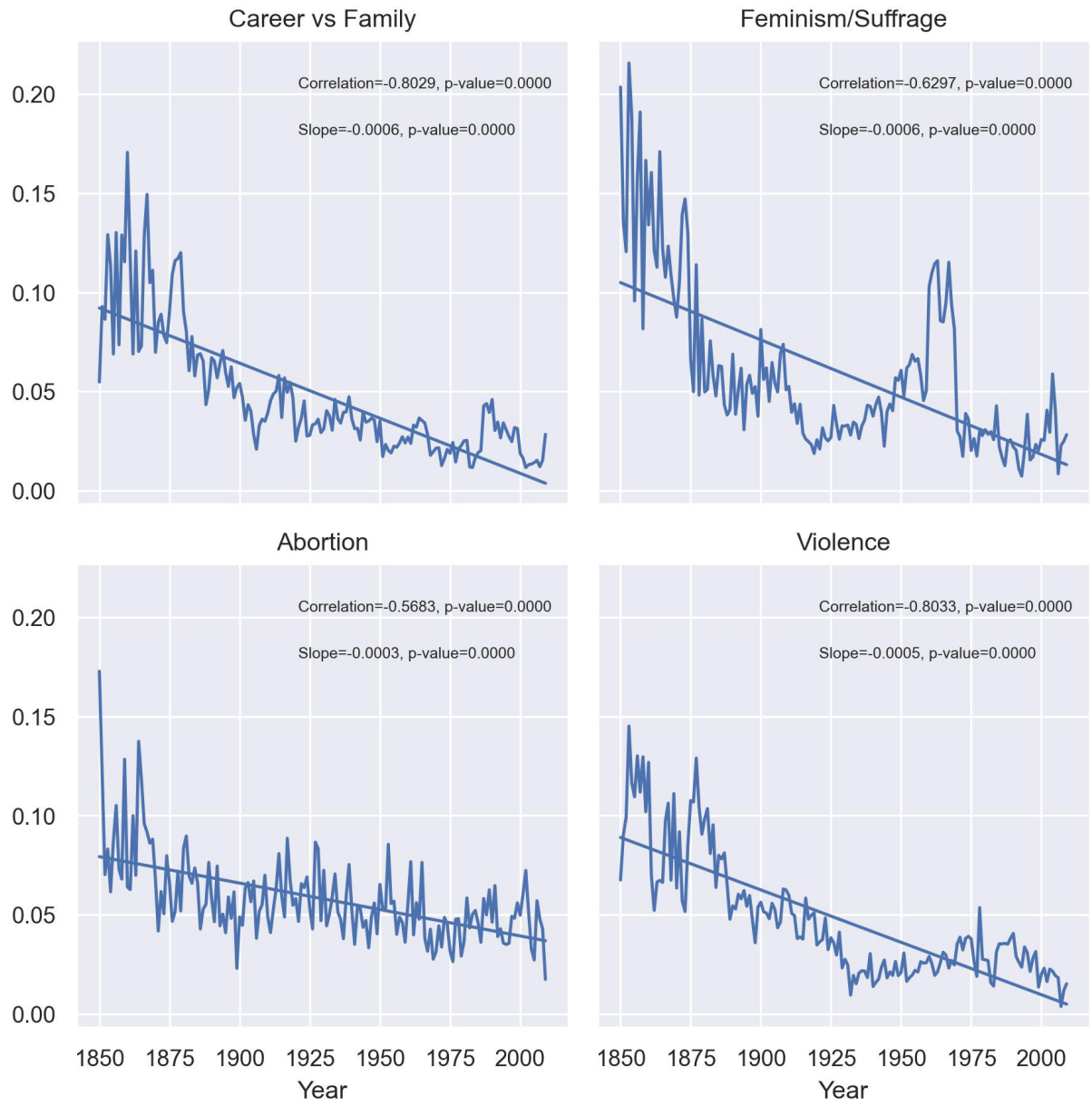
This figure plots the autoregressive coefficient of each lag using Equation 1 after excluding states that do not pass the ADF test (at the 10% level). We use Newey-West standard errors. The dot indicates the coefficient and the line indicates the 95% confidence intervals. Grey and red lines indicate significant and insignificant coefficients, respectively. The horizontal black line is plotted at 0.

