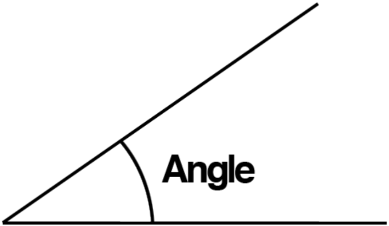
Building a Robot Judge: Data Science for the Law

5. Document Distance

Elliott Ash

Cosine Similarity: Idea



- each document is a non-negative vector in an *n*-space (size of the common dictionary) and it defines a *ray*
  - closer rays form smaller angles
  - the furthest rays are orthogonal
- $\cos(0) = 1$ and $\cos(\pi/2)=0$
- distance monotonically increases on $\{0,\pi/2\}$ ->
  cosine or similarity monotonically decreases on $\{1,0\}$

Cosine similarity: Formula

$$\text{cos\_sim}(v_1, v_2) = \frac{v_1 \cdot v_2}{||v_1||||v_2||}$$

where v_1 and v_2 are vectors, representing documents (e.g., tf-idf weighted word counts).

- $+1$ means identical documents; 0 means no words in common.
- Note that for n rows, this gives you $n \times (n-1)$ similarity scores.
- tf-idf similarities will down-weight terms that appear in many documents and usually give better results.

Other distance metrics

- Euclidean distance, $||v_1 - v_2||$
- Jensen-Shannon Divergence
- etc.
- hopefully empirical results are not sensitive to choice of metric.

Clustering

- k-means clustering separates documents into k groups:
  - Given document vectors $\{\vec{q}_1, \vec{q}_2, ..., \vec{q}_P\}$, the algorithm chooses clusters $Q = \{Q_{1,} Q_2, ... Q_k\}$, $k > 1$, to minimize the within-cluster sum of squares:

    $$\arg\min_Q \sum_{i=1}^{k} \sum_{\vec{q} \in Q_i} ||\vec{q} - \mu_i||^2$$

    where μ_i is centroid (mean vector) for cluster Q_i.
- Can also compute "k-medoid" clusters using the 1-norm:

  $$\arg\min_Q \sum_{i=1}^{k} \sum_{\vec{q} \in Q_i} ||\vec{q} - \mu_i||$$

  and μ_i would give the medoid (the median vector) for the cluster.
- Agglomerative (hierarchical) clustering makes nested clusters.
- DBSCAN doesnt require specification of number of clusters.

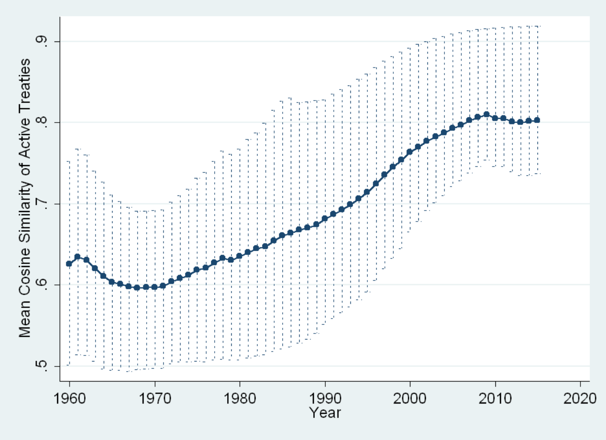
Clusters vs. Topics

- Each cluster is a set of documents that are close to each other in the vector space (normally, they will be topically related)
- The advantage of clusters, rather than topics or embeddings, is that they provide discrete groups.
  - This might be useful depending on your research task.

Tax Treaties have converged in language

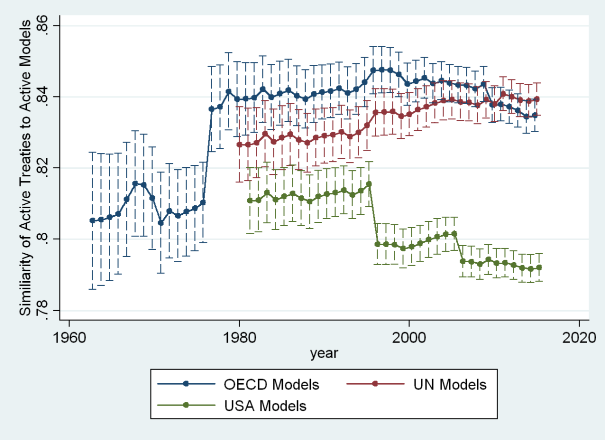
Ash and Marian (2018)

,



Average cosine similarity between active treaties by year. Error spikes give 25th and 75th percentiles.

