Measuring ideological segregation/polarization from text data (US Congress Records 1873-2016)

Lucas Girard (CREST-ENSAE)

SICSS-Paris – Institut Polytechnique de Paris (June 28, 2022)

Last week and yesterday: Conclusion

- obtain text data (scrapping it on the Web or elsewhere while complying with ethical rules of research)
- study some frameworks and models in natural language processing (NLP)

Today: application to a specific field, namely the measure of speech polarization: *to what extent (quantification) two (exogenous) groups use different words when they speak?*

- *General points about NLP and text analysis:*
  - Which representation/structuring of texts is the most relevant to answer a given question?
  - What impact of pre-processing?
- *Specific points about speech polarization:*
  - Methods and main results following mostly Gentzkow, Shapiro, Taddy; Econometrica 2019 (GST)
  - Comparison with another methodology: D'Haultfœuille, Girard, Rathelot; mimeo 2021 (DGR)

Motivation

Data and processing

Naive index and methodological issues (small-unit bias)

GST's method

DGR's method

Results and impact of processing/dictionary selection

Extrapolated estimators to avoid dictionary restrictions?

Including covariates: conditional polarization

Conclusion

Appendices

Outline

- Opposite views on a topic often translate into different phrases to talk about it:
  - "illegal aliens" – "undocumented workers"
  - "death tax" – "progressive wealth tax"
  - "witch hunt" – "impeachment hearing"
  - "Isis vs. US" – "Thanks, NRA" Headlines on Orlando shooting
  - French legislative election Headlines of some French newspapers
- Politics aim at defining shared goals and means: is it still possible if adverse parties speak different languages?
- Common sentiment that political polarization has been increasing $\rightarrow$ *role of political discourse?* Literature
- Which causes behind? Impact of social networks?
- **Problem**: obtain reliable measures of speech polarization (using text data)
  - **Application**: Republicans versus Democrats in US Congress debates between 1873 and 2016

Outline

Why text as data here?

- Surveys tell us a lot about political preferences, but:
  - The answers you get depend on the questions you ask
  - Subject to framing effects and other biases
  - Limited time coverage
- Text corpora present several advantages:
  - Very long time coverage (in some cases over centuries)
  - "in situ" reactions, articles, and speeches
  - Topics are not imposed beforehand

Data used: Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. Congressional Record for the 43rd-114th Congresses: Parsed Speeches and Phrase Counts. Palo Alto, CA: Stanford Libraries [distributor], 2018-01-16. `https://data.stanford.edu/congress_text`.

Representation used: counts of *N*-grams (bag-of-words)

- Speech is complex data: different approaches are relevant
- GST use *counts of phrases (bigrams)* pronounced by Republicans or Democrats

	phrase	K_D	K_R	K
	<chr>	<int>	<int>	<int>
1	legisl expertis	1	0	1
2	pledg protect	6	6	12
3	rulemak feder	2	1	3
4	unpreced threat	5	2	7
5	produc coal	0	12	12
6	produc cocain	2	0	2
7	back vulner	10	2	12
8	back washington	23	34	57
9	back white	0	4	4
10	privatepubl partnership	2	0	2
11	submit legsl	1	12	13
12	submit later	2	3	5
13	look job	23	36	59
14	quick senat	11	0	11
15	list suggest	0	3	3

Figure: Selected phrases from several hundreds of thousands of distinct phrases pronounced during 114th US Congress session (January 3, 2015 – January 3, 2017)

N-gram models: from texts to phrases

Details

A fictitious debate:

R1: *Dear colleagues, machines and artificial intelligence will destroy jobs in the future. Protectionism and regulation are necessary.*
D1: *I disagree with you despite your legislative expertise. Work is disutility. Technologies to produce more with less labor cannot be bad news.*
R2: *Protectionism is bad. Artificial intelligence will enable us to produce more and increase GDP.*
D2: *GDP is abstract; protecting people's job is concrete and pop up in the ballot.*
D3: *Earth is not an abstraction, neither accounted for in GDP.*

- Coercion to lowercase, delimitation of tokens, stemming
- *Choice of dictionary*: suppression of "stopwords" and bigrams with "bad syntax" or deemed "procedural"
- Count occurrences by R and by D for each phrase

N-gram models: from texts to phrases (e.g. 1-gram)

R~~1~~: *~~Dear colleagues~~, machines ~~and~~ artificial intelligence ~~will~~ destroy jobs ~~in the~~ future. Protectionism ~~and~~ regulation ~~are~~ necessary.*
D~~1~~: *~~I~~ disagree ~~with you~~ despite ~~your~~ legislative expertise. Work ~~is~~ disutility. Technologies ~~to~~ produce more with less labor cannot be bad news.*
R~~2~~: *Protectionism ~~is~~ bad. Artificial intelligence ~~will~~ enable us ~~to~~ produce more ~~and~~ increase GDP.*
D~~2~~: *GDP ~~is~~ abstract; protecting people's job ~~is~~ concrete ~~and~~ pop up in ~~the~~ ballot.*
D~~3~~: *Earth ~~is~~ not ~~an~~ abstract~~ion~~, neither account~~ed for in~~ GDP.*

Phrase (1-gram)	# occurrences by R (K^R)	# occ. by both R or D (K)
abstract	0	2
protectionism	2	2
Earth	0	1
GDP	1	3
expertise	0	1
...	...	...

Details of the processing and an example

See details of the processing operations
→ Appendix C.1 of working paper DGR
→ Check and analyses (see Powerpoint presentations describing empirical analyses in details)

- Original sentence

  'In fact the U.S. Government may want to examine the advantages of these products to lower its funding costs and thereby reduce the budget deficit.'

- Processed bigram representation (GST's processing)

  ('fact', 'govern'), ('want', 'examin'), ('examin', 'advantag'), ('advantag', 'product'), ('product', 'lower'), ('lower', 'fund'), ('fund', 'cost'), ('reduc', 'budget'), ('budget', 'deficit')

- Representation based on semantics

  ('government','want'), ('government','examine','advantage'), ('government','lower','cost'), ('government','reduce','deficit')

To keep in mind from this particular example (1)

(1) Think *ex ante* about **sensible representation/structuring of texts** to answer your question of interest

- Bag-of-words can be interesting in some cases
  - Example: Wu 2020, Review of Economics and Statistics, "Gender Bias in Rumors Among Professionals: An Identity-based Interpretation"
  - Study and compare words after "men are ..." / "women are ..."
- But less in others! Regarding speech polarization? Probably less appropriate ...
  - Bag-of-words $\Longrightarrow$ there is *no* notion of distance or proximity between phrases (just the same or different; 0 or 1)
  - $\neq$ words as vectors
  - In your opinion, what would be a good choice for studying speech polarization?
  - Henceforth, following GST, we keep with the bigram approach

To keep in mind from this particular example (2)

(2) Once given a representation, there is **no natural road** from raw texts to numerical inputs of the statistical analyses
$\longrightarrow$ a lot of room for more or less ad hoc choices in processing

- This does not mean such choices are always illegitimate
- But it is appropriate to study carefully their impact
  - Behind, more general question of how representative is a dataset of the target population of interest?
  - Constrained restrictions: missing data to link texts to speaker, in order to alleviate computational burden, etc.
  - Deliberate restrictions: remove procedural speech with low semantic content, etc.
- In particular, what impact on the final results?
  - Robustness checks by varying some pre-processing thresholds
  - Other approaches? (see the proposed extrapolated estimator later)

Outline

Notations and first idea to construct a polarization index

- For *phrase* $j \in \mathcal{J} := \{1, \ldots, J\}$, the *dictionary*, K_j^R (K_j^D) is the number of *occurrences* pronounced by R (D)
  - $K_j := K_j^D + K_j^R$: total # of occurrences US bipartite system
  - K_j^R / K_j : share of times phrase j is pronounced by R
- Conditional on the occurrence of phrase j, define ρ_j as the probability that it is pronounced by a R
  - K_j^R / K_j (observed empirical share) is an estimator of ρ_j (unobserved underlying probability)
- **Idea**: the more variation in the shares $\{K_j^R / K_j\}_{j \in \mathcal{J}}$ across phrases, the higher speech polarization, i.e. R and D use different words
  - Polarization index (v1): $\{K_j^R / K_j\}_{j \in \mathcal{J}} \xrightarrow{\text{map}}$ bounded scalar
  - Which map? Common maps from segregation indices

Small-unit bias

- **Problem**: K_j is often small Descriptive statistics for K

  $\rightarrow$ the variation in $\{K_j^R/K_j\}_j$ across phrases could be due to *small-sample variability* although the $\{\rho_j\}_j$ are all $\approx$ equal

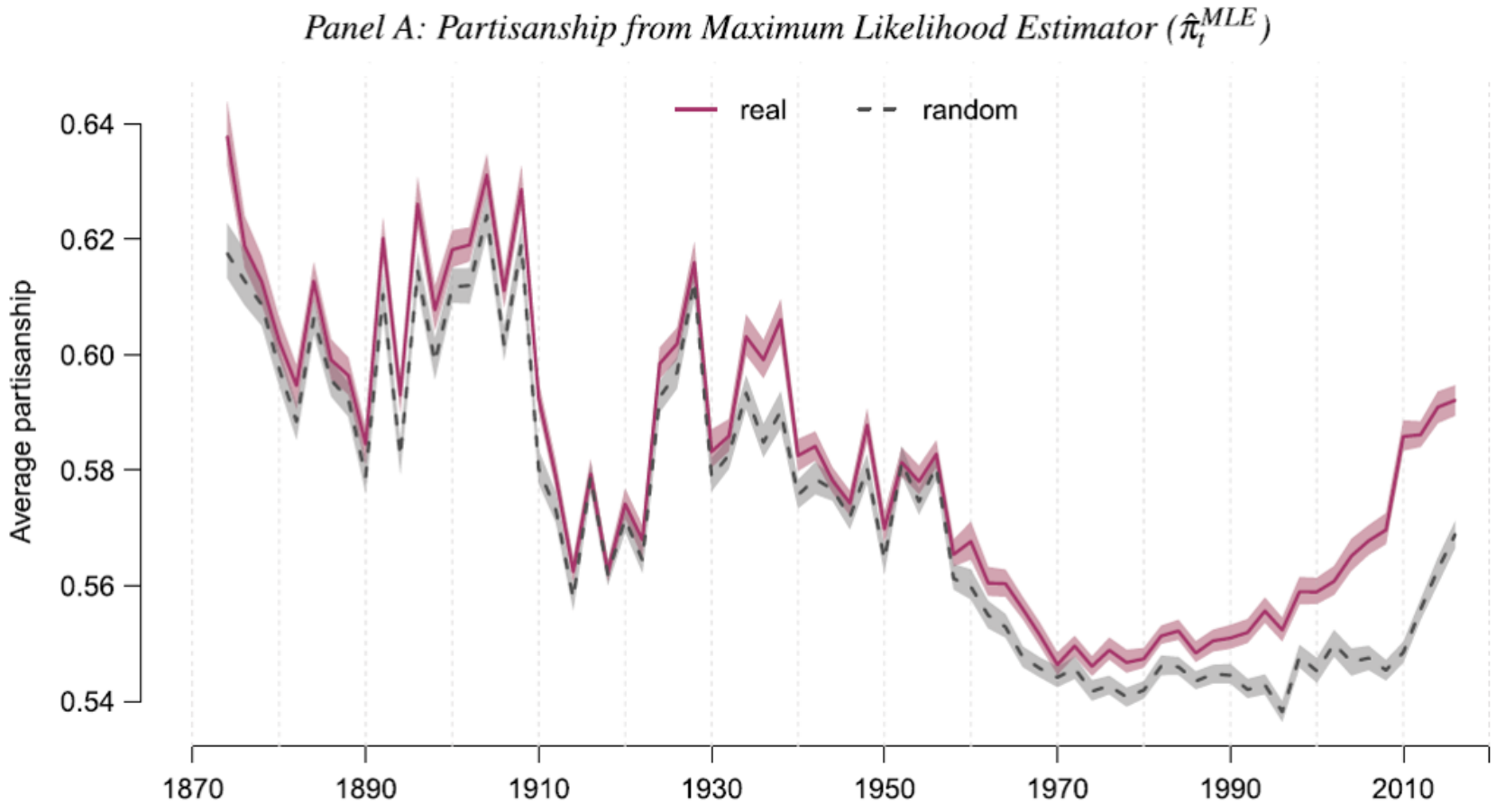
  - K_j^R/K_j consistently estimates ρ_j *provided* K_j tends to $+\infty$

- Partisanship indices based on $\{K_j^R/K_j\}_j$ are biased:
  - They overestimate the *real systematic level of polarization* defined by the variation in the $\{\rho_j\}_j$
  - Distinction between **evenness** v. **randomness** benchmarks, both may be interesting but differ (consequences v. measure)
  - Furthermore, they are not reliably comparable over time or across settings as the bias might change Longer speeches over time

$\rightarrow$ Need for better measures that account for the small-unit bias

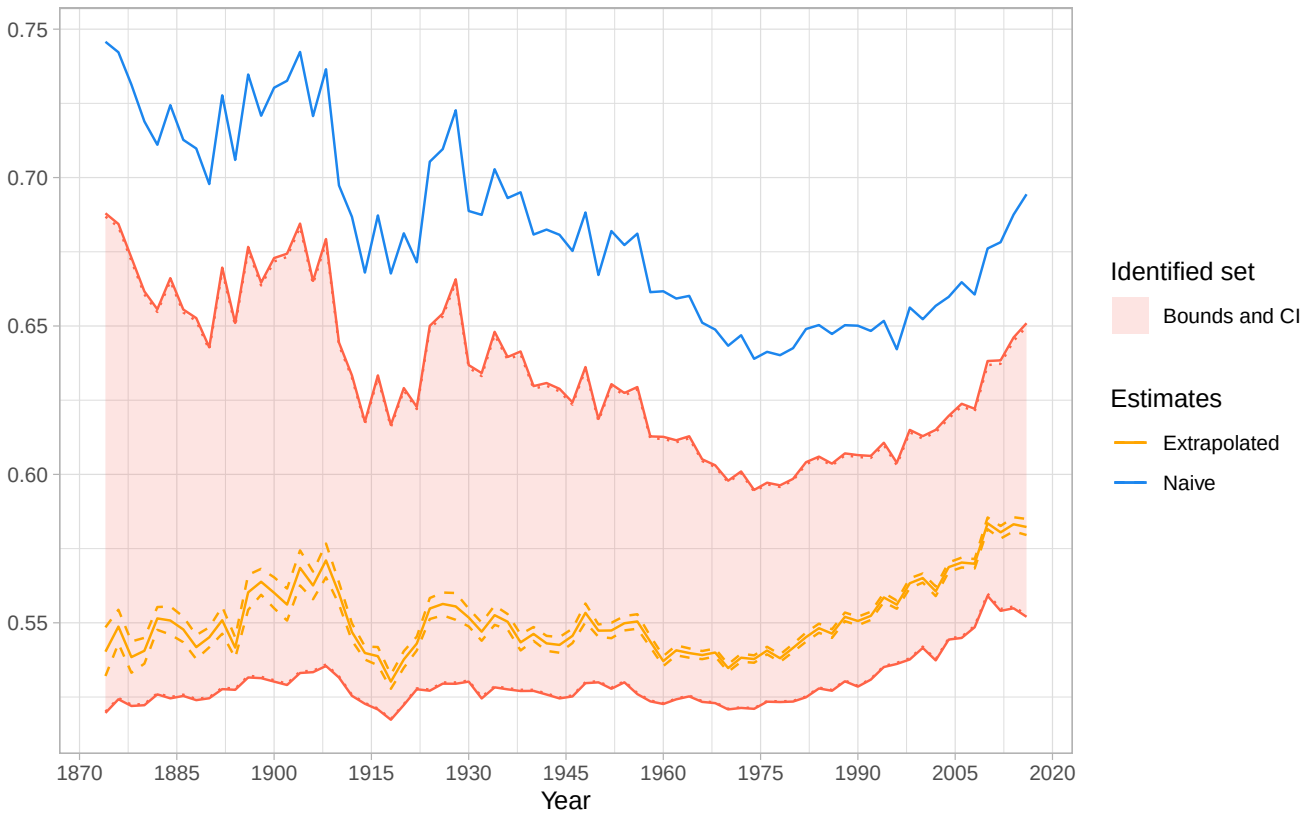
GST's estimation of partisanship by MLE (naive)

Figure: Figure 1, Panel A of GST



Small-unit bias: a "naive" measure is biased

Figure: Evolution of polarization over time: identification set, extrapolated index and naive index: preferred specification



Note: Naive polarization index (blue line), identified set of our polarization index (red area delimited by red dotted lines), 95% confidence interval of our polarization index (plain red lines), extrapolated index (orange plain line) and its 95% confidence interval (orange dashed lines). Each point corresponds to a Congressional session. The parameters of the extrapolation are $\overline{k} = 8$ (max number of occurrences) and $r = 3$ (polynomial degree). We do not include covariates in this analysis.
Sample: Bigrams after spelling corrections and exclusion of invalid words (no exclusion based on frequency).

Methodological issues

- The measure of polarization can be seen as a particular case of a more general problem: quantifying the differences in the choices made by two group in large choice sets
  - That problem encompasses (residential, school, occupational) segregation indices

    $\longrightarrow$ link with that (older) literature Segregation literature

- For text data, the issue of "small-unit bias" does matter

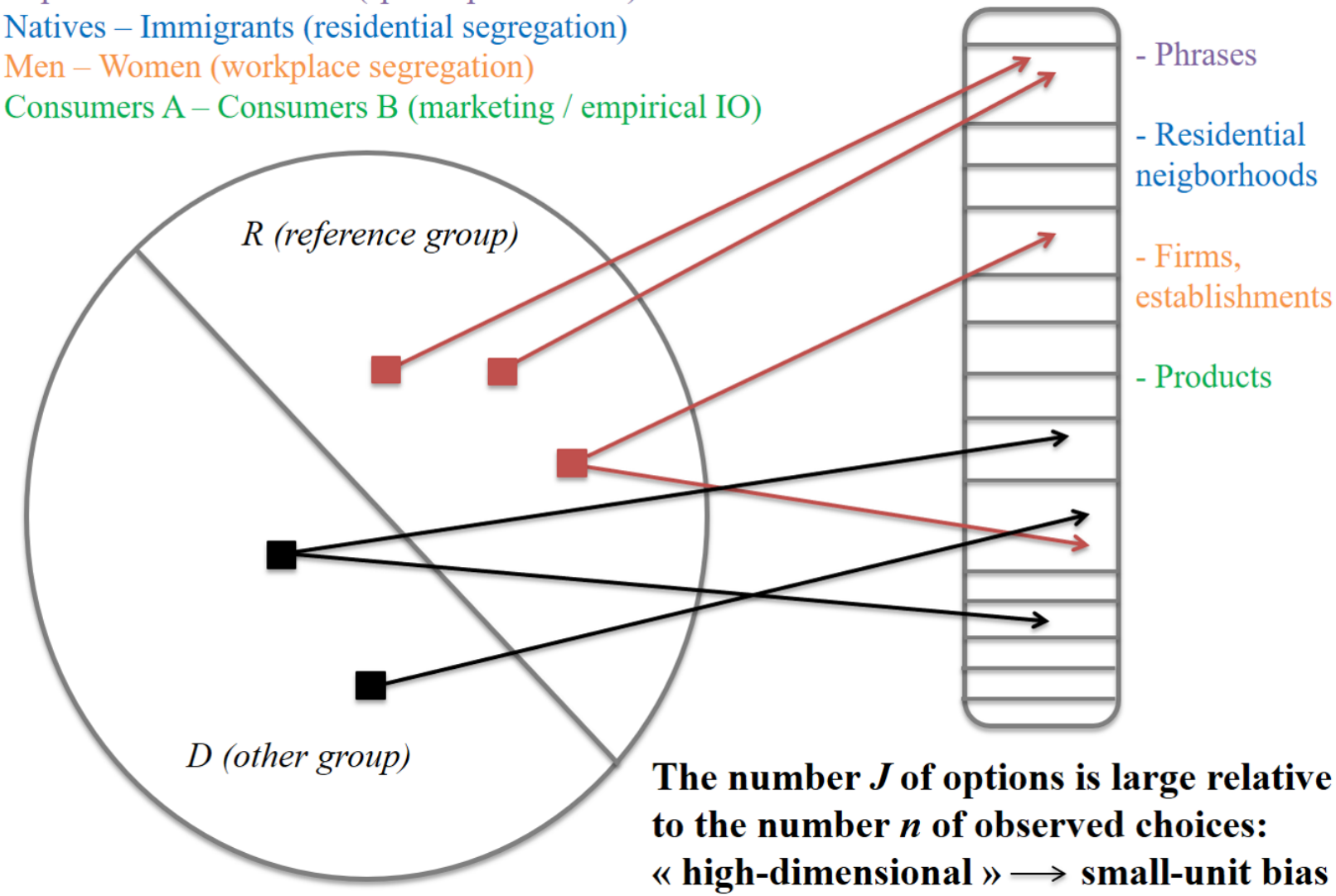
  $\longrightarrow$ need for methods that account for the bias
  - Gentzkow, Shapiro, Taddy (2019 Econometrica)
  - D'Haultfœuille, Girard, Rathelot (2021 mimeo)

General problem

A population split into two exogenous groups

- Republicans – Democrats (speech polarization)
- Natives – Immigrants (residential segregation)
- Men – Women (workplace segregation)
- Consumers A – Consumers B (marketing / empirical IO)

Individuals make choices among a set of options:



The number *J* of options is large relative to the number *n* of observed choices:
« high-dimensional » ⟶ small-unit bias

Outline

Generative model of speech

- Starting point: for a given dictionary $\mathcal{J}$, a generative model of speech through a multinomial distribution ($\Longrightarrow$ unrelated words, "bag-of-words")

The observed outcome is a J-vector $\mathbf{c}_{it}$ of phrase counts for speaker i, which we assume comes from a multinomial distribution

$$\mathbf{c}_{it} \sim \mathrm{MN}\big(m_{it}, \mathbf{q}_t^{P(i)}(\mathbf{x}_{it})\big), \tag{1}$$

with $m_{it} = \sum_j c_{ijt}$ denoting the total amount of speech by speaker i in session t, $P(i) \in \{R, D\}$ denoting the party affiliation of speaker i, $\mathbf{x}_{it}$ denoting a K-vector of (possibly time-varying) speaker characteristics, and $\mathbf{q}_t^P(\mathbf{x}_{it}) \in (0, 1)^J$ denoting the vector of choice probabilities. We let $R_t = \{i : P(i) = R, m_{it} > 0\}$ and $D_t = \{i : P(i) = D, m_{it} > 0\}$ denote the set of Republicans and Democrats, respectively, active in session t. The speech-generating process is fully characterized by the verbosity m_{it} and the probability $\mathbf{q}_t^P(\cdot)$ of speaking each phrase.

- For given covariates $\mathbf{x}$, session t, and word j in the dictionary, $q_{jt}^{\mathrm{R}}(\mathbf{x})$ and $q_{jt}^{\mathrm{D}}(\mathbf{x})$ are the probabilities of occurrence for the two parties

Measure of partisanship

- For each session t, a divergence between $\{q_{jt}^{\mathrm{R}}(\mathbf{x})\}_{j\in\mathcal{J}}$ and $\{q_{jt}^{\mathrm{D}}(\mathbf{x})\}_{j\in\mathcal{J}}$ averaged across speakers

For given characteristics $\mathbf{x}$, we define partisanship of speech to be the divergence between $\mathbf{q}_t^R(\mathbf{x})$ and $\mathbf{q}_t^D(\mathbf{x})$. When these vectors are close, Republicans and Democrats speak similarly and we say that partisanship is low. When these vectors are far from each other, the parties speak differently and we say that partisanship is high.

We choose a particular measure of this divergence that has a clear interpretation in the context of our model: the posterior probability that an observer with a neutral prior expects to assign to a speaker's true party after hearing the speaker utter a single phrase.

DEFINITION: The *partisanship* of speech at $\mathbf{x}$ is

$$\pi_t(\mathbf{x}) = \frac{1}{2}\mathbf{q}_t^R(\mathbf{x}) \cdot \boldsymbol{\rho}_t(\mathbf{x}) + \frac{1}{2}\mathbf{q}_t^D(\mathbf{x}) \cdot \big(1 - \boldsymbol{\rho}_t(\mathbf{x})\big), \tag{3}$$

where

$$\rho_{jt}(\mathbf{x}) = \frac{q_{jt}^R(\mathbf{x})}{q_{jt}^R(\mathbf{x}) + q_{jt}^D(\mathbf{x})}. \tag{4}$$

Average partisanship in session t is

$$\overline{\pi}_t = \frac{1}{|R_t \cup D_t|} \sum_{i \in R_t \cup D_t} \pi_t(\mathbf{x}_{it}). \tag{5}$$

ML (naive) and Leave-one-out estimators

- GST's polarization index is defined from the probabilities $\{q_{jt}^{\mathrm{R}}\}_{j\in\mathcal{J}}$, $\{q_{jt}^{\mathrm{D}}\}_{j\in\mathcal{J}}$, $\{\rho_{jt}\}_{j\in\mathcal{J}}$ (randomness benchmark)
- But they are unobserved $\longrightarrow$ estimation?
- Idea: replace the probabilities by the observed proportions

Maximum likelihood estimation is straightforward in our context. Ignoring covariates $\mathbf{x}$, the maximum likelihood estimator (MLE) can be computed by plugging in empirical analogues for the terms that appear in equation (3).