Figure 12: Robustness Test for Presence of Unit Roots: State Variation in Persistence



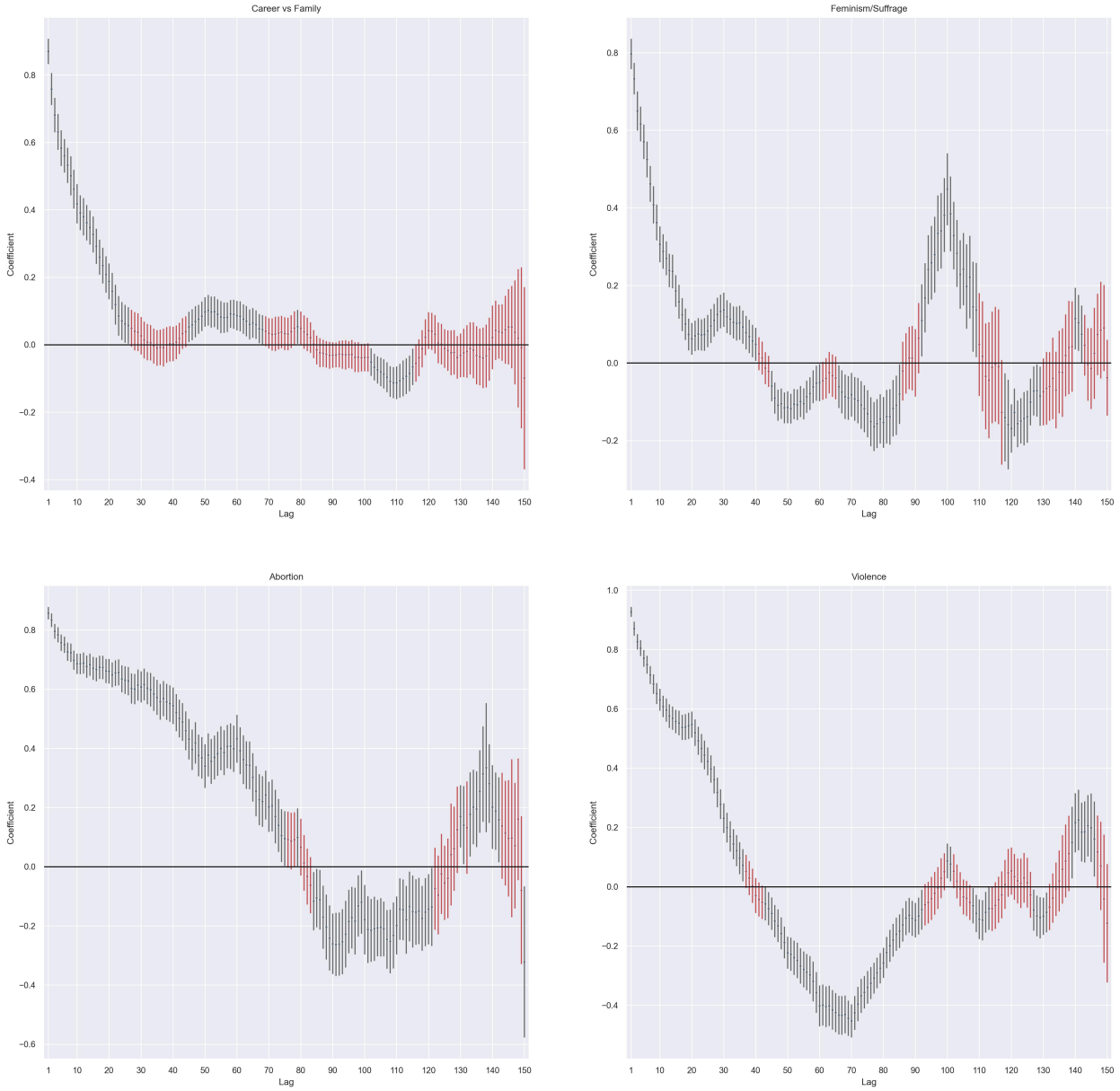
This figure plots the distribution of the autoregressive coefficient across states for $lags = 1, 4, 7, ..., 148$ using Equation 2 after excluding states that do not pass the ADF test (at the 10% level). We only plot the distribution of coefficients for every third lag for ease of visualisation. We use Newey-West standard errors. Insignificant coefficients at the 5% level are given a value of 0. The lower edge of the box, horizontal line inside the box and upper edge of the box represent Q1, median, and Q3, respectively, where Q1 is the first quartile and Q3 is the third quartile. The whiskers extend from Q1 - 1.5 × IQR to Q3 + 1.5 × IQR where IQR is the interquartile range. Please note for lags for which only a horizontal line is visible, the five points (Q1 - 1.5 × IQR, Q1, median, Q3, and Q3 + 1.5 × IQR) are all the same.

