Lecture 9: Texual Analysis Research Applications

Textual analysis

- Textual analysis becoming increasingly popular in asset pricing, macro, and other fields.
- Most successful applications use text to measure economic concepts that are otherwise hard or impossible to measure.
- So-far, simplest applications have been the most successful.
- Many cutting-edge methods of machine learning are not necessary or even counter-productive, similar to kitchen-sink regressions, prone to over-fitting.
- Advice:
    - keep it simple.
    - stay close to the text, read a lot.
    - frontier is more in learning from new data than in fancy techniques.

Conference Call Transcripts

- Transcripts of 326,247 earnings conference calls of 11,943 firms headquartered in 84 different countries, available 2002-20 from EIKON.
- Typically four calls per year, after earnings releases.
- Management presentation followed by Q&A with firm's analysts (0-70 questions, average duration 45 min).

Hassan, Hollander, van Lent, and Tahoun (2019): Firm-level political risk

- What share of the conversation between management and participants centers on political risks?
- Extract all two-word combinations (“bigrams”) from training libraries indicative of discussion of political topics, P, and non-political topics

Measure of Political Risk

- Count the number of occurrences of (exclusively) political bigrams in conjunction with a synonym for risk or uncertainty and divide by the total number of bigrams in the transcript:

$$PRisk_{it} = \frac{1}{B_{it}} \sum_{b}^{B_{it}} \{1[b \in \mathbb{P} \backslash \mathbb{N}] \times 1[|b - r| < 10] \times f_{b,\mathbb{P}}/B_{\mathbb{P}}\},$$

where r is the position of the nearest synonym of risk or uncertainty and $b = 0, 1, ...B_{it}$ are the bigrams contained in call of firm i at time t. (Application of "$tf \times idf$.")

Synonyms for “risk” or “uncertainty”

Synonym	Frequency	Synonym	Frequency	Synonym	Frequency	Synonym	Frequency
risk	413,925	sticky	4,325	unforeseeable	466	equivocation	55
risks	106,858	dangerous	4,297	halting	453	indecisive	43
uncertainty	91,775	tentative	4,018	wager	446	chancy	40
variable	68,138	hazardous	3,155	torn	437	menace	38
chance	60,863	queries	2,676	precarious	362	qualm	35
possibility	57,599	danger	2,465	undetermined	349	vacillating	33
pending	53,318	fluctuating	2,462	insecurity	348	gnarly	32
uncertainties	51,092	unstable	2,440	debatable	346	disquiet	30
uncertain	39,191	vague	2,427	undecided	341	ambivalence	30
doubt	39,022	erratic	1,876	dicey	330	imperil	28
prospect	30,926	query	1,826	indecision	324	vacillation	22
bet	21,279	jeopardize	1,821	wavering	266	incalculable	17
variability	21,215	unsettled	1,664	iffy	235	untrustworthy	17
exposed	19,553	unpredictability	1,563	faltering	212	equivocating	15
likelihood	19,280	dilemma	1,547	endanger	205	diffident	15
threat	19,021	skepticism	1,502	quandary	204	fickleness	11
probability	15,791	hesitancy	1,491	insecure	189	misgiving	11
unknown	12,050	riskier	1,352	changeable	189	changeability	11
varying	9,442	unresolved	1,214	riskiest	183	undependable	9
unclear	9,036	unsure	1,151	hairy	177	incertitude	8
unpredictable	8,467	irregular	1,123	ambivalent	169	fitful	8
speculative	8,132	jeopardy	1,077	dubious	158	parlous	8
fear	7,939	suspicion	1,027	riskiness	135	unconfident	6
reservation	7,026	risking	863	treacherous	130	defenseless	5
hesitant	6,275	peril	660	oscillating	112	unsureness	3
gamble	6,065	hesitating	628	perilous	92	fluctuant	3
risky	5,227	risked	577	tentativeness	85	niggle	3
instability	4,762	unreliable	550	unreliability	72	diffidence	3
doubtful	4,736	unsafe	486	wariness	70	precariousness	1
hazard	4,626	hazy	472	vagueness	59	doubtfulness	1
tricky	4,359	apprehension	466	dodgy	58		

Single-word synonyms of ‘risk’, ‘risky’, ‘uncertain’, and ‘uncertainty’ from Oxford Dictionary, excluding ‘question’, ‘questions’, and ‘venture’.