Influence of Model Treaties over Time



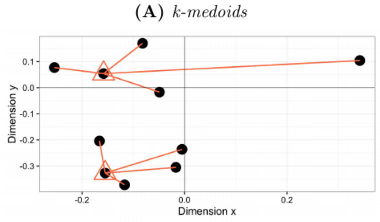
- OECD and UN are most influential on tax treaties.

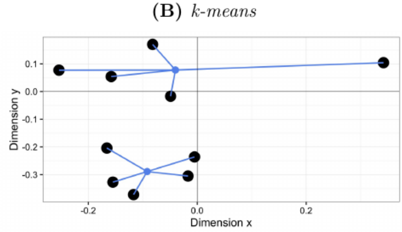
Customization of Debt Contracts

Ganglmair and Wardlaw, "Complexity, Standardization, and the Design of Loan Agreements"

- Substantive question:
  - what explains customization and complexity in debt contracts?
- Methodological question:
  - Can we use contract text to analyze customization and complexity?
    - previous work relies on expensive hand-coding

Measuring customization





- Measuring **customization** of contracts:
  - distance to the k-medoid for all debt contracts drafted within a two-year window.

Descriptive findings

- Contracts are not boilerplate – there are important differences between contracts.
    - Text differences are driven by borrowers, rather than lenders
- More standardization:
    - larger deals, less renegotiation

Abrahamson and Barber

The Evolution of National Constitutions (QJPS 2019)

- Corpus: Comparative Constitutions Project:
  - A repository of current and historical constitutions across countries and provinces.
  - 1297 constitutions, 185 countries, 1789-2010
- Annotations (1329 features):
  - e.g. structure of executive, amendment process, election process, legislative composition

Colonial Path Dependence

Table 4: Between estimates of colonial history and constitutional similarity.

Distance from:	(1) UK	(2) France	(3) Spain
Former British colony	**−0.48**	−0.36	0.41
	(0.12)	(0.07)	(0.10)
Former French colony	−0.14	**−0.40**	0.02
	(0.11)	(0.07)	(0.10)
Former Spanish colony	0.31	0.31	**−0.33**
	(0.13)	(0.09)	(0.10)
Other colonies	−0.03	−0.17	0.08
	(0.17)	(0.11)	(0.14)
N	190	190	190

In each model the dependent variable is the average absolute distance of each country's constitution from the country listed at the top of the column. For example, Model 1 shows the average distance from the UK constitution. Negative coefficients indicate more similarities. The omitted category in each model is countries that were never colonized. Robust standard errors shown below OLS coefficients.

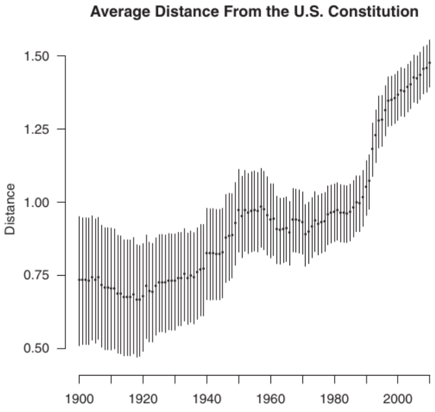


Figure 5: Similarity of constitutional systems to the United States over time.

Text analysis of patent innovation

"Measuring technological innovation over the very long run," Kelly, Papanikolau, Seru, and Taddy (2018)

- Data:
  - 9 million patents since 1840, from U.S. Patent Office and Google Scholar Patents.
  - date, inventor, backward citations
  - text (abstract, claims, and description)
- Text pre-processing:
  - drop HTML markup, punctuation, numbers, capitalization, and stopwords.
  - remove terms that appear in less than 20 patents.
  - 1.6 million words in vocabulary.