More precisely, let $\hat{\mathbf{q}}_{it} = \mathbf{c}_{it}/m_{it}$ be the empirical phrase frequencies for speaker i. Let $\hat{\mathbf{q}}_t^P = \sum_{i\in P_t} \mathbf{c}_{it} / \sum_{i\in P_t} m_{it}$ be the empirical phrase frequencies for party P, and let $\hat{\rho}_{jt} = \hat{q}_{jt}^R/(\hat{q}_{jt}^R + \hat{q}_{jt}^D)$, excluding from the choice set any phrases that are not spoken in session t. Then the MLE of $\overline{\pi}_t$ when $\mathbf{x}_{it} := \mathbf{x}_t$ is

$$\hat{\pi}_t^{\mathrm{MLE}} = \frac{1}{2}(\hat{\mathbf{q}}_t^R) \cdot \hat{\boldsymbol{\rho}}_t + \frac{1}{2}(\hat{\mathbf{q}}_t^D) \cdot (1 - \hat{\boldsymbol{\rho}}_t). \tag{6}$$

- Problem: small-unit bias $\longrightarrow$ GST proposes a leave-one-out estimator and their novel penalized "preferred" estimator

GST's penalized estimator – discrete choice model

- Observed proportions cause bias in small-unit settings
  $\longrightarrow$ change strategy: add structure with an underlying discrete choice model defined at speaker $\times$ session level

We suppose further that the choice probabilities are

$$q_{jt}^{P(i)}(\mathbf{x}_{it}) = e^{u_{ijt}} \Big/ \sum_{l} e^{u_{ilt}}, \tag{2}$$

$$u_{ijt} = \alpha_{jt} + \mathbf{x}_{it}'\boldsymbol{\gamma}_{jt} + \varphi_{jt}\mathbf{1}_{i\in R_t}.$$

Here α_{jt} is a scalar parameter capturing the baseline popularity of phrase j in session t, $\boldsymbol{\gamma}_{jt}$ is a K-vector capturing the effect of characteristics $\mathbf{x}_{it}$ on the propensity to use phrase j in session t, and φ_{jt} is a scalar parameter capturing the effect of party affiliation on the propensity to use phrase j in session t. If $\mathbf{x}_{it} := \mathbf{x}_t$, any phrase probabilities $(\mathbf{q}_t^R(\cdot), \mathbf{q}_t^D(\cdot))$

- Recover the index from the estimated parameters of the model

GST's penalized estimator – approximation and penalization

- Issue: size J of the dictionary = choice set (over 500,000)
- Solutions: (1) Poisson approximation of the likelihood
  $\longrightarrow$ enables distributed computing
- Solutions: (2) lasso L1 penalization
  $\longrightarrow$ correct the small-unit bias

minimization of the following penalized objective function:

$$\sum_j \Bigg\{ \sum_t \sum_i \big[m_{it} \exp\big(\alpha_{jt} + \mathbf{x}_{it}' \boldsymbol{\gamma}_{jt} + \varphi_{jt} \mathbf{1}_{i \in R_t}\big) - c_{ijt} \big(\alpha_{jt} + \mathbf{x}_{it}' \boldsymbol{\gamma}_{jt} + \varphi_{jt} \mathbf{1}_{i \in R_t}\big) + \psi \big(|\alpha_{jt}| + \|\boldsymbol{\gamma}_{jt}\|_1\big) + \lambda_j |\varphi_{jt}| \big] \Bigg\}. \tag{9}$$

We form an estimate $\hat{\pi}_t^*$ of $\overline{\pi}_t$ by substituting estimated parameters into the probability objects in equation (5).

The minimand in (9) encodes two key decisions. First, we approximate the likelihood of our multinomial logit model with the likelihood of a Poisson model (Palmgren (1981),

The second key decision is the use of an L_1 penalty $\lambda_j |\varphi_{jt}|$, which imposes sparsity on the party loadings and shrinks them toward zero (Tibshirani (1996)). Sparsity and shrinkage limit the effect of sampling error on the dispersion of the estimated posteriors ρ_{jt}, which is the source of the bias in $\hat{\pi}_t^{\mathrm{MLE}}$. We determine the penalties $\boldsymbol{\lambda}$ by regularization

- The underlying speaker-level discrete choice model
  $\longrightarrow$ strengths and weaknesses of GST's approach
- **Pros:** correct the bias while incorporating any individual covariates (even continuous)
- But at some **costs:**
  - Computationally demanding, all the more so as inference is done by subsampling
  - Somewhat black-box: how exactly does the lasso penalty correct the bias? What impact of the selection of the dictionary?
  - Requires data at speaker-level $\neq$ aggregated counts
  - Theoretical results: only for the non-approximated problem and with fixed dictionary $\mathcal{J}$ and asymptotics in the number of speaker $\neq$ Herdan's law Details on GST's asymptotics | Herdan's law

  $\longrightarrow$ does the method apply equally to "measure group differences in high-dimensional choices" in other applications?

Outline

- In contrast with GST's weaknesses, **pros:**
  - Formal identification result in a testable statistical model with an asymptotic in the dictionary size J ($=$ Herdan's law)
    ($\neq$ asymptotic in the number of speakers for a fixed dictionary)
  - Consistent and tractable estimators and confidence intervals
    ($\neq$ somewhat black-box and computationally demanding)
  - Our identification results shed light on the impact of the selection of the dictionary: the length of our identification interval is proportional to $\mathbb{P}(K = 1)$, the fraction of one-occurrence phrase
    $\rightarrow$ **trade-off** btw identification power v. dictionary's richness
  - Uses aggregate data: counts K_j^R, K_j^D
    ($\neq$ no need to record speakers' identities, for unconditional analyses)
- In contrast with GST's strengths, **cons:**
  - The conditional analysis is very limited compared to GST
  - In particular, complicated to control for continuous covariates
  - Issue behind: composition variance of the index

Statistical model

Formal links with GST

Assumption (DGP)

We observe an i.i.d. sample $(K_j, K_j^R)_{j=1,\dots,J}$ *with* (K_j, K_j^R, ρ_j) *having the same distribution as* (K, K^R, ρ)*, which satisfies:*

$$\mathbb{E}[K] > \mathbb{E}[K\rho] > 0, \text{ and } K^R \mid K, \rho \sim \text{Binomial}(K, \rho)$$

- i.i.d.: simplification of N-gram models
- Conditional distribution of K^R: no interaction across the occurrences of a given phrase (as Bernoulli trials)
  - Possibly restrictive but *testable* — Test of binomial assumption
  - Not rejected in our application
- Asymptotic in $J \to +\infty$: $J \approx$ hundreds of thousands and motivated by Herdan's law — Details and illustration

Partisanship index

Composition (in)variance

- To address the small-unit bias, we consider the variations in $\{\rho_j\}_j$ across phrases $\rightarrow$ index π defined as a function of P^ρ
- First idea: $\pi = 1 - 2\mathbb{E}[\rho(1-\rho)] \in [1/2, 1]$
  - If $\rho \sim$ Dirac$(1/2)$, $\pi = 1/2$ (*no polarization*)
  - If $\rho \sim$ Bernoulli$(1/2)$, $\pi = 1$ (*complete polarization*)
  - If $\rho \sim$ Dirac(p), $\pi = 1 - 2p(1-p)$, higher as p is far from $1/2$
- *Consider* $K \rightarrow \pi$ as a function of $P^{(\rho,K)}$ (weight phrases according to their frequencies): $\pi = 1 - 2\mathbb{E}[K\rho(1-\rho)]/\mathbb{E}[K]$
- *Consider* $p := \mathbb{E}[K^R]/\mathbb{E}[K]$: the share of speech from R
  p is not always $1/2$ Changes in p over time $\rightarrow$ generalized index:

$$\pi := 1 - \frac{\mathbb{E}\left[K\rho(1-\rho)\right]}{2\,\mathbb{E}\left[K\right]\,p(1-p)}$$

- $\forall p \in (0,1)$, $\rho \sim \delta_p \implies \pi = 1/2$; $\rho \sim$ Bern$(p) \implies \pi = 1$

- **(v1)**: dispersion of the observed empirical shares

$$\{K_j^R/K_j\}_{j\in\mathcal{J}} \in [0,1]^J \xrightarrow{\text{map}} \text{bounded scalar } \pi$$

  - *Evenness* benchmark
  - No notion of sampling, no formal population estimand
  - Subject to small-unit bias

- **(v2)**: dispersion of the underlying unobserved probabilities with fixed dictionary $\mathcal{J}$

$$\{\rho_j\}_{j\in\mathcal{J}} \in [0,1]^J \xrightarrow{\text{map}} \text{bounded scalar } \pi$$

  - *Randomness* benchmark
  - Asymptotics? Fixed dictionary but text length $\to +\infty$, $\neq$ Herdan's law, small-unit bias vanishes asymptotically

- **(v3)**: idem with i.i.d. model and asymptotics in $J \to +\infty$

Common maps from segregation indices

$$P^\rho \in \{\text{distributions in } [0,1]\} \xrightarrow{\text{map}} \text{bounded scalar } \pi$$

- Those are global indices $\neq$ measures defined at the level of a word/phrase or at the level of a speaker

Identification result

Theorem (Identification)

Suppose that Assumption (DGP) holds and the distribution of (K^R, K) *is identified. Then* $\pi \in [\underline{\pi}, \overline{\pi}]$*, that are sharp bounds, with*

$$\underline{\pi} := 1 - \frac{\mathbb{E}(K)}{2\mathbb{E}(K^R)\mathbb{E}(K^D)} \left\{ \mathbb{E}\left[\frac{K^R K^D}{K-1} \mathbb{1}\{K > 1\} \right] + \frac{\mathbb{E}(K^R \mathbb{1}\{K=1\})\mathbb{E}(K^D \mathbb{1}\{K=1\})}{\mathbb{P}(K=1)} \right\},$$

$$\overline{\pi} := 1 - \frac{\mathbb{E}(K)}{2\mathbb{E}(K^R)\mathbb{E}(K^D)} \mathbb{E}\left[\frac{K^R K^D}{K-1} \mathbb{1}\{K > 1\} \right]$$

Proof

- $\vdots = \mathbb{E}[K^R|K=1]\mathbb{E}[K^D|K=1]\mathbb{P}(K=1)$, hence partial identification comes from phrases pronounced only once
- The length of the identified set is proportional to $\mathbb{P}(K=1)$
- $\rightarrow$ definition of dictionary (processing)?

Estimation of the bounds on π

- Replace expectations by empirical counterparts (method of moments) $\rightarrow$ simple and computationally light estimators

$$\widehat{\underline{\pi}} := 1 - \frac{\sum_{j=1}^{J} K_j}{2\sum_{j=1}^{J} K_j^R \sum_{j=1}^{J} K_j^D} \left[\sum_{j=1}^{J} \frac{K_j^R K_j^D}{K_j - 1} \mathbb{1}\{K_j > 1\} + \frac{\sum_{j=1}^{J} K_j^R \mathbb{1}\{K_j = 1\} \sum_{j=1}^{J} K_j^D \mathbb{1}\{K_j = 1\}}{\sum_{j=1}^{J} \mathbb{1}\{K_j = 1\}} \right]$$

$$\widehat{\overline{\pi}} := 1 - \frac{\sum_{j=1}^{J} K_j}{2\sum_{j=1}^{J} K_j^R \sum_{j=1}^{J} K_j^D} \sum_{j=1}^{J} \frac{K_j^R K_j^D}{K_j - 1} \mathbb{1}\{K_j > 1\}$$

Convention: take $(\sum_{j:K_j=1}^{J} K_j^R)/\sum_{j=1}^{J} \mathbb{1}\{K_j = 1\} = 0$ if $\sum_{j=1}^{J} \mathbb{1}\{K_j = 1\} = 0$

- Assumption (DGP), Law of Large Numbers, and Continuous Mapping Theorem yield consistency when $J \rightarrow +\infty$

Inference: confidence interval for the polarization index π

Theorem (Inference)

Suppose that Assumption (DGP) is satisfied. Then:

$$\sqrt{J}\left[\begin{pmatrix}\widehat{\underline{\pi}}\\ \widehat{\overline{\pi}}\end{pmatrix}-\begin{pmatrix}\underline{\pi}\\ \overline{\pi}\end{pmatrix}\right]\xrightarrow{d}\mathcal{N}\left(0,\begin{pmatrix}a & b\\ b & c\end{pmatrix}\right),$$

with $a := V(\underline{\delta}_j)$, $b := Cov(\underline{\delta}_j, \overline{\delta}_j)$ *and* $c := V(\overline{\delta}_j)$, *where*

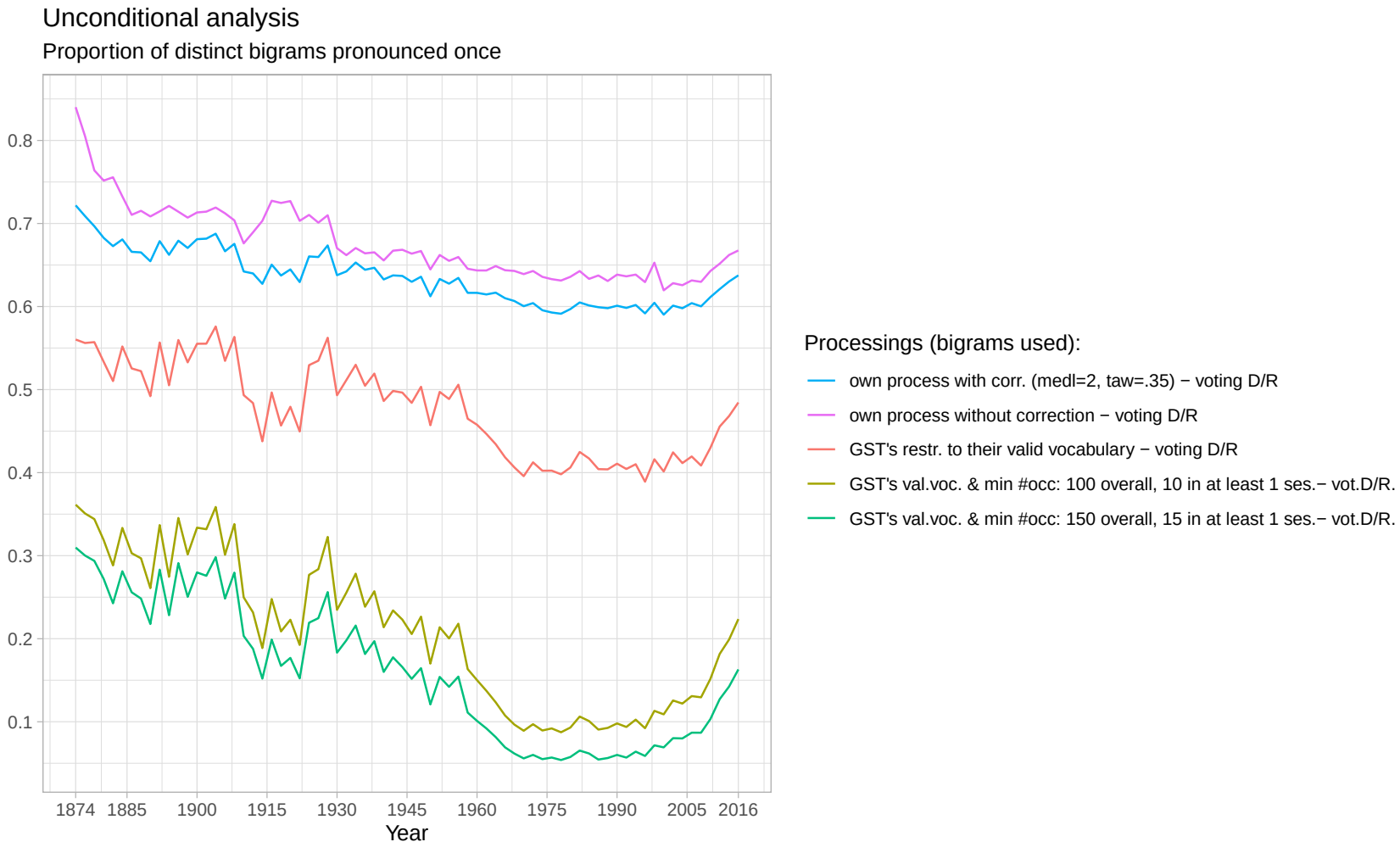
$$\overline{\delta}_j := (\overline{\pi}-1)\left\{\frac{K_j}{E(K)}+\frac{K_j^R K_j^D \mathbb{1}\{K_j>1\}/(K_j-1)}{\mathbb{E}[K^R K^D \mathbb{1}\{K>1\}/(K-1)]}-\frac{K_j^R}{E(K^R)}-\frac{K_j^D}{E(K^D)}\right\}$$

$$\underline{\delta}_j := \overline{\delta}_j - \frac{E(K)E(K^R\mathbb{1}\{K=1\})E(K^D\mathbb{1}\{K=1\})}{2E(K^R)E(K^D)\mathbb{P}(K=1)}\left\{\frac{K_j}{E(K)}+\frac{K_j^R\mathbb{1}\{K_j=1\}}{E(K^R\mathbb{1}\{K=1\})}\right.$$
$$\left.+\frac{K_j^D\mathbb{1}\{K_j=1\}}{E(K^D\mathbb{1}\{K=1\})}-\frac{K_j^R}{E(K^R)}-\frac{K_j^D}{E(K^D)}-\frac{\mathbb{1}\{K_j=1\}}{\mathbb{P}(K=1)}\right\}$$

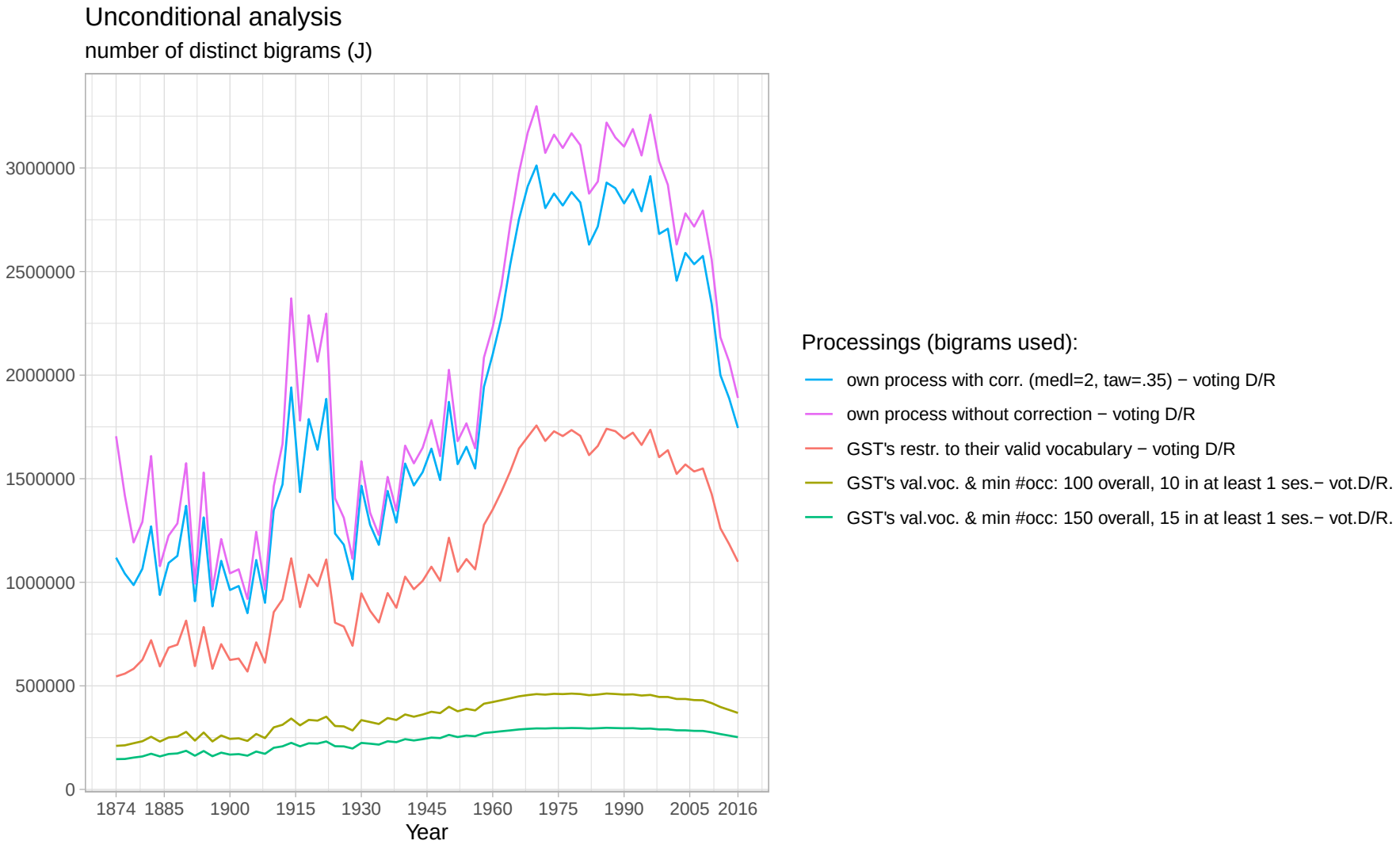
- CI based on Imbens and Manski (2004) and Stoye (2009) to account for partial identification of π CI details | Monte-Carlo simulations

Outline

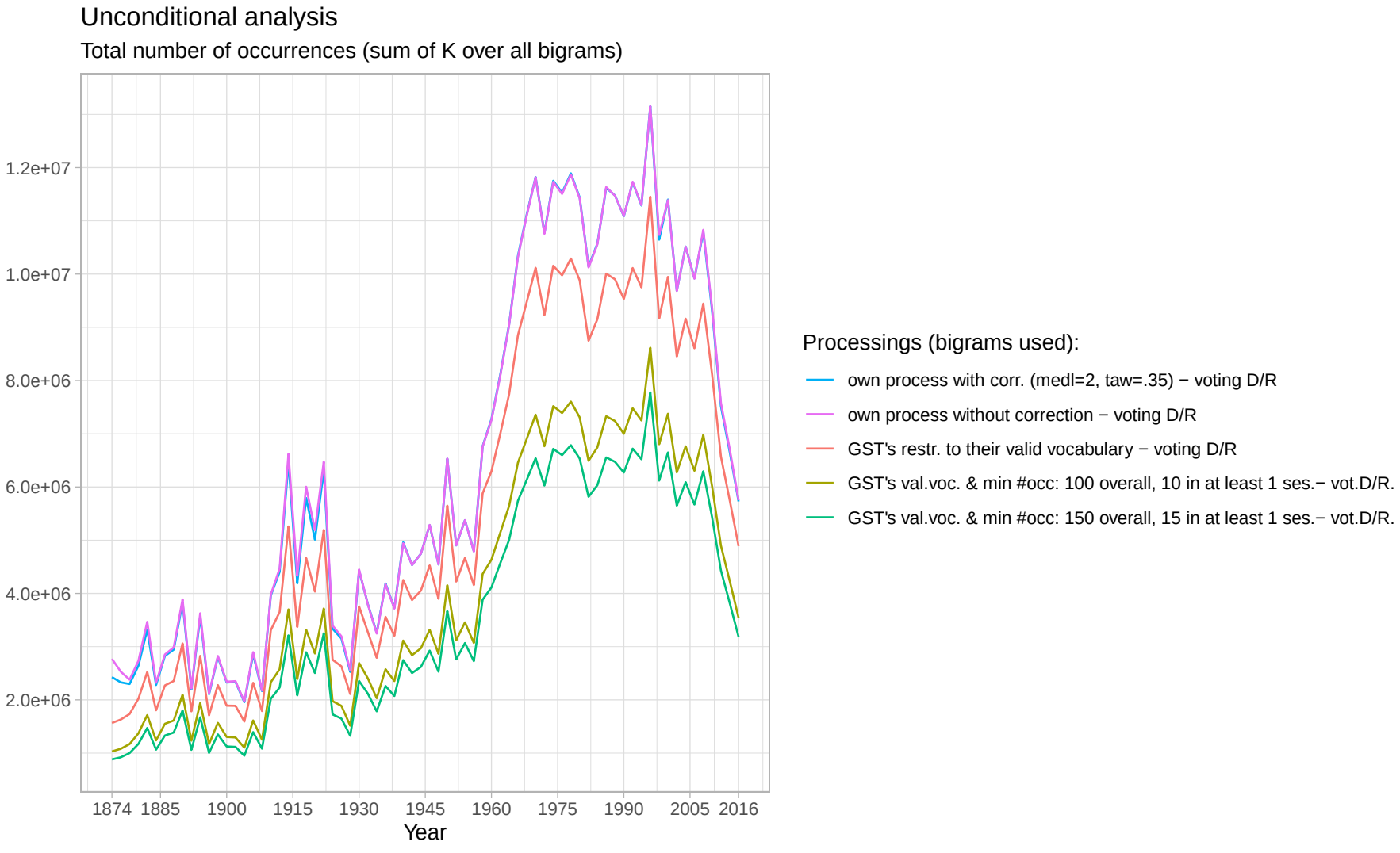
Impact of dictionary selection on $\mathbb{P}(K = 1)$



Impact of dictionary selection on J



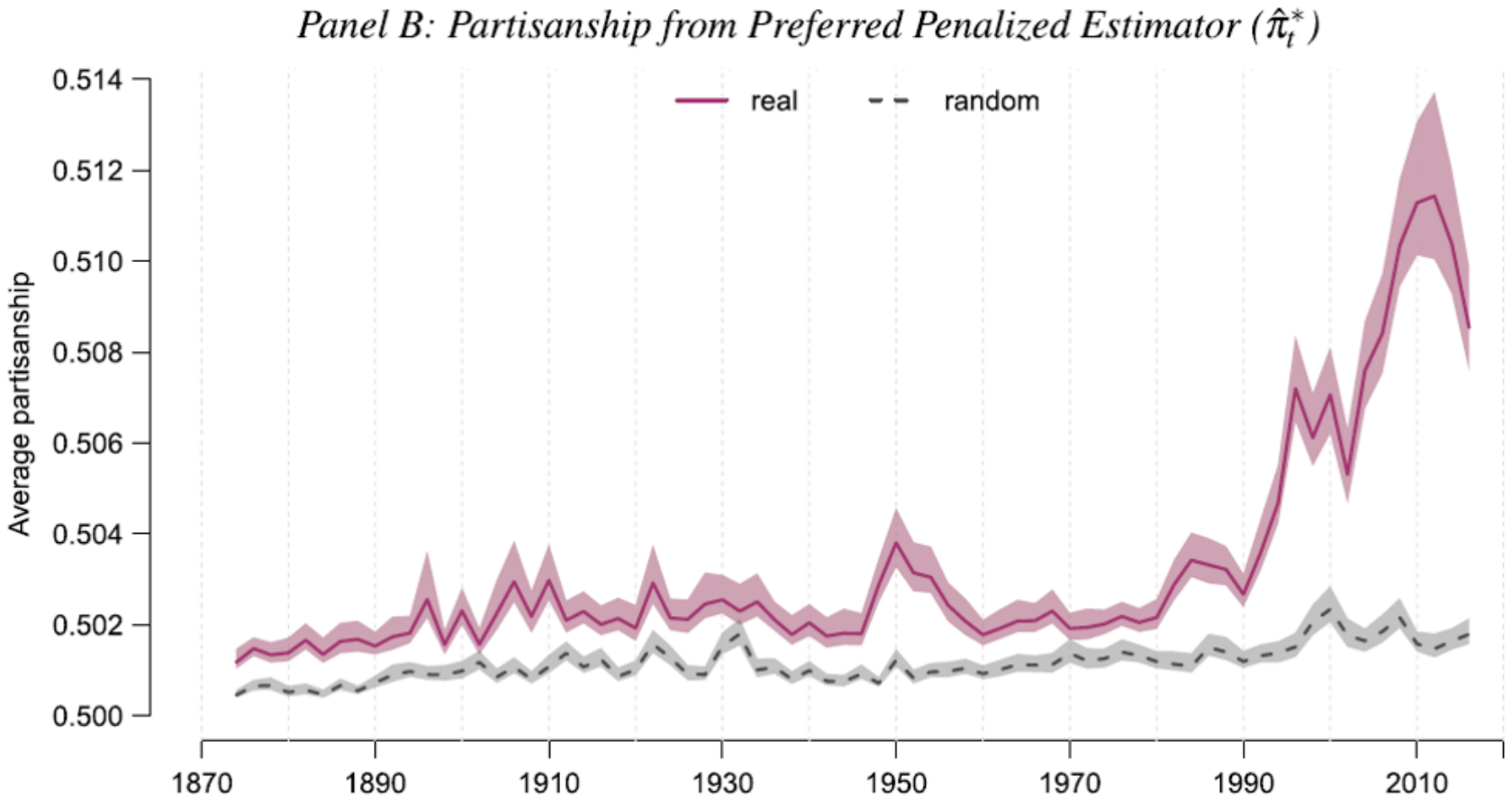
Impact of dictionary selection on $n := \sum_{j=1}^{J} K_j$



