Measuring Segregation in High Dimensions

Matthew Gentzkow, Chicago Booth and NBER
Jesse M. Shapiro, Brown and NBER
Matt Taddy, Chicago Booth

There's a big lit on measuring change in polarization/segregation.

For example:

- The d-vector $\mathbf{c}_{tr} = [c_{tr1} \dots c_{trd}]$ counts the number of members of race r in each neighborhood j at time t.
  Or, compare across cities/countries/school districts/etc.
- Segregation measure maps counts $\{\mathbf{c}_{tr}\}_{r=1}^{R}$ into scalar s_t
- Use $s_1 \dots s_T$ to answer questions like

  "Has this city become more segregated over time?"

Today's talk is about how to build s_t when d is really big.

Current indices are derived from sets of stated desirable properties.

Example: Atkinson index (Frankel and Volij 2010)

$$s_t = 1 - \sum_j \sqrt{\frac{c_{t0j}}{m_{t0}} \frac{c_{t1j}}{m_{t1}}}$$

for two groups (e.g., races) $r = 0$ and $r = 1$, where $m_{tr} = \sum_j c_{trj}$.

This builds on an earlier literature that provides isolation, dissimilarity, mutual information, gini, and other indices.

Are we after properties of the sample or of the DGP?

That is, if the data are consistent with race-blind assignment do we want to say that segregation is low?

e.g., Cortese et al. (1976), Carrington and Troske (1997) both do.

Think of Atkinson as

$$\hat{s}_t = 1 - \sum_j \sqrt{\hat{p}_{t0j}\ \hat{p}_{t1j}}$$

where $\mathbf{p}_{trj}$ is the true probability that a member of r lives in j at t.

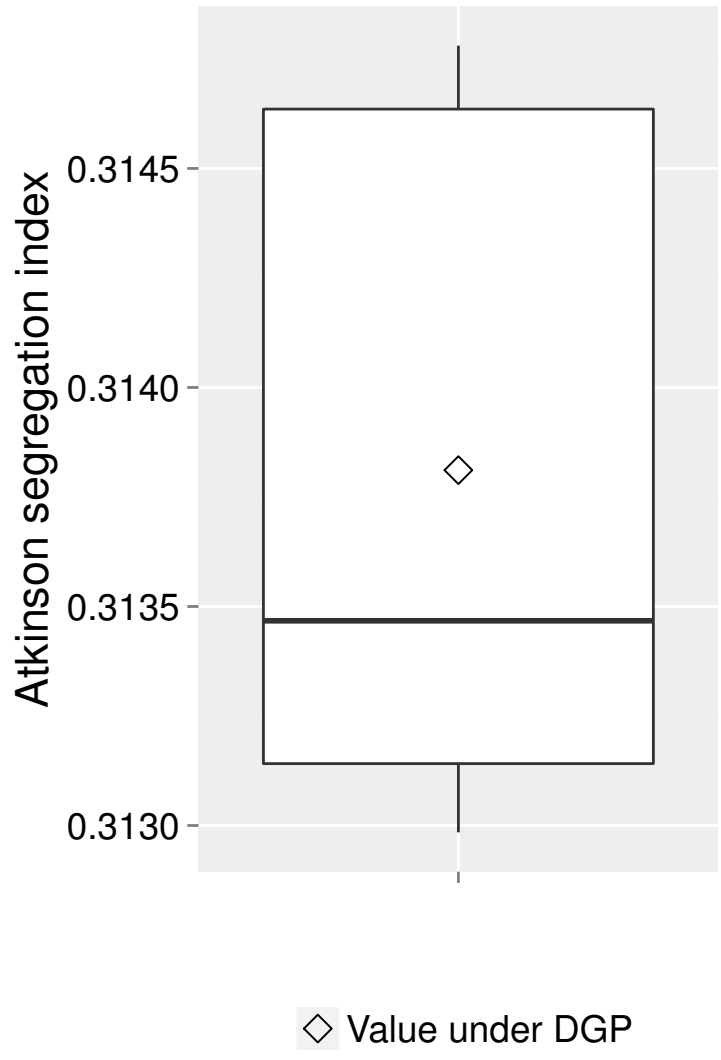
This distinction is unimportant if $\hat{s}_t \approx s_t$.

It is so for residential segregation because zipcodes are large.

But estimation bias becomes very important as units get smaller...

Number of Districts Small

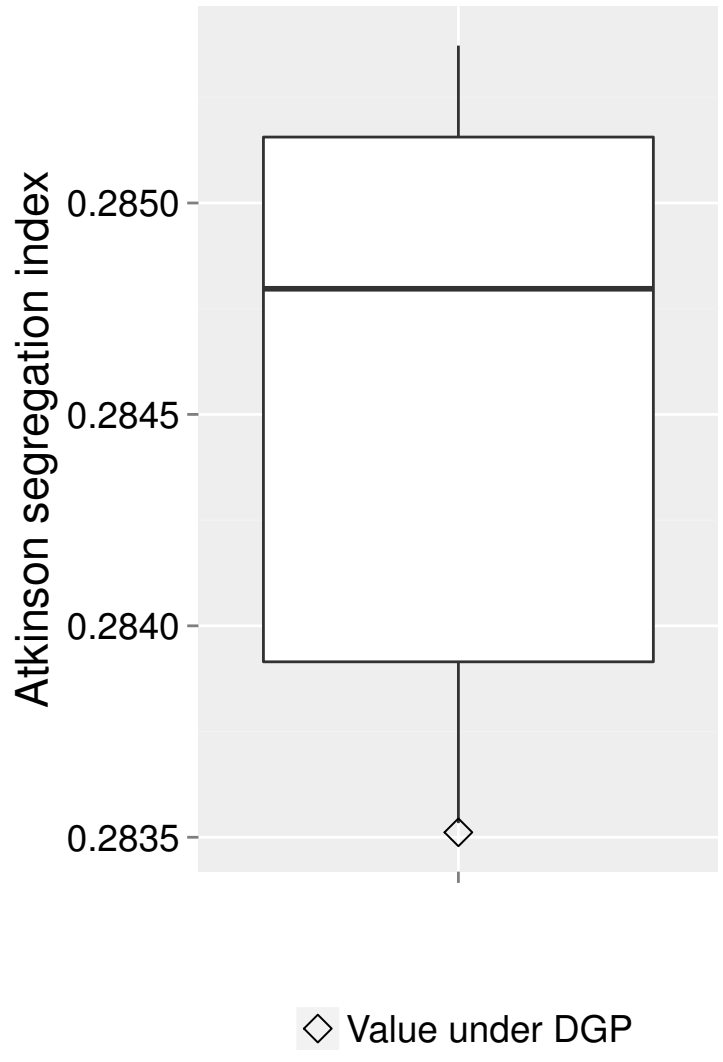
1 million students



100 districts.

