Supplementary Information Appendix for - "Evidence for a Scale Invariant Relationship Between the Incumbency Advantage and the Nationalization of U.S. House Elections 1866–2014"

Contents

Introduction

This supplementary information appendix contains all of the code necessary to replicate all of the analyses reported in the main manuscript, as well as a number of robustness checks. It was written using Rmarkdown (http://rmarkdown.rstudio.com/), which means that compiling the Rmd file associated with this PDF simultaneously creates this document and reproduces all of the analyses in R.

Preparing the Data for Analysis

First, read in the CQ election data. This data was collected from the CQ Voting and Elections Collection (http://library.cqpress.com/elections/). Data was obtained for every U.S. House of Representatives election from 1866 - 1944. We subset the data to include only years that were not included in Jacobson (2015)'s dataset. Therefore, our new dataset contains election results from 1866 to 1944 but follows the exact same variable coding scheme as Jacobson (2015). We also subset the data to eliminate uncontested elections where the Democratic vote share was either 0 or 100. Finally, we also remove the elections that followed redistricting years and elections in off years.

```
rm(list=ls())
# load necessary libraries
library(grDevices)
library(cowplot)
library(grid)
library(dplyr)
library(ggplot2)
library(tidyr)
library(broom)


########################################
# read in data and clean for analysis #
########################################

cq_data <- tbl_df(read.csv("CQdata_elections_master.csv", stringsAsFactors = F)) %>%
  select(year, statecd, ptynow, inc3, dv, dvp, state) %>%
  rename(stcd = statecd) %>%
  filter(year%%2 ==0, # remove off year elections (to replace dead mc's etc.)
         year< 1946 & year >1864, # remove elections pre Civil War and post-Jacobson data
         dv>0 & dvp>0 &  dv<100 & dvp <100, # remove uncontested elections
         !year%in%seq(1862, 1912, by=10)
         & !year%in%seq(1932, 2012, by=10)) # remove post-redistricting elections
```

Next, read in data from Jacobson (2015) "It's Nothing Personal: The Decline of the Incumbency Advantage in US House Elections." This data is available from the JOP dataverse (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/29559).

```
load("JOPrepfile1.RData") # data available from the JOP dataverse.
gary_j_data <- x %>%
  select(year, stcd, ptynow, inc3, dv, dvp, state, south, dpres) %>%
  filter(dv>0 & dvp>0 &  dv<100 & dvp <100,
         !year%in%seq(1862, 1912, by=10) & !year%in%seq(1932, 2012, by=10))
```

Because our analysis makes use of the distinction between South and non-South states, we must code the Southern states in the new CQ data to match Jacobson's definition of South states. The definition of "South" is the 11 states that were part of the Confederacy.

```
south_by_state <- data.frame(south = x$south, state=x$state) %>%
  group_by(state) %>%
  summarise(south = max(south, na.rm=T)) %>%
  filter(!is.na(state)) %>%
  ungroup
cq_data <- left_join(cq_data, south_by_state, by="state") # merge the south code to the
                                                          # new dataset.
```

Next, bind the datasets together into one large dataset called cq_data.

```
cq_data <- rbind(cq_data, select(.data=gary_j_data, -dpres) ) # dpres is not included in
                                                              # the CQ dataset, so we drop
                                                              # it here.
```

Calculating the Measure of Nationalization

As described in the main text, our analysis uses the standard deviation in Democratic vote swings as a measure of how localized elections are. In the code chunk below, we define a function that calculates bootstrapped standard errors for our nationalization measure.

```
boot_se <- function(n_boot, dv, dvp){
  observations <- na.omit(cbind(dv,dvp))
  boot_straps <- vector(length=n_boot, mode="numeric")
  for(i in 1:n_boot){
    sampled_ids <- sample(1:dim(observations)[1],
                          size=length(observations),
                          replace=T)
    boot_straps[i] <- sd(dv[sampled_ids]-dvp[sampled_ids])
  }
  return(sd(boot_straps))
}
```