Figure 13: Robustness Test for Internal and External Migration: Dispersion



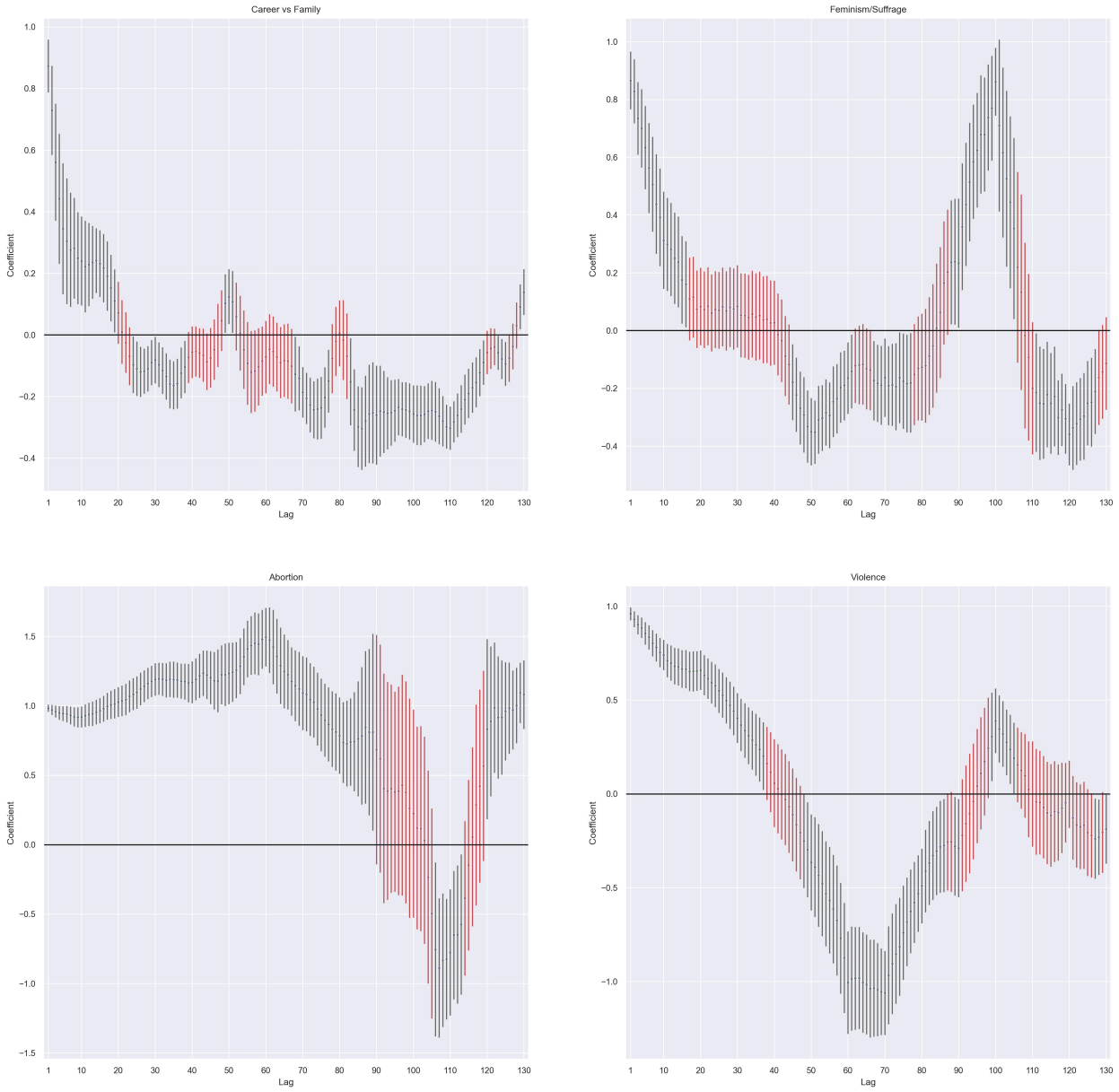
This figure plots the interquartile range of the state level measures over time using only those state-year pairs where at least 70% of the population is born in-state. The correlation with the year variable and the slope of the regression from Equation 3 are in the top right corner. We use robust standard errors. The line of best fit is plotted.