Measuring Innovation

- Backward IDF weighting of word w in patent p:

$$\text{BIDF}(w,p) = \frac{\#\text{ of patents prior to } p}{\log\left(1 + \#\text{ documents prior to } p \text{ that include } w\right)}$$

  - down-weights words that appeared frequently before a patent, but up-weights new words.
- For each patent:
  - compute cosine similarity to all future patents, using BIDF of earlier patent.
- 9m$\times$9m similarity matrix = 30TB of data.
  - enforce sparsity by setting similarity $< .05$ to zero (93.4% of pairs).

Novelty, Impact, and Quality

- "Novelty" is defined by (negative) similarity to previous patents:

$$\text{Novelty}_j = -\sum_{i \in B(j)} \rho_{ij}$$

where $B(j)$ is the set of previous patents (in, e.g., last 20 years).

- "Impact" is defined as similarity to subsequent patents:

$$\text{Impact}_i = \sum_{i \in F(j)} \rho_{ij}$$

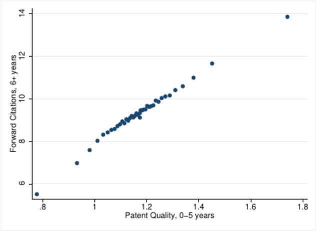
where $F(j)$ is the set of future patents (in, e.g., next 100 years).

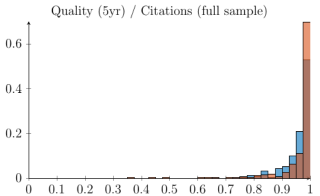
- A patent has high quality if it is novel and impactful:

$$\text{Quality}_i = \frac{\text{Impact}_i}{-\text{Novelty}_i}$$

Validation

- For pairs with higher $\rho_{i,j}$, patent j is more likely to cite patent i.
- Patent office assigns 3-digit technology class code; similarity is significantly higher within class compared to across class.
- Higher quality patents get more cites:

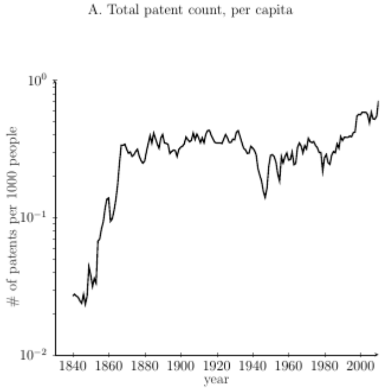


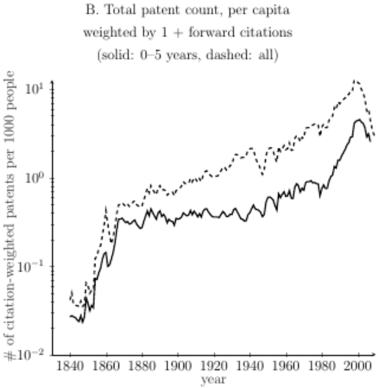


Most Innovative Firms

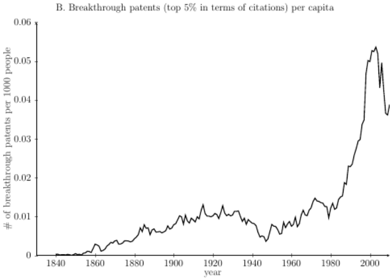
Assignee	First Year	# Breakthroughs
General Electric	1872	3,457
Westinghouse Electric Co.	1889	1,762
Eastman Kodak Co.	1890	2,244
Western Electric Co.	1899	1,222
AT&T (includes Bell Labs)	1899	5,645
Standard Oil Co.	1900	1,212
Dow Chemical Co.	1902	1,235
Du Pont	1905	3,353
International Business Machines	1908	14,913
American Cyanamid Co.	1909	690
Universal Oil Products Co.	1919	590
RCA	1920	3,222
Monsanto Company (inc. Monsanto Chemicals)	1921	902
Honeywell International, inc.	1928	872
General Aniline & Film Corp.	1929	1,181
Massachusetts Institute of Technology	1935	504
Philips	1939	1145
Texas Instruments	1960	2,088
Xerox	1961	2,198
Applied Materials	1971	510
Digital Equipment	1971	1,101
Hewlett-Packard Co.	1971	2,661
Intel	1971	2,629
Motorola, inc.	1971	4,129
Regents of the University of California	1971	823
United States Navy	1945	791
NCR	1973	737
Advanced Micro Devices	1974	1,195
Apple Computer	1978	864