Now, we calculate the nationalization measure by year. Our measure of nationalization is calculated by taking the standard deviation of the difference between the Democratic vote share in an election from the Democratic vote share in the previous election, $sd(demvote_t - demvote_{t-2})$. This standard deviation is then made negative so that we have a measure of nationalization and not "less nationalization."

```
swing_variance <- cq_data %>%
  filter(inc3 !=0) %>%
  group_by(year) %>%
  summarise(nationalization = -sd(dv - dvp, na.rm=T),
            nationalization_se = boot_se(n_boot=3000, dv, dvp)
            ) %>%
  ungroup()
```

Below, we calculate the correlation between Jacobson (2015)'s measure of nationalization and our measure of nationalization. We find that the correlation between the two measures is $r = .79$ ($p < 2.39 \times 10^{-6}$).

```
measure_comparison <- gary_j_data %>%
  filter(year>=1954) %>%
  group_by(year) %>%
  summarise(nationalization1 = -sd(dv - dvp, na.rm=T),
         nationalization2 = cor(dv,dpres,use = "pairwise.complete.obs"))

tidy(cor.test(measure_comparison$nationalization1, measure_comparison$nationalization2))
```

```
##    estimate statistic      p.value parameter  conf.low conf.high
## 1 0.7920573  6.222612 2.386109e-06        23 0.5777394 0.9042031
```

Calculating the Incumbency Advantage Coefficients

Now, we calculate the incumbency advantage coefficients by year. The incumbency advantage is calculated using ordinary least squares regression for every year except the years following redistricting (Gelman and King, 1990). The formula is:

$$d_{it} = \beta_0 + \beta_1 * d_{it-2} + \beta_2 * inc_{it} + \beta_3 * party_{it} + e_{it}$$

Where:

- d_{it} is the Democratic vote share in district i and year t.
- d_{it-2} is the Democratic vote share in district i for the previous election.
- inc_{it} is coded 1 if the incumbent is a Democrat, -1 if they are a Republican, and 0 if neither candidate is an incumbent.
- $party_{it}$ is coded 1 if the winning party was a Democrat and -1 if the winning party was the non-Democratic candidate (almost always a Republican).

The incumbency advantage is estimated by the coefficient β_2.

```
b_incumbency_adv <- cq_data %>%
group_by(year) %>%
  do(., tidy(lm(dv ~ dvp + ptynow + inc3, data = . ))) %>%
  filter(term == "inc3") %>%
  ungroup() %>%
  left_join(., swing_variance, by="year") %>% # join the nationalization measure to the
                                              # incumbency advantage coefficients.
  mutate(post_1952 = as.numeric(year>1952)) # dummy varaible for whether the year is after
                                            # 1952 or not. 1 = post 1952, 0 = otherwise.
```

Creating the Graphs (Figure 1)

Three graphs are included in Figure 1. In panel A, we plot our measure of nationalization over time. In panel B, we plot the incumbency advantage coefficients over time. To create panel C, we divide the data into two eras, before 1952 and after 1952. We then calculate the correlation between nationalization and the incumbency advantage by era.

We then plot the relationship between nationalization and the incumbency advantage (panel C). This is a path diagram that allows us to see changes over time as a process. It connects all of the observations and takes advantage of shading to denote years passing. The earliest years are seen in dark blue and as years in the data pass, the blue line becomes lighter. This diagram plots nationalization on the x-axis and the incumbency advantage coefficients on the y-axis. Each point represents the nationalization level and the incumbency advantage for that year.

```
## Source: local data frame [2 x 7]
## Groups: post_1952 [2]
##
##   post_1952 estimate statistic p.value parameter conf.low conf.high
##       (dbl)    (dbl)     (dbl)   (dbl)     (dbl)    (dbl)     (dbl)
## 1         0    -0.46     -2.98    0.01        34    -0.68     -0.15
## 2         1    -0.54     -3.05    0.01        23    -0.77     -0.18
```

Analyzing the Eras (Pre/Post 1952)

Next, we use the non-parametric bootstrap to look at whether there is a difference in the correlations between the eras (pre/post 1952). `data_1` contains observations that are post-1952 while `data_2` contains observations pre-1952. We find that we cannot reject the null-hypothesis, suggesting that there is no difference between the two correlations.