Figure 14: Robustness Test for Internal and External Migration: Persistence



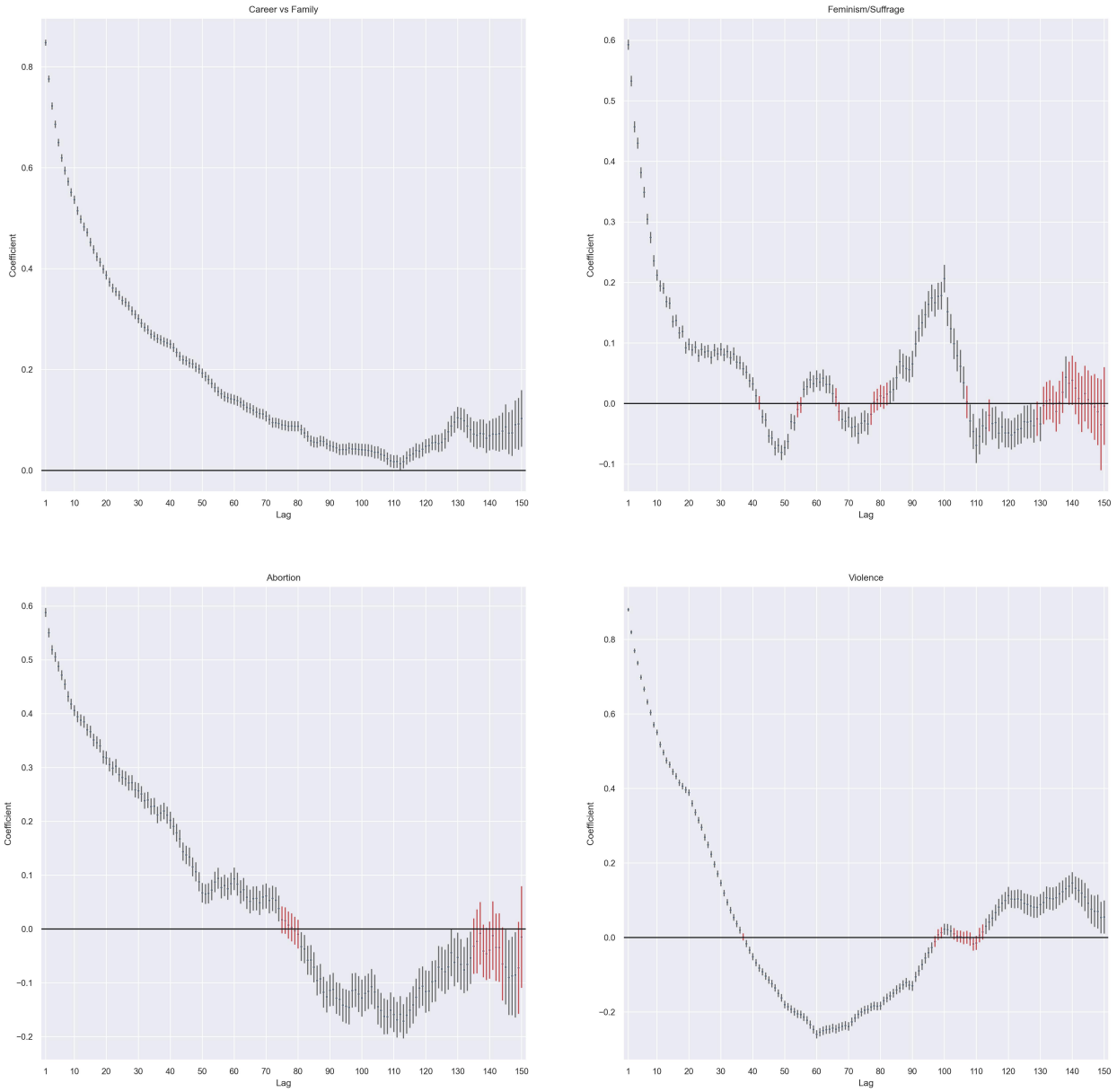
This figure plots the autoregressive coefficient of each lag using Equation 1 and only those state-year pairs where at least 70% of the population is born in-state. We use robust standard errors. The dot indicates the coefficient and the line indicates the 95% confidence intervals. Grey and red lines indicate significant and insignificant coefficients, respectively. The horizontal black line is plotted at 0.