Number of Districts Medium

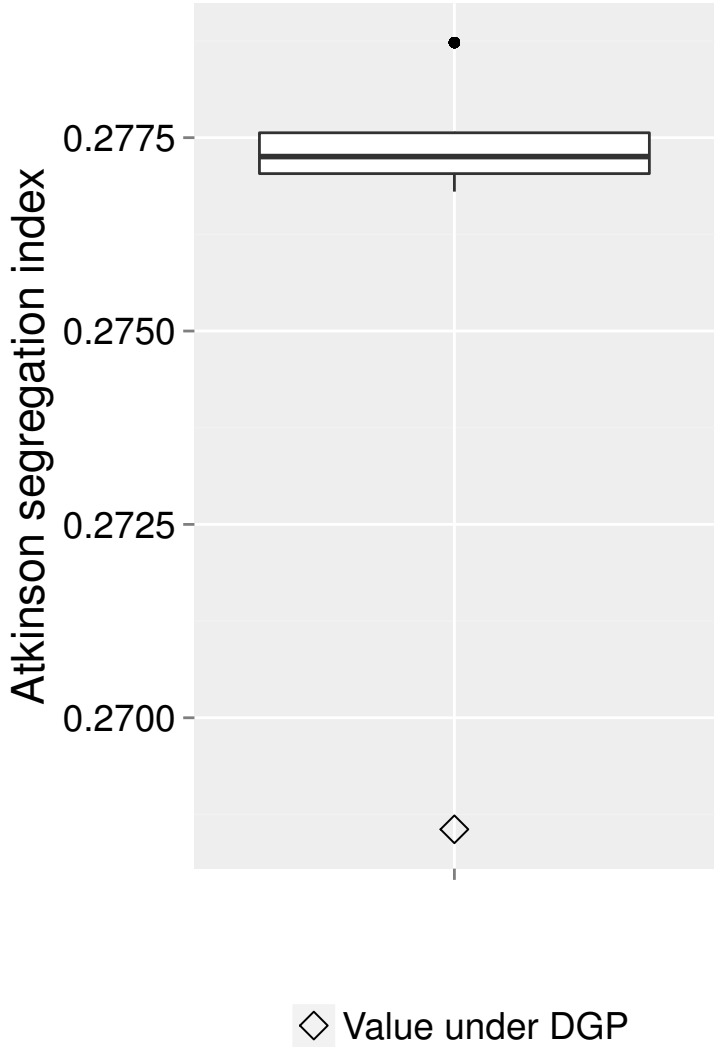
1 million students



1000 districts

Number of Districts Large

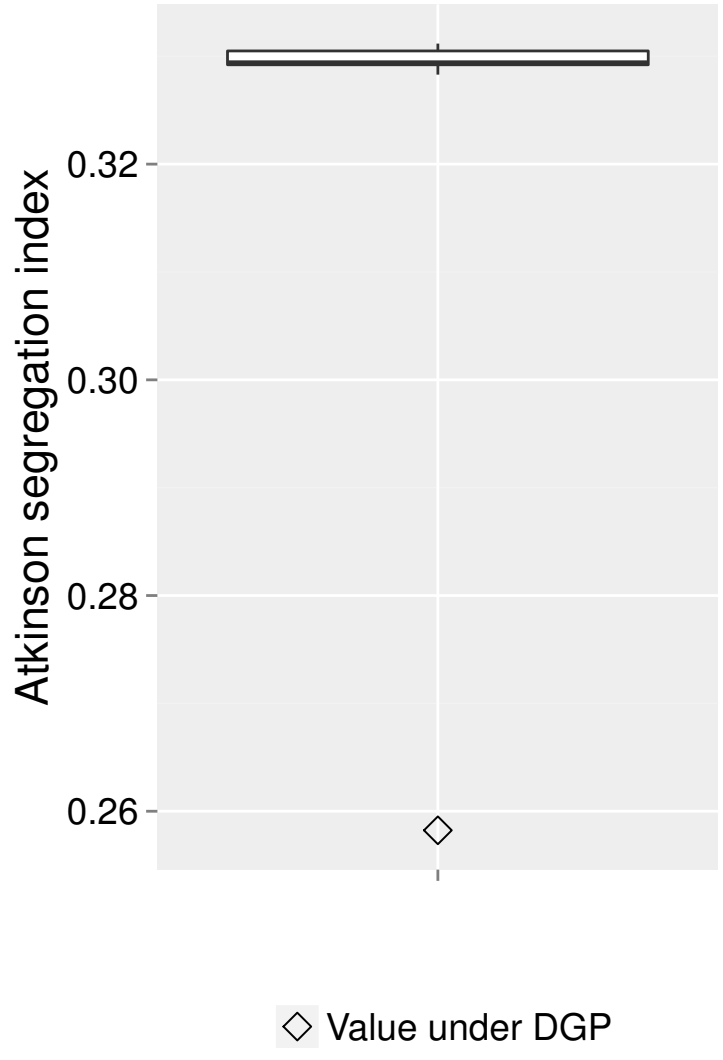
1 million students



10,000 districts

Number of Districts Very Large

1 million students



100,000 districts

This paper

We build a utility-based model of assignment to groups,
then define a segregation measure within this model.

$\Rightarrow$ a Big-response-dimension multinomial logistic regression.

- Use penalization to control finite-sample behavior
- Implement distributed estimation for scalability

We control for 100s of covariates and 1000s of random effects.

Applications to polarization/segregation in

- congressional text (gop vs. dem, south vs. north)
- internet browsing (white vs. black)
- grocery store purchases (college educated vs. not)

Multinomial logit model: Each individual i in period t makes m_{it} choices over units j to maximize utility

$$\eta_{itj} + \varepsilon_{itj} = \alpha_{jt} + \mathbf{u}_{it}'\boldsymbol{\gamma}_{jt} + \boldsymbol{\varphi}_{jt}' r_{it} + \varepsilon_{ijt}$$

where:

- α_{jt} is unit-specific utility intercept
- $\mathbf{u}_{it}$ are covariates and $\boldsymbol{\gamma}_{jt}$ are associated loadings
- $r_{it} \in \{0, 1\}$ is an indicator for group membership and φ_{jt} are associated loadings.
- ε_{ijt} is T1EV random utility component

Alternatively, $\mathbf{c}_{it} \sim \mathrm{MN}(\mathbf{p}_{it}, m_{it})$ with $p_{itj} = e^{\eta_{itj}} / \sum_l e^{\eta_{itl}}$.

Motivated by congressional example define the partisanship z_{it} of individual i at time t (after observing $m_{it} = \sum_j c_{itj}$ choices) as

$$z_{it} = \boldsymbol{\varphi}_t' \mathbf{c}_{it} / m_{it},$$

utility gain to $r = 1$ relative to an $r = 0$ from individual i's choices.

z_{it} is also a model-based sufficient statistic for group membership:

$$r_{it} \perp\!\!\!\perp \mathbf{c}_{it} \mid z_{it}, \mathbf{u}_{it}, m_{it}.$$

'Sufficient projection' z_{it} contains all info in $\mathbf{c}_{it}$ relevant to r_i.

Finally, we can measure segregation (polarization) as difference in mean partisanship between those with $r = 1$ and those with $r = 0$.

This builds on work in text analysis where $\mathbf{c}_i$ are word counts and $\mathbf{r}_i$ is any multivariate set of document attributes of interest.

- negative/neutral/positive, :-) vs. :-(
- political affiliation, ideology, vote-share
- review ratings on business quality, food, service
- usefullness, funniness, coolness ratings by others

The big logits are an alternative to latent factor models for text.

SPs $\mathbf{z}_i$ are supervised factors to be used inside prediction and inference systems. See Taddy MNIR (2013) and DMR (2015).

Like in any regression, the φ are measuring *partial correlation*.

A regression like any other, except the response is super HD.

- Computationally impractical to estimate exactly MLE
- Overfit without a regularization penalty (like Atkinson)

We'll discuss each challenge in turn.

Distributed Multinomial Regression

We approximate the MN likelihood with *independent* Poissons:

$$c_{itj} \sim \text{Po}(\ m_{it} e^{\eta_{itj}}\ )$$

$\Rightarrow$ you can estimate each regression fully independently!

This works because MN dependence is *only induced by totals*.

DMR is equivalent to MN logit in a variety of simple examples, and is shown empirically to perform well in more complex settings.

Everything in distribution: estimation, penalization, selection ...

More precisely, start from the Poisson:

$$c_{ij} \overset{ind}{\sim} \mathrm{Pois}\left(\exp\left[\mu_i + \eta_{ij}\right]\right)$$

where μ_i is a 'verbosity' nuisance parameter.

This model leads to

$$\Pr\left(\mathbf{c}_i \mid m_i\right) = \frac{\prod_j \mathrm{Po}\left(c_{ij}; \exp\left[\mu_i + \eta_{ij}\right]\right)}{\mathrm{Po}\left(m_i; \sum_l \exp\left[\mu_i + \eta_{il}\right]\right)} = \mathrm{MN}(\mathbf{c}_i;\ \mathbf{q}_i, m_i)$$

Thus, given m_i, Poisson and MN imply the same model.

DMR fixes $\hat{\mu}_i = \log m_i$, so LHD factorizes to independent Poissons.

Big Data: focus computation on the bits that are hard to measure.

Penalization

We also place L_1 estimation penalties on key parameters.

Partisanship loadings are decomposed

$$\varphi_{jt} = \bar{\varphi}_j + \sum_{k=1}^{T} \tilde{\varphi}_{jk} \mathbb{1}_{t>k}$$

And the coefficients in this spline are penalized

$$c(\varphi_{tj}) = \lambda_j \left(|\bar{\varphi}_j| + \sum_k |\tilde{\varphi}_{jk}| \right)$$

We select λ_j using a BIC within each Poisson regression.

When $\mathbf{u}_{it}$ also gets really HD, we'll penalize elements of γ_{jt}.

Data example: US Congressional Record, 1872-2009

Full record of everything said on the floor of the House or Senate

Use automated script to identify speaker and tag with metadata (party, etc.)

Use some rules of thumb to remove procedural phrases

- "I yield the remainder of my time..."

Turn into counts of two-word phrases less stems and stopwords

- "war on terrorism" and "war on terror" become "war terror"

How has polarization in political speech changed over time?