GST's penalized estimation of partisanship

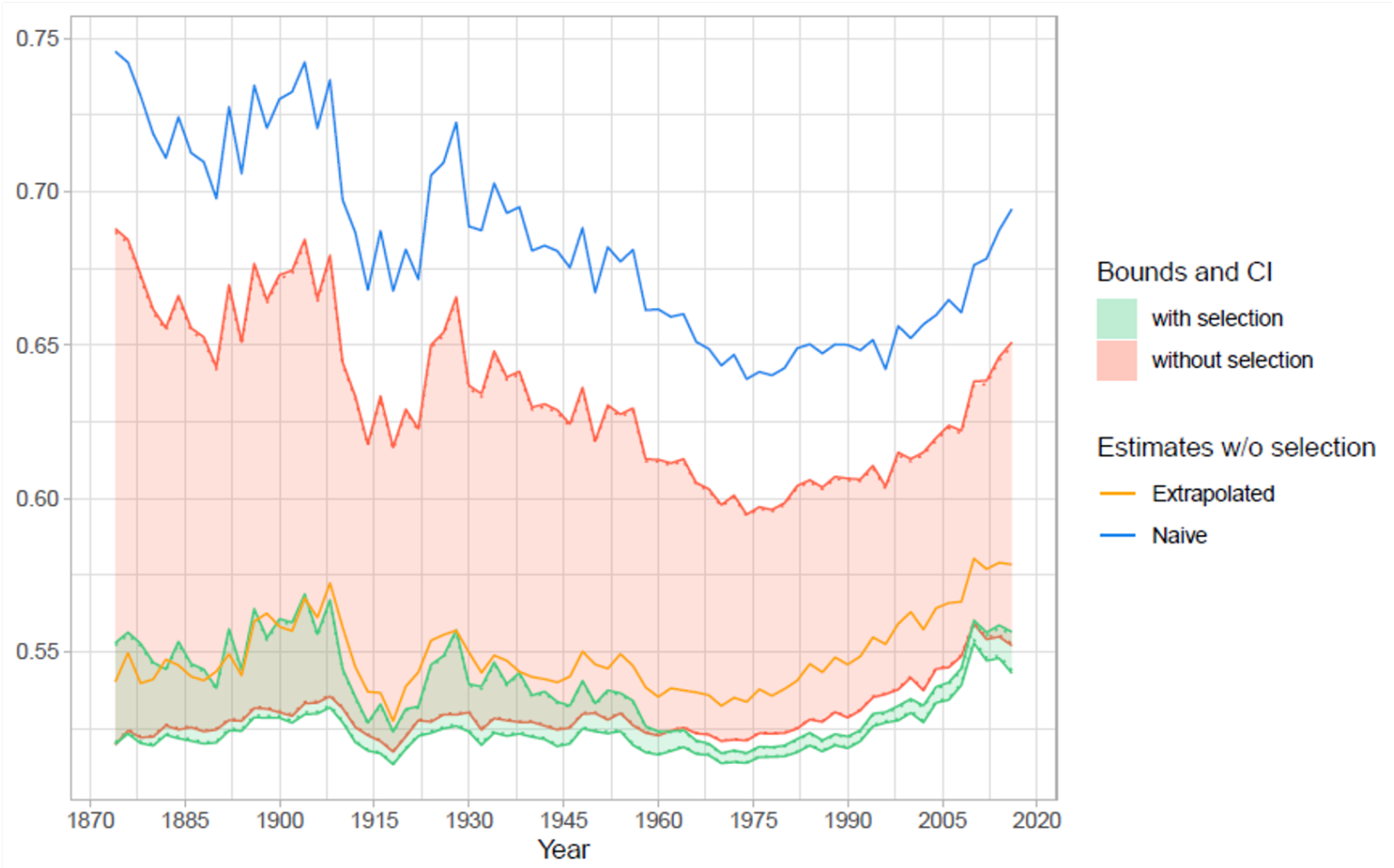
Figure: Figure 2, Panel B of GST

DGR's main result



Link between identification power and dictionary selection

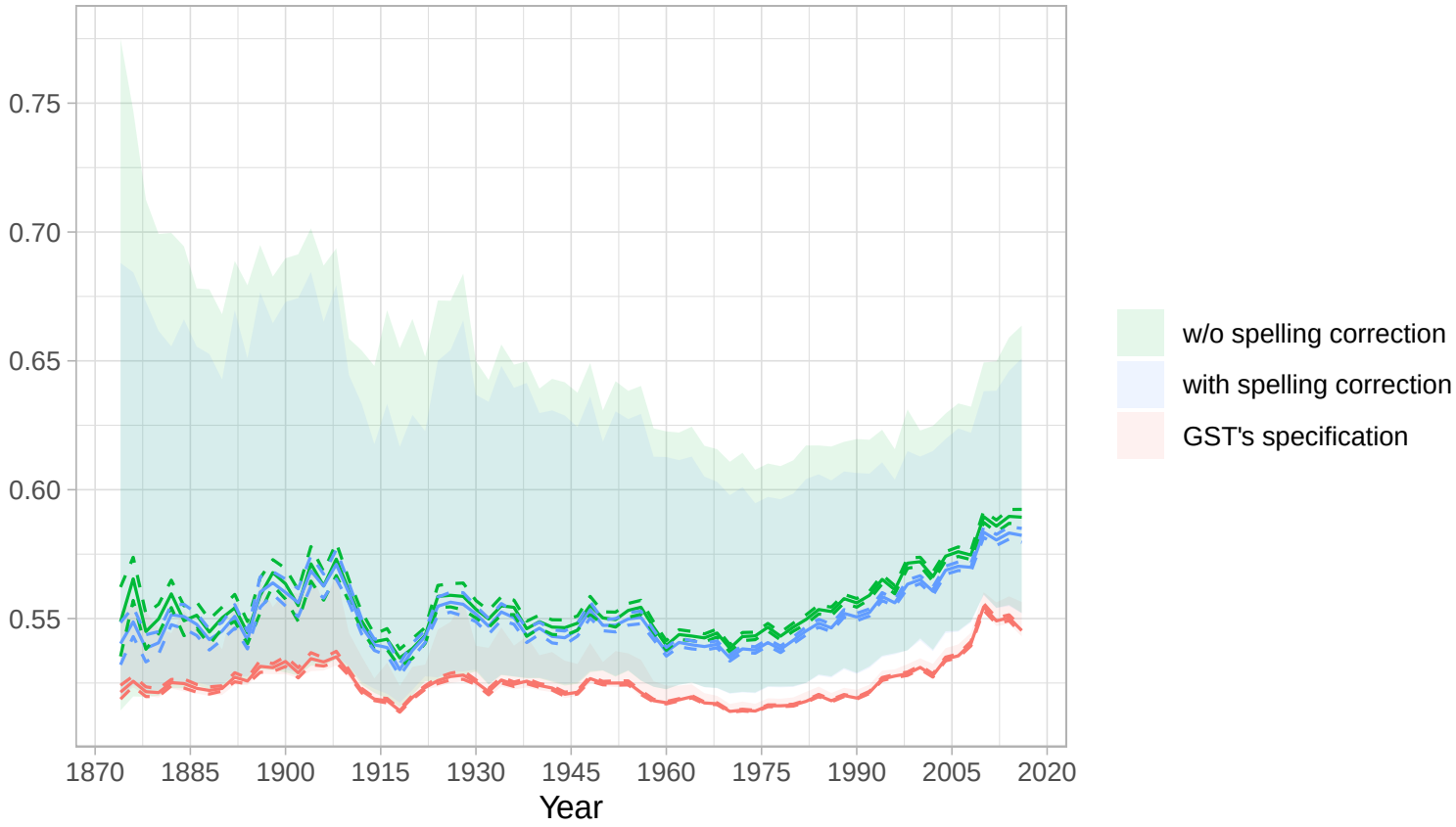
Figure: Evolution of polarization over time



Note: with(out) selection means (not) applying the bigram frequency restrictions used by GST: at least 100 occurrences overall across sessions, and at least 10 occurrences in at least one session. In both cases: correction and suppression of "bad syntax" or "procedural" phrases.

Moderate impact of the spelling correction step

Figure: Evolution of polarization over time



Note: GST's specification means applying the bigram frequency restrictions used by GST: at least 100 occurrences overall across sessions, and at least 10 occurrences in at least one session. Confidence interval (areas) on π and extrapolated estimators (solid lines) (see next Section) with confidence intervals through delta-method (dashed lines).

Outline

How to avoid restrictions on K to define dictionary $\mathcal{J}$?

- It happens that naturally in the data ($\neq$ encoding errors with OCR or misspelling) a lot of phrases are rarely pronounced
  $\rightarrow$ high $\mathbb{P}(K = 1)$: what do to with them?
  - Drop them: rare phrases not related to polarization?
  - But how to know ex ante?
  - On the other hand, bounds without restrictions on K are hardly informative
- We propose two solutions to avoid restricting the dictionary based on K while sharpening the identification of π
  - Under the additional assumption of an "independence" between K and ρ, π is point-identified and we have simple, consistent, and asymptotically normal estimator Details
  - Extrapolation

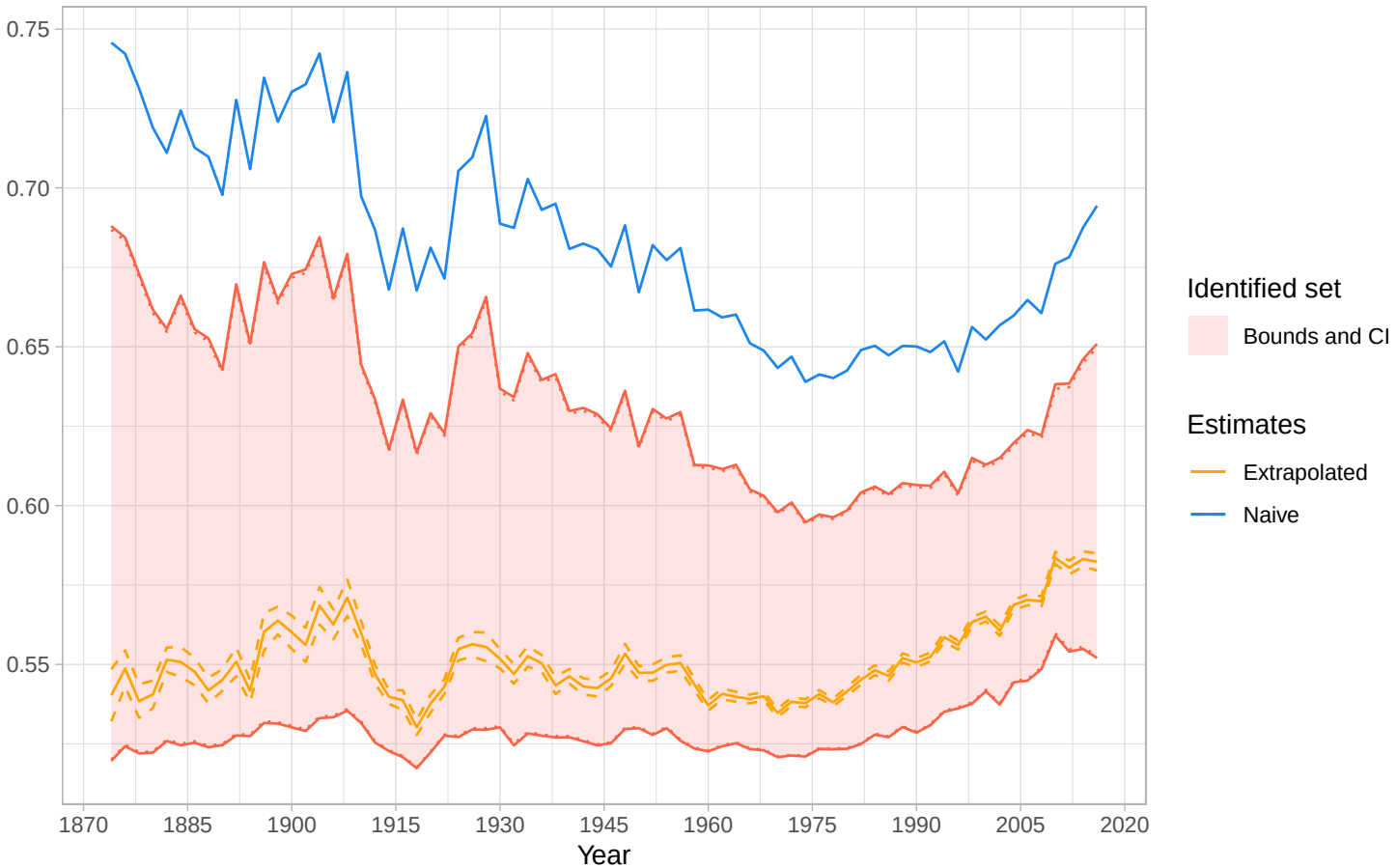
- We have partial identification because we cannot point identify $m(1) := \mathbb{E}\left[\rho^2 \mid K = 1\right]$
- However, for $k > 1$, we can identify $m(k) := \mathbb{E}\left[\rho^2 \mid K = k\right]$
  - Indeed, in the proof of our identification theorem, we obtain $\forall k > 1, m(k) = \mathbb{E}\left[\frac{K^R(K^R-1)}{K(K-1)} \mid K = k\right]$
- A strategy to get a point-estimate is to extrapolate $m(1)$ $\rightarrow \widehat{m(1)}$ based on $\{\widehat{m(k)}\}_{k=2,\ldots,\bar{k}} \longrightarrow \hat{\pi}_{\text{extrapolated}}$
- Underlying hyp.: $m(k)$ is sufficiently regular close to $k = 1$
  - It can be tested by over-identification tests ("Classical Minimum Distance Estimation" set-up; see Wooldridge, Section 14.5), and, besides, assessed graphically
  - Seems to hold in our application for flexible enough functions

    Illustration: regularity of m(k) | Over-identification tests
  - The extrapolated indices for different choices of $\bar{k}$ or r (the order of the polynomial function used) are quite similar

    Robustness checks

Figure: Evolution of polarization over time: identification set, extrapolated index and naive index: preferred specification

GST's main result



Note: Naive polarization index (blue line), identified set of our polarization index (red area delimited by red dotted lines), 95% confidence interval of our polarization index (plain red lines), extrapolated index (orange plain line) and its 95% confidence interval (orange dashed lines). Each point corresponds to a Congressional session. The parameters of the extrapolation are $\overline{k} = 8$ (max number of occurrences) and $r = 3$ (polynomial degree). We do not include covariates in this analysis.
Sample: Bigrams after spelling corrections and exclusion of invalid words (no exclusion based on frequency).

Outline

Covariates can be included at two levels:

- As *phrase characteristics* (e.g., topic) $\rightarrow$ essentially the same analysis on a subset of the data
- As speaker characteristics (e.g., gender) $\rightarrow$ more complicated
- Let $Z \in \{1, \ldots, \overline{Z}\}$ be the type of the phrase (e.g., topic)
- We can define (and identify and estimate as in the unconditional case) π_z as the partisanship restricted to *phrases* such that $Z = z$
- Then, we can consider (estimation and inference by plug-in) the conditional index:

$$\pi_{\text{cond}(Z)} := \sum_{z=1}^{\overline{Z}} \mathbb{P}(Z = z)\pi_z$$

- Examples: by topic, (see page 1328 of GST), by chamber

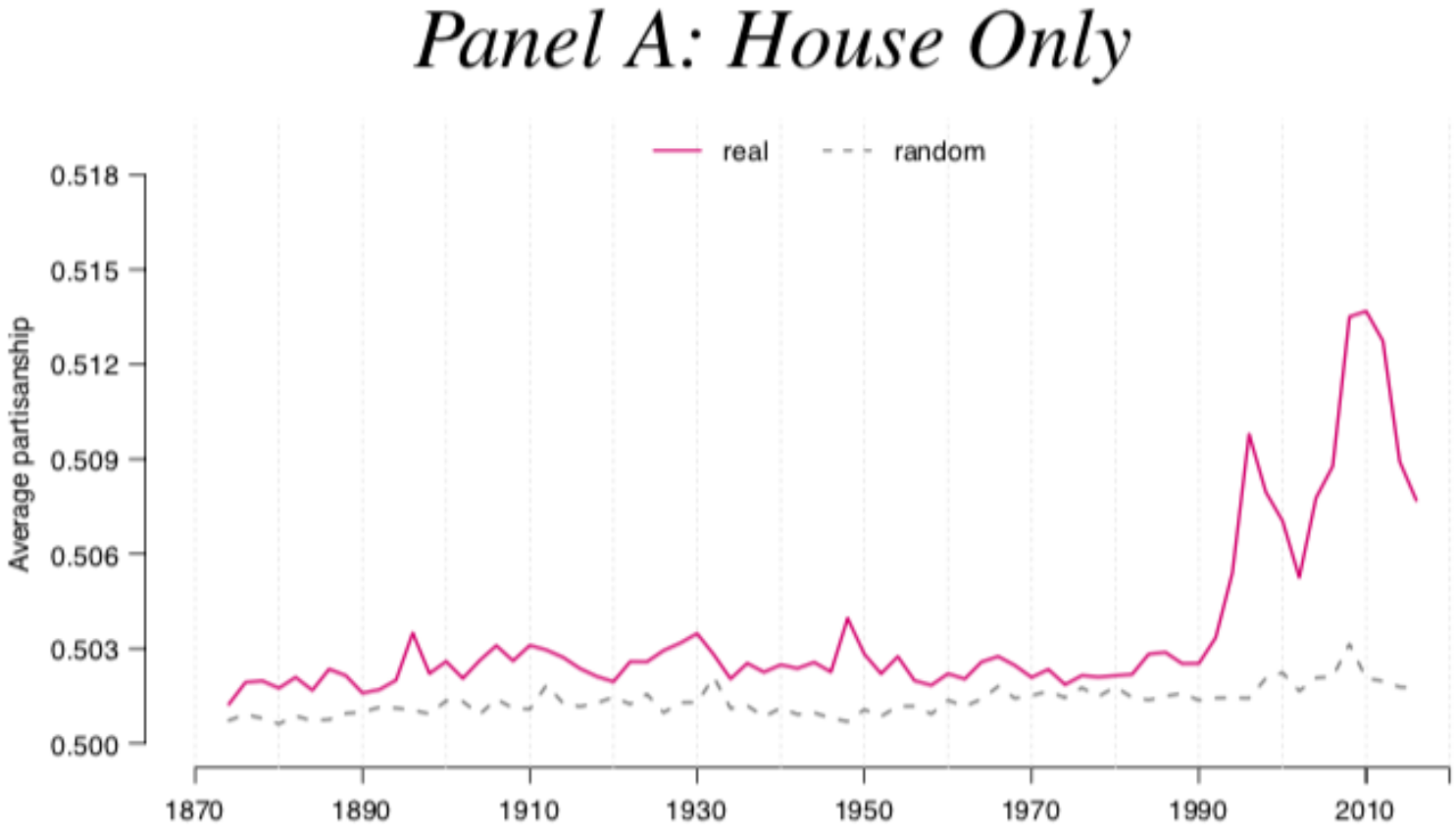


Figure: GST (ECTA 2019) – Figure 1, Online Appendix

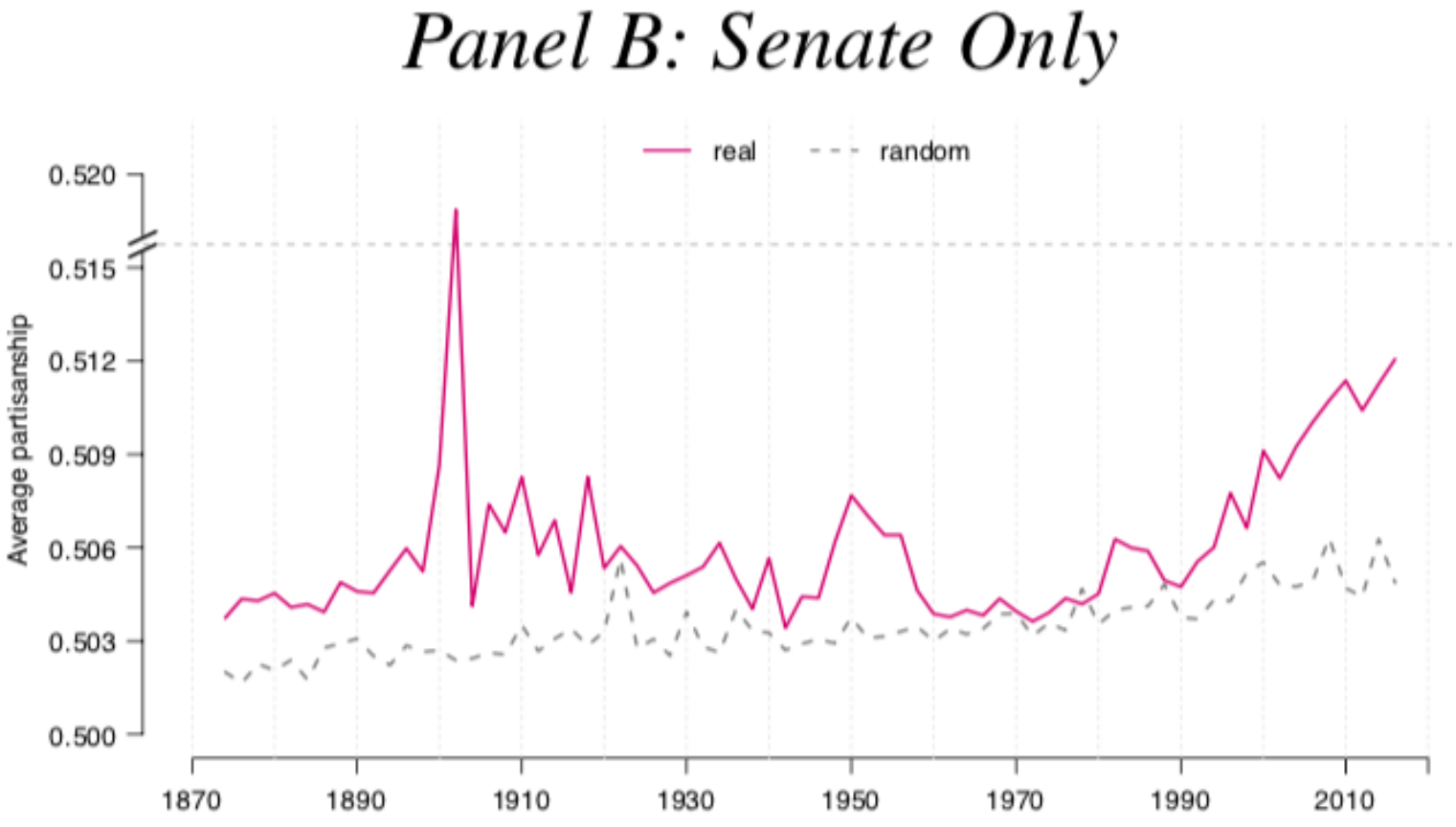


Figure: GST (ECTA 2019) – Figure 1, Online Appendix

Speaker characteristics in GST :)

- Partisanship is defined as a function of the characteristics
- Baseline specification: indicators for state, chamber, gender, Census region, and whether the party is in the majority for the entirety of the session ("characteristics that are likely to be related both to party and to speech but whose relationship with speech would not generally be thought of as a manifestation of party differences")
- The underlying discrete choice models for phrases enable to account for individual characteristics (included continuous covariates possibly) in their penalized estimator ($\neq$ no covariates in their leave-one-out estimator) Graphs
- Minimal impact of the covariates for their penalized estimates

We suppose further that the choice probabilities are

$$q_{jt}^{P(i)}(\mathbf{x}_{it}) = e^{u_{ijt}} \Big/ \sum_{l} e^{u_{ilt}}, \qquad (2)$$

$$u_{ijt} = \alpha_{jt} + \mathbf{x}_{it}' \boldsymbol{\gamma}_{jt} + \varphi_{jt} \mathbf{1}_{i \in R_t}.$$

Here α_{jt} is a scalar parameter capturing the baseline popularity of phrase j in session t, $\boldsymbol{\gamma}_{jt}$ is a K-vector capturing the effect of characteristics $\mathbf{x}_{it}$ on the propensity to use phrase j in session t, and φ_{jt} is a scalar parameter capturing the effect of party affiliation on the propensity to use phrase j in session t. If $\mathbf{x}_{it} := \mathbf{x}_t$, any phrase probabilities $(\mathbf{q}_t^R(\cdot), \mathbf{q}_t^D(\cdot))$

Speaker characteristics in DGR :(

- Let $W \in \{1, \ldots, \overline{W}\}$ be the type of speakers (here defined as interactions of gender, state, and chamber)
- We can define π_w as the partisanship restricted to *occurrences* pronounced by speakers such that $W = w$, i.e. intra type w
- Then, we can consider the conditional index:

$$\pi_{\text{cond}(W)} := \sum_{w=1}^{\overline{W}} \mathbb{P}(W = w)\pi_w$$

- Issue: it compounds the small-unit bias
  - The number of times each phrase is pronounced within each type is smaller than for the full sample

Speaker characteristics: heuristic method

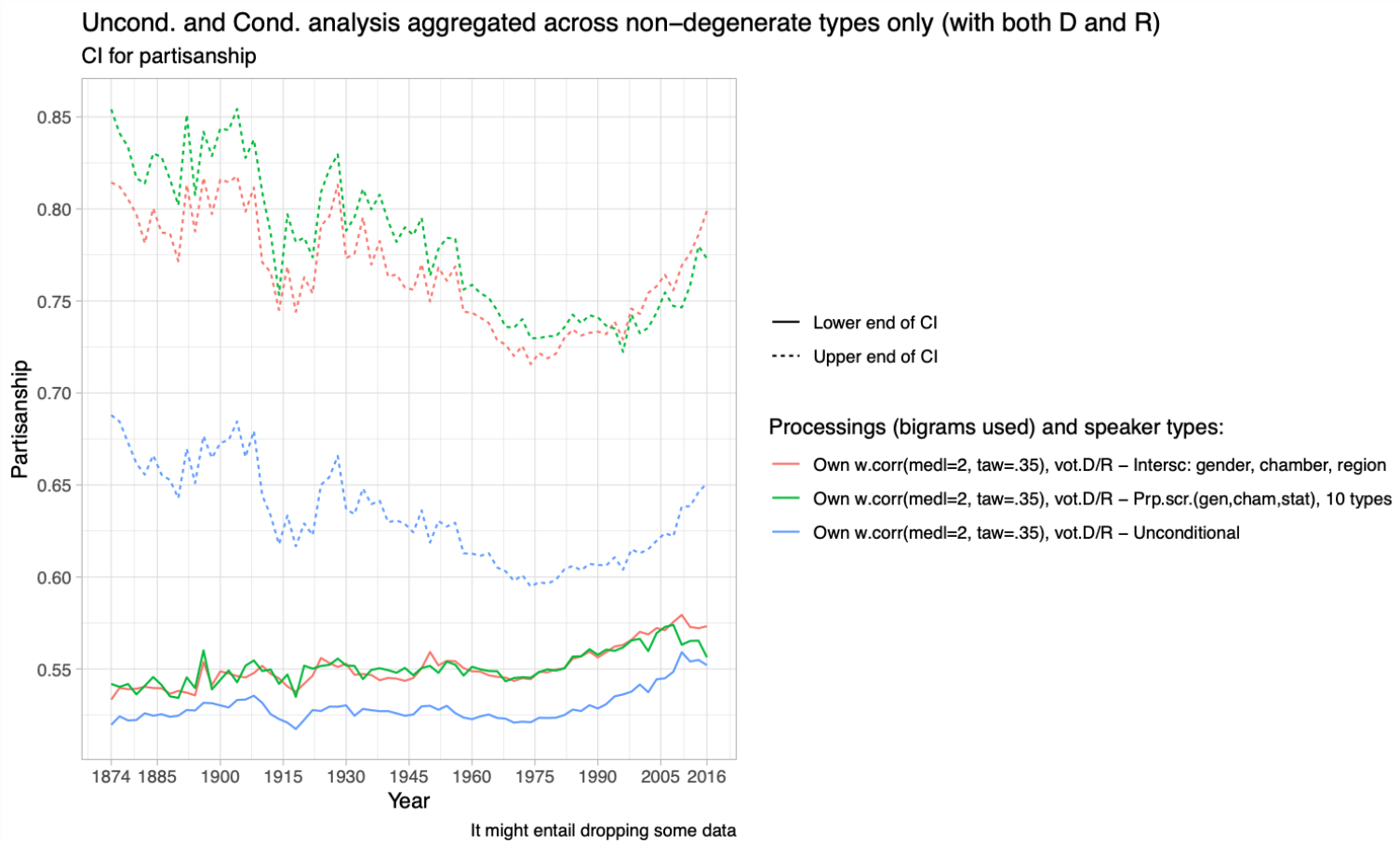
- When $\overline{W}$ becomes large, or when we want to include continuous characteristics, the former approach fails
- We propose the following heuristic method (requires to observe identities of speakers and their characteristics):
  1. Let V_s denote the vector of characteristics of speaker s
  2. Estimate the β corresponding to:

$$\mathbb{1}\{s \text{ is Republican}\} = \mathbb{1}\{V_s\beta + \varepsilon > 0\}$$