Figure 1-1.pdf

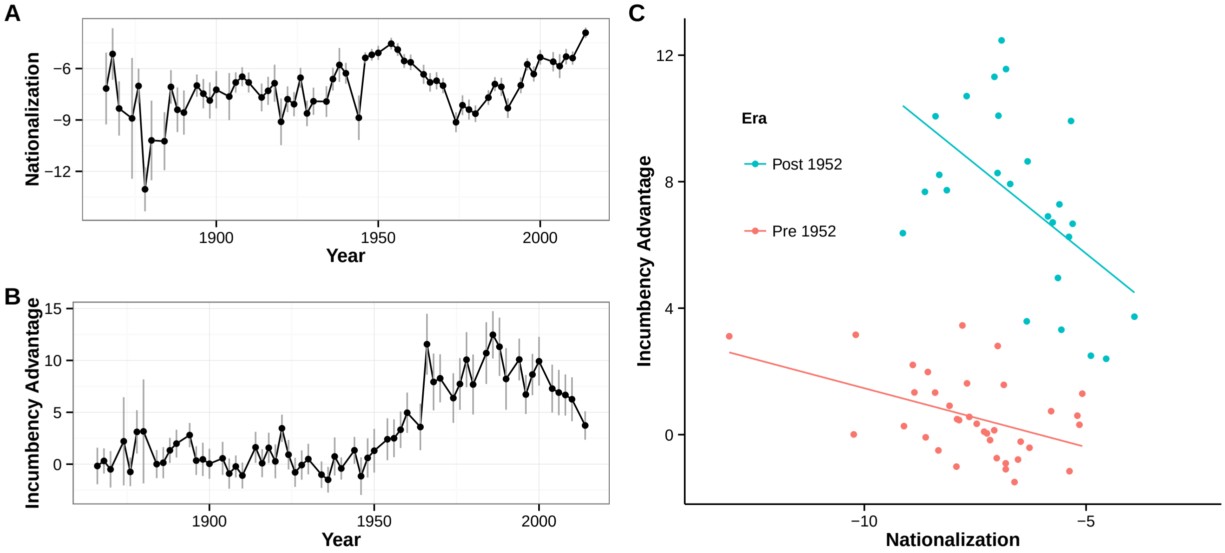

Figure 1: Reproduction of Figure 1 from the main manuscript

```
# This function uses a non-parametric bootstrap to look
# at whether there is a difference in correlations between the eras (pre/post 1952).
# data_1 contains observations that are post-1952. data_2 contains observations pre-1952.
boot_cor_diff <- function(data_1, data_2, n_boot=5000, cor_method = "pearson"){
  bootstraps <- vector(length=n_boot)
  for(i in 1:n_boot){
    sample_index1 <- sample(1:dim(data_1)[1],
                           size = dim(data_1)[1],
                           replace = T)
    sample_index2 <- sample(1:dim(data_1)[1],
                            size = dim(data_1)[1],
                            replace = T)
    boot_cor1 <- tidy(cor.test(data_1[sample_index1,1],
                          data_1[sample_index1,2],
                          method = cor_method))
    boot_cor2 <- tidy(cor.test(data_2[sample_index2,1],
                          data_2[sample_index2,2],
                          method = cor_method))
    bootstraps[i] <- boot_cor1$estimate - boot_cor2$estimate
  }
  return(bootstraps)
}

# Here, we separate the observations by the dummy variable "post_1952"
# to test if the correlations are different.
# We incorporate the above bootstrap function into this test.
post_1952_data <- b_incumbency_adv %>%
  filter(post_1952==1) %>%
  select(nationalization, estimate)
pre_1952_data <- b_incumbency_adv %>%
  filter(post_1952==0) %>%
  select(nationalization, estimate)

pearson_r_diff_bootstrap <- boot_cor_diff(
  data_1 = as.data.frame(post_1952_data),
  data_2 = as.data.frame(pre_1952_data),
  n_boot=10000
)

# Calculate boot-strapped p-values
# from the distribution boot-strapped estimates.
cor_diff_pval <- sum(pearson_r_diff_bootstrap > 0)/10000
cor_diff_pval
```

```
## [1] 0.401
```