Figure 15: Robustness Test for Measurement Error: Persistence



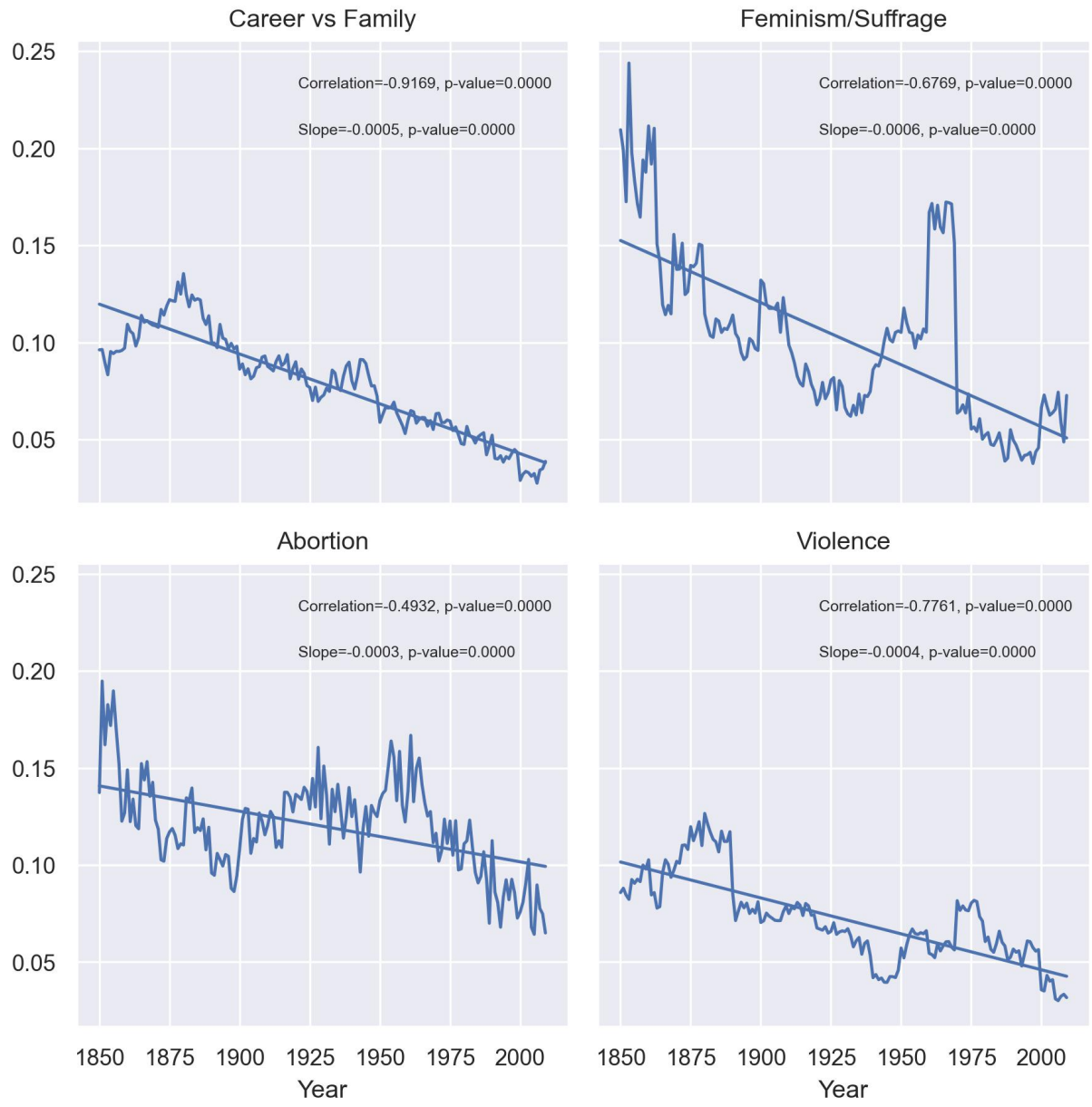
This figure plots the autoregressive coefficient of each lag using Equation 1 using only one national level time series. We use robust standard errors. The dot indicates the coefficient and the line indicates the 95% confidence intervals. Grey and red lines indicate significant and insignificant coefficients, respectively. The horizontal black line is plotted at 0.