Measuring news about the mean: $PSentimentit_{it}$

- Use same approach to measure mean of political news:
  - Count positive and negative words ("sentiment") used in conjunction with a political bigram:

$$PSentiment_{i,t} = \frac{1}{B_{it}} \sum_{b}^{B_{it}} \left(1[b \in \mathbb{P} \backslash \mathbb{N}] \times \frac{f_{b,\mathbb{P}}}{B_{\mathbb{P}}} \times \sum_{c=b-10}^{b+10} S(c) \right),$$

where S assigns sentiment to each c (Loughran & McDonald 2011)

$$S(c) = \begin{cases} +1 \text{ if } c \in \mathbb{S}^{+} \\ -1 \text{ if } c \in \mathbb{S}^{-} \\ 0 \text{ otherwise} \end{cases}$$

  - Find that $Corr(PRisk_{it}, PSentiment_{it}) = -0.095^{***}$

$PRisk_{it}$ identifies conversations about risks associated with political topics.

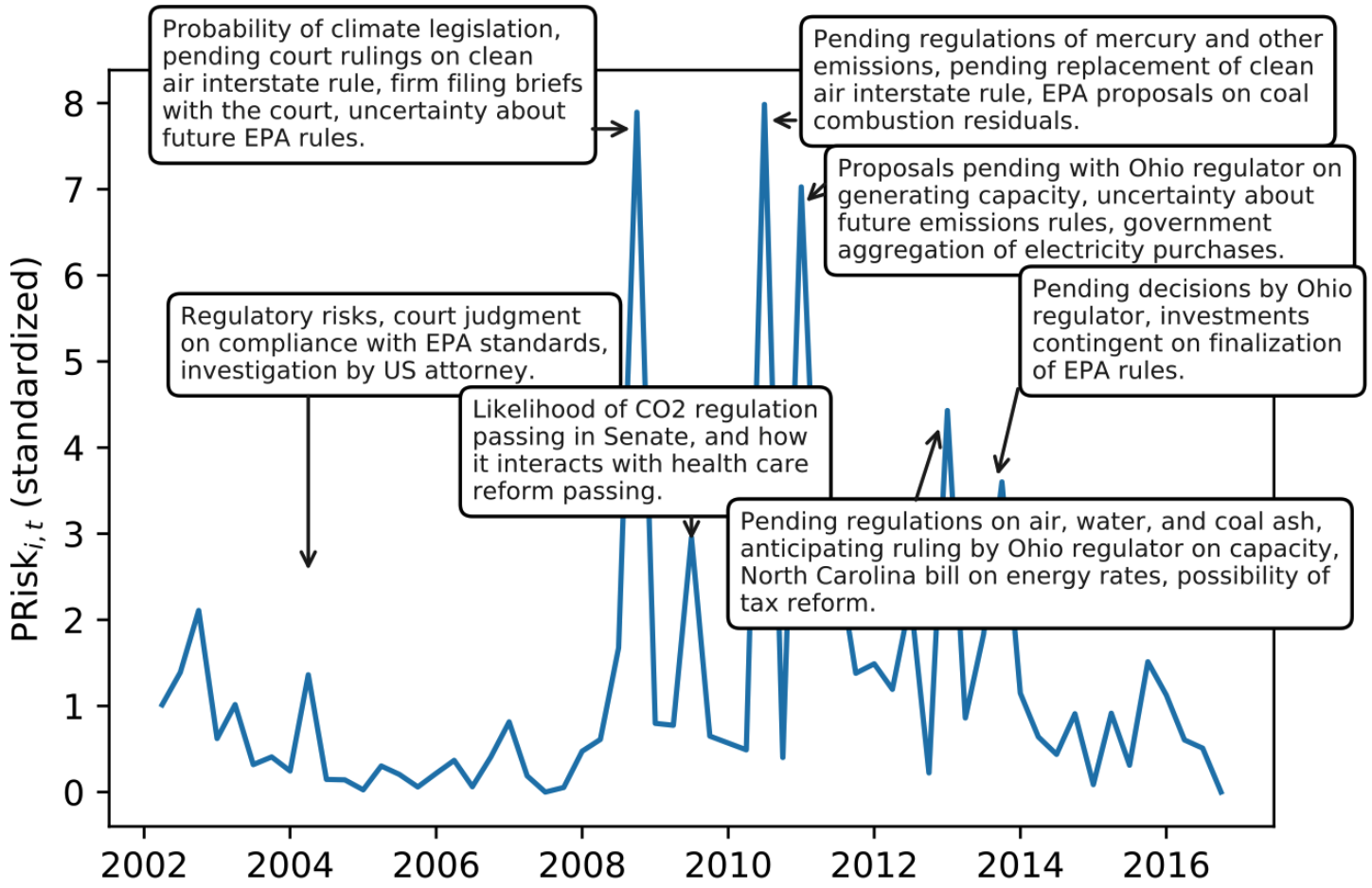
- Bigrams with highest scores intuitively linked to politics (‘the constitution,’ ‘public opinion,’ ‘interest groups,’ ‘the FAA’ …)
  Transcripts with highest PRisk it indeed center around discussions about ballot initiatives, legislation, regulation, government expenditure,…