  3. For each s, define $\widetilde{W}_s$ as a categorical variable from the quantiles of $V_s\widehat{\beta}$, the estimated propensity score of being R
     - In our application: $\overline{\widetilde{W}} = 10$, that is: $\widetilde{W}_s = 1$ if the index $V_s\widehat{\beta}$ is in the first decile, $\widetilde{W}_s = 2$ in the second decile, etc.
  4. Use $\widetilde{W}$ to compute $\pi_{\text{cond}(\widetilde{W})}$

Comparing unconditional and conditional analyses

Figure: Processing with correction, suppression of "bad syntax" or "procedural" phrases, without restrictions based on the number of occurrences

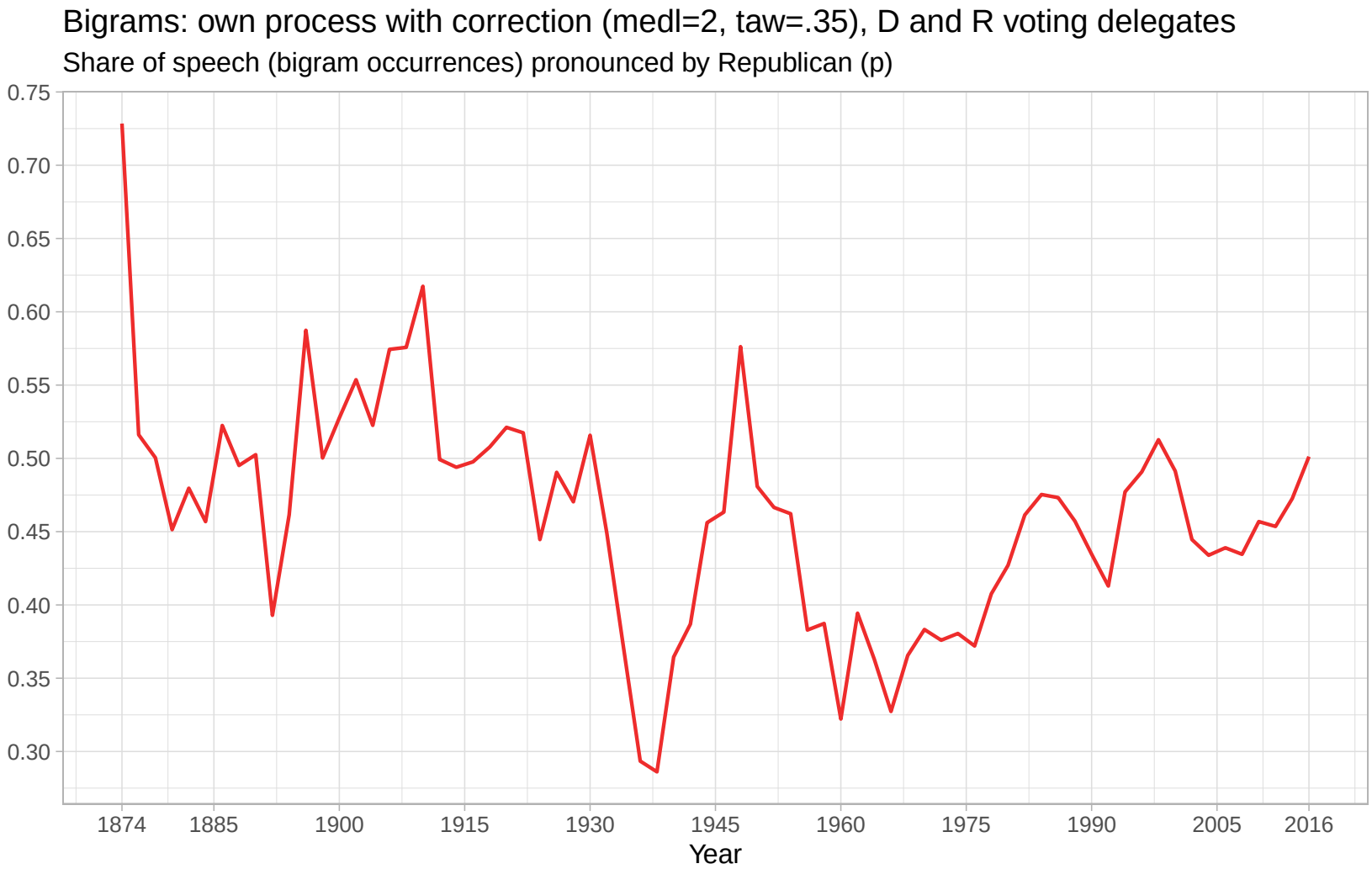


Higher polarization conditional on characteristics?

Warning: composition (in)variance!

Evolution of p over time

Figure: Processing with correction, suppression of "bad syntax" or "procedural" phrases, without restriction based on the number of occurrences



GST uses a neutral (half-half) prior between R and D

DEFINITION: The *partisanship* of speech at $\mathbf{x}$ is

$$\pi_t(\mathbf{x}) = \frac{1}{2}\mathbf{q}_t^R(\mathbf{x}) \cdot \boldsymbol{\rho}_t(\mathbf{x}) + \frac{1}{2}\mathbf{q}_t^D(\mathbf{x}) \cdot \big(1 - \boldsymbol{\rho}_t(\mathbf{x})\big), \tag{3}$$

where

$$\rho_{jt}(\mathbf{x}) = \frac{q_{jt}^R(\mathbf{x})}{q_{jt}^R(\mathbf{x}) + q_{jt}^D(\mathbf{x})}. \tag{4}$$

Average partisanship in session t is

$$\overline{\pi}_t = \frac{1}{|R_t \cup D_t|} \sum_{i \in R_t \cup D_t} \pi_t(\mathbf{x}_{it}). \tag{5}$$

To understand these definitions, note that $\rho_{jt}(\mathbf{x})$ is the posterior belief that an observer with a neutral prior assigns to a speaker being Republican if the speaker chooses phrase j in session t and has characteristics $\mathbf{x}$. Partisanship $\pi_t(\mathbf{x})$ averages $\rho_{jt}(\mathbf{x})$ over the possible parties and phrases: if the speaker is a Republican (which occurs with probability $\frac{1}{2}$), the probability of a given phrase j is $q_{jt}^R(\mathbf{x})$ and the probability assigned to the true party after hearing j is $\rho_{jt}(\mathbf{x})$; if the speaker is a Democrat, these probabilities are $q_{jt}^D(\mathbf{x})$ and $1 - \rho_{jt}(\mathbf{x})$, respectively. Average partisanship $\overline{\pi}_t$, which is our target for estimation, averages $\pi_t(\mathbf{x}_{it})$ over the characteristics $\mathbf{x}_{it}$ of speakers active in session t. Average partisanship is defined with respect to a given vocabulary of J phrases.

- When the proportions are not half-half, a neutral prior might be sensible in sharp bipartite oppositions(?) But in general?

An old debate in the segregation literature

- Should a polarization/segregation measure depend on the proportions of the minority and majority groups?
- Or only depend on the dispersion of the choices without consideration for differences in the number of choices made by the two groups?
- We would rather advocate the latter

  Hence, the generalized partisanship index
- Nonetheless, the index π is still mildly composition-variant

  (see Appendix B.1 of DGR for details about the link with the Coworker segregation index)

$\Longrightarrow$ *Warning*: aggregated conditional indices are fallacious!

- By construction, the covariates are predictors of party membership, hence proportions far from half-half in the different cells

Figure: Strong composition-variance of the half-half index

Details on the moment space

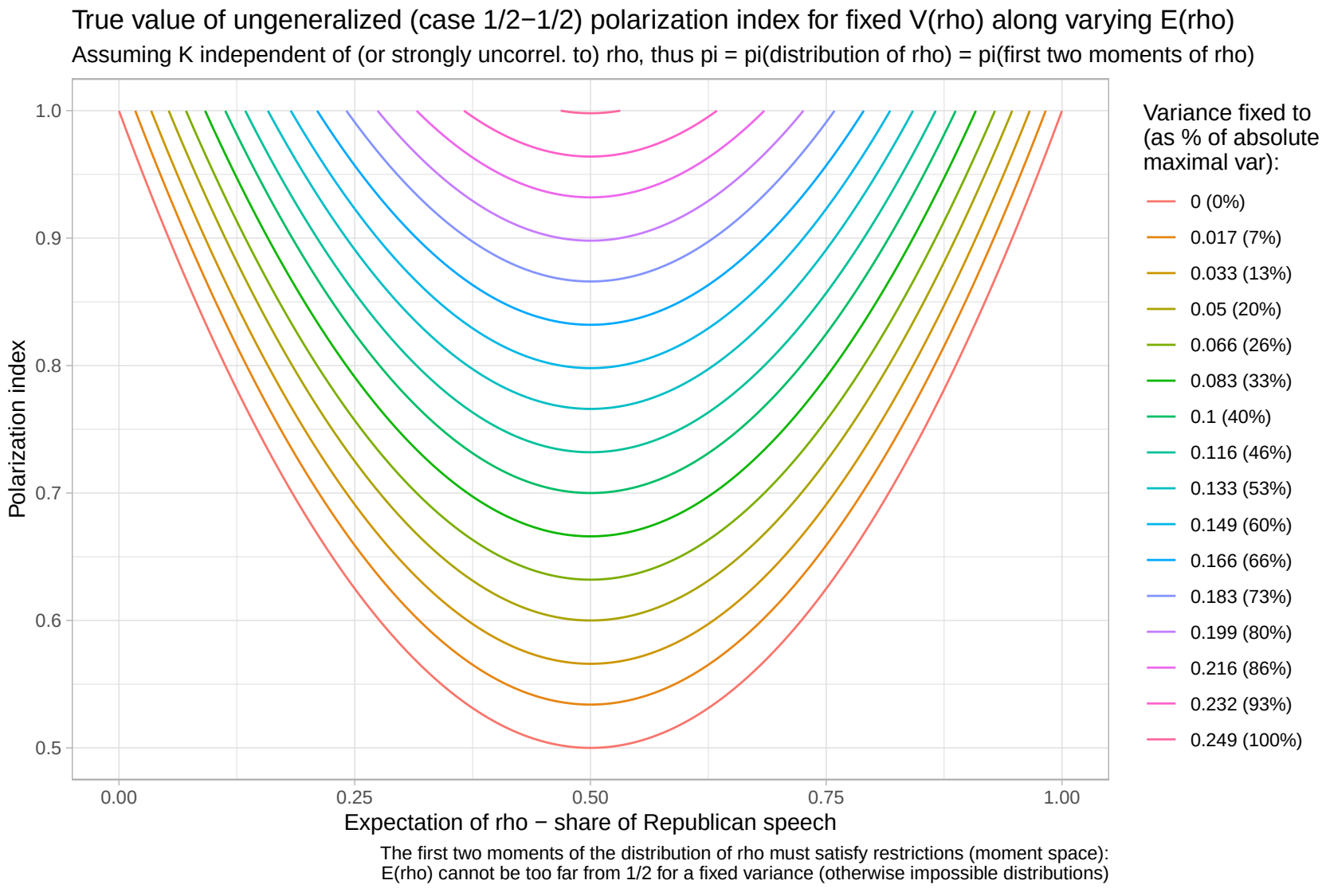
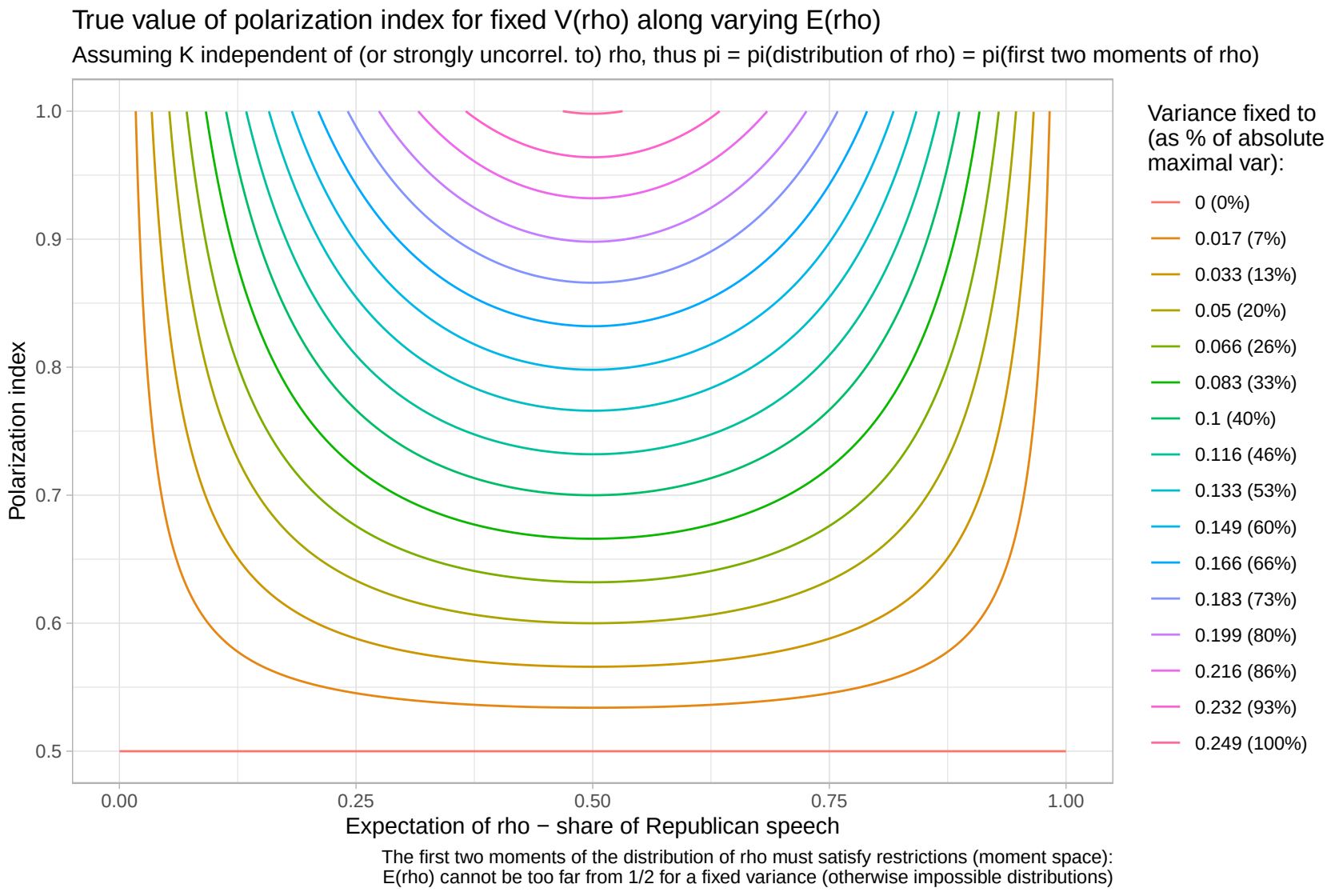


Figure: Mild composition-variance of the generalized index

Details on the moment space



Outline

Conclusion and questions

Overview

- An example of application of text analysis: measuring political speech polarization
- Can be seen as a special case of a more general methodological issue: "measuring group differences in high-dimensional choices":
  - The number n of observed *choices* (occurrences) is small relative to the number J of *options* (phrases)
    $\rightarrow$ few occurrences observed per option $\rightarrow$ small-unit bias
- Two methods with somewhat complementary pros and cons
- **Several interrogations remain:**
  - Is a "bag-of-words" approach sufficient for this application?
  - Reliability of the conventional i.i.d modeling? Details
  - Impact of upstream text processing and dictionary selection?
  - Relevant index (issue of composition-variance)?

Thank you for your attention.

In case of questions: `lucas.girard[at]ensae.fr`

Bonne continuation !

Outline

Illustration of speech polarization (1)

Back

Figure: Headlines of three New York newspapers on June 13th, 2016, in the aftermath of the Orlando shooting

Illustration of speech polarization (2)

Back

Figure: Headlines of some French newspaper on June 20th, 2022 following the legislative election

Influence of political discourse

Back

- Gentzkow and Shapiro (2010 Econometrica): *What Drives Media Slant? Evidence from U.S. Daily Newspaper*
  - Use US congressional speech data to construct an index of media slant that measures the similarity of a newspaper's language to that of a Republican or Democrat congressperson: political discourse diffuses into other domains of public debate
  - Use the measure into a structural model of newspaper demand that incorporates slant and find that readers do have a preference for like-minded news
- Chong and Druckman (2007 American Political Science Review): *Framing Public Opinions in Competitive Democracies*
  - Political discourse can have framing effects on public opinion

Segregation literature (1)

Back

Seminal papers defining common segregation indices

- Duncan, O. D., & Duncan, B. (1955). A methodological analysis of segregation indexes. *American sociological review*, 20(2), 210-217.
  $\longrightarrow$ Duncan or dissimilarity index, link of scalar indices with the "segregation curve" representation
- James, D. R., & Taeuber, K. E. (1985). Measures of segregation. *Sociological methodology*, 15, 1-32.
  $\longrightarrow$ Review of several proportion-based indices
- Massey, D. S., & Denton, N. A. (1988). The dimensions of residential segregation. *Social forces*, 67(2), 281-315.
  $\longrightarrow$ Five dimensions or axes of geographical segregation: evenness, exposure, concentration, centralization, and clustering
  - We look only at "evenness" here when considering speech polarization

Segregation literature (2)

Back

Papers dealing with the small-unit bias

- Jahn, J., C. F. Schmid, & C. Schrag (1947). The measurement of ecological segregation. *American Sociological Review*, 12 (3), 293-303.
  $\longrightarrow$ first occurrence of the idea of randomness benchmark
- Cortese, C. F., Falk, R. F., & Cohen, J. K. (1976). Further considerations on the methodological analysis of segregation indices. *American sociological review*, 630-637.
  $\longrightarrow$ first evidence of and correction of the small-unit bias
- Winship, C. (1977). A revaluation of indexes of residential segregation. *Social Forces*, 55 (4), 1058-1066.
  $\longrightarrow$ Distinction btw evenness and randomness benchmarks
- Carrington, W. J., & Troske, K. R. (1997). On measuring segregation in samples with small units. *Journal of Business Economic Statistics*, 15(4), 402-409.
  $\longrightarrow$ The commonly used correction of the bias for indices based on empirical proportions (v1)

Segregation literature (3)

Back

More recent papers dealing with the small-unit bias

- Aslund, O. & Skans, O. N. (2009). How to measure segregation conditional on the distribution of covariates. *Journal of Population Economics*, 22(4), 971-981.
  $\longrightarrow$ Extension of Carrington & Troske (1997) method for conditional indices
- Rathelot, R. (2012). Measuring segregation when units are small: a parametric approach. *Journal of Business Economic Statistics*, 30(4), 546-553.
  $\longrightarrow$ Introduction of segregation indices defined as functional of the distribution of ρ (v3), parametric estimation by mixtures of beta
- D'Haultfœuille, X., & Rathelot, R. (2017). Measuring segregation on small units: A partial identification analysis. *Quantitative Economics*, 8(1), 39-73.
  $\longrightarrow$ Same set-up, non-parametric partial identification and estimation

Segregation literature (4)

Back

- Allen, R., Burgess, S., Davidson, R., & Windmeijer, F. (2015). More reliable inference for the dissimilarity index of segregation. *The econometrics journal*, 18(1), 40-66.
  $\longrightarrow$ bootstrap-based correction of the bias, asymptotics with fixed number of units (= fixed dictionary)
- D'Haultfœuille, X., Girard L., & Rathelot, R. (forthcoming). segregsmall: A command to estimate segregation in the presence of small units. *The Stata Journal*.
  $\longrightarrow$ brief literature review and implements the methods of Carrington & Troske (1997), Rathelot (2012), D'Haultfœuille & Rathelot (2017)
- Kalter, F. (2000). Measuring segregation and controlling for independent variables. (unpublished)
  $\longrightarrow$ Idea to use a discrete choice model with fixed choice set

Segregation literature (5)

Back

Axiomatic properties of (proportion-based) indices

- Frankel, D. M., & Volij, O. (2011). Measuring school segregation. *Journal of Economic Theory*, 146(1), 1-38.
  $\longrightarrow$ Axiomatic properties of various indices
  Remark: the old sociological papers also discuss the respective and competing properties of indices

Methodological issues to extend to multi-group, i.e. > 2, settings

- Reardon, S. F., & Firebaugh, G. (2002). 2. Measures of Multigroup Segregation. *Sociological methodology*, 32(1), 33-67.
  $\longrightarrow$ with proportion-based indices v1
  $\longrightarrow$ problem is not evident, one possibility: use convex combinations of two-group indices across all the pairs

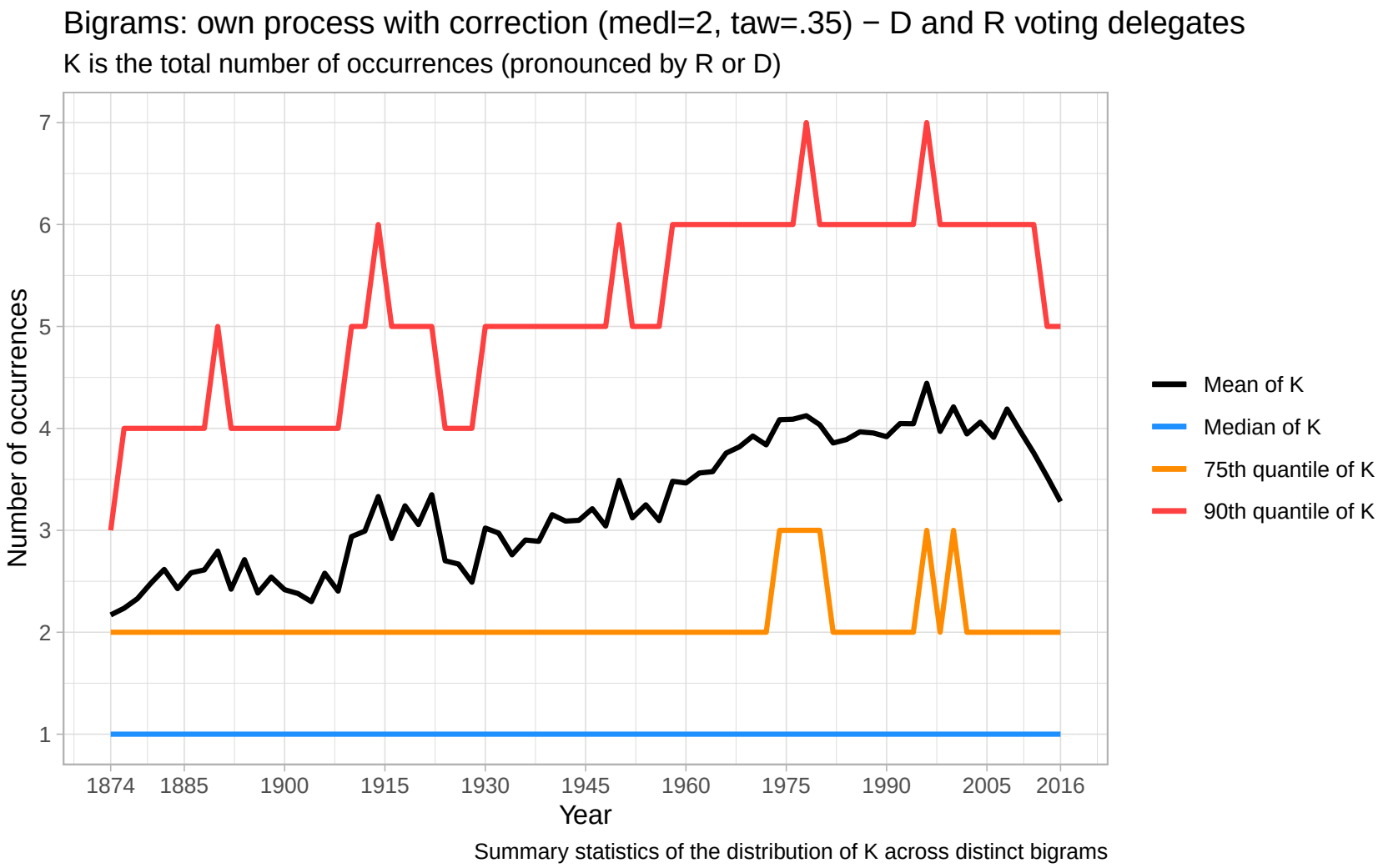
Political sciences literature

- Jensen J. et al. (2012 Brooking Papers on Economic Activity): *Political Polarization and the Dynamics of Political Language: Evidence from 130 years of Partisan Speech*
  - Use US congressional speech and Google Books corpus data
  - Identify partisan phrases from congressional speech and use them to measure partisanship and political polarization in Google Books corpus between 1873 and 2000
  - Polarization of discourse in books seems to predict legislative gridlock but polarization of congressional speech does not
- Peterson A. and Spirling A. (2016 mimeo): *Parliamentary Polarization: Cohort Effects and Ideological Dynamics in the UK House of Commons 1935-2013*
- Lauderdale B.E. and Herzog A. (2016 Political Analysis): *Measuring Political Positions from Legislative Speech*