Robustness Checks

The 1878 election

As noted in the main text, we were concerned that the pre-1952 correlation between nationalization and the incumbency advantage could be driven by the 1878 election, which is a bit of an outlier. To address this possibility, we examined whether our result changes when we remove that year from our data. The correlation shrinks, but is still significantly different from 0 in the same direction. We also get a similar result if we use the Spearman rank correlation, which is relatively insensitive to outliers.

```
# where we remove the 1878 election.
pearson_cor_pre_1952_minus_1878 <- b_incumbency_adv %>%
  filter(year != 1878 & year < 1952) %>%
  do(tidy(cor.test(.$nationalization, .$estimate)))
pearson_cor_pre_1952_minus_1878
```

```
##     estimate statistic    p.value parameter   conf.low    conf.high
## 1 -0.3357106 -2.047327 0.04865523        33 -0.6016481 -0.002774458
```

```
# Below is a robustness check using the Spearman rank correlation for the elections after
# 1952.
spearman_cor_by_era <- b_incumbency_adv %>%
  group_by(post_1952) %>%
  do(tidy(cor.test(.$nationalization, .$estimate, method="spearman")))
spearman_cor_by_era
```

```
## Source: local data frame [2 x 4]
## Groups: post_1952 [2]
##
##   post_1952   estimate statistic     p.value
##       (dbl)      (dbl)     (dbl)       (dbl)
## 1         0 -0.3842986     10756 0.021308355
## 2         1 -0.5761538      4098 0.003047957
```

Presidential Elections vs. Midterms

It is possible that the results reported here also depend on whether an election is a presidential or midterm election.[1] To investigate this possibility, we use an ordinary least squares regression, where the dependent variable is the incumbency advantage in each year, and the main independent variable is the nationalization of House elections. We also include a binary indicator for whether the election was post-1952 to account for the vertical jump in the post-1952 incumbency advantage (shown in Figure 1c).

In Model 1 below, we control for the effect of presidential vs. midterm elections by using a binary indicator for whether the election was a presidential election. Using this model, we find that the association between our *presidential election* variable and the incumbency advantage is statistically insignficant at conventional levels.

In Model 2 below, we examine whether our nationalization variable works differently in presidential elections (vs. midterms) by interacting *nationalization* with our *presidential election* variable. Using this model, we also find that the interaction between *nationalization* and *presidential election* is statistically insignficant. This suggests that the association between nationalization and the incumbency advantage does not strongly depend on whether or not an election is a presidential election.

```
b_incumbency_adv$pres_election <- 0
b_incumbency_adv$pres_election[b_incumbency_adv$year%in%seq(from = 1868,
                                                              to = 2014,
                                                              by = 4)] <- 1

model1 <- lm(estimate ~ nationalization + pres_election + post_1952,data=b_incumbency_adv)
summary(model1)
```

```
##
## Call:
## lm(formula = estimate ~ nationalization + pres_election + post_1952,
##     data = b_incumbency_adv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9744 -1.0598 -0.0366  1.1164  4.6110
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -4.0129     1.2961  -3.096 0.003040 **
## nationalization  -0.6286     0.1638  -3.837 0.000314 ***
## pres_election    -0.3828     0.4816  -0.795 0.429998
## post_1952         7.5283     0.5199  14.480  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.878 on 57 degrees of freedom
## Multiple R-squared:  0.7876, Adjusted R-squared:  0.7765
## F-statistic: 70.47 on 3 and 57 DF,  p-value: < 2.2e-16
```

```
model2 <- lm(estimate ~ nationalization * pres_election + post_1952 ,data=b_incumbency_adv)
summary(model2)
```

```
##
```