Transcripts with highest PRisk

Firm Name	Call Date	$\text{PRisk}_{i,t}$ (std)	Text surrounding bigram with highest weight ($f_{b,\mathbb{P}} / B_{\mathbb{P}}$)
Insurance Australia Group Ltd	23-Feb-2012	38.70	leadership i just wondered if you **had concerns** about how **the political** —INSTABILITY— might affect **policies that** have **ramifications for** the industry
FPIC Insurance Group, Inc.	30-Oct-2008	38.69	a —CHANCE— for national **tort reform and** i dont see **the constitution of** congress **changing in** such a way after **this election**
BANKFINANCIAL CORP	4-Nov-2008	38.33	was **an accurate** metaphor and really given all the —UNCERTAINTIES— **of government involvement** in operations and business activities and given the capital
Nanogen, Inc.	8-Aug-2007	37.20	**a dip** in revenues during q related to the —UNCERTAINTY— **of government approval for** the phase **funding of** the cdc contract additionally
World Acceptance Corporation	25-Jul-2006	36.90	management analyst i wanted to followup on the regulatory **front the states** that you had **mentioned the** —POSSIBILITY— of some positive **legislation**
United Refining Company	23-Jul-2010	35.32	shape on asphalt **the funding is** very —IFFY— in all **the states so** and the private work is very slow operator operator
Magellan Health Services	29-Jul-2010	35.26	future so this is a time of quite —UNCERTAINTY— for **the states** they are not **sure what** the fmap will be if

Example: Duke Energy Corporation

A coal company's $PRisk_{it}$



Sources and Transmission of Country Risk: Hassan, Schreger, Schwedeler, and Tahoun (2021)

For each of 56 countries assemble a training library, $\mathbb{T}^C$

- All "Country Commerce" reports published by the *Economist Intelligence Unit* 2002-2016
- All names of the country, names of towns with more than 15,000 inhabitants in 2018, and administrative subdivisions from geonames.org and *CIA World Factbook*

Use $tf \times idf$ to identify two-word combinations ("bigrams") that that are indicative of discussion of each country. For example, a mention of "Angela Merkel" is indicative of discussions of country C if:

- The bigram is frequent in C's training library ($f_{b,C}/B_C$).
- It is rarely used in other countries' libraries ($\log(C/f_{b,C})$).