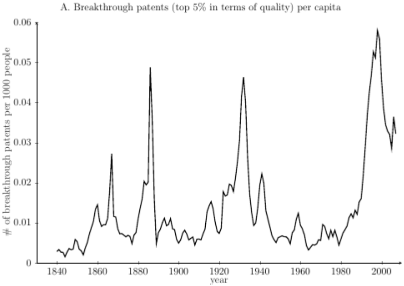
Patents per capita



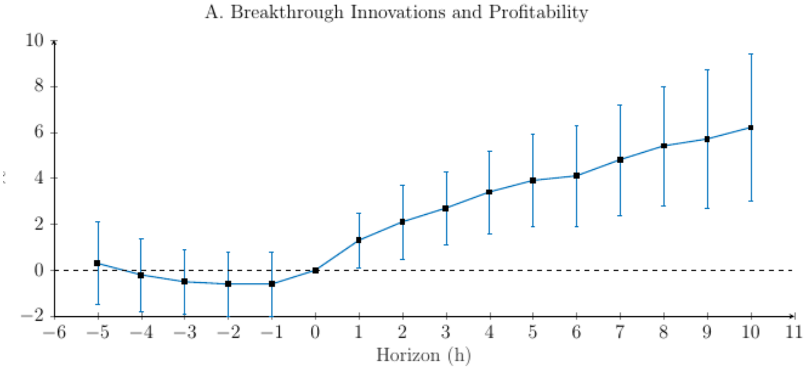


Breakthrough patents per capita





Breakthrough patents and firm profits



Text analysis of corporate filings

“Text-Based Network Industries and Endogenous Product Differentiation” (2016)

- Data
  - 10-K annual filings from EDGAR, 1996-2008
  - Extract **“business description”** section, where firms are **legally required** to “describe the significant products they offer to the market” for the current fiscal year.
- Text features:
  - nouns (including proper nouns), except location names (state, county, city)
  - drop words appearing in more than 25% of documents.
  - binary for whether word appears (rather than counts)
- Similarity:
  - cosine similarity between these vectors of binaries

Text-Based Industries

- The paper constructs "industries" as sets of firms with similar lists of nouns in their business descriptions.
  - they use an unusual clustering algorithm that probably ends up being close to k-means.

- Qualitative validation: Example sub-markets for "Business Services" (SIC code 737):
  1. entertainment (42), video (42), television (38), royalties (35), internet (34), content (33), creative (31), promotional (31), copyright (31), game (30), sound (29), publishing (29)
  2. client (59), database (54), solution (49), patient (47), copyright (47), secret (47), physician (47), hospital (46), health care (46), server (45), resource (44), func- tionality (44), billing (44)
  3. internet (236), telecommunications (211), interface (194), communication (188), solution (187), platform (184), architecture (182), call (177), infrastructure (173), voice (173), functionality (173), server (173)