K is small for most phrases

Back

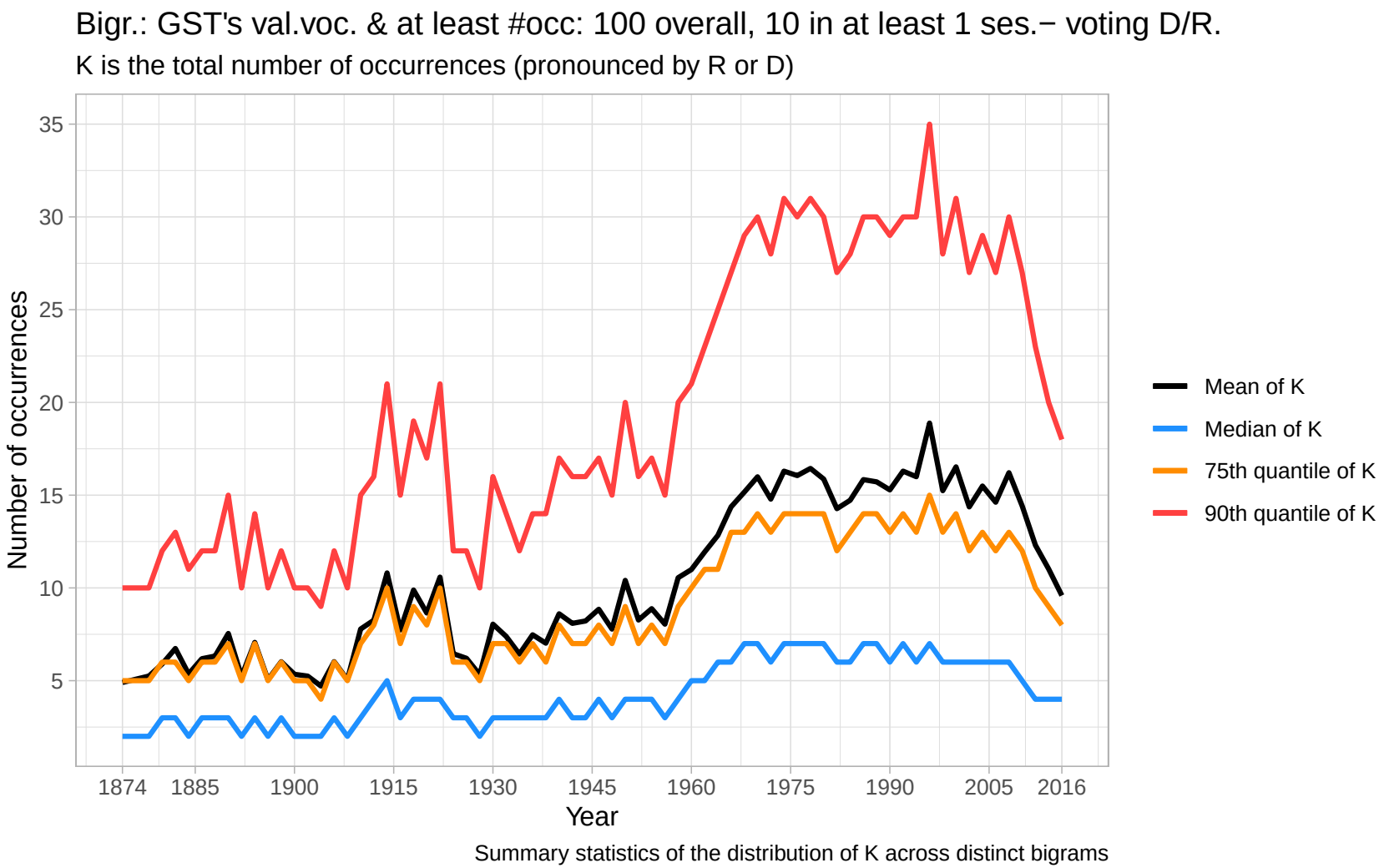
Figure: Processing with correction, suppression of "bad syntax" or "procedural" phrases, without restriction based on the number of occurrences



K is small for most phrases

Back

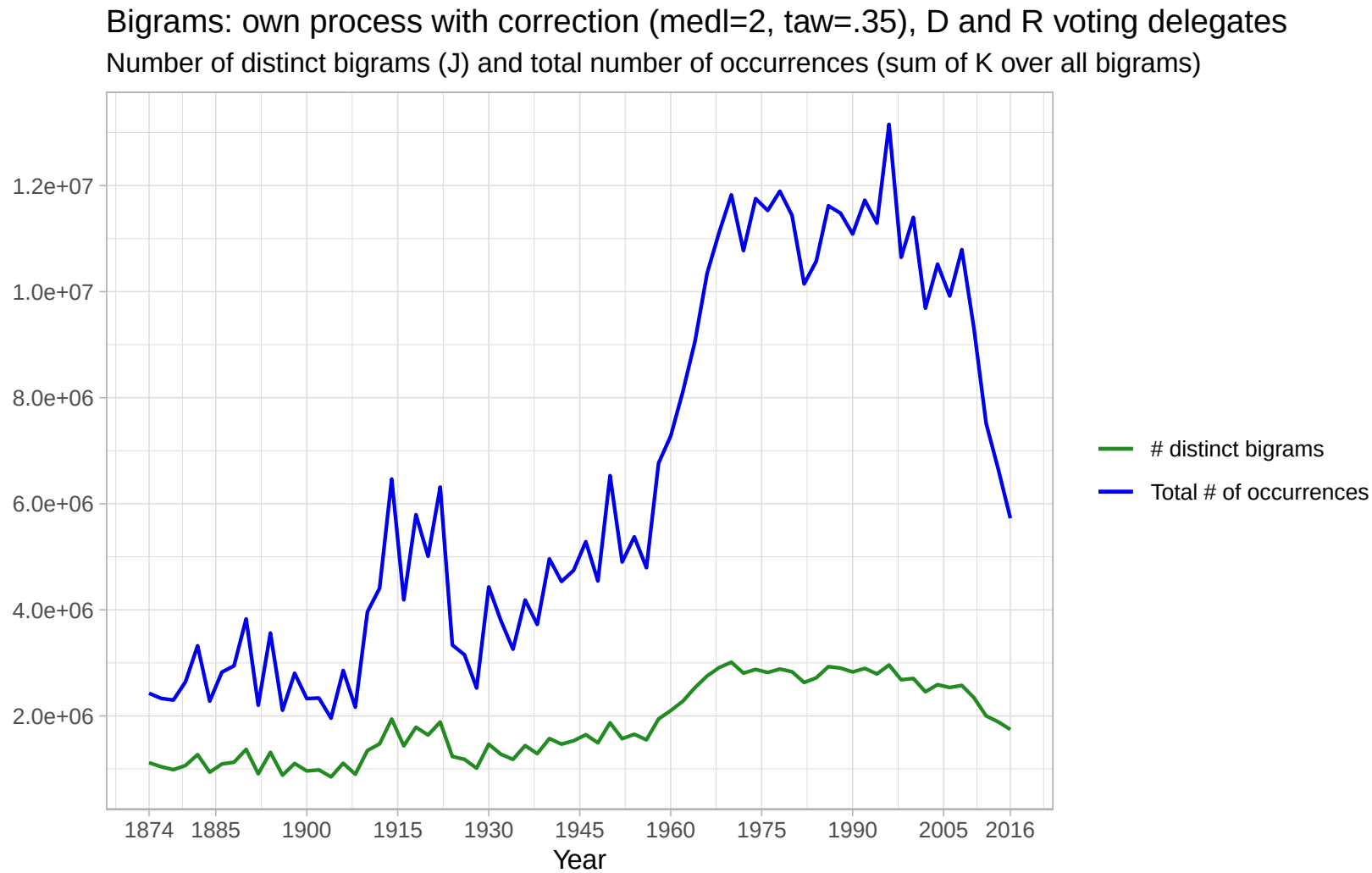
Figure: Processing with suppression of "bad syntax" or "procedural" phrases, with restrictions based on the number of occurrences



Longer speeches over time: small-unit bias ↓

Back

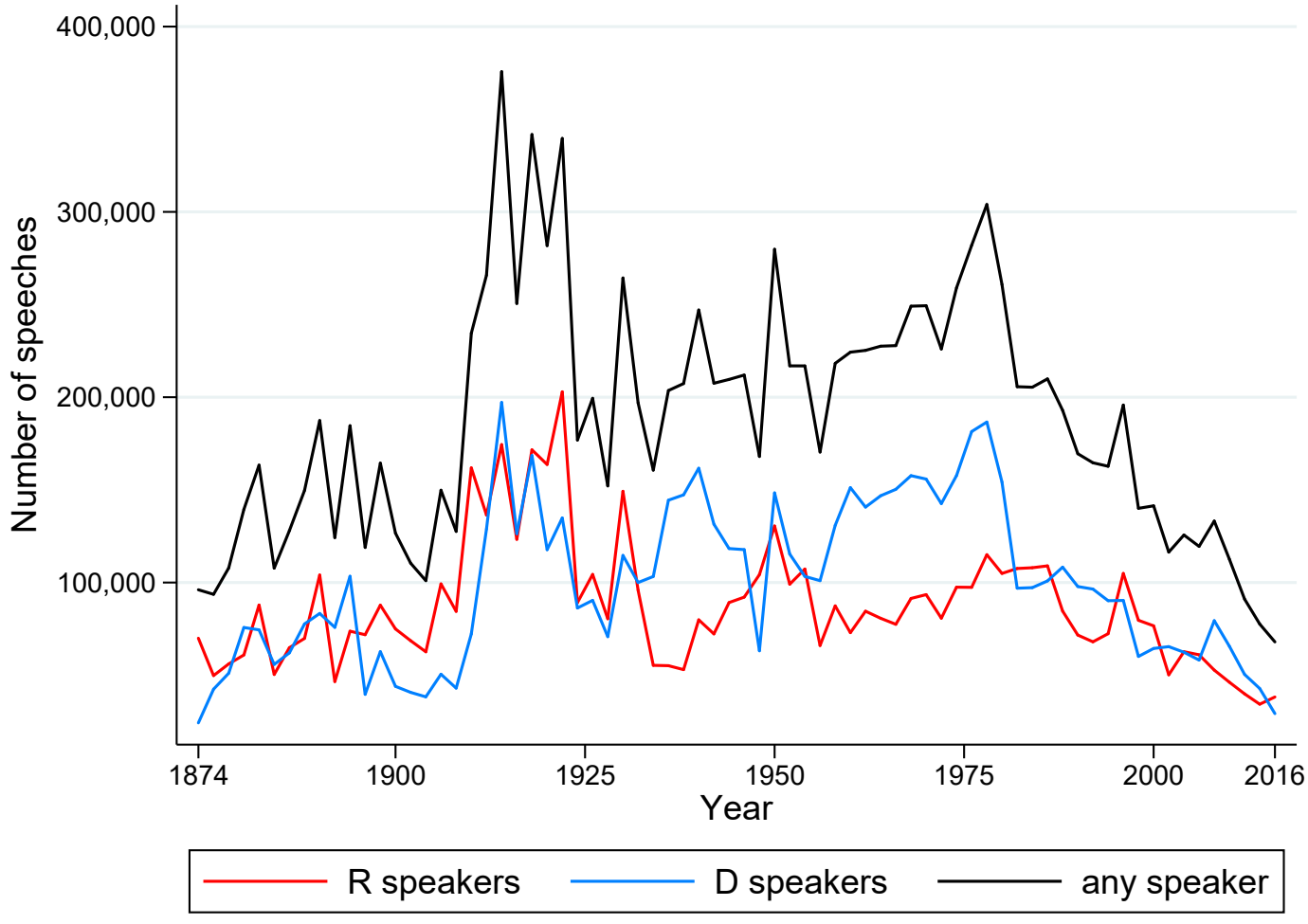
Figure: Processing with correction, suppression of "bad syntax" or "procedural" phrases, without restriction based on the number of occurrences



US bipartite political system

Back

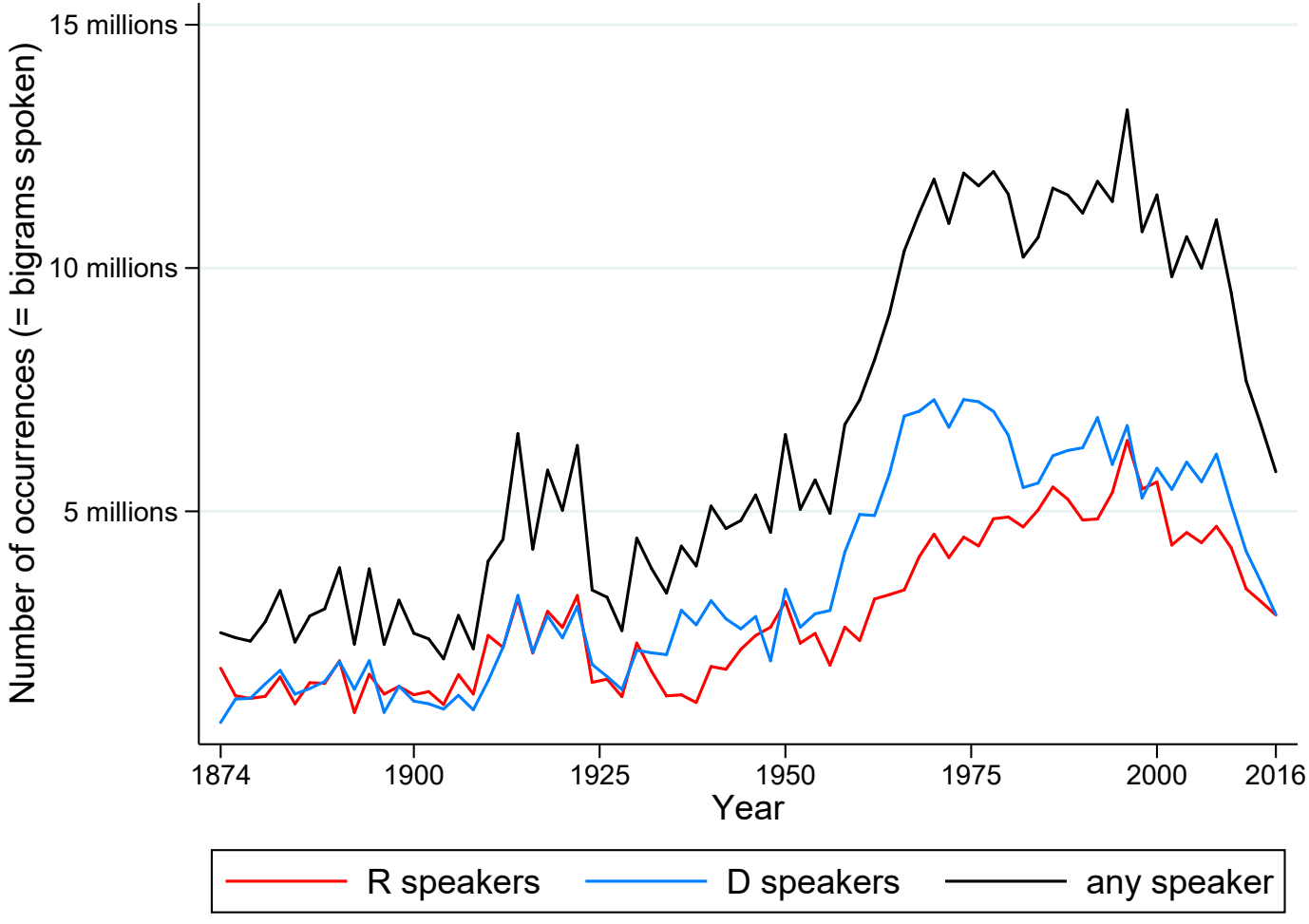
Figure: Processing with correction, suppression of "bad syntax" or "procedural" phrases, without restriction based on the number of occurrences



US bipartite political system

Back

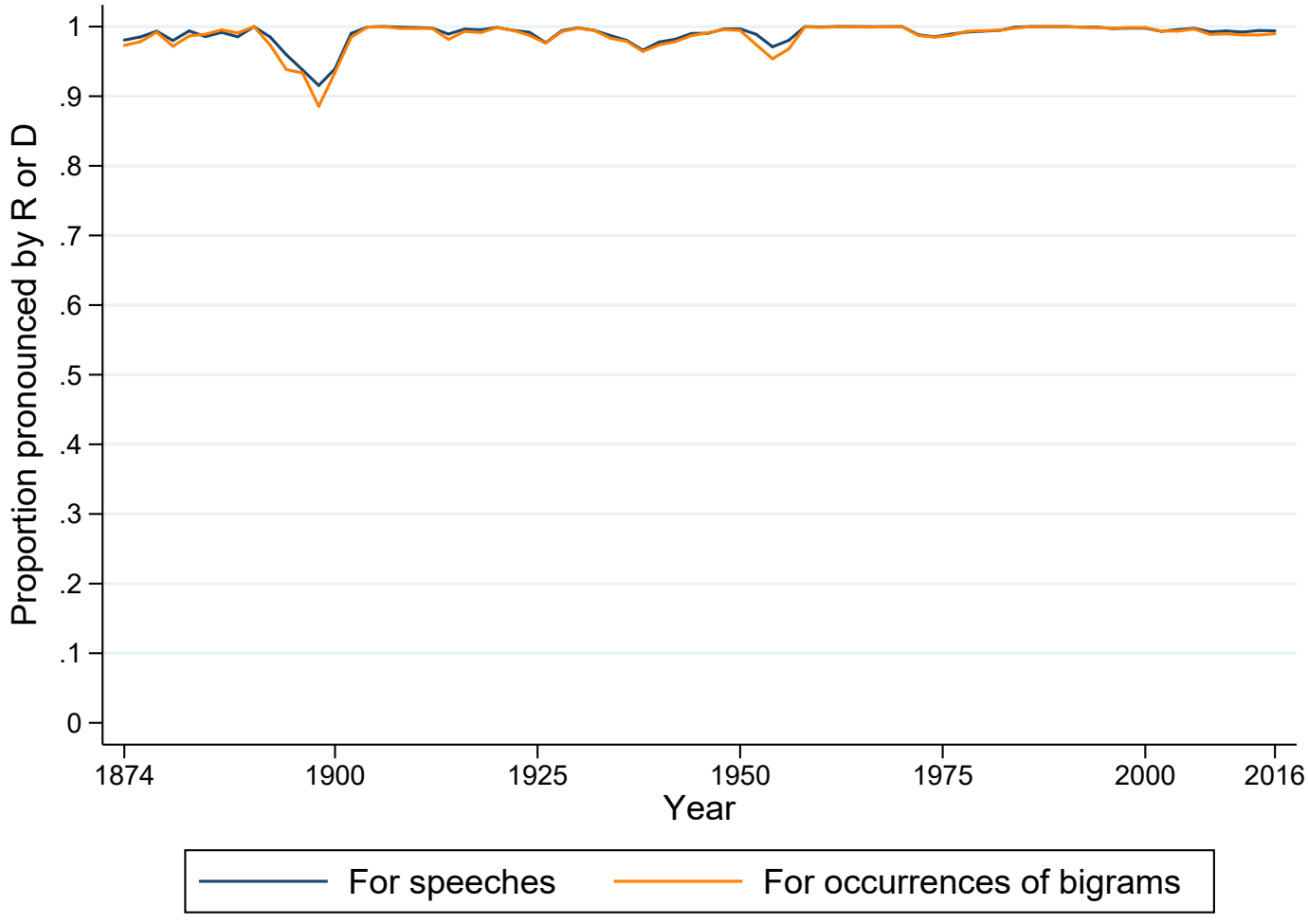
Figure: Processing with correction, suppression of "bad syntax" or "procedural" phrases, without restriction based on the number of occurrences



US bipartite political system

Back

Figure: Processing with correction, suppression of "bad syntax" or "procedural" phrases, without restriction based on the number of occurrences



Common proportion-based indices

Back

Define $n := \sum_{j=1}^{J} K_j$ the total number of occurrences,

$\widehat{q} := \frac{\sum_{j=1}^{J} K_j^R}{\sum_{j=1}^{J} K_j}$ the empirical overall share of speech said by R,

$\forall j \in \{1, \ldots, J\}$, $s_j := K_j^R / K_j$, *idem* for the j-th word

- Duncan: $D := \frac{\sum_{j=1}^{J} K_j |s_j - \widehat{q}|}{2\, n\, \widehat{q}\,(1-\widehat{q})}$
- Gini: $G := \frac{\sum_{j=1}^{J} \sum_{\ell=1}^{J} K_j\, K_\ell\, |s_j - s_\ell|}{2\, n^2\, \widehat{q}\,(1-\widehat{q})}$
- Theil: $T := \frac{\sum_{j=1}^{J} K_j\, (E - E_j)}{n\, E}$, with
  - $E := \widehat{q} \log_2\left(\frac{1}{\widehat{q}}\right) + (1-\widehat{q}) \log_2\left(\frac{1}{1-\widehat{q}}\right)$, whole population's entropy
  - $E_j = s_j \log_2\left(\frac{1}{s_j}\right) + (1-s_j) \log_2\left(\frac{1}{1-s_j}\right)$, j-th option/unit's entropy
- Atkinson: for $b \in (0,1)$,

$$A_b := 1 - \frac{\widehat{q}}{1-\widehat{q}} \left(\frac{1}{n\,\widehat{q}} \sum_{j=1}^{J} K_j\, s_j^b\, (1-s_j)^{1-b} \right)^{\frac{1}{1-b}}$$

- Coworker: $CW := \frac{\sum_{j=1}^{J} K_j (s_j - \widehat{q})^2}{n\, \widehat{q}\,(1-\widehat{q})}$

Probability-based indices

Back

Define $m_{0k} := \mathbb{E}_{\rho\sim P^\rho}\left[\rho^k\right]$ and F_ρ the c.d.f. of P^ρ, the distribution of ρ

- Duncan:
$$D^d := D(P^\rho) := \frac{1}{2}\mathbb{E}\left[\left|\frac{\rho}{\mathbb{E}(\rho)} - \frac{1-\rho}{1-\mathbb{E}(\rho)}\right|\right] = \frac{\int_0^1 |u-m_{01}|dF_\rho(u)}{2m_{01}(1-m_{01})}$$
- Gini: $G^d := G(P^\rho) := \frac{1-m_{01}-\int_0^1 F_\rho(u)^2(u)du}{m_{01}(1-m_{01})}$
- Theil: $T^d := T(P^\rho) := 1 - \frac{\int_0^1 u\ln(u)dF_\rho(u)+\int_0^1(1-u)\ln(1-u)dF_\rho(u)}{m_{01}\ln(m_{01})+(1-m_{01})\ln(1-m_{01})}$
- Atkinson: for $b \in (0,1)$,
$$A_b^d := A_b(P^\rho) := 1 - \frac{m_{01}^{-\frac{b}{1-b}}}{1-m_{01}}\left(\int_0^1(1-u)^{1-b}u^b dF_\rho(u)\right)^{\frac{1}{1-b}}$$
- Coworker:
$$CW^d := CW(P^\rho) := \frac{\int_0^1(u-m_{01})^2 dF_\rho(u)}{m_{01}-m_{01}^2} = \frac{Var(\rho)}{m_{01}-m_{01}^2} = \frac{m_{02}-m_{01}^2}{m_{01}-m_{01}^2}$$

GST's method – inference and asymptotics

Back

- Inference by subsampling
- Asymptotics: fixed dictionary while the number of distinct speakers grows to infinity (hence the length of the text grows to infinity too)

Politis, Romano, and Wolf (1999, Theorem 2.2.1) showed that this procedure yields valid confidence intervals under the assumption that the distribution of the estimator converges weakly to some non-degenerate distribution at a $\sqrt{n}$ rate. In the Appendix, we extend a result of Knight and Fu (2000) to show that this property holds, with fixed vocabulary and a suitable rate condition on the penalty, for the penalized maximum likelihood estimator of our multinomial logit model. This is the estimator that we approximate with the Poisson distribution in equation (9). Though we do not pursue formal results for the case where the vocabulary grows with the sample size, we note that such asymptotics might better approximate the finite-sample behavior of our estimators.