Four Dimensions of $CountryRisk_{i,c,t}$

1. Risk a given set of firms K associates with country c:

$$CountryRisk^{K}_{c,t} = \frac{1}{N_K} \sum CountryRisk_{i,c,t}$$

2. Foreign risks perceived by firm i at time t:

$$ForeignRisk_{i,t} = \sum_{c \neq d(i)} CountryRisk_{i,c,t}$$

Four Dimensions of $CountryRisk_{i,c,t}$

3. Transmission of risk from o to d at time t:

$$TransmissionRisk_{o \to d,t} = \frac{1}{N_d} \sum CountryRisk_{i,o,t}$$

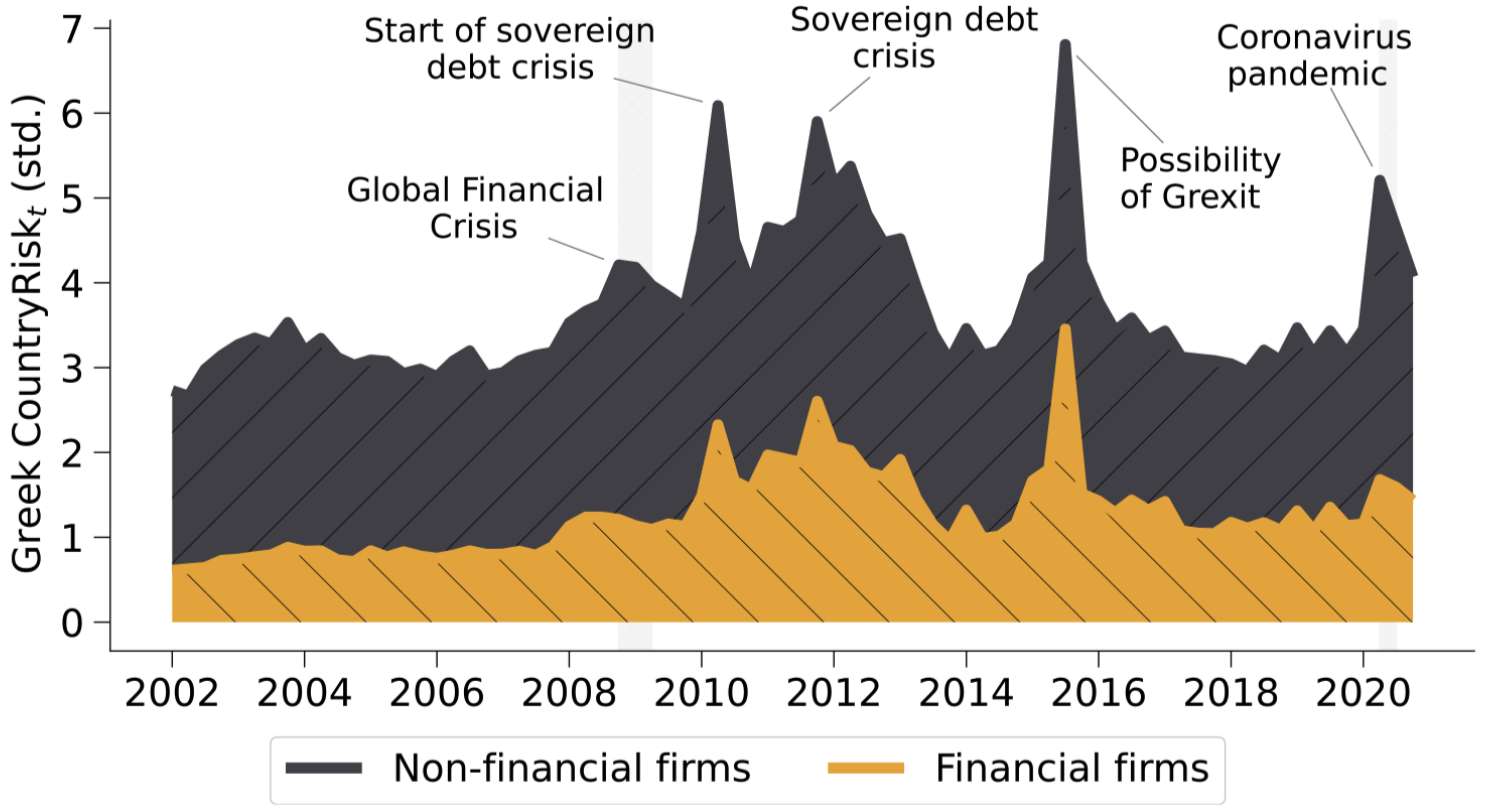
4. Global Risk at time t:

$$GlobalRisk_t = \frac{1}{N_I} \frac{1}{N_C} \sum_{i \in I} \sum_{c \in C} CountryRisk_{i,c,t}$$

Measuring Exposure, Sentiment, and Firm Risk

- $CountryExposure_{i,c,t}$: tf × idf weighted share of words related to country c
- $CountrySentiment_{i,c,t}$: tf × idf weighted sum of tone words toward country c (Loughran & McDonald 2011) (Proxy for positive/negative news about country c)
- $FirmRisk_{i,t}$: Unweighted count of risk words. (Proxy for overall risk faced by the firm)