Text industries explain outcomes better than standard codings

TABLE 3
FIRM CHARACTERISTICS AND INDUSTRY CLASSIFICATIONS

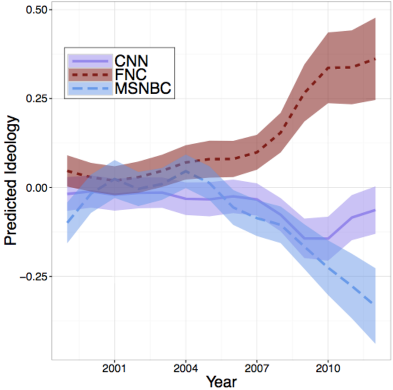
Industry Controls	OI/Sales	OI/ Assets	Sales Growth	Market Beta	Asset Beta
	A. Across-Industry Standard Deviations: Firm-Weighted Results; All Industry Classifications				
1. SIC-3 fixed effects	.204	.111	.126	.283	.271
2. NAICS-4 fixed effects	.205	.112	.136	.289	.276
3. 10-K-based 300 fixed effects	.231	.128	.157	.298	.285
4. TNIC equal-weighted average	.248	.142	.163	.332	.324
5. TNIC similarity-weighted average (excluding the focal firm)	.267	.153	.199	.384	.369
	B. Across-Industry Standard Deviations: Industry-Weighted Results; Transitive Industry Classifications Only				
1. SIC-3 fixed effects	.156	.111	.179	.347	.308
2. NAICS-4 fixed effects	.169	.126	.210	.414	.362
3. 10-K-based 300 fixed effects	.202	.139	.224	.469	.432

NOTE.—For a given variable indicated in the left-hand column, across-industry standard deviations are computed as the standard deviation of the industry average of the given variable across all firms in our sample (panel A) and across all industries (panel B). TNIC refers to text-based network industries.

Cable News and Political Discourse

- Context:
  - U.S. congressional districts ($N = 435$), years 2005-2008
- Data:
  - transcripts for prime time shows in major cable channels
  - transcripts for congressional speeches in U.S House
  - geographical data on U.S House representatives
  - channel position and viewership for cable news channels

Fox News Channel is Politically Conservative



Martin and Yurukoglu (2017): Estimated ideology based on phrase usage for CNN, Fox News Channel (FNC), and MSNBC. Higher is more conservative.

Text Features

- Featurization:
  - Convert to lower case, remove punctuation, stopwords, numbers
  - Do stemming
  - Construct 3-grams for the observations
  - Remove rare 3-grams
- Create frequency matrices for congressional speakers by year, and for cable news transcripts for each channel by year.

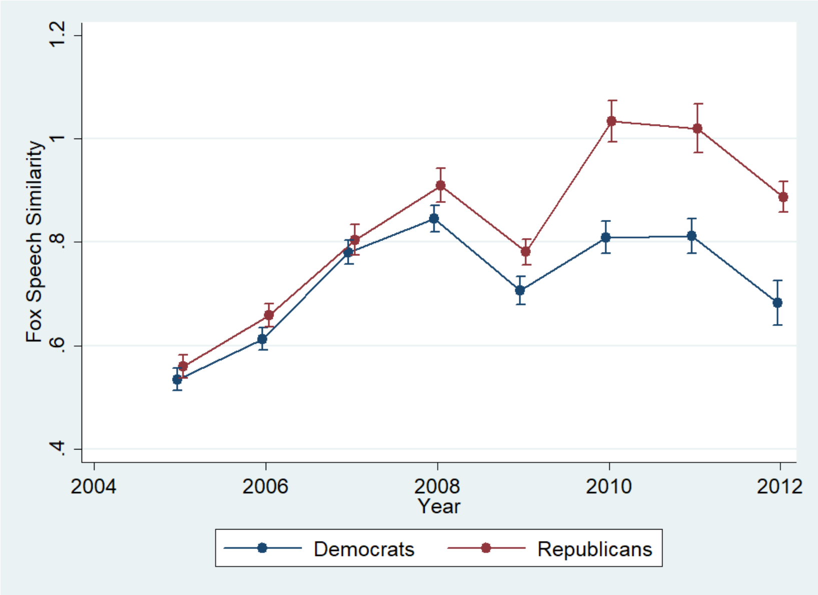
Compute similarity of each speech to cable channels

What we have	What we want
frequency matrices	similarity columns
$M_{congress},\ M_{Fox}$	$S_{congressFox}$
$M_{congress},\ M_{CNN}$	$S_{congressCNN}$
$M_{congress},\ M_{MSNBC}$	$S_{congressMSNBC}$