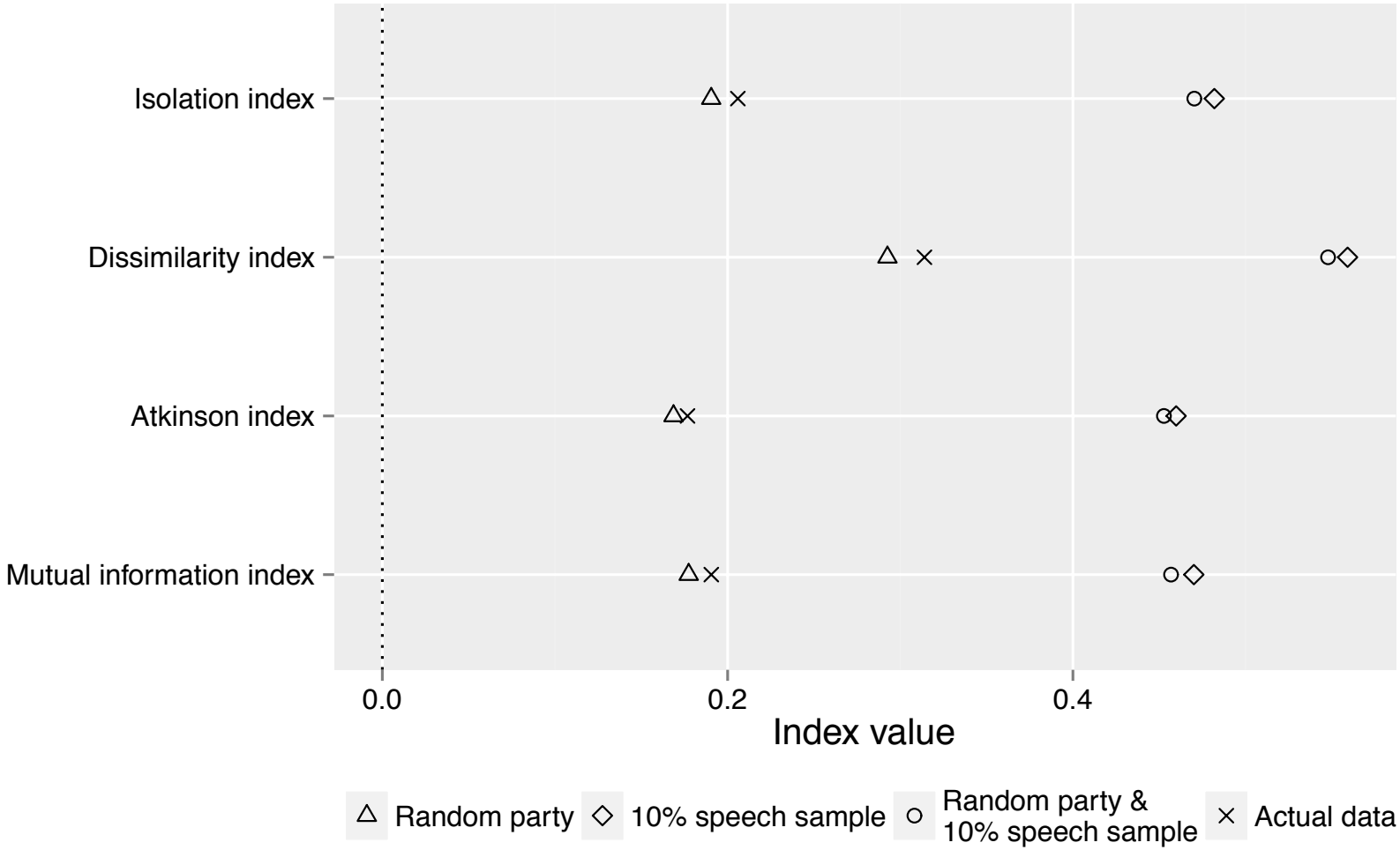
The concept here is of speakers across parties:

- using different words to describe the same thing
  `tax.cut`/`tax.break`, `war.on.terror`/`war.in.iraq`
- choosing to focus on different substantive topics
  `stem.cell`, `african.american`, `soldier.sailor`.

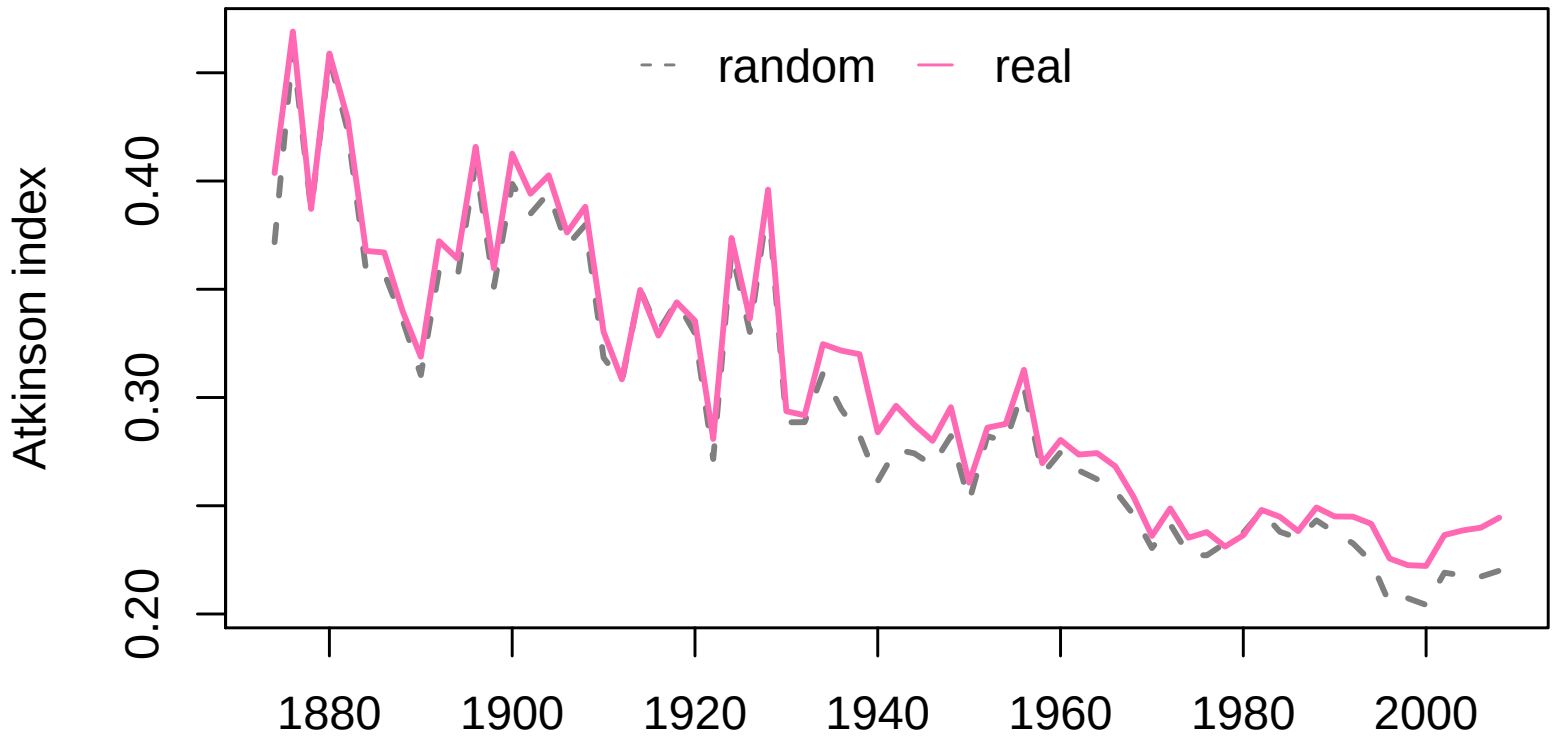
And doing this because of party membership or ideology.

Say $r_{it} = 1$ if speaker i is republican at t, 0 if democrat.