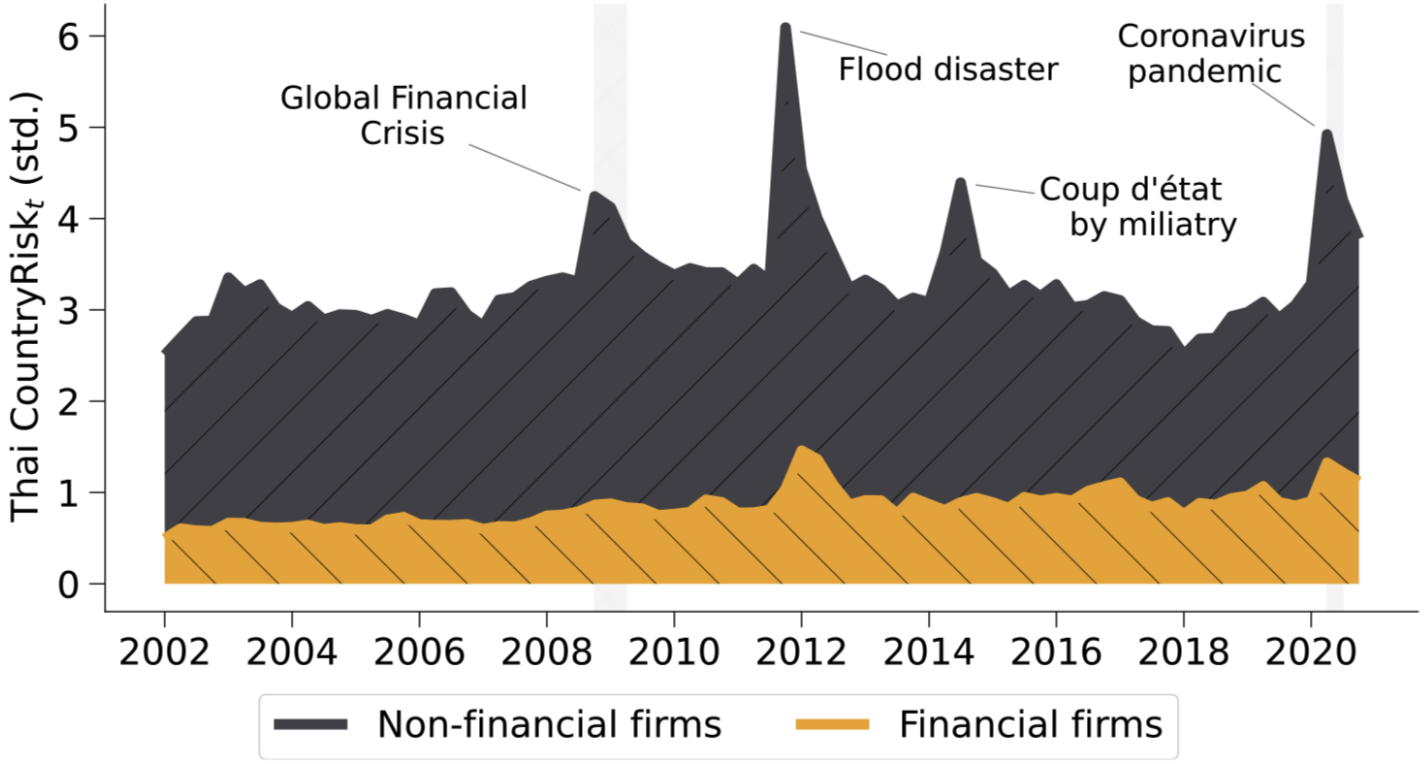
Financial and Non-Financial Risk: Greece



Example text snippet about **possiblity of Grexit** (2015q3):

> *"[...] concern related to the possible impact of a Greek euro-zone exit has led to persistent volatility [...]" (BlackRock Inc, July 15, 2015)*

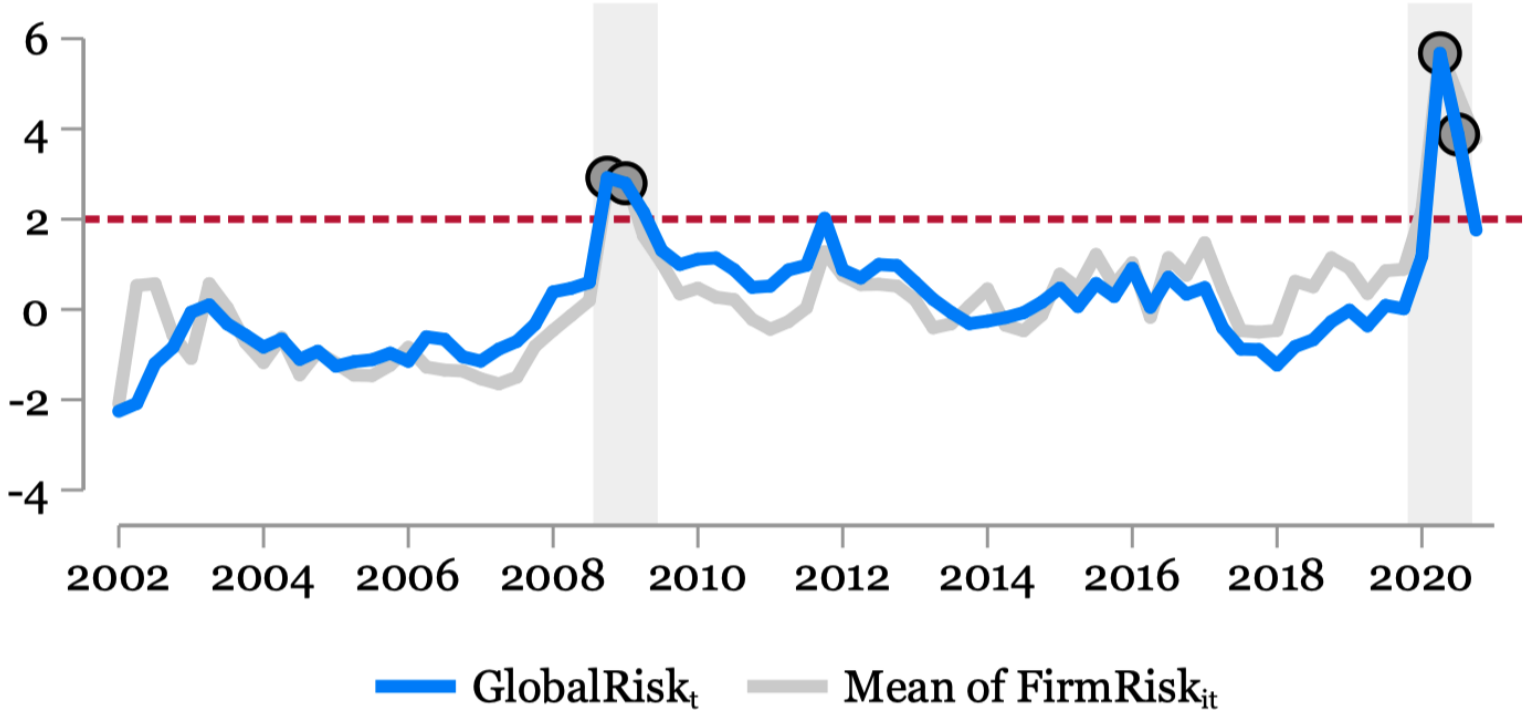
Financial and Non-Financial Risk: Thailand