- cosine similarity captures linguistic similarity between TV shows and congress speeches
- need to normalize to get the similarity specific for Fox News Channel:

$$foxsim = \frac{2similarity(fox, congress)}{similarity(cnn, congress) + similarity(msnbc, congress)}$$

Similarity toward Fox News



Regression Model

$$Y_{i,t} = \alpha + \rho V_{i,t} + X_{i,t}\beta + \epsilon_{i,t}$$

- congressional district i, year t
- $Y_{i,t}$, a speeches' similarity to Fox News variable
- $V_{i,t}$, measure of Fox News viewership
- $X_{i,t}$, covariates
  - state-time fixed effects
  - demographic covariates
- $\varepsilon_{i,t}$, unobservable factors and randomness
- ρ, effect of Fox News on House speeches similarity to Fox

Instrumental Variables: Main Intuition

- OLS Regression based on observables:
  - The consistency of the estimate relies on the "hope" that any unobserved factor that might affect the outcome variable is balanced across the treatment and the control group.
  - Therefore, any difference in outcomes between the control and the treatment group can be attributed to the treatment.
- Instrumental variables:
  - We identify some source of variation in the assignment to the treatment which, for some reason, we know that it is orthogonal to any relevant unobserved variable which might be affecting the outcome variables.
  - We compare group of individuals that, due to the instrument, are assigned to the control and the treatment group. Any difference in outcomes between these two groups is attributed to the treatment.

Cable television channel positions

- In 2000s, majority of American households had paid cable television.
- lineup of channels varies across local cable systems.
- channel positions set in mid to late 1990s, haphazardly, based on order of joining systems, and what channels were being replaced.
    - once channels are set, providers rarely change them.
- Martin and Yurukoglu (2017) show that when Fox News has a lower channel number, that increases viewership.
    - **use channel position as instrumental variable**.

What is a valid instrumental variable?

Instrumental variable (IV) is a variable that:

1. Is correlated with causal variable of interest, V_i:

$$\text{Cov}[Z_i, V_i] \neq 0$$

2. Is uncorrelated with any other determinants of Y_i:

$$\text{Cov}[Z_i, \epsilon_i] = 0$$

- The second requirement can be decomposed in two:
  - 2.1: Exogeneity: None of the unobserved factors affects the instrument:

$$\epsilon_i \nrightarrow Z_i$$

  - 2.2 Exclusion restriction: Instrument only affects outcome through treatment variable:

$$Z_i \nrightarrow \epsilon_i$$

- With a valid instrumental variable we can consistently estimate ρ in
$$Y_i = \alpha + \rho V_i + \epsilon_i$$
- Write the covariance of Z_i and Y_i as:
$$\text{Cov}[Z_i, Y_i] = \rho \text{Cov}[Z_i, V_i] + \text{Cov}[Z_i, \epsilon_i]$$
- The **exogeneity/exclusion assumption** is $\text{Cov}[Z_i, \epsilon_i] = 0$.
- Thus:
$$\rho = \frac{\text{Cov}[Z_i, Y_i]}{\text{Cov}[Z_i, V_i]}$$
is a consistent population estimate.

Weak Instruments

- The bias of 2SLS can be written as:

$$\text{plim}\hat{\rho} = \rho + \frac{\text{Corr}[Z,\epsilon]}{\text{Cov}[V,Z]} \cdot \frac{\sigma_\epsilon}{\sigma_V}$$

  - When the instrument is weakly correlated with the endogenous regressor, the bias increases.

- Can check for a weak instrument with first-stage F-statistic: it should be higher than 10.

Matrix Notation, and Comparison to OLS

With model $Y = X'\beta + U$ and instrument Z, we have

$$\beta_{OLS} = (X'X)^{-1}(X'Y)$$
$$\beta_{IV} = (Z'X)^{-1}(Z'Y)$$

$$\begin{aligned}\mathbb{E}[\beta_{OLS}] &= \mathbb{E}[(X'X)^{-1}(X'Y)] = \mathbb{E}[(X'X)^{-1}(X'(X'\beta + U))] \\ &= \beta + \mathbb{E}[(X'X)^{-1}(X'U)]\end{aligned}$$