Existing Approaches

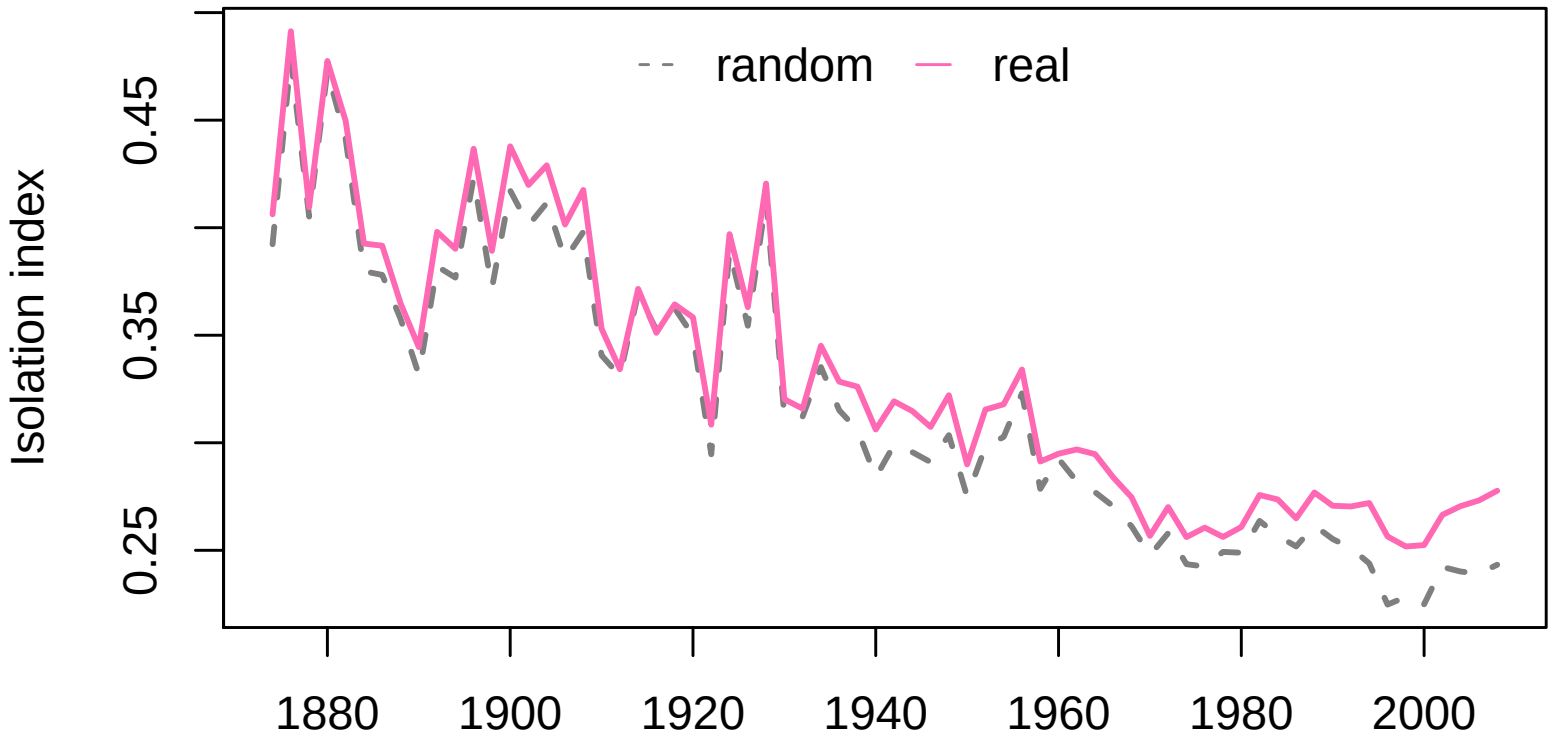


Atkinson



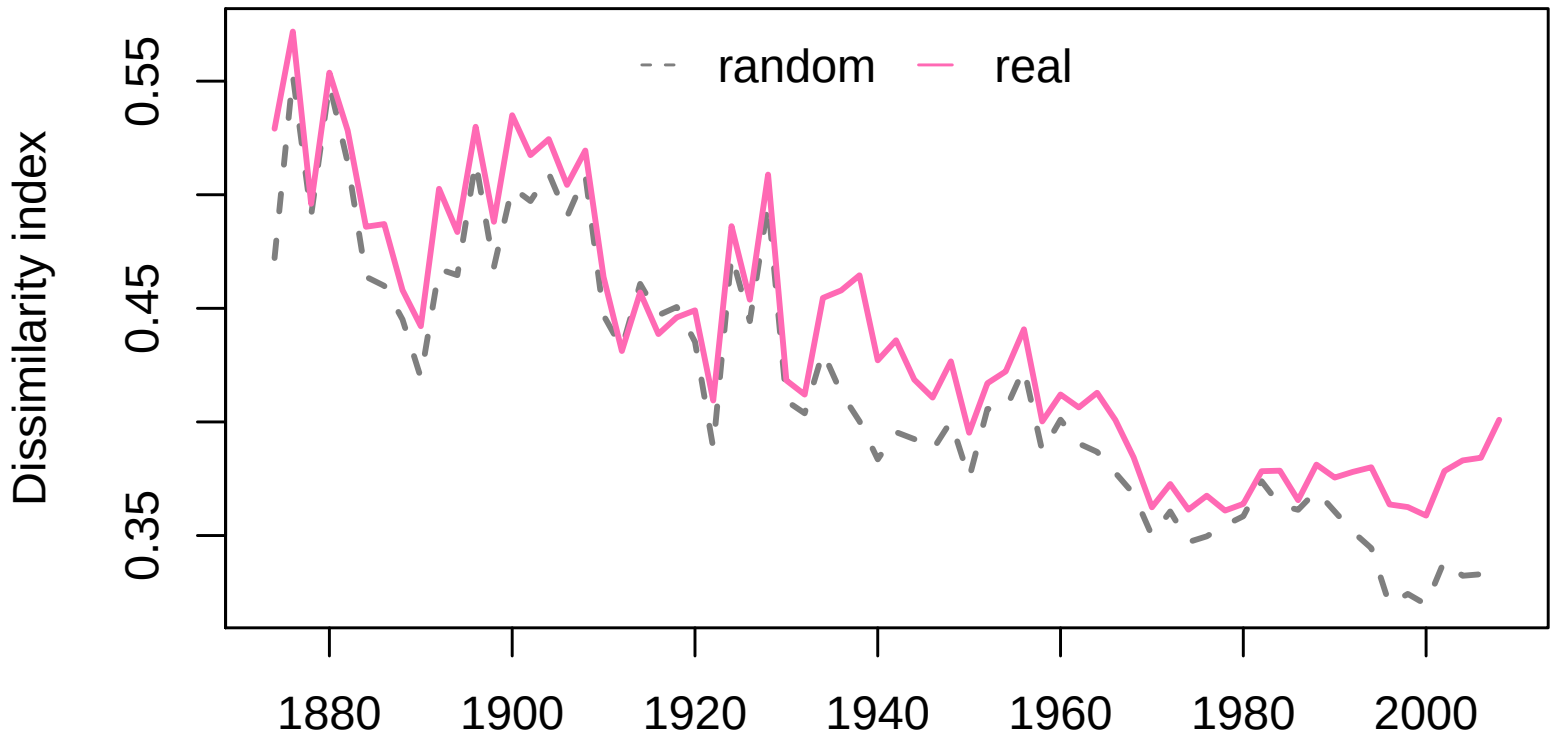
$$\hat{s}_t = 1 - \sum_j \sqrt{\hat{p}_{t1j}\ \hat{p}_{t0j}}$$

Isolation



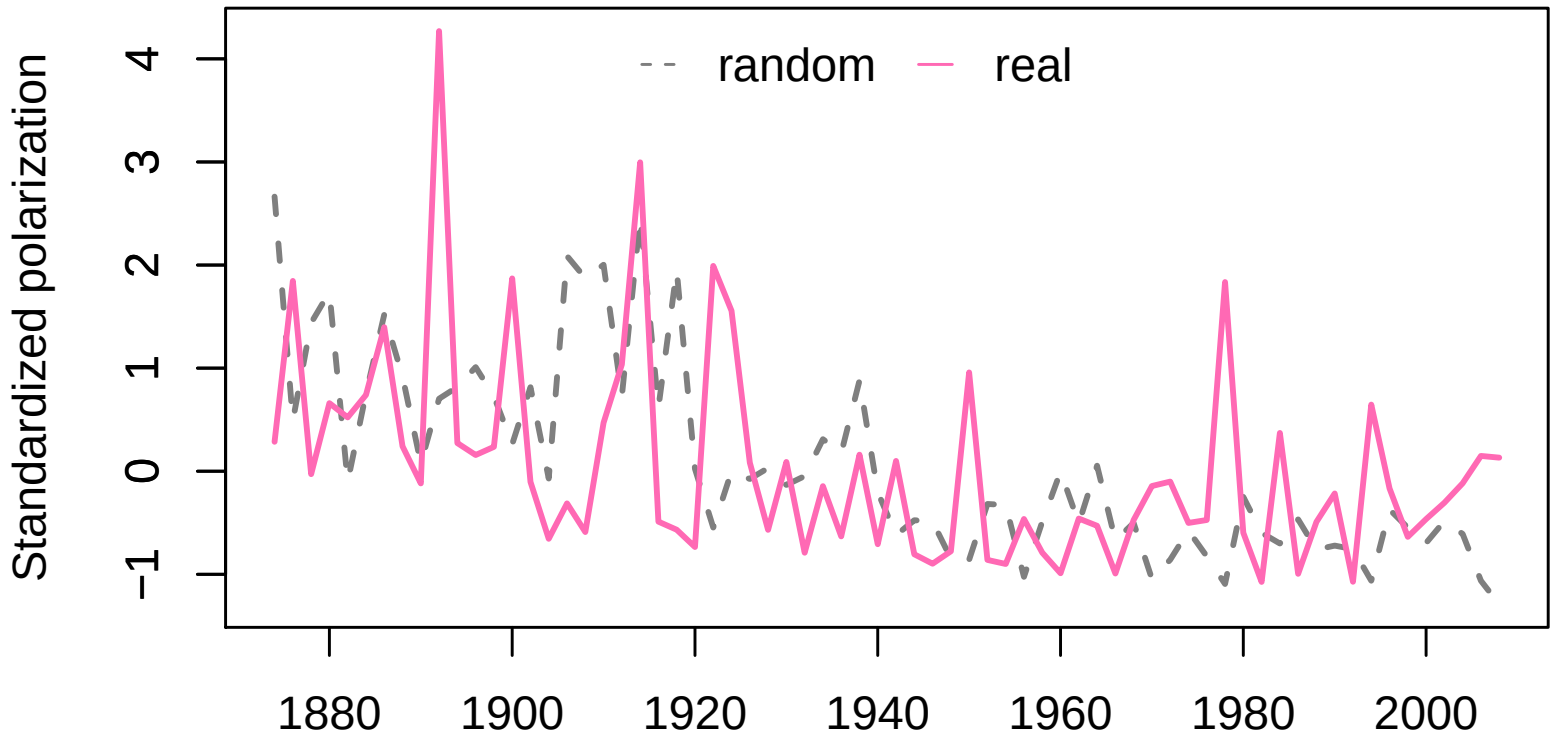
$$\hat{s}_t = \sum_j (\hat{p}_{t1j} - \hat{p}_{t0j}) \frac{c_{t0j}}{c_{t1j} + c_{t0j}}$$