Example text snippet about **flood disaster** (2011q4):

> *"[...] risk of supply constraints resulting from the recent flood-ing in Thailand Working capital decreased by approximately million to million during the first [...]" (March Networks Corp, December 9, 2011)*

$GlobalRisk_t$



Hoberg and Philips (2016)

- Use similarity between product descriptions in 10Ks to identify industry clusters.
- How similar are two firms' products?
- The most popular way of calculating similarity is cosine similarity.

$$S\_i, j = c_i \cdot c_j$$

where c_i is the normalized representative vector of words for document i.

- A creative way of figuring out who is competing with whom!

Hoberg Phillips 2016 - Prod. Desc. to vector

- Only keep nouns (webster.com) and proper nouns. Drop most commonly used nouns.
- Vector (c_i) is binary values for included words.
- Cosine similarity between all firm-year pairs results in a huge matrix with firm-year as rows and columns.
- Cluster firm-pairs year by year to form yearly industry clusters.

Hoberg Phillips 2016 - Result

Sample industry that changed a lot

**** Industry Surrounding CACI International in 1997 ***

SIC CODES OF 60 RIVALS: COMPUTER PROGRAMMING AND DATA PROCESSING [SIC3=737] (48 RIVALS), ENGINEERING AND ARCHITECTURAL [SIC3=871] (2 RIVALS) PERSONNEL SUPPLY SERVICES [SIC3=736] (2 RIVALS), PROFESSIONAL AND COMMERCIAL EQUIPMENT [SIC3=504] (2 RIVALS), MISC OTHER (6 RIVALS)

Core Words: CLIENT (56), SERVER (54), INTERNET (53), SOLUTION (51), ARCHITECTURE (51), DATABASE (51), ENTERPRISE (50), CLIENTS (48), DATABASES (48), PROGRAMMING (47), MICROSOFT (47), ENVIRONMENTS (46), PRODUCTIVITY (43), COPYRIGHT (43), SECRET (43), INTERFACE (42), WINDOWS (42), FUNCTIONALITY (40), TOOL (40), BACKGROUND (39), DOCUMENTATION (39), INTRANET (39), TELECOMMUNICATIONS (38), OBJECT (38), CYCLE (36), LEGACY (36), SUITE (36), VENDOR (36), ...

**** Industry Surrounding CACI International in 2008 ***

SIC CODES OF 18 RIVALS: COMPUTER PROGRAMMING AND DATA PROCESSING [SIC3=737] (8 RIVALS), SEARCH, DETECTION, NAVIGATION, GUIDANCE, AND AERONAUTICAL [SIC3=381] (5 RIVALS), COMMUNICATIONS EQUIPMENT [SIC3=366] (2 RIVALS), MISC OTHER (3 RIVALS)

Core Words: DEFENSE (19), MILITARY (18), MISSION (18), CONTRACTOR (17), HOMELAND (17), PROCUREMENT (17), PRIME (17), QUANTITY (16), INTELLIGENCE (16), ENVIRONMENTS (15), AWARD (15), BUDGET (14), COMMAND (14), ARCHITECTURE (13), SPECTRUM (13), UNDERSTANDING (13), WARFARE (13), SURVEILLANCE (13), TASK (12), LOCKHEED (12), MARTIN (12), SUBCONTRACTOR (12), PROPOSAL (12), PROCUREMENTS (12), RECONNAISSANCE (12), ARMY (11), ...

- Use this to track who is competing with whom, form industry definitions.
- One drawback of 10K's: only available for US firms.