Herdan's or Heaps's law

Back

- Herdan (1960) in linguistics, Heaps (1978) in NLP
- Empirical relationship between the number J of words/phrases ("types") and the number n of occurrences ("tokens") in a corpus of texts:

$$J = \alpha n^{\beta}$$

with $\alpha > 0$ and $0 < \beta < 1$ constants that depend on the corpus genre

  - Several classical English corpora (Shakespeare, Brown corpus, telephone conversations, Google N-grams) display β from around 0.67 to 0.75
  - *Interpretation*: the dictionary size J for a corpus of texts typically increases faster than the square root of its length $n := \sum_{j=1}^{J} K_j$ (total # of observed occurrences)

Test of binomial assumption

Back

- Reference: D'Haultfœuille and Rathelot (Quantitative Economics 2017)
- The binomial assumption implies that, conditional on $K = k$, there exists a known one-to-one mapping between P^{K^R} (defined by k probabilities) and m_k the vector of first k moments of P^ρ: m_k is identified from P^{K^R}
- But m_k cannot lie anywhere in $[0,1]^k$ (e.g. variance ≥ 0)
- Idea: test whether $\widehat{m}_k$ obtained under the binomial assumption is a valid vector of moments i.e. belongs to the moment space $\mathcal{M}_k$
- The null hypothesis of binomial distribution is not rejected in the text Congress data of our application

Moment space (Illustration)

Back (model) Back (composition variance)

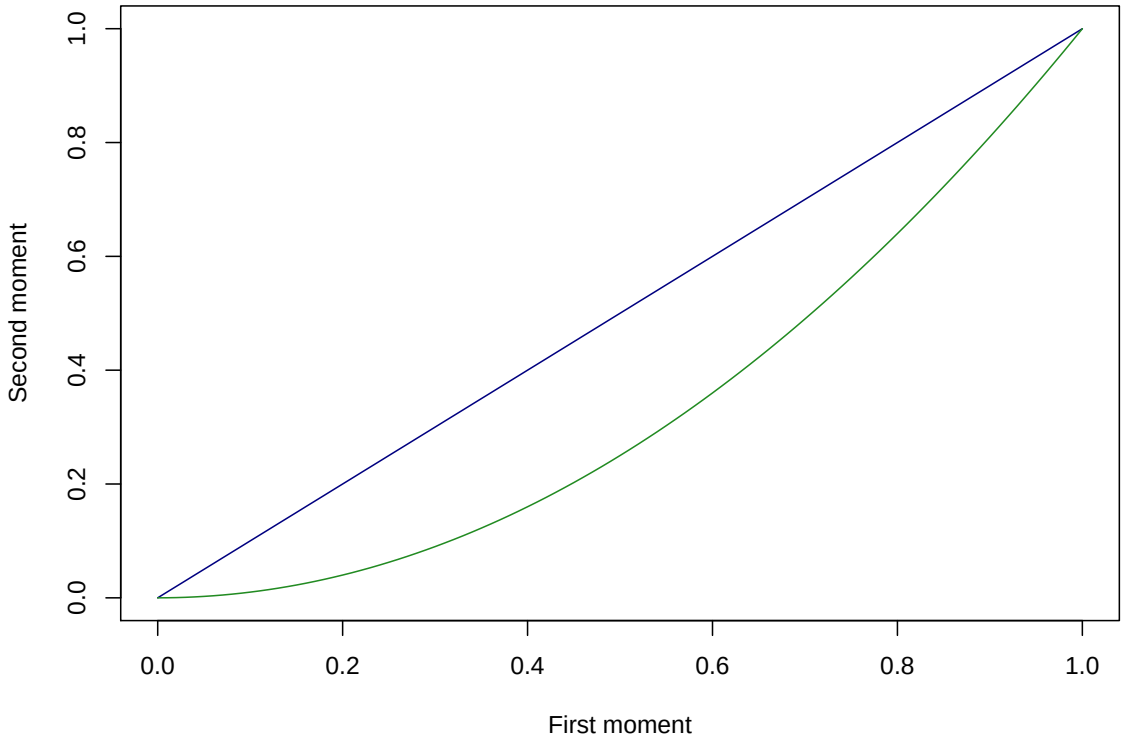
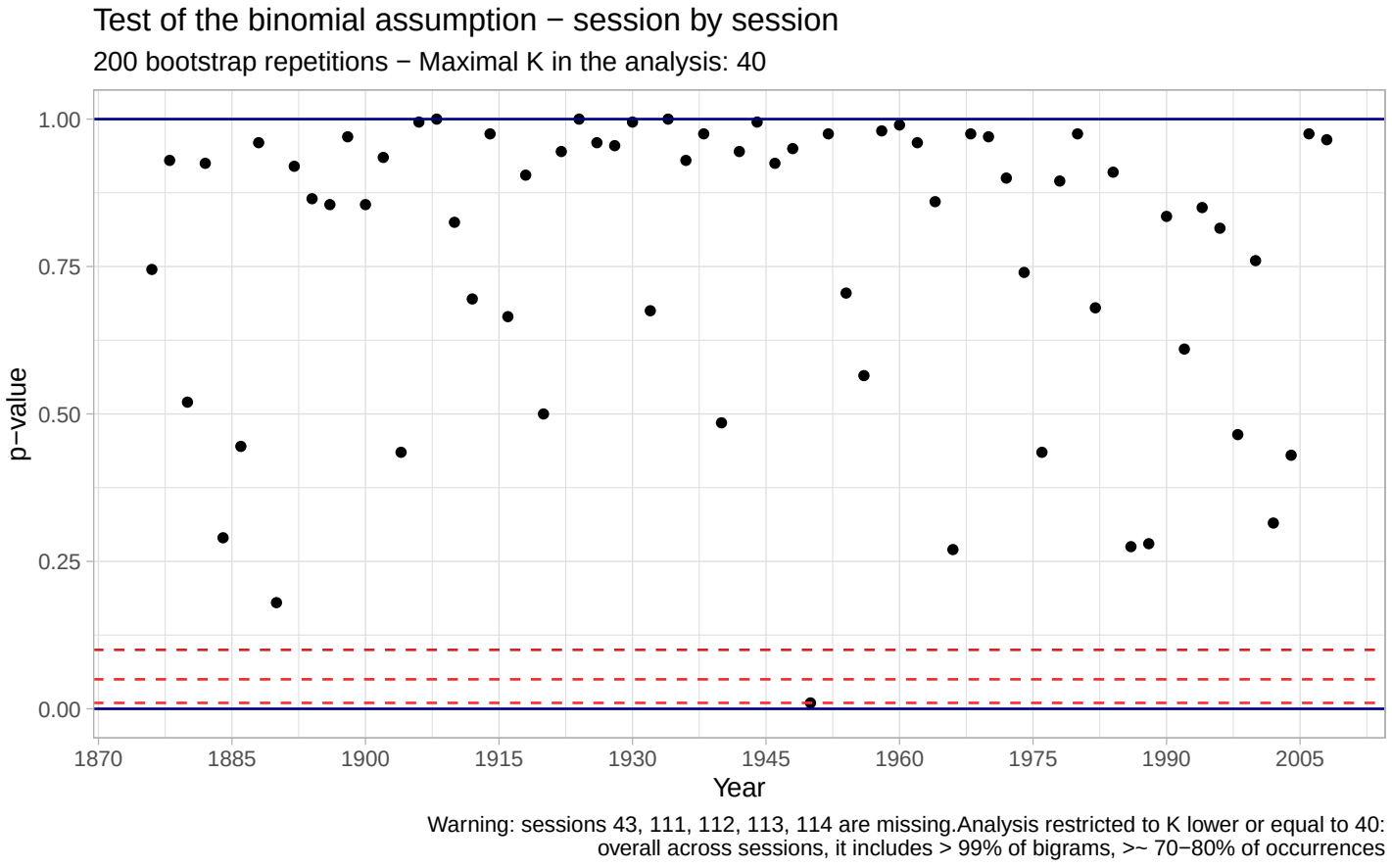


Figure: Example of moment space (case $K = 2$).
The vector of first two moments $m = (m_1, m_2)$ cannot lie anywhere in $[0, 1]^2$:
non-negativity of variance: $m_2 \geq m_1^2$ and support in $[0, 1]$: $0 \leq m_2 \leq m_1 \leq 1$.
If $m_1 = m_2$, Bernoulli(m_1) (blue frontier).
If $m_2 = m_1^2$ (i.e., null variance), Dirac(m_1) (green frontier).

The binomial assumption is (overall) not rejected

Back

Figure: p-values for the test whose null hypothesis is the conditional distribution of K^R knowing K and ρ – 200 bootstrap repetitions – N.B.: analysis limited to $\{j \in \mathcal{J} : K_j \leq 40\}$



Herdan's or Heaps's law

Back

- Herdan (1960) in linguistics, Heaps (1978) in NLP
- Empirical relationship between the number J of words/phrases ("types") and the number n of occurrences ("tokens") in a corpus of texts:

$$J = \alpha n^{\beta}$$

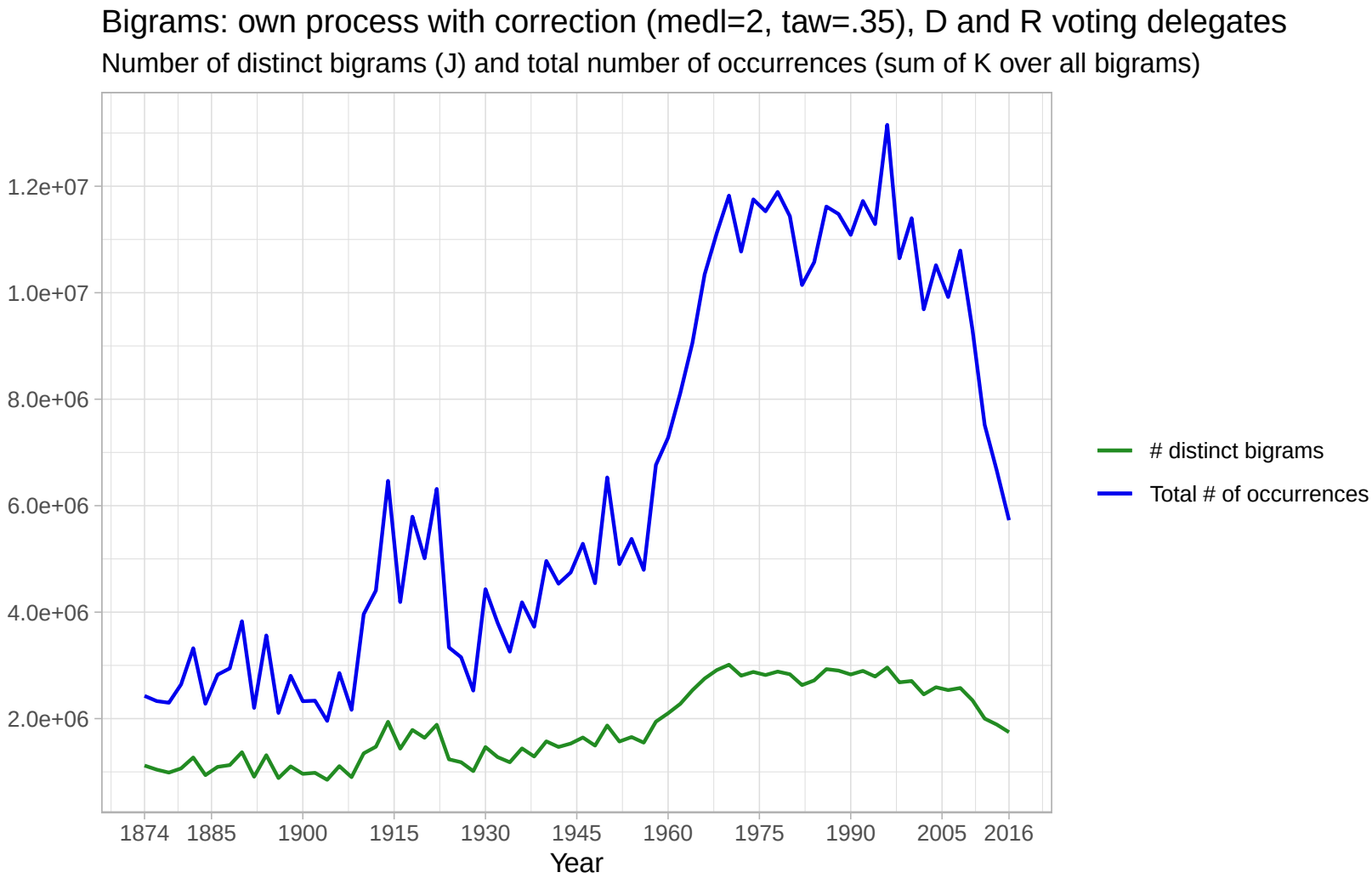
with $\alpha > 0$ and $0 < \beta < 1$ constants that depend on the corpus genre

  - Several classical English corpora (Shakespeare, Brown corpus, telephone conversations, Google N-grams) display β from around 0.67 to 0.75
  - *Interpretation*: the dictionary size J for a corpus of texts typically increases faster than the square root of its length n

Herdan's or Heaps's law: illustration

Back

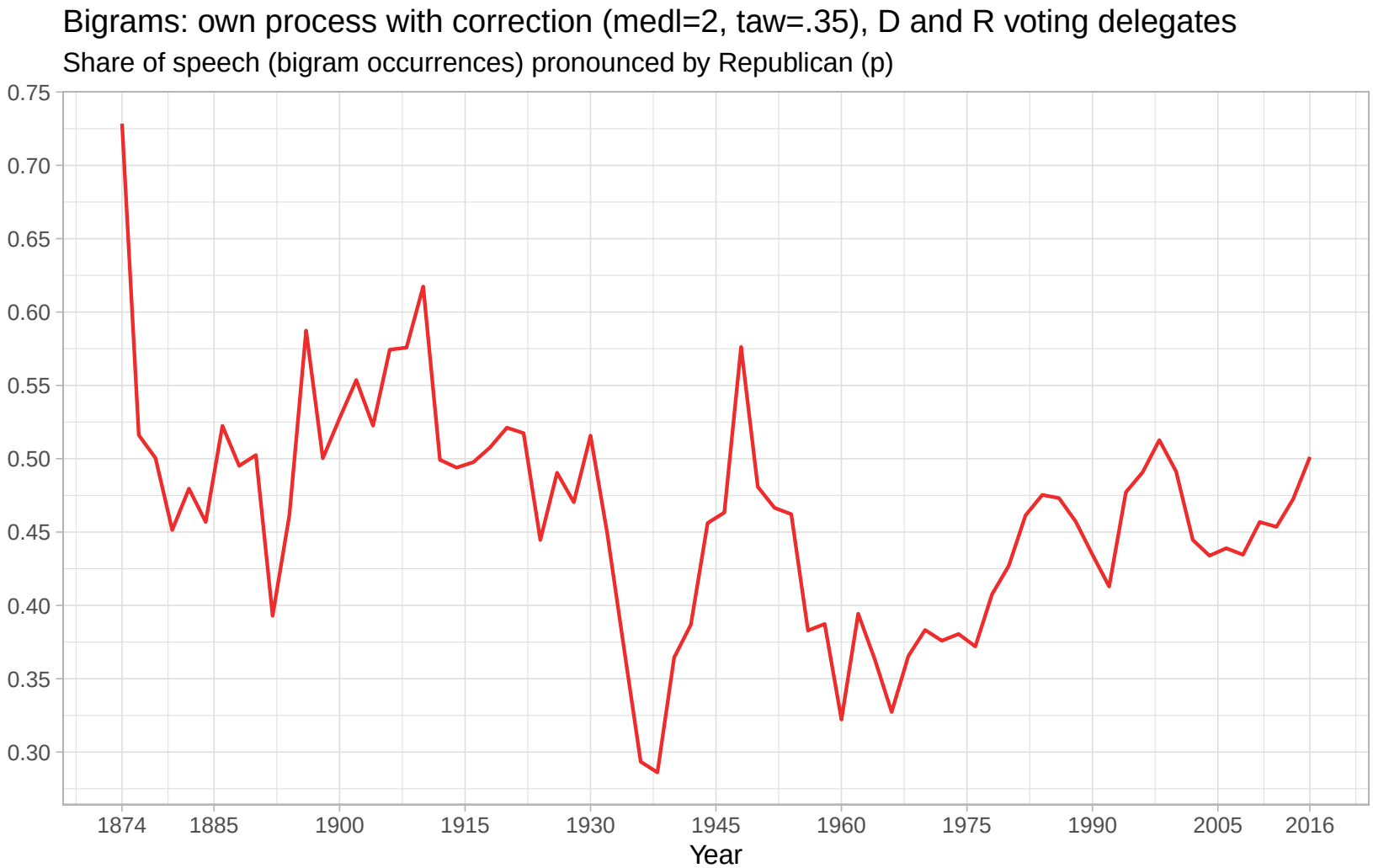
Figure: Processing with correction, suppression of "bad syntax" or "procedural" phrases, without restriction based on the number of occurrences



Evolution of p over time

Back

Figure: Processing with correction, suppression of "bad syntax" or "procedural" phrases, without restriction based on the number of occurrences



Proof of identification (1)

Back

- We want to identify and estimate:

$$\pi := 1 - \frac{\mathbb{E}[K\rho(1-\rho)]}{2\mathbb{E}[K]\,p(1-p)} = 1 - \frac{\mathbb{E}[K\rho] - \mathbb{E}[K\rho^2]}{2\mathbb{E}[K]\frac{\mathbb{E}[K\rho]}{\mathbb{E}[K]}\left(1 - \frac{\mathbb{E}[K\rho]}{\mathbb{E}[K]}\right)}$$

- Use DGP assumption:

$$\mathbb{E}[K\rho] = \mathbb{E}\Big[\mathbb{E}\left(K^R | K, \rho\right)\Big] = \mathbb{E}[K^R]. \quad \mathbb{E}\big[K\rho^2\big] = ?$$

- And algebraic simplifications using $K = K^R + K^D$ to obtain the expression as stated in the theorem
- Besides, we use the convention: $\mathbb{E}[K^R \mathbb{1}\{K = 1\}]/\mathbb{P}(K = 1) = 0$ when $\mathbb{P}(K = 1) = 0$

Proof of identification (2)

Back

- Use proportions instead of conditional probabilities?
  $\Longrightarrow$ positive bias: illustration of the small-unit bias

If we replace ρ by K^R/K:

$$\mathbb{E}\left[K\rho^2\right] \overset{\text{replaced by}}{\longrightarrow} \mathbb{E}\left[K\left(\frac{K^R}{K}\right)^2\right] = \mathbb{E}\left[K^{-1}\,\mathbb{E}\Big((K^R)^2 \mid K, \rho\Big)\right]$$

$$\overset{\text{Jensen}}{\geq} \mathbb{E}\left[K^{-1}\,\mathbb{E}\Big(K^R \mid K, \rho\Big)^2\right]$$

$$= \mathbb{E}\left[K^{-1}\,(K\rho)^2\right] = \mathbb{E}\left[K\rho^2\right]$$

- Idea is rather to use the binomial assumption

Proof of identification (3)

Back

- Define $m(k) := \mathbb{E}\left[\rho^2 \,|\, K = k\right], \forall k \in \mathbb{N}$

$$\mathbb{E}[K\rho^2] = \mathbb{E}\left[\mathbb{E}\left(K\rho^2 \mid K\right)\right] = \mathbb{E}\left[K \times \mathbb{E}\left(\rho^2 \,|\, K\right)\right] = \mathbb{E}[Km(K)]$$

$$\mathbb{E}\left[K^R(K^R-1) \,|\, K, \rho\right] = \mathbb{E}_{B\sim\mathcal{B}(K,\rho)}\left[B^2 - B\right] = K(K-1)\rho^2$$

$$\implies \mathbb{E}\left[K^R(K^R-1) \,|\, K\right] = K(K-1)\mathbb{E}\left[\rho^2 \,|\, K\right] = K(K-1)m(K)$$

$$\implies \mathbb{E}[K\rho^2] = \mathbb{E}\left[\frac{K^R(K^R-1)}{K-1}\mathbb{1}\{K>1\}\right] + m(1)\mathbb{P}(K=1)$$

as $Km(K) = Km(K)\mathbb{1}\{K>1\} + Km(K)\mathbb{1}\{K=1\}$

- $m(1) = \mathbb{E}\left[\rho^2 \,|\, K=1\right]$ is only partially identified:

$$\mathbb{E}\left[K^R \,|\, K=1\right]^2 \overset{\text{Jensen}}{\leq} m(1) \overset{\rho\in[0,1]}{\leq} \mathbb{E}\left[K^R \,|\, K=1\right]$$

using: $\mathbb{E}[\rho \,|\, K=1] = \mathbb{E}\left[\mathbb{E}\left(K^R \,|\, K=1, \rho\right) \,|\, K=1\right] = \mathbb{E}\left[K^R \,|\, K=1\right]$

Construction of CI for π (1)

Back

- Interest in a partially identified parameter $\pi \in [\underline{\pi}, \overline{\pi}]$
  $\longrightarrow$ interest in coverage of π rather than $[\underline{\pi}, \overline{\pi}]$
- Classical inference (point identification, asymptotic normality):

$$\left[\hat{\pi} - \Phi^{-1}(1-\alpha/2)\,\frac{\text{se}(\hat{\pi})}{\sqrt{J}}\;,\; \hat{\pi} + \Phi^{-1}(1-\alpha/2)\,\frac{\text{se}(\hat{\pi})}{\sqrt{J}}\right]$$

- Idea: asymptotically $\Delta := \overline{\pi} - \underline{\pi}$ is large relative to sampling error, thus non-coverage risk is one-sided
  $\longrightarrow$ use $\Phi^{-1}(1-\alpha)$ instead of $\Phi^{-1}(1-\alpha/2)$
- Yet, the intuition does not work uniformly: it fails when Δ does not diverge relative to sampling error
  - CI would shrink as a parameter moves from point identification to slight under-identification
- Solution: adapt the quantile used to the value of $\hat{\Delta} := \widehat{\overline{\pi}} - \widehat{\underline{\pi}}$ and the precision of the estimated bounds

Construction of CI for π (2)

Back

- Adapting Imbens and Manski (2004) and Stoye (2009) in our setting, we construct CI with asympotic size $1-\alpha$:

$$\mathsf{CI}^{\pi}_{1-\alpha} = \left[\underline{\widehat{\pi}} - \frac{c_\alpha \sqrt{\hat{a}}}{\sqrt{J}}, \widehat{\overline{\pi}} + \frac{c_\alpha \sqrt{\hat{c}}}{\sqrt{J}}\right],$$

where c_α solves

$$\Phi\left(c_\alpha + \frac{\sqrt{J}\,\hat{\Delta}}{\max\left\{\sqrt{\hat{a}}, \sqrt{\hat{c}}\right\}}\right) - \Phi(-c_\alpha) = 1-\alpha\,.$$

- Remark that:

$$\Phi(c_\alpha) - \Phi(-c_\alpha) = 1-\alpha \iff c_\alpha = \Phi^{-1}(1-\alpha/2)$$

$$\Phi(+\infty) - \Phi(-c_\alpha) = 1-\alpha \iff c_\alpha = \Phi^{-1}(1-\alpha)$$

Construction of CI for π (3)

Back

- Provided some additional regularity condition, Lemma 3 and Proposition 1 of Stoye (2009) ensure the asymptotic uniform coverage of $CI^{\pi}_{1-\alpha}$ since basically:
  - $(\widehat{\underline{\pi}}, \widehat{\overline{\pi}})$ are jointly asymptotically normal estimators
  - $\widehat{\overline{\pi}} \geq \widehat{\underline{\pi}}$ almost surely, by construction

Theorem (Asymptotic CI for π)

Under Assumption (DGP) and provided $\underline{\sigma}^2 \leq d \leq \overline{\sigma}^2$, $d \in \{a, b, c\}$ *for some positive and finite constants* $\underline{\sigma}^2$ *and* $\overline{\sigma}^2$, *for any* $\alpha \in (0, 1)$,

$$\lim_{J \to +\infty} \mathbb{P}\left(\pi \in CI^{\pi}_{1-\alpha}\right) = 1 - \alpha$$

Simulations: implementation

Back

- We take $J = 1{,}000$
- First layer: draw $\{(\lambda_j, \rho_j)\}_{j=1,\ldots,J}$ i.i.d. as follows:
  - λ_j drawn from a Gamma with parameters fixed so that $(E(K), Pr(K = 1)) = (5, 10\%)$ or $(15, 5\%)$
  - ρ_j drawn from a Beta with parameters fixed so that $(p, \pi) \simeq (0.5, 0.545)$ or $(0.2, 0.530)$
  - (λ_j, ρ_j) have Gaussian copula with correlation equal to 0 or 0.5
- Second layer: draw $K_j \sim \mathcal{P}(\lambda_j)$ and $K_j^R \sim \text{Binomial}(K_j, \rho_j)$
- Following tables report means over 5,000 draws, except for π which is an MC approximation of the true π over a very large sample ($J = 10^7$)
- Confidence intervals are at 5% nominal level

Simulations: bounds and coverage rate

Back

$E(K)$	p	$\text{Cor}(\rho, \lambda)$	$\%K = 1$	π	bounds *(CI)*	cov. rate
5	0.5	0	10%	0.5453	[0.5447, 0.5546] *(0.5345, 0.5649)*	0.965
5	0.5	0.5	10%	0.5452	[0.5445, 0.5539] *(0.5343, 0.5641)*	0.962
5	0.2	0	10%	0.5303	[0.5296, 0.5396] *(0.5196, 0.5500)*	0.967
5	0.2	0.5	10%	0.5306	[0.5302, 0.5371] *(0.5204, 0.5472)*	0.967
15	0.5	0	5%	0.5455	[0.5453, 0.5471] *(0.5390, 0.5534)*	0.955
15	0.5	0.5	5%	0.5417	[0.5415, 0.5431] *(0.5353, 0.5493)*	0.949
15	0.2	0	5%	0.5303	[0.5300, 0.5318] *(0.5246, 0.5373)*	0.955
15	0.2	0.5	5%	0.5295	[0.5293, 0.5303] *(0.5234, 0.5361)*	0.950

Simulations: extrapolated and (RMSE)

Back

$E(K)$	p	$\mathrm{Cor}(\rho, \lambda)$	$\%K = 1$	π	extrapolation	naive
5	0.5	0	10%	0.5453	0.5455	0.6316
					(0.0066)	*(0.086)*
5	0.5	0.5	10%	0.5452	0.5532	0.6308
					(0.010)	*(0.086)*
5	0.2	0	10%	0.5303	0.5301	0.7562
					(0.0062)	*(0.23)*
5	0.2	0.5	10%	0.5306	0.5336	0.7507
					(0.0068)	*(0.22)*
15	0.5	0	5%	0.5455	0.5455	0.5743
					(0.0036)	*(0.029)*
15	0.5	0.5	5%	0.5417	0.5424	0.5703
					(0.0036)	*(0.029)*
15	0.2	0	5%	0.5303	0.5301	0.7184
					(0.0032)	*(0.19)*
15	0.2	0.5	5%	0.5295	0.5295	0.7140
					(0.0033)	*(0.18)*

Simulations: details (1)

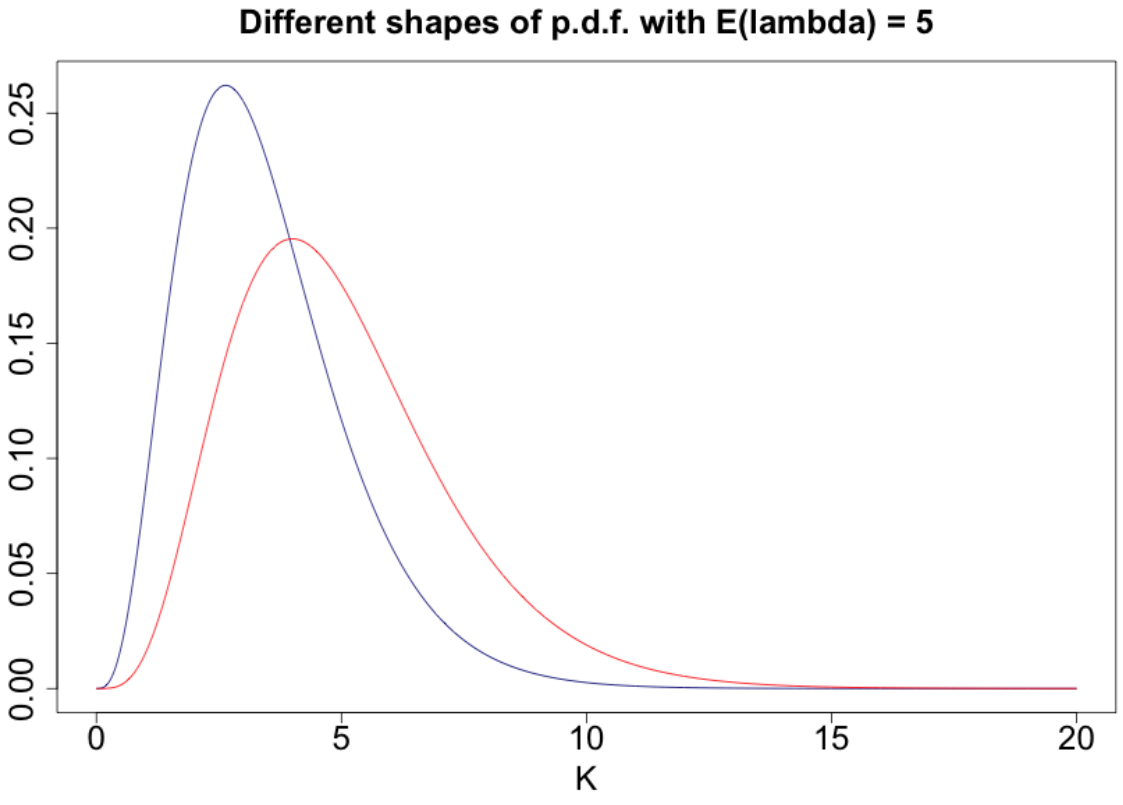
Back

- We use the inversion method to introduce correlation between ρ and λ while choosing arbitrary marginal distributions
  - draw bivariate normal $Z = (Z_1, Z_2)$, possibly with correlation
  - apply Φ the c.d.f. of $\mathcal{N}(0,1)$ to each component to get uniforms: $(U_1, U_2) = (\Phi(Z_1), \Phi(Z_2))$
  - apply adequate generalized inverse of c.d.f., here: $(\lambda, \rho) = (F^{-1}_{\text{Gamma}}(U_1), F^{-1}_{\text{Beta}}(U_2))$
  - correlation between Z_1 and Z_2 translates into correlation between λ and ρ