Dissimilarity



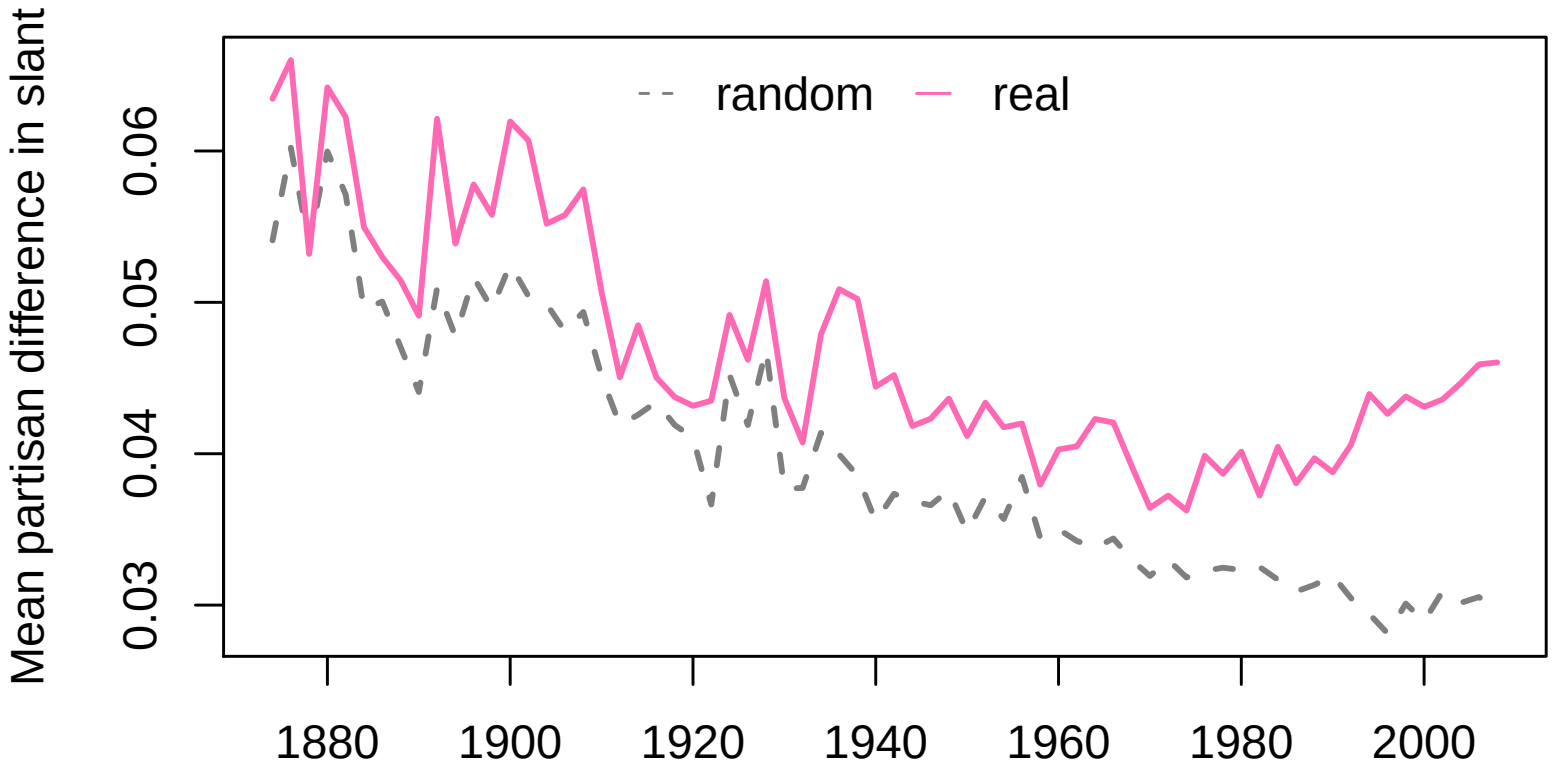
$$\hat{s}_t = \frac{1}{2} \sum_j \left| \hat{p}_{t1j} - \hat{p}_{t0j} \right|$$

Brookings



Jensen et al. (2010)

Slant



Gentzkow and Shapiro (2010)

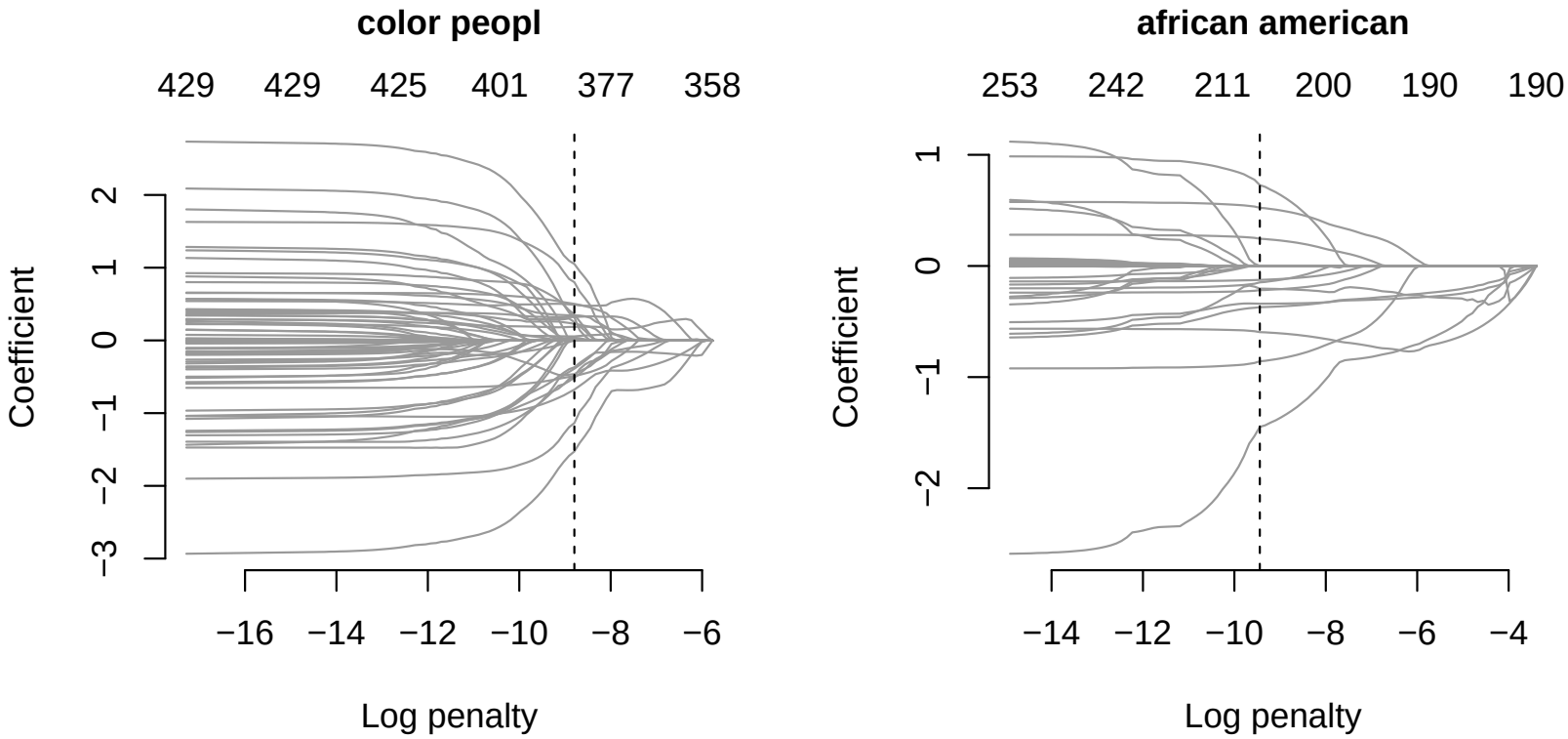
Instead, we'll fit our big logit model.

r_{it} indicates gop/dem and φ_{tj} moves 'smoothly' in t.

Partisanship is defined as segregation in speech by party that is not explainable by our set of controls.