[1] We thank a helpful reviewer for bringing this to our attention.

```
## Call:
## lm(formula = estimate ~ nationalization * pres_election + post_1952,
##     data = b_incumbency_adv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9055 -0.8876 -0.0923  1.1911  4.6370
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -4.4525     1.5817  -2.815  0.00672 **
## nationalization                -0.6917     0.2089  -3.311  0.00163 **
## pres_election                   0.7222     2.2988   0.314  0.75454
## post_1952                       7.5066     0.5253  14.291  < 2e-16 ***
## nationalization:pres_election   0.1550     0.3152   0.492  0.62481
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.891 on 56 degrees of freedom
## Multiple R-squared:  0.7886, Adjusted R-squared:  0.7734
## F-statistic: 52.21 on 4 and 56 DF,  p-value: < 2.2e-16
```

The South

One might also wonder if our analysis was driven by the South, especially since some of the least nationalized elections occur around the end of Reconstruction. To examine this possibliy, we subsetted our data to only include Northern states. Doing this we obtain very similar results, which are shown in Figure 2 below.

```
cq_data <- filter(cq_data, south==0)

swing_variance <- cq_data %>%
  filter(inc3 !=0) %>%
  group_by(year) %>%
  summarise(nationalization = -sd(dv - dvp, na.rm=T),
            nationalization_se = boot_se(n_boot=3000, dv, dvp)
            ) %>%
  ungroup()

b_incumbency_adv <- cq_data %>%
group_by(year) %>%
  do(., tidy(lm(dv ~ dvp + ptynow + inc3, data = . ))) %>%
  filter(term == "inc3") %>%
  ungroup() %>%
  left_join(., swing_variance, by="year") %>% # join the nationalization measure to the
                                              # incumbency advantage coefficients.
  mutate(post_1952 = as.numeric(year>1952)) # dummy varaible for whether the year is after
                                            # 1952 or not. 1 = post 1952, 0 = otherwise.
```

```
# The following code is used to create Figure 1.

# Add in markers for the years 1950 and 1954
# This will be used in figure 1c
b_incumbency_adv$marker_break <- NA
```

```
b_incumbency_adv$marker_break[b_incumbency_adv$year==1950] <- 1950
b_incumbency_adv$marker_break[b_incumbency_adv$year==1954] <- 1954

# Plot our nationalization measure over time.
nationalization_year_plot <- ggplot(data=swing_variance,
                                    aes(x= year,
                                        y= nationalization)) +
  geom_line() +
  geom_pointrange(aes(ymax=nationalization + 1.96*nationalization_se,
                      ymin=nationalization - 1.96*nationalization_se),
                  color="darkgrey") +
  geom_point() +
  theme_bw() + ylab("Nationalization") + xlab("Year")


# Plot the incumbency advantage coefficients over time.
incumbency_adv_year_plot <- ggplot(data=b_incumbency_adv,
                                   aes(x=year, y=estimate)) +
  geom_line() +
  geom_pointrange(aes(ymax=estimate + 1.96 * std.error,
                      ymin=estimate - 1.96* std.error),
                  color="darkgrey") +
  geom_point() +
  theme_bw() + ylab("Incumbency Advantage") + xlab("Year")


# Calculate the correlation between nationalization and the incumbency advantage by era.
pearson_cor_by_era <- b_incumbency_adv %>%
  group_by(post_1952) %>%
  do(tidy(cor.test(.$nationalization, .$estimate)))
round(pearson_cor_by_era, 2)
```

```
## Source: local data frame [2 x 7]
## Groups: post_1952 [2]
##
##   post_1952 estimate statistic p.value parameter conf.low conf.high
##       (dbl)    (dbl)     (dbl)   (dbl)     (dbl)    (dbl)     (dbl)
## 1         0    -0.43     -2.76    0.01        34    -0.66     -0.12
## 2         1    -0.48     -2.60    0.02        23    -0.73     -0.10
```

```
# Create a list with the correlation coefficients to be used
# in the plot.
l <- list(r1 = format(pearson_cor_by_era$estimate[2], digits = 2),
          pval1 = format(pearson_cor_by_era$p.value[2], digits= 1),
          r2 = format(pearson_cor_by_era$estimate[1], digits = 2),
          pval2 = format(pearson_cor_by_era$p.value[1], digits= 1)
)

# Make text labels for the legend for figure 1c.
eq1 <- substitute(italic(r) == r1*","~~italic(p) < pval1,l)
eqstr1 <- as.character(as.expression(eq1))
eq2 <- substitute(italic(r) == r2*","~~italic(p) < pval2,l)
eqstr2 <- as.character(as.expression(eq2))
```

```r
# Plot the relationship between nationalization and the incumbency advantage.
nationalization_incumbency_path_plot <- ggplot(data=b_incumbency_adv,
                                               aes(x=nationalization,
                                               y=estimate,
                                               group=post_1952))+
  geom_point(aes(colour=as.factor(post_1952))) +
  geom_smooth(aes(colour=as.factor(post_1952), group=post_1952), method="lm", se=F)+
  theme_classic() +
  xlab("Nationalization") +
  ylab("Incumbency Advantage") +
  xlim(c(-13.5,-2.5)) +
  scale_color_discrete(name="Era", labels = c("Pre 1952", "Post 1952"),
                       guide = guide_legend(reverse=TRUE))+
  theme(legend.key.height = unit(1.5, "cm"),
        legend.title.align=0,
        legend.position=c(.2, .65))

figure1 <- ggdraw() +
  draw_plot(nationalization_year_plot,   x= 0.01, y= 1/2, width = .50, height = .5) +
  draw_plot(incumbency_adv_year_plot, x= 0.01, y= 0  , width = .50, height = .5) +
  draw_plot(nationalization_incumbency_path_plot, x=.50, y=0, .50, 1) +
  draw_plot_label(c("A", "B", "C"), c(0, 0, 0.50), c(1, 0.5, 1), size = 15)
figure1
```