Simulations: details (2)

Back

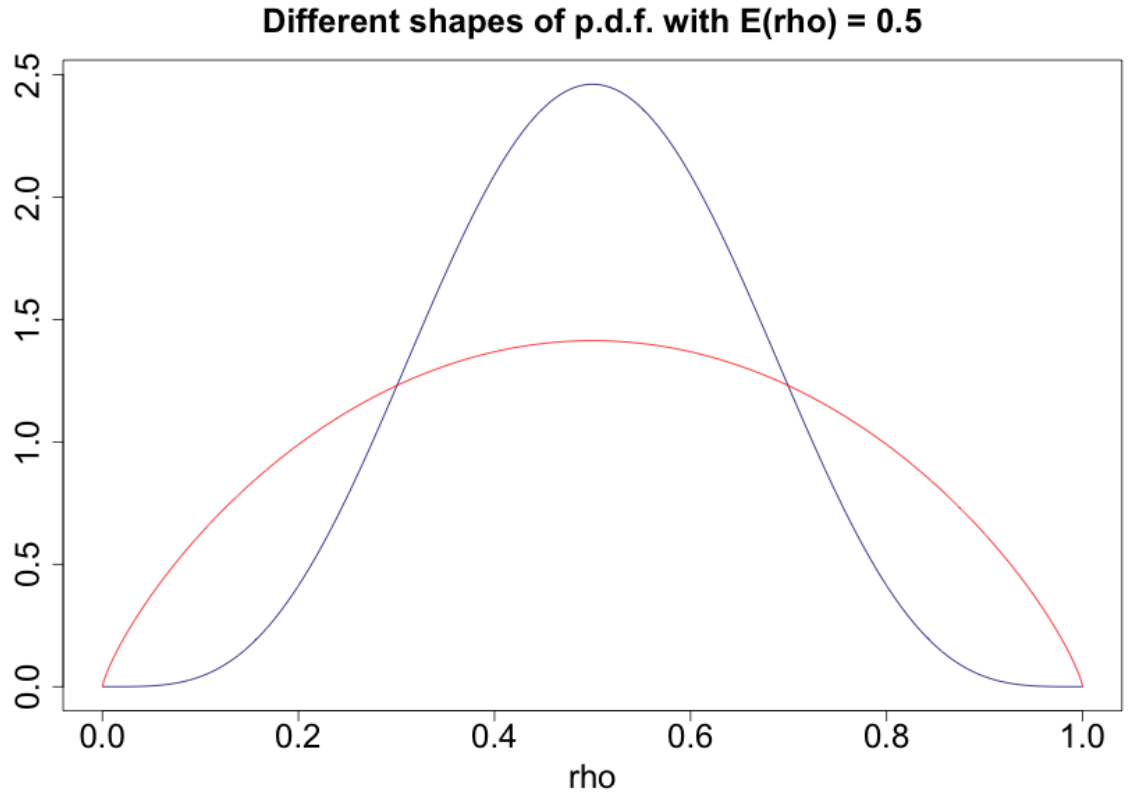
- The two parameters of the Gamma distribution are set to control
  - $E(\lambda)$, hence $E(K)$ in our DGP
  - the shape in order to monitor the proportion of words with only one occurrence



Simulations: details (3)

Back

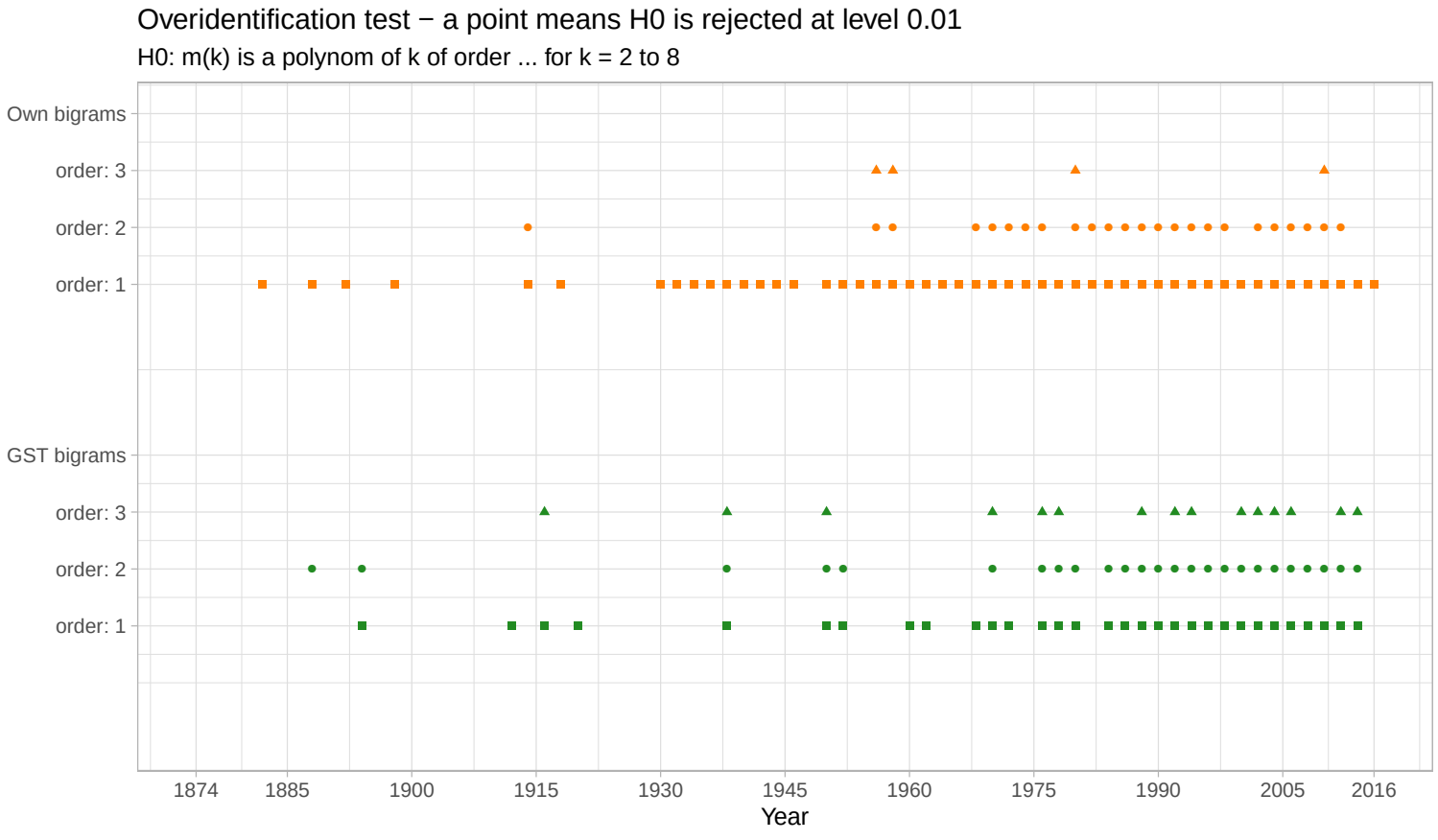
- The two parameters of the Beta distribution are set to control
  - $E(\rho)$ and therefore p - although potential correlation between λ and ρ also impacts p
  - the shape in order to monitor the magnitude of polarization - although potential correlations between λ and (ρ, ρ^2) also impact π



Over-identification tests for different polynomials

Back

Figure: Rejection of the hypothesis underlying the extrapolation of $m(1)$ at 1% level for different polynomial orders r and $\overline{k} = 8$



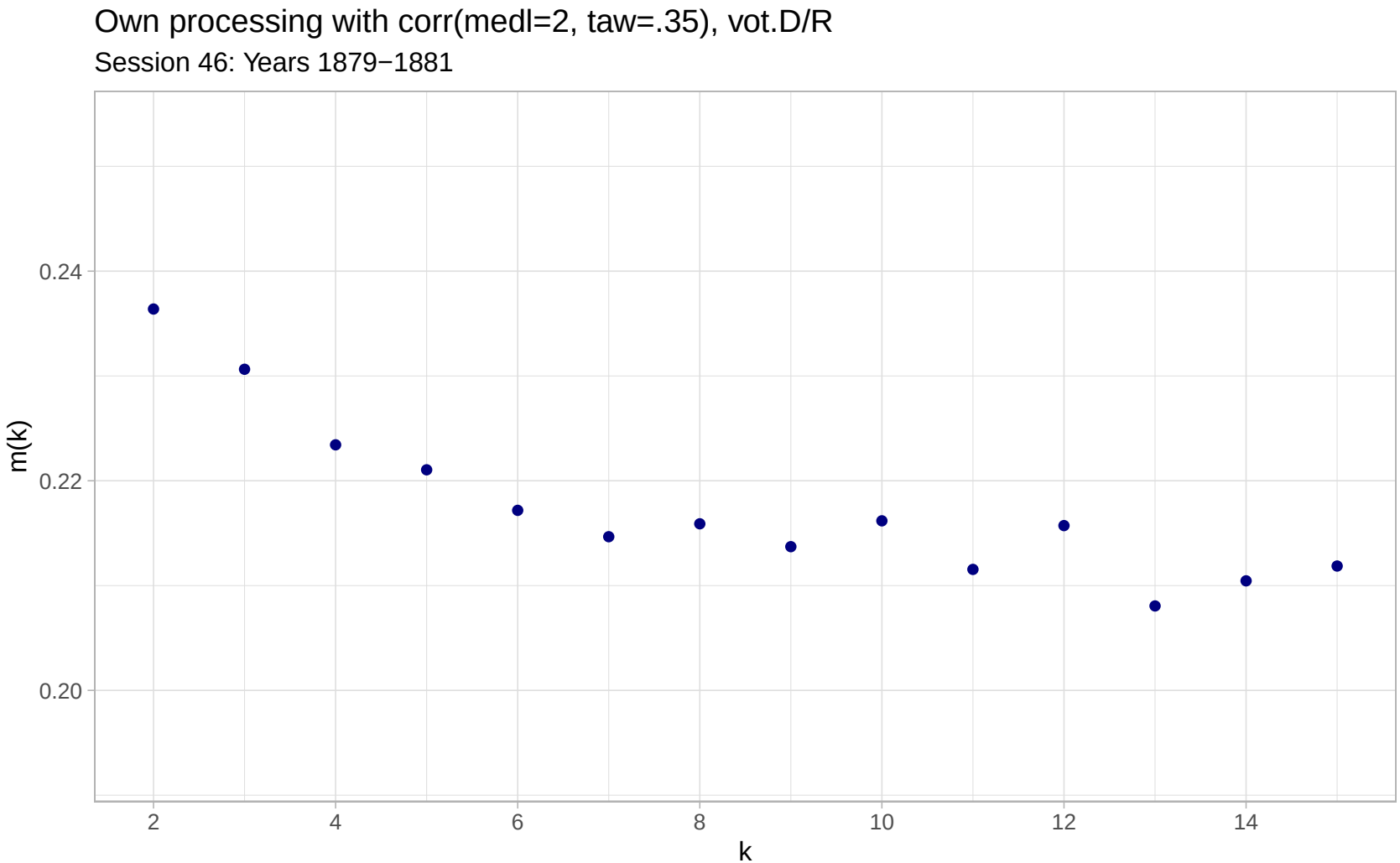
Test from CMD estimation to assess extrapolation's hypothesis. Bigrams restricted to vot. D and R.

Note: beware of multiple testing. Also, in a Bayesian perspective, interrogations about the relevance of such "ponctual" tests in a setting with millions of observations (see Abadie, "Statistical Nonsignificance in Empirical Economics", American Economic Review 2020).

Regularity of m(k) – session: 46

Back

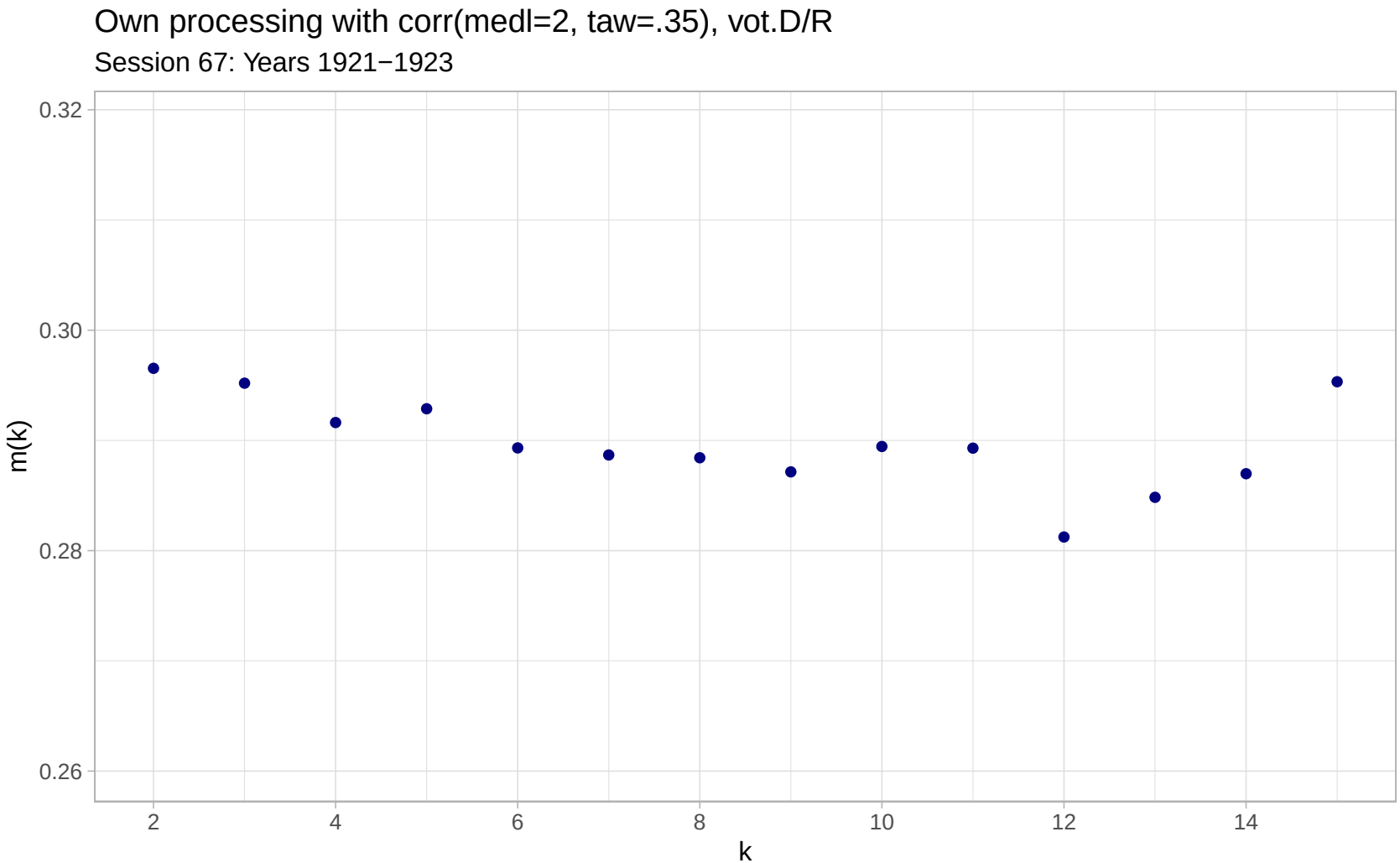
Figure: Processing with correction, suppression of "bad syntax" or "procedural" phrases, without restrictions based on the number of occurrences



Regularity of m(k) – session: 67

Back

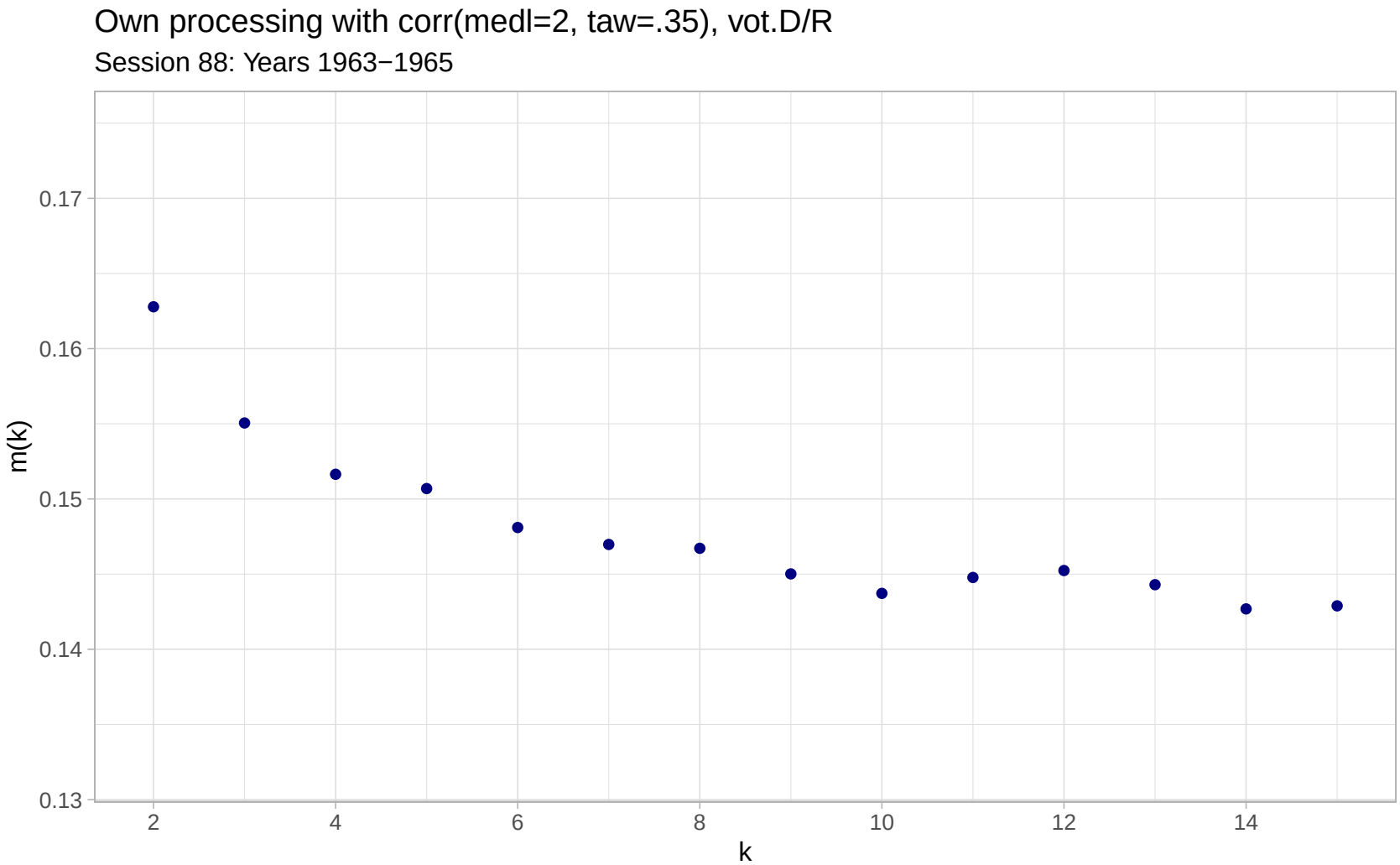
Figure: Processing with correction, suppression of "bad syntax" or "procedural" phrases, without restrictions based on the number of occurrences



Regularity of m(k) – session: 88

Back

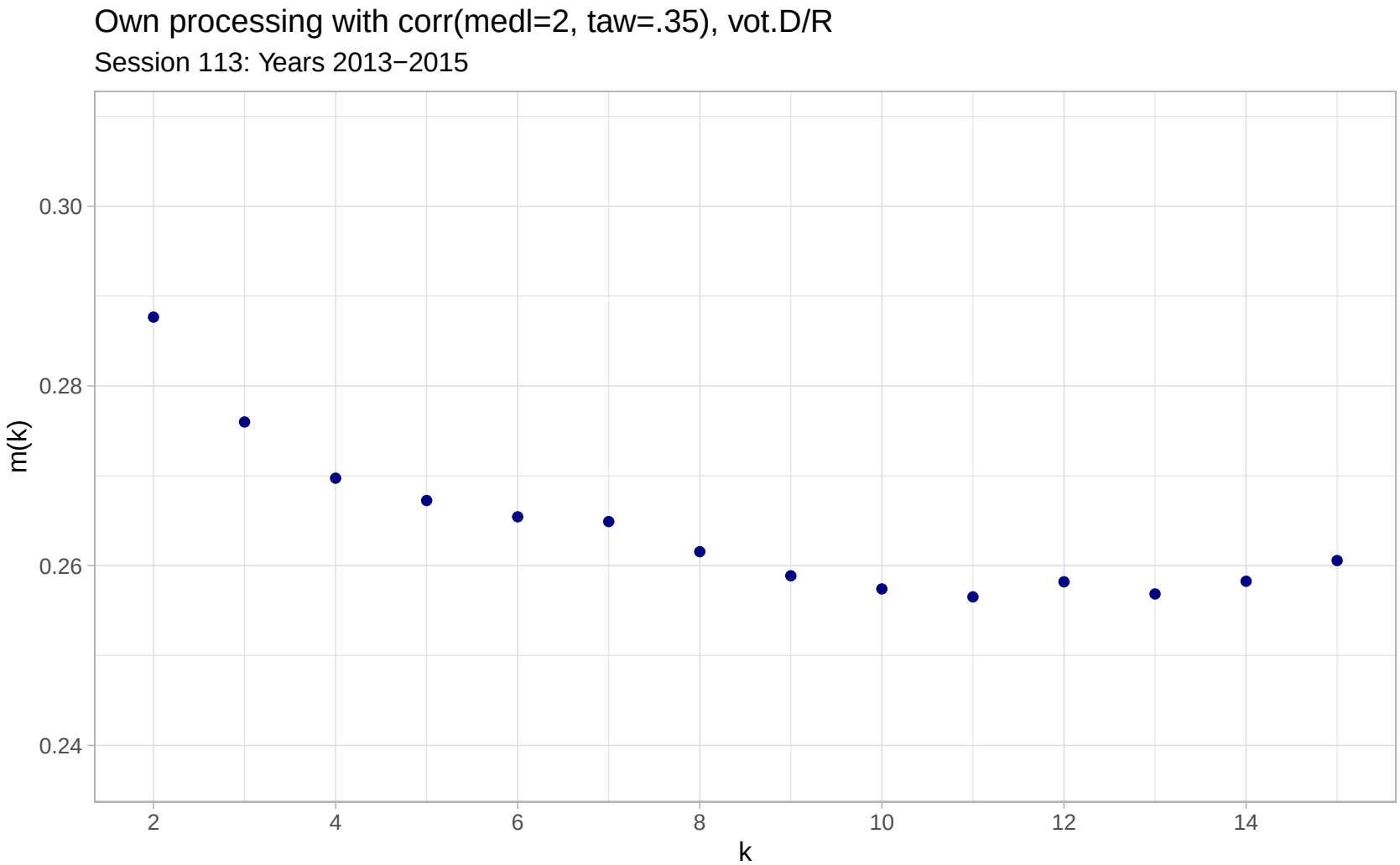
Figure: Processing with correction, suppression of "bad syntax" or "procedural" phrases, without restrictions based on the number of occurrences



Regularity of m(k) – session: 113

Back

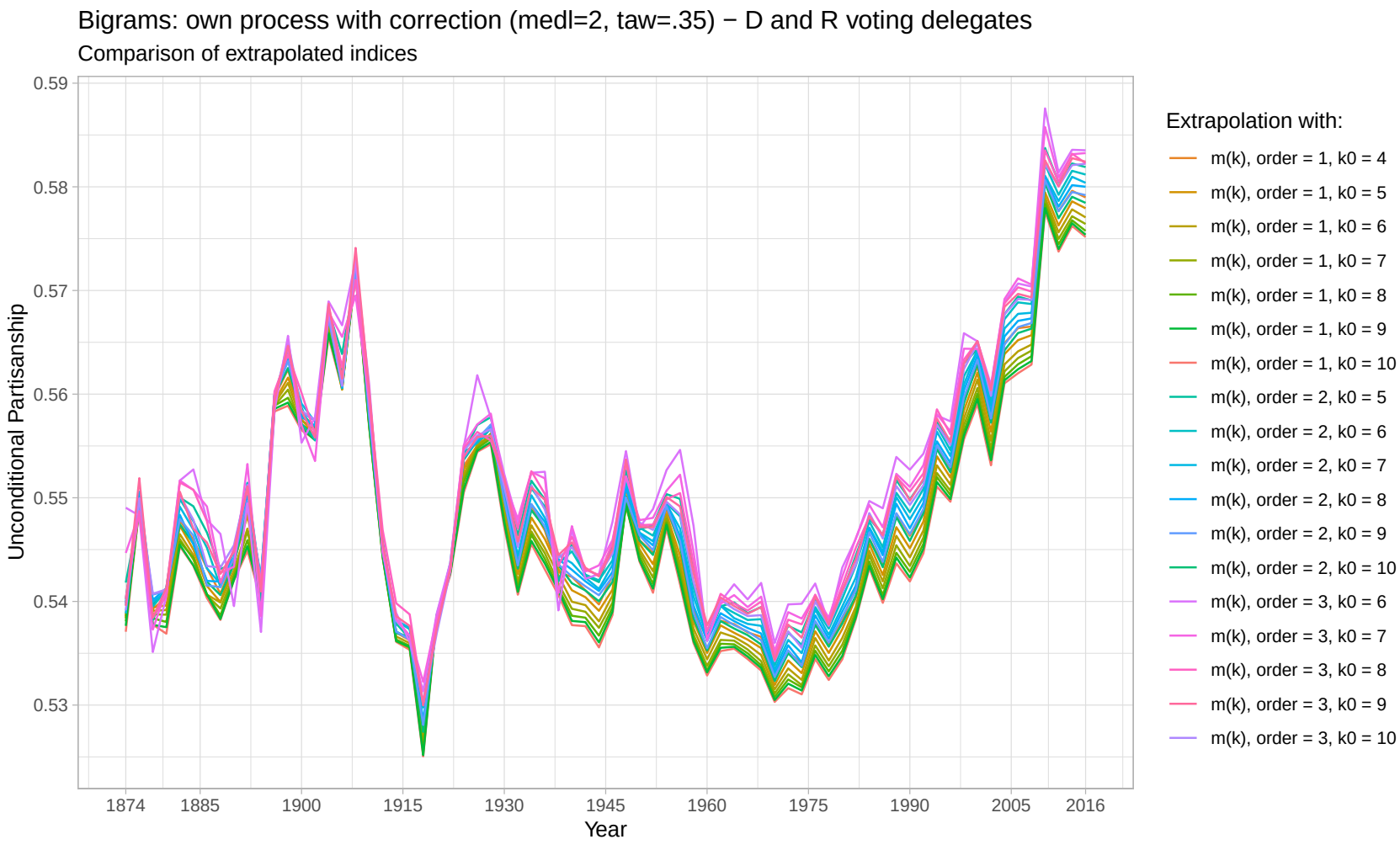
Figure: Processing with correction, suppression of "bad syntax" or "procedural" phrases, without restrictions based on the number of occurrences



Robustness of extrapolation

Back

Figure: Extrapolated estimators for different choices of $\overline{k}$ and r



Point-identification (1) – Assumption

Back

Assumption (DGP + "independence" between K and ρ)

Assumption (DGP) holds and in addition:

$$Cov(K, \rho) = Cov(K, \rho^2) = Cov(K^2, \rho^2) = 0.$$

- Interpretation: the popularity of a phrase is uncorrelated with its partisanship
- Under this additional assumption, π is point-identified and again we have a simple, computationally light estimator, consistent, and asymptotically normal

Point-identification (2) – Identification

Back

Theorem (Point identification with additional restriction)

Suppose that Assumption (DGP + "independence" between K and ρ) holds and the distribution of (K^R, K) is identified. Then, provided $E[K(K-1)] \neq 0$,

$$\pi = 1 - \frac{\mathbb{E}[\rho] - \mathbb{E}\left[\rho^2\right]}{2\mathbb{E}[\rho]\left(1 - \mathbb{E}[\rho]\right)},$$

and $\mathbb{E}[\rho]$, $\mathbb{E}\left[\rho^2\right]$ are identified, hence π is identified.