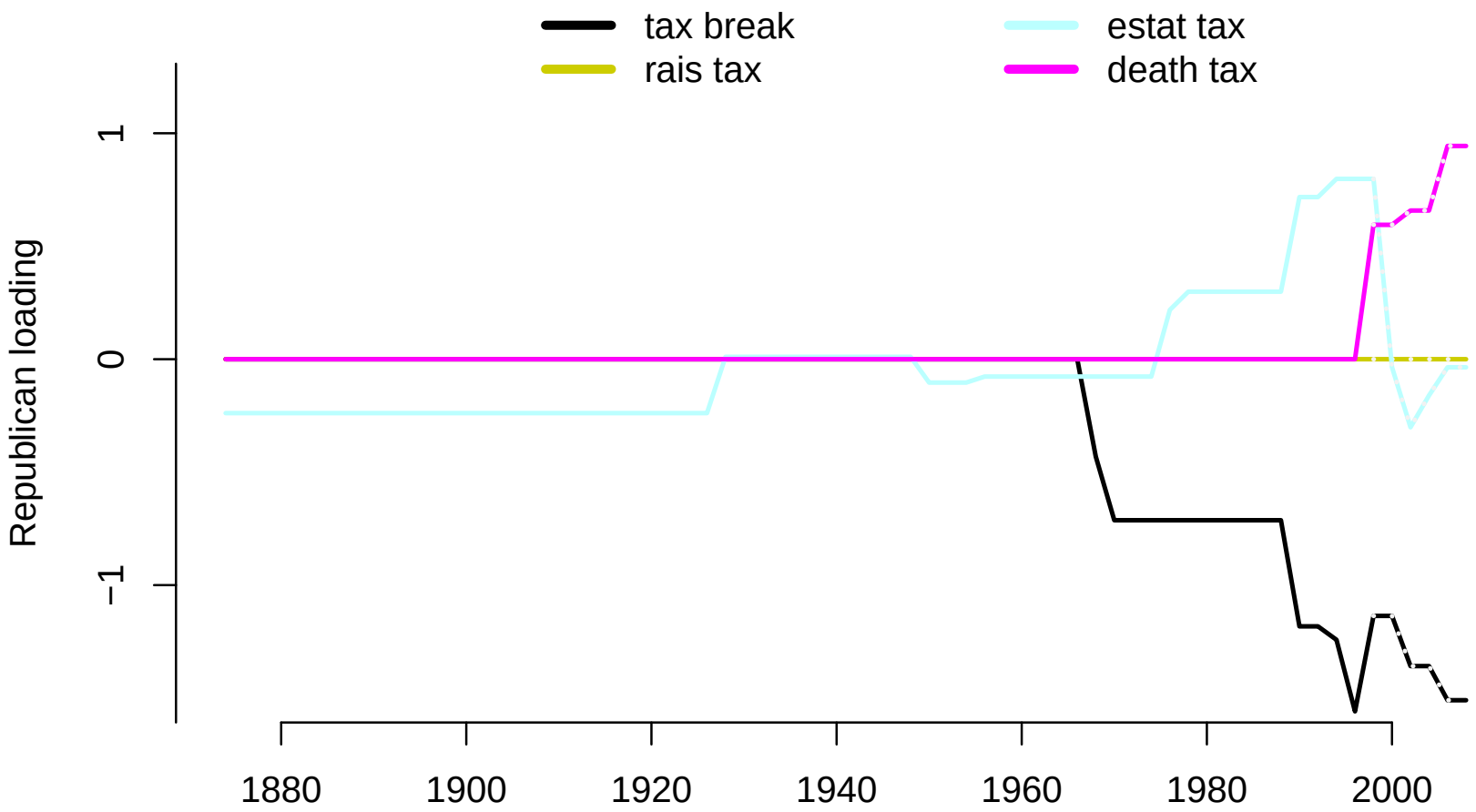
- $\mathbf{u}_{it}$ contain state, chamber, indicator for majority party.
- We allow region effects to vary with time
- check robustness to speaker random effects.

Poisson regression regularization paths



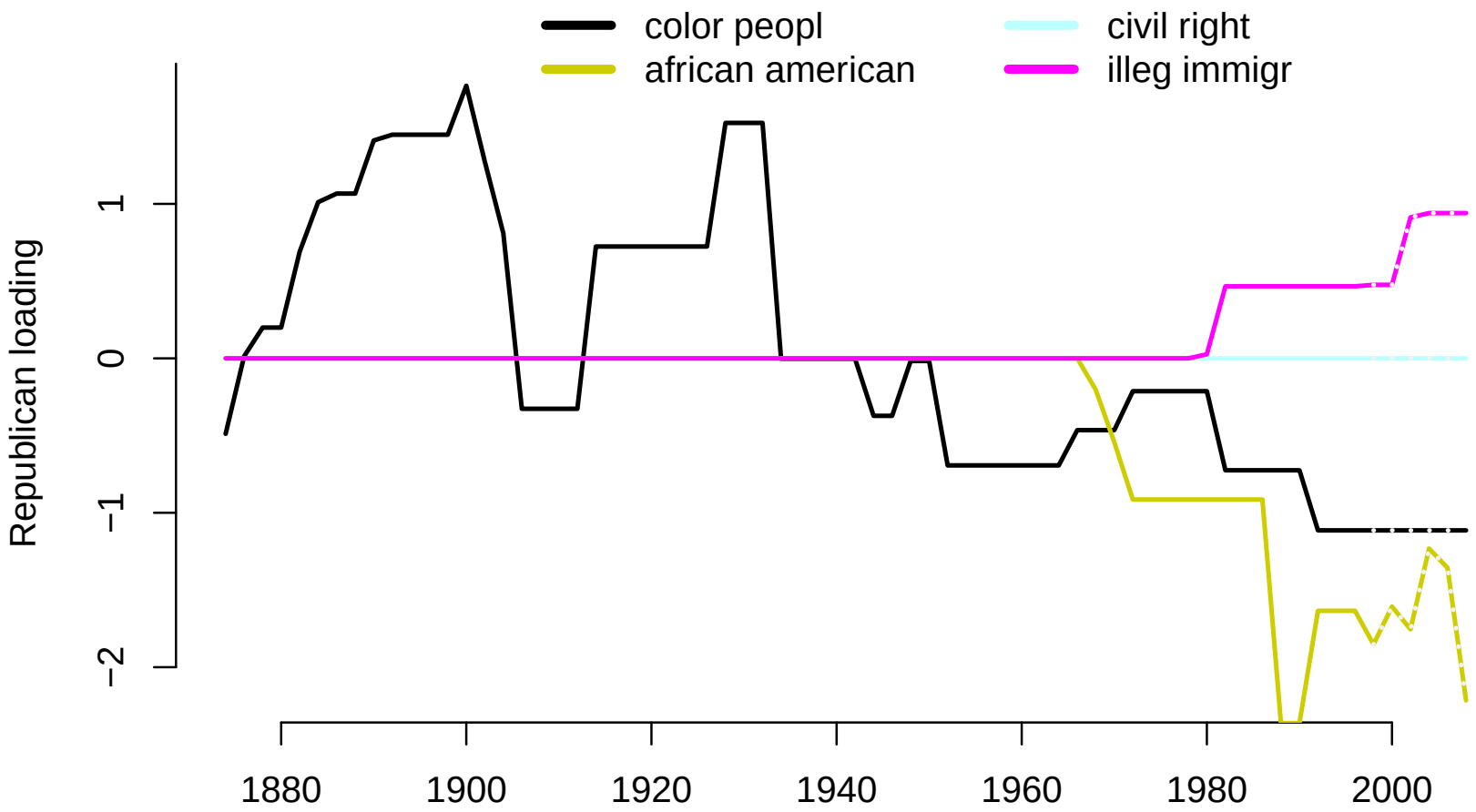
We've run a separate Poisson regression for each phrase.
The code runs a MapReduce routine using `dmr` for R.
BIC selection occurs within reducer, and is marked here.

Dynamic Phrase-Party Loadings: Tax



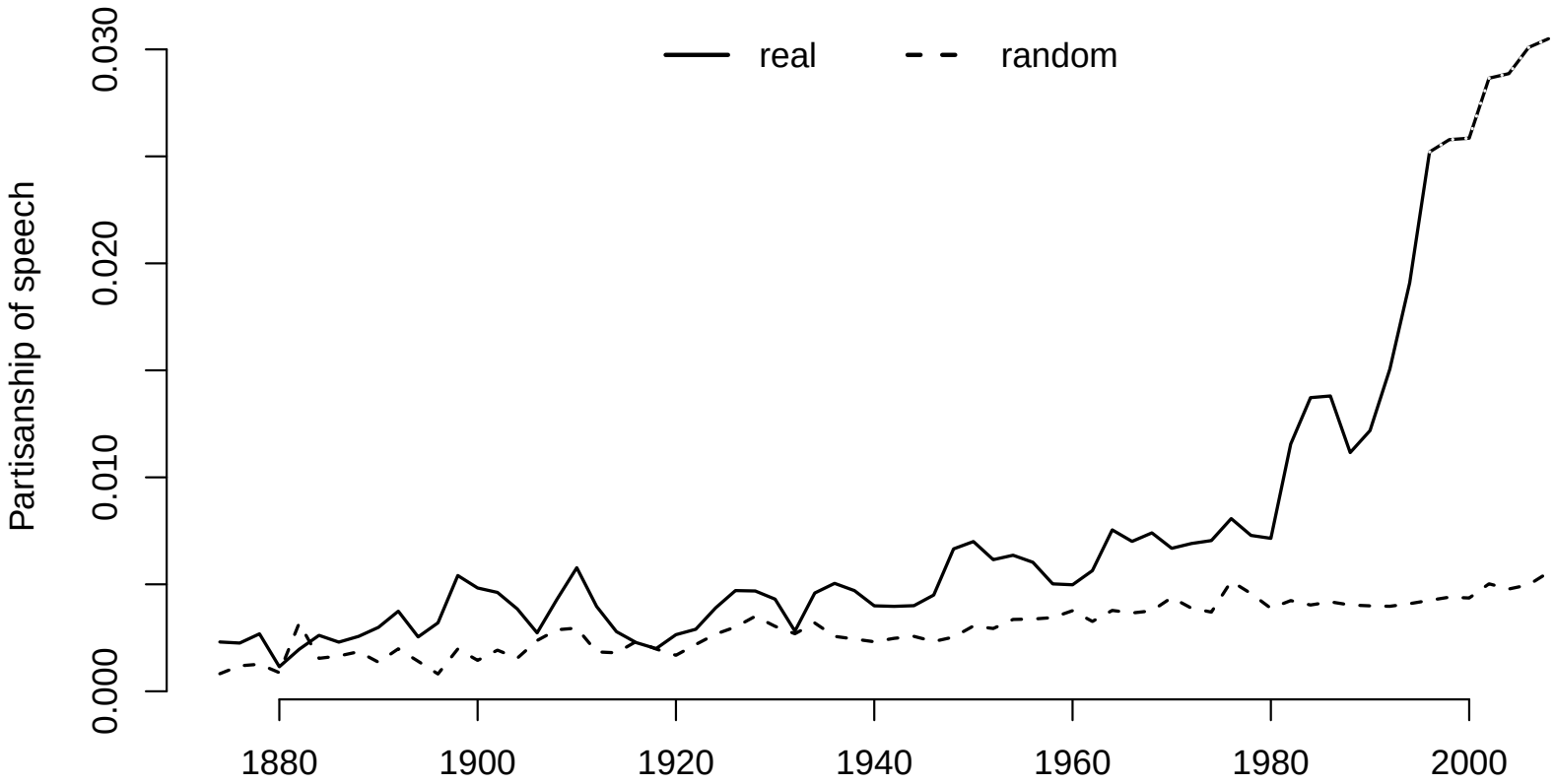
The resulting fit has φ_{tj} changing as a step function in t.