Measuring Economic Policy Uncertainty (EPU)

THE
QUARTERLY JOURNAL
OF ECONOMICS

Vol. 131 November 2016 Issue 4

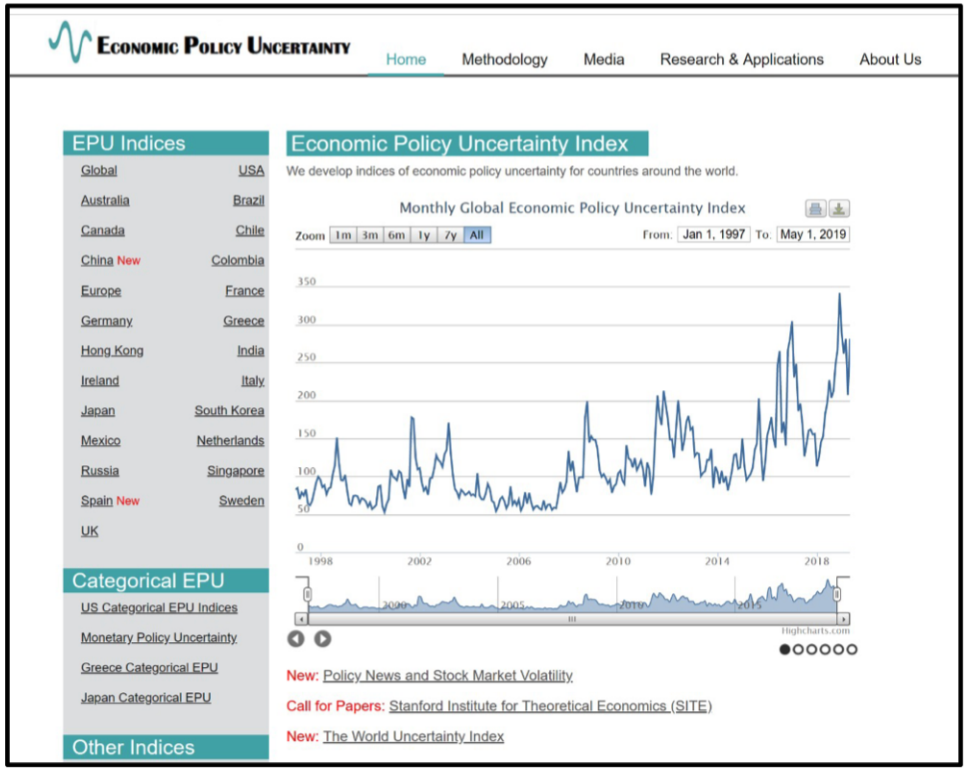
MEASURING ECONOMIC POLICY UNCERTAINTY*

SCOTT R. BAKER
NICHOLAS BLOOM
STEVEN J. DAVIS

We develop a new index of economic policy uncertainty (EPU) based on newspaper coverage frequency. Several types of evidence—including human readings of 12,000 newspaper articles—indicate that our index proxies for movements in policy-related economic uncertainty. Our U.S. index spikes near tight presidential elections, Gulf Wars I and II, the 9/11 attacks, the failure of Lehman Brothers, the 2011 debt ceiling dispute, and other major battles over fiscal policy. Using firm-level data, we find that policy uncertainty is associated with greater stock price volatility and reduced investment and employment in policy-sensitive sectors like defense, health care, finance, and infrastructure construction. At the macro level, innovations in policy uncertainty foreshadow declines in investment, output, and employment in the United States and, in a panel vector autoregressive setting, for 12 major economies. Extending our U.S. index back to 1900, EPU rose dramatically in the 1930s (from late 1931) and has drifted upward since the 1960s. *JEL Codes:* D80, E22, E66, G18, L50.

*We thank Adam Jorring, Kyle Kost, Abdulla Al-Kuwari, Sophie Biffar, Jörn Boehnke, Vladimir Dashkeyev, Olga Deriy, Eddie Dinh, Yuto Ezure, Robin Gong, Sonam Jindal, Ruben Kim, Sylvia Klosin, Jessica Koh, Peter Lajewski, David Nebiyu, Rebecca Sachs, Ippei Shibata, Corinne Stephenson, Naoko Takeda, Melissa Tan, Sophie Wang, and Peter Xu for research assistance and the National Science Foundation, MacArthur Foundation, Sloan Foundation, Becker Friedman Institute, Initiative on Global Markets, and Stigler Center at the University of Chicago for financial support. We thank Ruedi Bachmann, Sanjai Bhagat, Vincent Bignon, Yongsung Chang, Vladimir Dashkeyev, Jesus Fernandez-Villaverde, Laurent Ferrara, Luis Garicano, Matt Gentzkow, Yuriy Gorodnichenko, Kevin Hassett, Takeo Hoshi, Greg Ip, Anil Kashyap, Patrick Kehoe, John Makin, Johannes Pfeifer, Meijun Qian, Itay Saporta, John Shoven, Sam Schulhofer-Wohl, Jesse Shapiro, Erik Sims, Stephen Terry, Cynthia Wu, and many seminar and conference audiences for comments. We also thank the referees and editors, Robert Barro and Larry Katz, for comments and suggestions.