- The restriction $\mathbb{E}[K(K-1)] \neq 0$ says that there are words that are pronounced more than once ($K > 1$)

Point-identification (3) – Proof

Back

Remember:

$$\pi := 1 - \frac{\mathbb{E}[K\rho(1-\rho)]}{2\mathbb{E}[K]\, p(1-p)}$$

For the denominator:

$$p = \frac{\mathbb{E}[K^R]}{\mathbb{E}[K]} = \frac{\mathbb{E}[K\rho]}{\mathbb{E}[K]} \overset{\text{Cov}(K,\rho)=0}{=} \frac{\mathbb{E}[K]\mathbb{E}[\rho]}{\mathbb{E}[K]} = \mathbb{E}[\rho]$$

For the numerator, using the non-correlation between K, ρ and ρ^2:

$$\mathbb{E}[K\rho(1-\rho)] = \mathbb{E}[K]\mathbb{E}[\rho(1-\rho)] = \mathbb{E}[K]\left(\mathbb{E}[\rho] - \mathbb{E}[\rho^2]\right)$$

Hence, the remaining term to be proven identified is $\mathbb{E}[\rho^2]$

Point-identification (4) – Proof

Back

The binomial assumption implies:

$$\begin{aligned}\mathbb{E}\big[K^R(K^R-1)\mid K,\rho\big] &= \mathbb{E}_{B\sim\mathcal{B}(K,\rho)}\big[B^2-B\big]\\ &= \mathbb{V}_{B\sim\mathcal{B}(K,\rho)}(B)+\mathbb{E}_{B\sim\mathcal{B}(K,\rho)}(B^2)-\mathbb{E}_{B\sim\mathcal{B}(K,\rho)}(B)\\ &= K\rho(1-\rho)+(K\rho)^2-K\rho = K(K-1)\rho^2\end{aligned}$$

Then taking the expectation over the joint distribution of K and ρ in the previous equality of random variables yields:

$$\mathbb{E}\big\{\mathbb{E}\big[K^R(K^R-1)\,|\,K,\rho\big]\big\} = \mathbb{E}\big\{K(K-1)\rho^2\big\}\ .$$

The left-hand side is equal to $E[K^R(K^R-1)]$ through the law of iterated expectations. As for the right-hand side, restrictions between the correlation of orders 1 and 2 of K and ρ entail:

$$\begin{aligned}\mathbb{E}\big\{K(K-1)\rho^2\big\} &= \mathbb{E}\big[K^2\rho^2\big]-\mathbb{E}\big[K\rho^2\big]\\ &= \mathbb{E}\big[K^2\big]\,\mathbb{E}\big[\rho^2\big]-\mathbb{E}[K]\,\mathbb{E}\big[\rho^2\big]\\ &= \mathbb{E}[K(K-1)]\,\mathbb{E}\big[\rho^2\big]\end{aligned}$$

Point-identification (5) – Estimation

Back

- Again, the proof of identification yields a natural Method of Moment estimator for π under Assumption (DGP + "independence" between K and ρ)

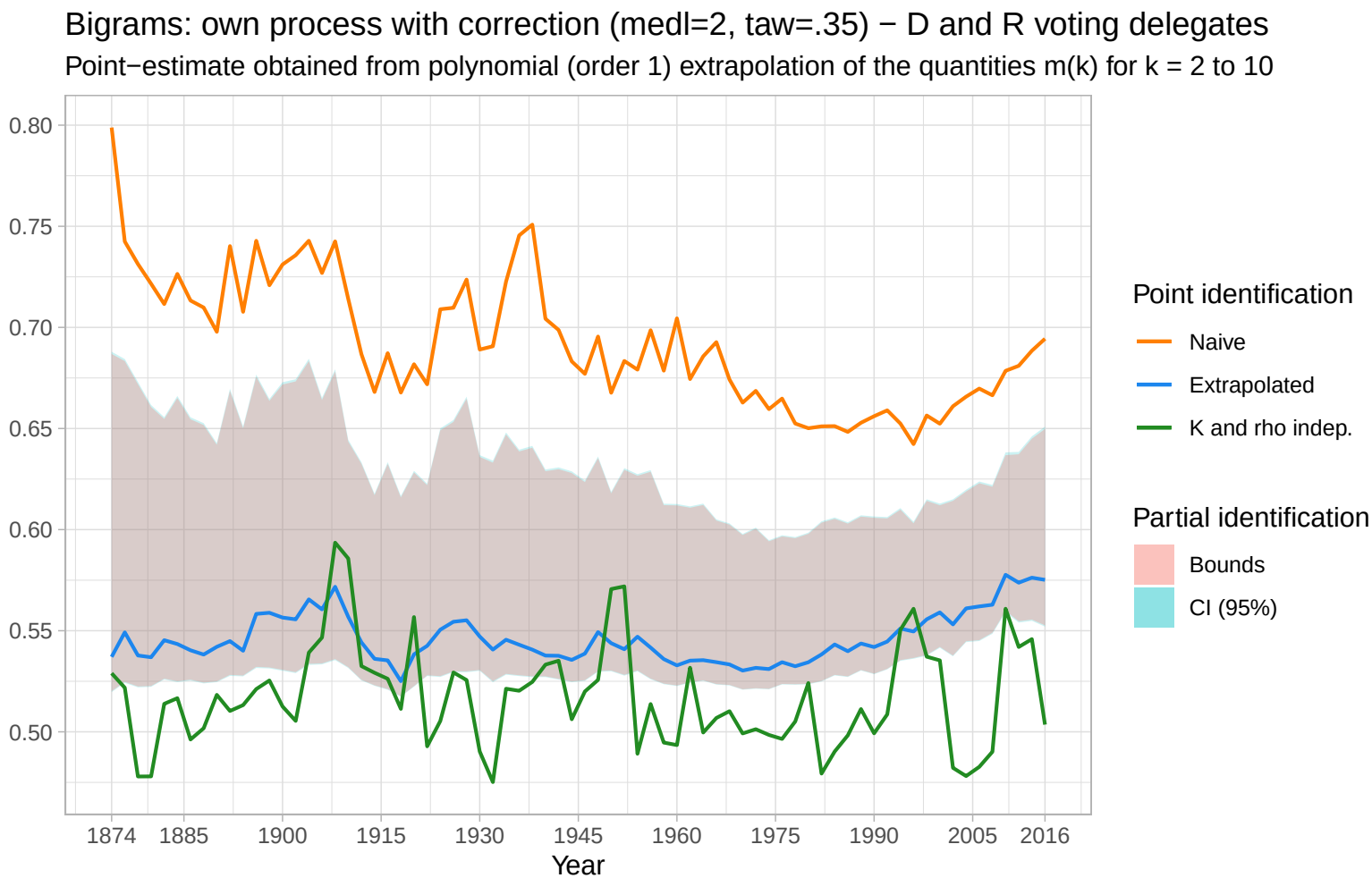
$$\widehat{\pi}_{\text{point}} := 1 - \frac{\frac{\sum_{j=1}^{J} K_j^R}{\sum_{j=1}^{J} K_j} - \frac{\sum_{j=1}^{J} K_j^R \left(K_j^R - 1\right)}{\sum_{j=1}^{J} K_j \left(K_j - 1\right)}}{2 \frac{\sum_{j=1}^{J} K_j^R}{\sum_{j=1}^{J} K_j} \frac{\sum_{j=1}^{J} K_j^D}{\sum_{j=1}^{J} K_j}}$$

- Under classic appropriate regularity conditions as regards existence of higher order moments, the estimator is consistent and asymptotically normal

Point-identification (6) – Result: the independence is not credible here

Back

Figure: Processing with correction, suppression of "bad syntax" or "procedural" phrases, without restrictions based on the number of occurrences



GST's penalized estimator with different sets of speaker covariates

Back

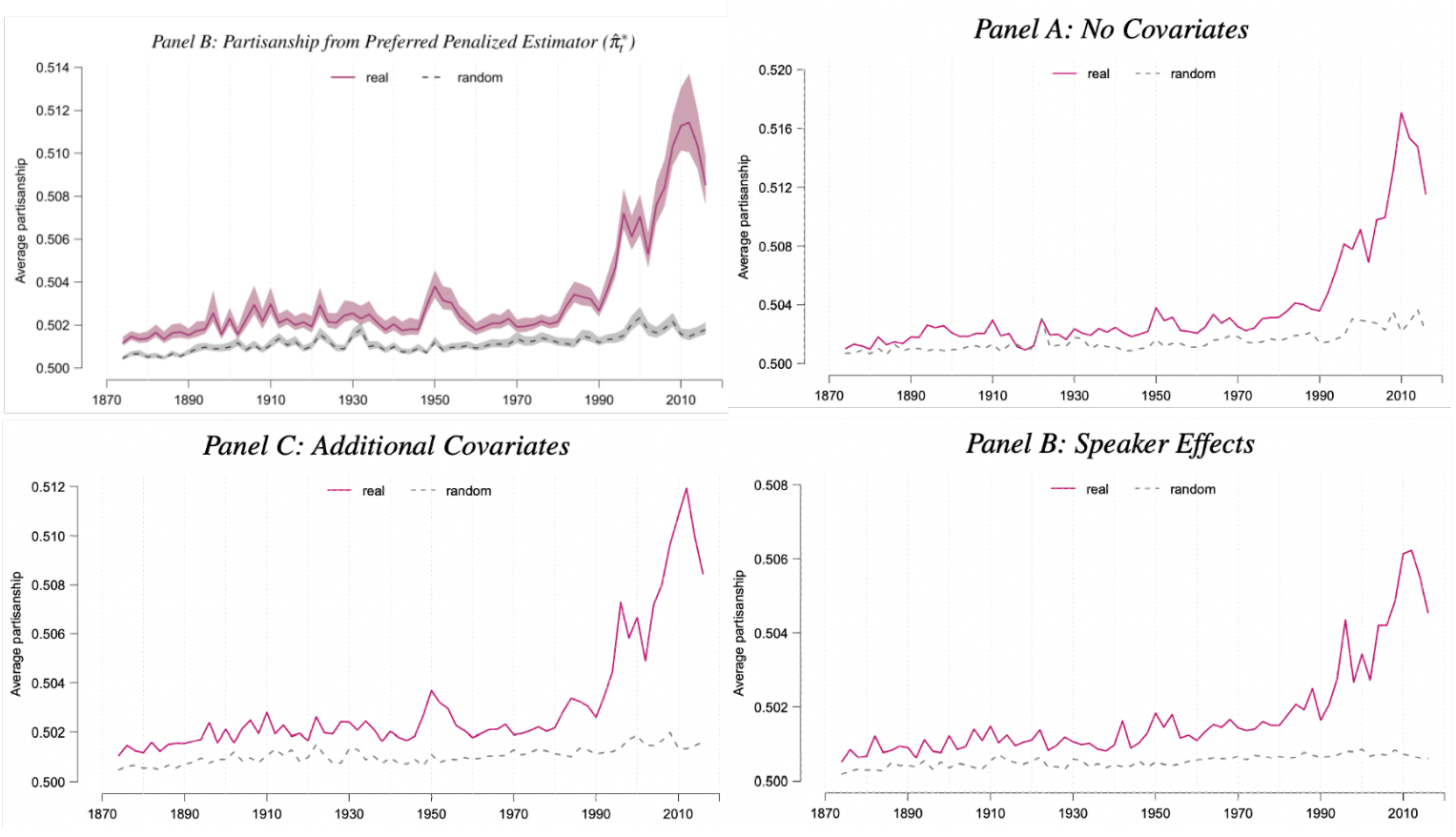


Figure: Gentzkow, Shapiro, and Taddy (ECTA 2019) – Panel B of Figure 2 (page 1321) and Online Appendix Figure 4 (page 24)

Data

Back

- Matthew Gentzkow, Jesse M. Shapiro, and Matt Taddy. Congressional Record for the 43rd-114th Congresses: Parsed Speeches and Phrase Counts. Palo Alto, CA: Stanford Libraries [distributor], 2018-01-16 `https://data.stanford.edu/congress_text`

- Congress speeches already parsed into stemmed bigrams
  - Restrict our analysis to Republican and Democrat
  - Following GST main specification, we use the "bound" source for sessions when it is available (43rd-111th = 1873-2011), then the "daily" record (112th-114th = 2012-2016)
  - Speaker-level information: gender, state, chamber

Vocabulary/dictionary choice

Back

- While these data are high-quality, there are still a fair amount of bigrams that are pronounced only once and never appear again: part of these words may be genuine rare words while some may be mistakes
- Some summary statistics about the distribution of K (mean over sessions 43-114 or over sessions 79-114)
  - *Raw data*:
    - number of distinct words present in data: 2.86 millions (3.65 millions)
    - average K by session: 3.67 (4.18)
    - proportion of words with only one occurrence by session: 0.664 (0.635)
  - *Raw data restricted to GST a priori dictionary*:
    - number of distinct words present in data: 1.13 millions (1.49 millions)
    - average K by session: 4.42 (5.31)
    - proportion of words with only one occurrence by session: 0.476 (0.432)

GST's dictionary

Back

- In their main specification, GST use different conditions for choosing the words (i.e. bigrams) included in the analysis:
  - **(1)** *Selection based on an a priori dictionary* (and usual pre-processing steps in text analysis)
    - delete hyphens, punctuation, etc.
    - drop some "stopwords": extremely common words + here phrases listed as procedural or with low semantic meaning
    - reduce words to their stems (Porter 2009 algorithm)
    - It yields a dictionary with $J = 4,525,242$ unique bigrams
  - **(2)** *Selection based on occurrences of words in data*: GST restrict attentions to bigrams that are sufficiently pronounced basically (cumulative criterion)
    - spoken at least 10 times in at least one session (local session)
    - spoken in at least 10 unique speakers-sessions (local session $\times$ speaker)
    - phrases spoken at least 100 times across all sessions ($= 1.39$ occurrence per session on average) (global across session)
- It yields a dictionary with $J = 508,352$ unique bigrams, spoken a total of 287 millions times (occurrences)

Getting closer to GST (1)

Back

- $\mathcal{J} = \{1 \dots J\}$ is the set of words and n is the number of occurrences
- Take the occurrence i as the individual observation/statistical unit (instead of the word j)
  - W_i, G_i are the word and the speaker's group for occurrence i
- GST model

Assumption (GST)

$(W_i, G_i)_{i=1\dots n}$ *are i.i.d.,* $P(G_i = R) = p$, *and*

$$W_1 | G_1 = g \sim \mathcal{M}(1, q_1^g, ..., q_J^g), \quad g \in \{D, R\}.$$

Getting closer to GST (2)

Back

- Probabilities $(q_j^g)_{j=1,\ldots,J}$ relate to the $(\rho_j)_{j=1,\ldots,J}$ (Bayes' rule)

$$\rho_j = \frac{pq_j^R}{pq_j^R + (1-p)q_j^D}$$

- The parameter considered by GST is

$$\pi_{GST} = \frac{1}{2}\sum_{j=1}^{J} q_j^R \rho_j + q_j^D(1-\rho_j)$$

- When p is allowed to differ from $1/2$:

$$\pi_J = 1 + \frac{1}{4p(1-p)}\left\{\sum_{j=1}^{J} pq_j^R\rho_j + (1-p)q_j^D(1-\rho_j) - 1\right\}$$

- How does this π_J relate to our π?

Getting closer to GST (3)

Back

- In line with the definition of K_j^g above, let

$$K_{j,n}^g = \sum_{i=1}^{n} 1\{G_i = g, W_i = j\}, \quad g \in \{R, D\}$$

- Under Assumption (GST), we have

$$K_{j,n}^R | K_{j,n} \sim B(K_{j,n}, \rho_j)$$

Hence, the conditional model on K_j^R is the same as ours.

Getting closer to GST (4)

Back

- But i.i.d. conditions in (DGP) and (GST) are different
  - Although difference becomes negligible when the number of words is large
- Asymptotics differ: GST assume $n \to \infty$ and J fixed

Assumption (Link-GST)

$\lim_{n\to\infty} J_n = +\infty$ *and there exist* $(\lambda_j^R)_{j\geq 1}$ *and* $(\lambda_j^D)_{j\geq 1}$ *i.i.d. such that:*

(i) $P(\lambda_1^g > 0) = 1$ *and* $E(\lambda_1^g) = 1$, *for* $g \in \{D, R\}$;

(ii) $q_{j,n}^g = \lambda_j^g / [\sum_{j=1}^{J_n} \lambda_j^g]$ *for all* $n \geq 1$ *and* $(j, g) \in \{1, ..., J_n\} \times \{D, R\}$.

- $\lim_{n\to\infty} J_n = +\infty$: consistent with Heaps's law

Getting closer to GST (5)

Back

- To make the link between π_J and π, let $\lambda_j = p\lambda_j^R + (1-p)\lambda_j^D$, $\rho_{j,\infty} = p\lambda_j^R/\lambda_j$ and let

$$\pi_\infty = 1 - \frac{\mathbb{E}[\lambda\rho_\infty(1-\rho_\infty)]}{2p(1-p)\mathbb{E}[\lambda]},$$

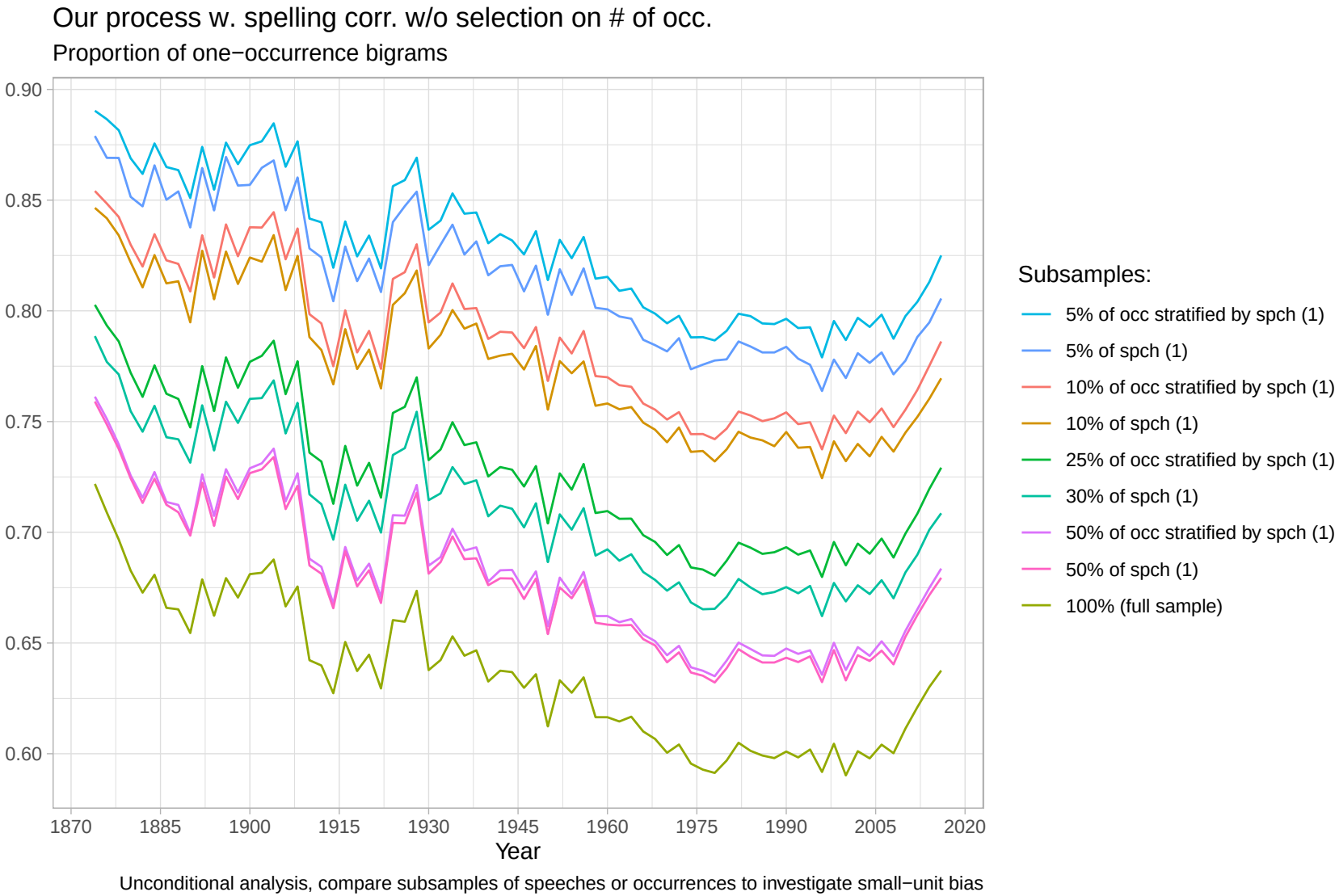
  where (λ, ρ_∞) has the same distribution as $(\lambda_1, \rho_{1,\infty})$

Theorem (Equivalence)

Suppose that Assumptions (GST) and (Link-GST) hold. Then as n tends to infinity $\pi_{J_n} \xrightarrow{p} \pi_\infty$.

Clue about positive correlation inter-occurrences intra-speech

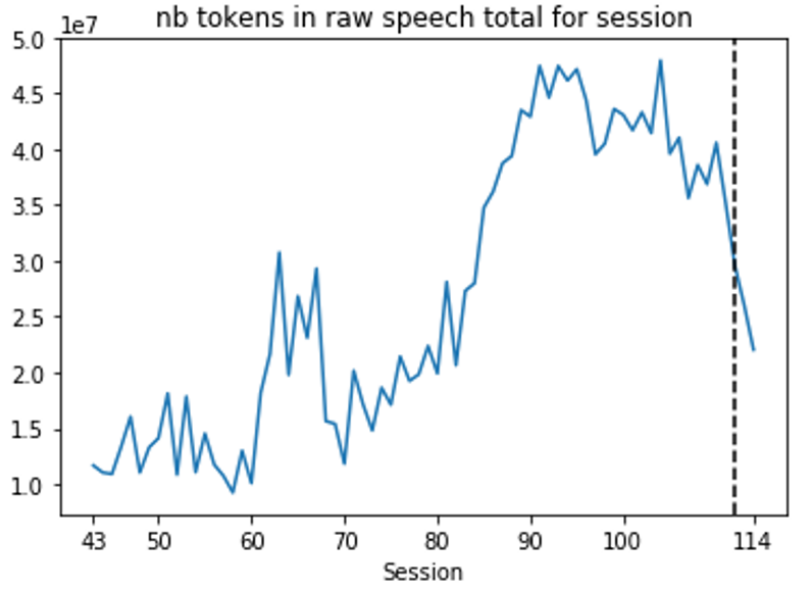
Back

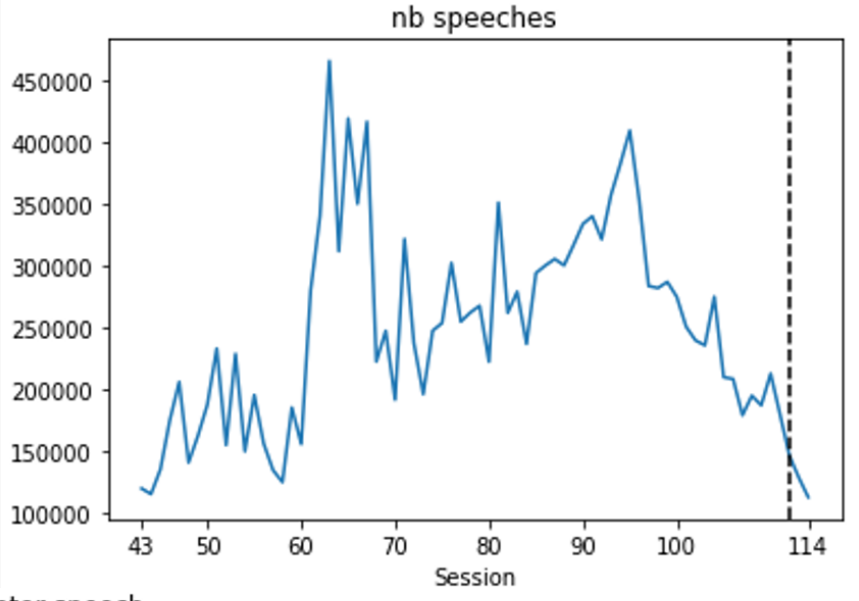


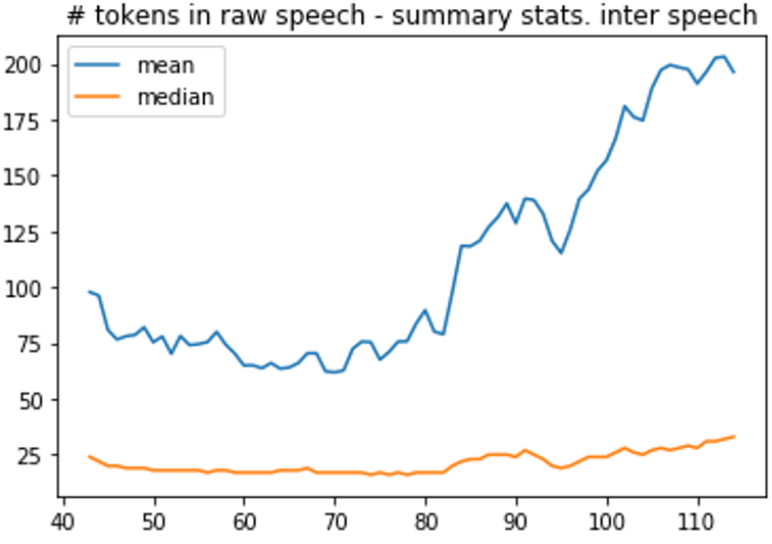
Unconditional analysis, compare subsamples of speeches or occurrences to investigate small−unit bias

If so, possibly higher correlation in recent years: possible impact on the results?

Back







43rd session: 1873-1875
50th session: 1887-1889
60th session: 1907-1909
70th session: 1927-1929
80th session: 1947-1949
90th session: 1967-1969
100th session: 1987-1989
110th session: 2007-2009
114th session: 2015-2017