Dynamic Phrase-Party Loadings: Race



For this example, partisanship is robust to fixing $\varphi_{tj} = \varphi_j$.

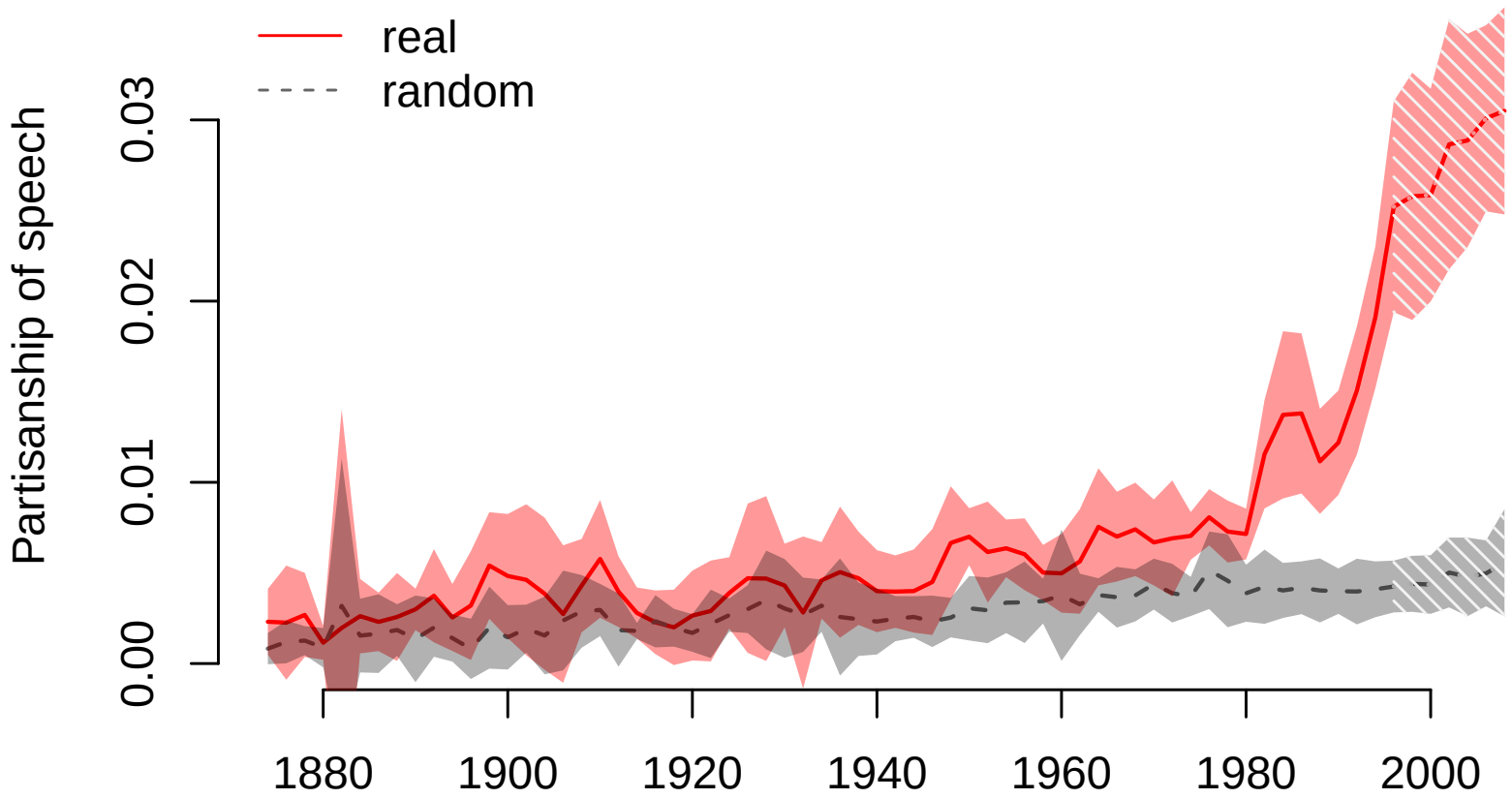
Polarization Results: Baseline Specification



We take $\bar{z}_{rt} = \frac{1}{n_{rt}} \sum_{i:r_i=r} z_{it}$ for each party in each session, and the difference is our partisanship index.

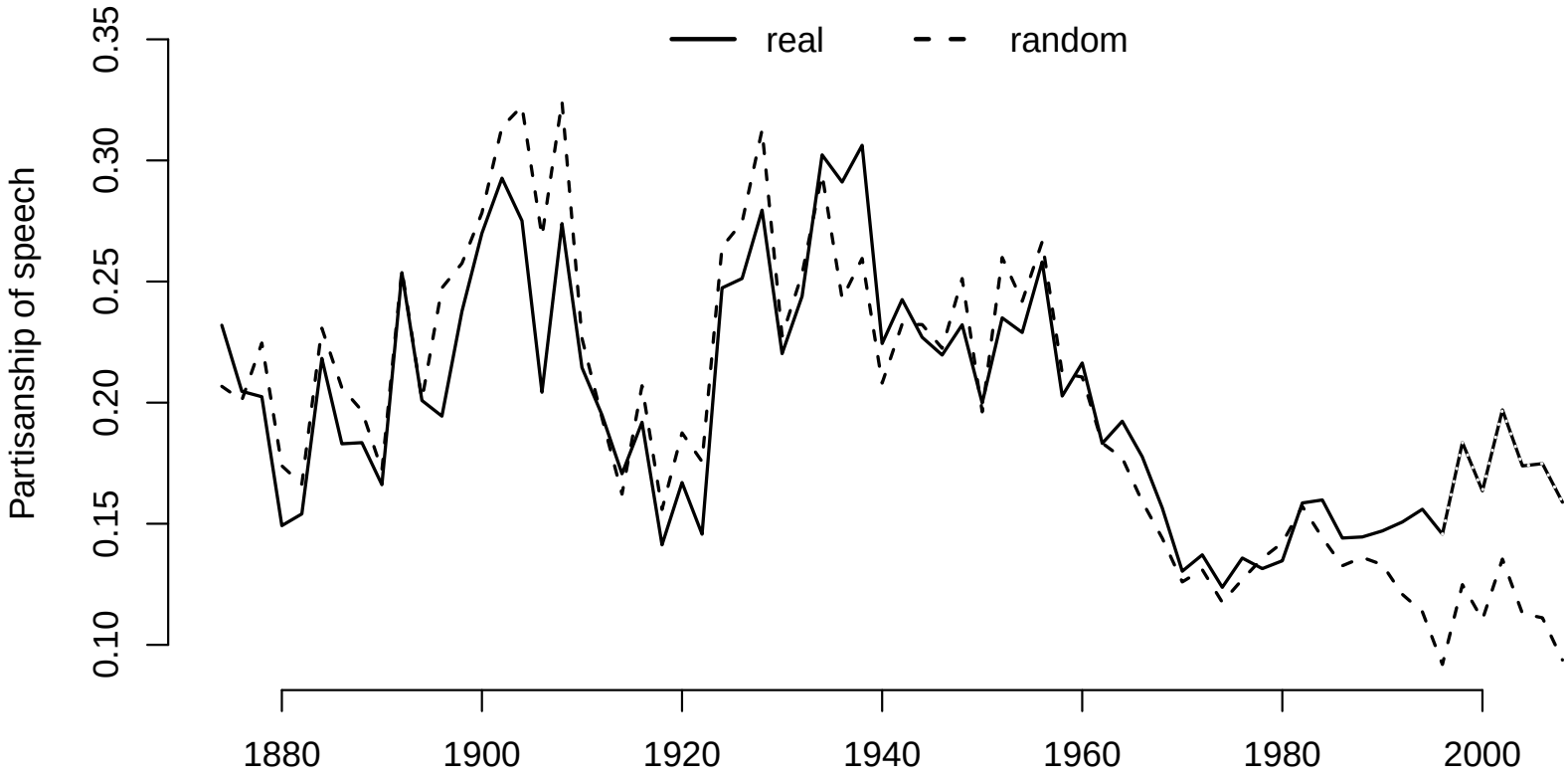
Partisanship of rhetoric has exploded since around 1980.