Figure 16: Robustness Test: County Level Persistence



This figure plots the autoregressive coefficient of each lag using Equation 1 using counties instead of states. We use robust standard errors. The dot indicates the coefficient and the line indicates the 95% confidence intervals. Grey and red lines indicate significant and insignificant coefficients, respectively. The horizontal black line is plotted at 0.

Figure 17: Robustness Test: County Level Dispersion



This figure plots the interquartile range of the county level measures over time. The correlation with the year variable and the slope of the regression from Equation 3 are in the top right corner. We use robust standard errors. The line of best fit is plotted.

Table 1: Methodology: Word2vec Evaluation Metrics

Test	COHA 1991-2000 Score	Our 1990-1999 Model Score
Categorisation Tests		
AP	0.33	**0.52**
BLESS	0.43	**0.71**
Battig	0.21	**0.34**
ESSLI-1a	0.64	**0.80**
ESSLI-2b	0.68	**0.78**
ESSLI-2c	0.67	**0.71**
Similarity Tests		
MEN	0.62	**0.67**
WS353	0.56	**0.62**
WS353R	0.52	**0.59**
WS353S	0.61	**0.68**
SimLex999	0.28	**0.32**
RW	**0.22**	0.20
RG65	0.28	**0.62**
MTurk	0.29	**0.43**
Analogy Tests		
Google	0.05	**0.12**
MSR	0.00	**0.07**
SemEval2012	0.10	**0.16**

This table shows our relative performance on word embedding benchmarks (Levy et al. (2015) and Jastrzebski et al. (2017)) compared to COHA model from Garg et al. (2018). Scores in red indicate the model which performs better.

Table 2: Methodology: Validation Regressions

	Survey Question	Labour Force Participation	Labour Force Participation	Educational Attainment	Educational Attainment	Per Capita Income	Survey Question	Democrat Voting Share	Survey Question	Crime
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Career vs Family	3.214**	-0.362***	-0.320***	-0.636***	-0.330***	-0.802***				
	(1.531)	(0.043)	(0.036)	(0.079)	(0.049)	(0.271)				
Abortion							-0.554**	-0.100***		
							(0.272)	(0.034)		
Feminism/Suffrage									-37.628***	
									(14.129)	
Violence										-0.007***
										(0.002)
Adj. R-squared	0.17	0.60	0.17	0.61	0.19	0.79	0.030	0.19	0.63	0.019
Observations	146	1,023	1,074	1,023	1,074	456	97	992	122	6,914
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