Figure 1 on the south only-1.pdf

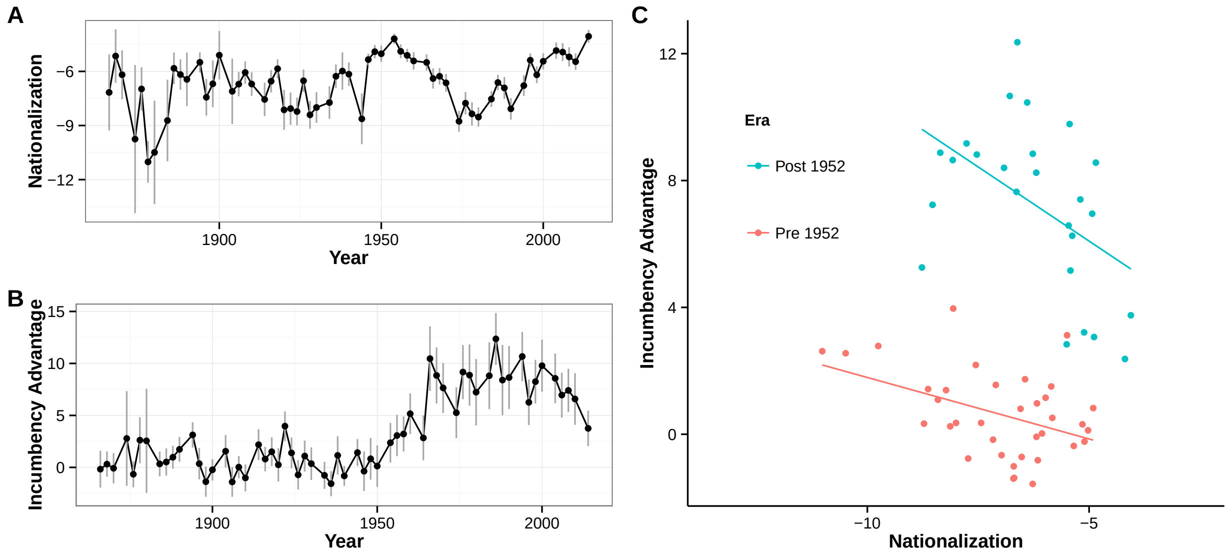

Figure 2: Reproduction of Figure 1 from the main manuscript, using only Northern states