Nonparametric Bootstrap



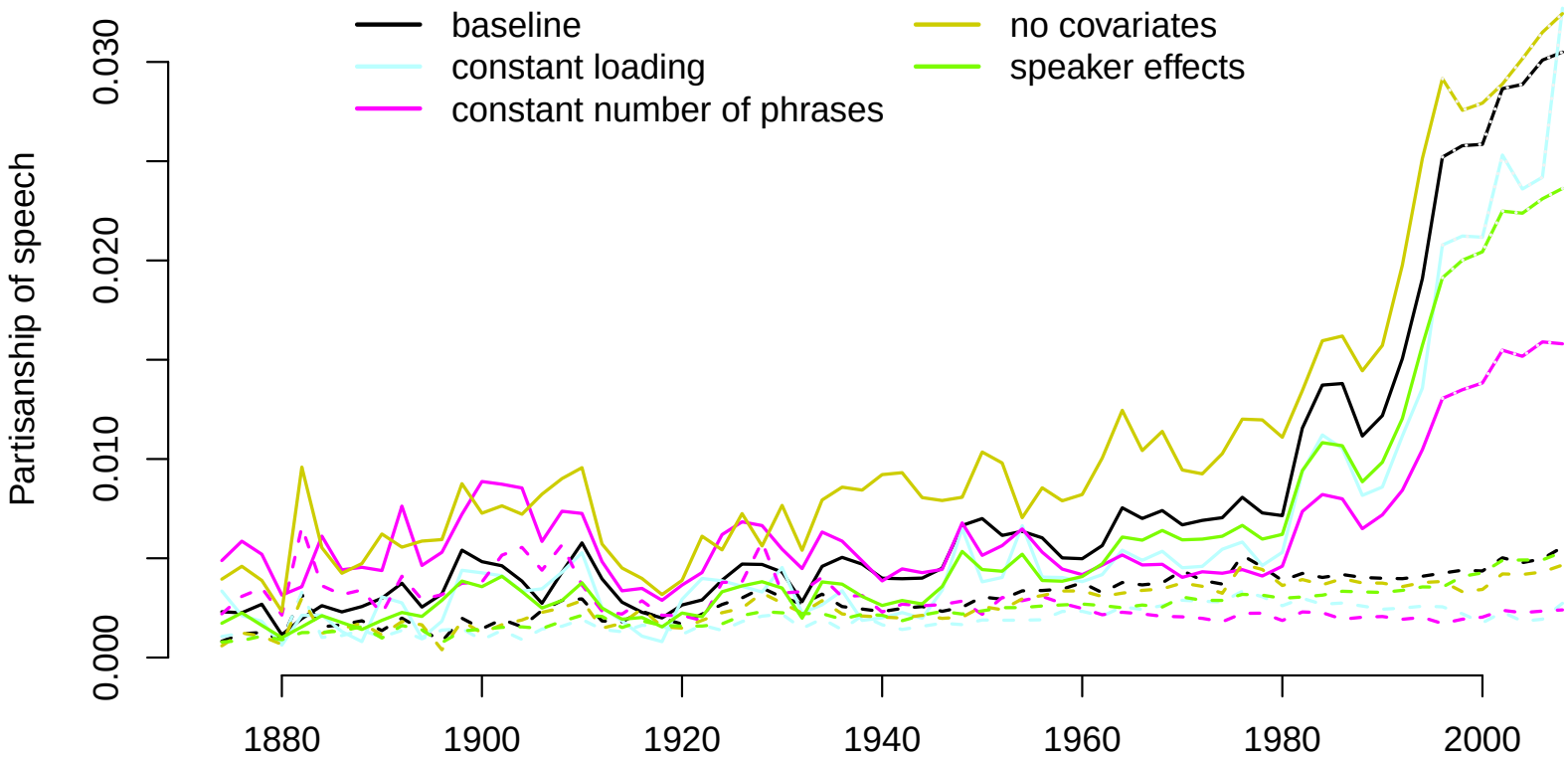
1980 is also when real partisanship becomes more than $2SE$ away from that for random permutations.

Penalization is a necessary ingredient



These are the results corresponding to the end of each lasso path (most complex model)

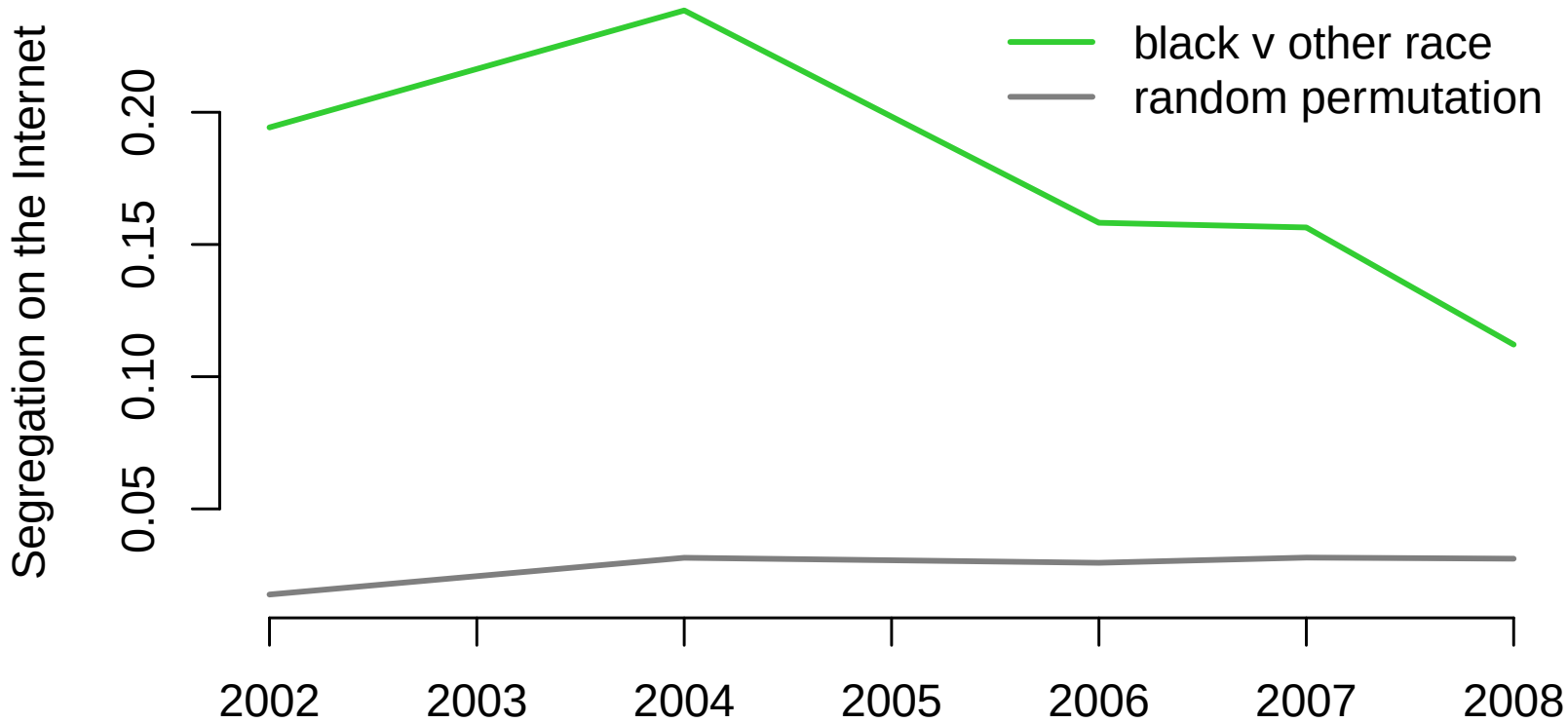
Robustness



With BIC penalty selection, the same shape holds under a variety of specifications. But notice: the controls do make a difference.

Another application: racial segregation on the internet

ComScore browser histories on the websites used in Gentzkow + Shapiro (2011), making visit-counts a function of r_{it} black/white.



Segregation online is high, but has been dropping since 2004.

Conclusion

We've written down a model for choices,
and defined segregation in terms of that model.

We can specify time dynamics and covariates inside this model.

The same techniques that allow machine learners to avoid overfit
in prediction can be used to recover representative model fits.

Thanks!