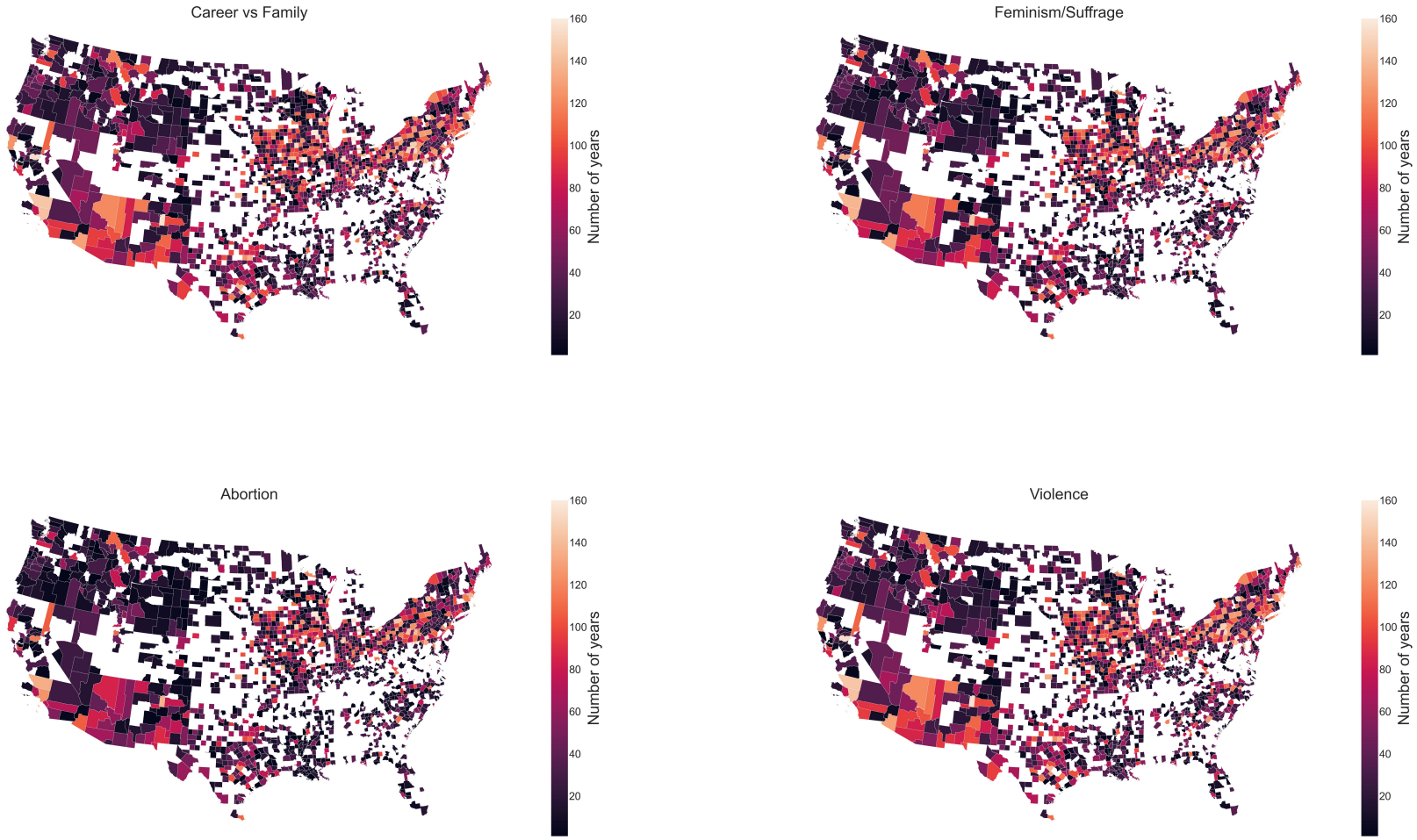
This table shows validation regressions for all our measures using Equation 5. Robust Standard errors are reported in parentheses below each coefficient. ***, **, * denote significance at the 1%, 5%, and 10% level, respectively.

Table 3: Methodology: Robustness Tests

Alternative Modelling Choice	Career vs Family	Feminism/Suffrage	Abortion	Violence
Most Associated Words				
Top 50	0.99	0.93	0.99	0.99
Top 200	0.99	0.93	0.99	0.99
Alternative Target Words for Women				
Target words from Garg et al. (2018)	0.99			0.90
Alternative Method to Generate Word Dictionary				
SentProp from from Hamilton et al. (2016a)	0.99	0.69	0.92	

This table shows results for the robustness checks for our methodology. The numerical values indicate correlations between our measure and measures calculated under the alternative modelling choice at the page level described in the left hand columns.

Figure 18: Methodology: County Coverage



This figure shows the coverage of all our measures. The colour scheme shows the number of years we can construct our measures for each county.