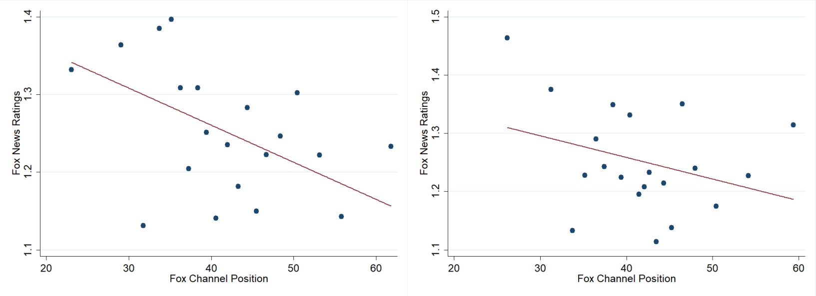
$$\begin{aligned}\mathbb{E}[IV] &= \mathbb{E}[(Z'X)^{-1}(Z'Y)] = \mathbb{E}[(Z'X)^{-1}(Z'(X'\beta + U))] \\ &= \beta + \mathbb{E}[(Z'X)^{-1}(Z'U)]\end{aligned}$$

$$\mathbb{E}[(X'X)^{-1}(X'U)] \gtrless \mathbb{E}[(Z'X)^{-1}(Z'U)]?$$

Instrumental Variables Approach

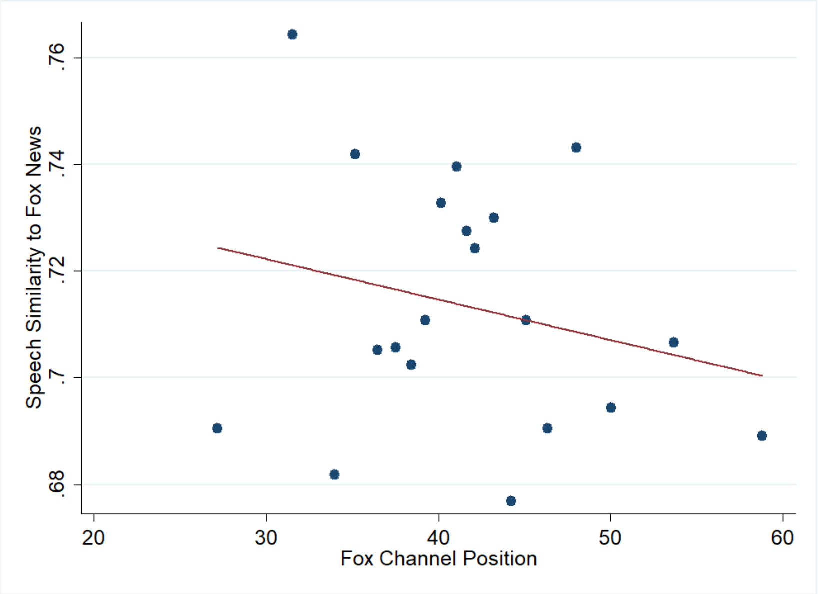
- Adopt IV method from Martin and Yurukoglu (AER 2017), with first stage
$$V_{i,t} = \alpha + \gamma Z_{i,t} + \eta_{i,t}$$
- Z_i, Fox News channel number in district i
  - constructed as the population-weighted average channel positions for each zip code in district i.
- Second stage is
$$Y_{i,t} = \alpha + \rho \hat{V}_{i,t} + X_{i,t}\beta + \epsilon_{i,t}$$
estimated with two-stage least squares (2SLS).

Low Fox Channel Number → High Fox Viewership



Average Fox News viewership share plotted against Fox News channel position (left panel, without state-year controls; right panel, with controls).

Reduced Form: Channel Position and Speech Similarity



2SLS Effect of Fox Exposure on similarity to Fox

	2SLS	
	(1)	(2)
Viewship % FNC	0.247*	0.308*
	(0.147)	(0.176)
N observations	1321	1321
State-time FE	YES	YES
Demographics	NO	YES

2SLS estimates of effect of FNC ratings on congress speech similarity to FNC; standard errors in parenthesis clustered by district; * p<.1.