The Quarterly Journal of Economics (2016), 1593–1636. doi:10.1093/qje/qjw024.
Advance Access publication on July 11, 2016.



www.policyuncertainty.com

This proxy for Economic Policy Uncertainty (EPU) comes from computer searches of newspapers

- US index: 10 major papers get monthly counts of articles with:
  - `E` {economic or economy}, and
  - `P` {regulation or deficit or federal reserve or congress or legislation or white house}, and
  - `U` {uncertain or uncertainty}
- Divide the count for each month by the count of all articles
- Normalize and sum 10 papers to get the U.S monthly index

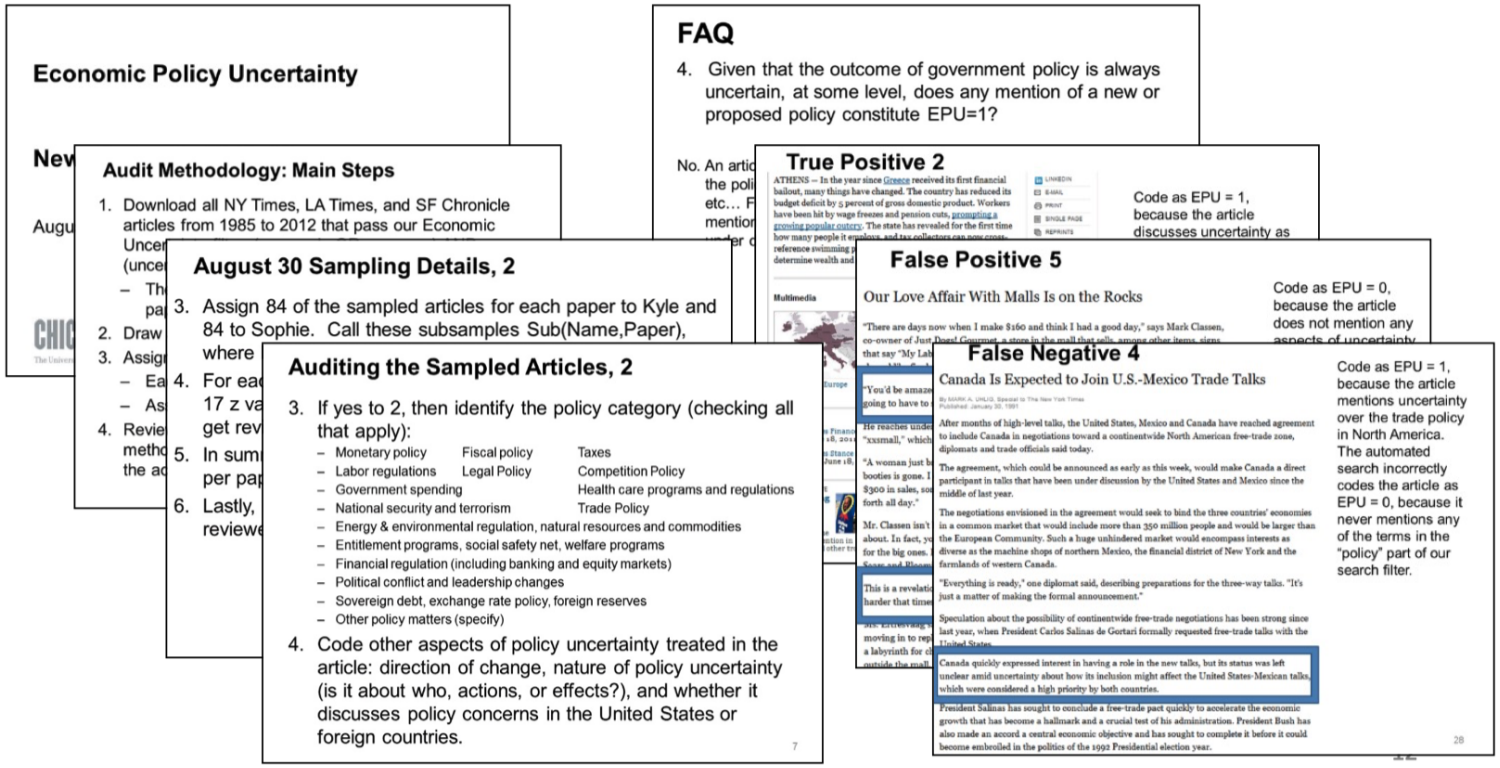
onstructing the US News-Based EPU Index

Newspapers:

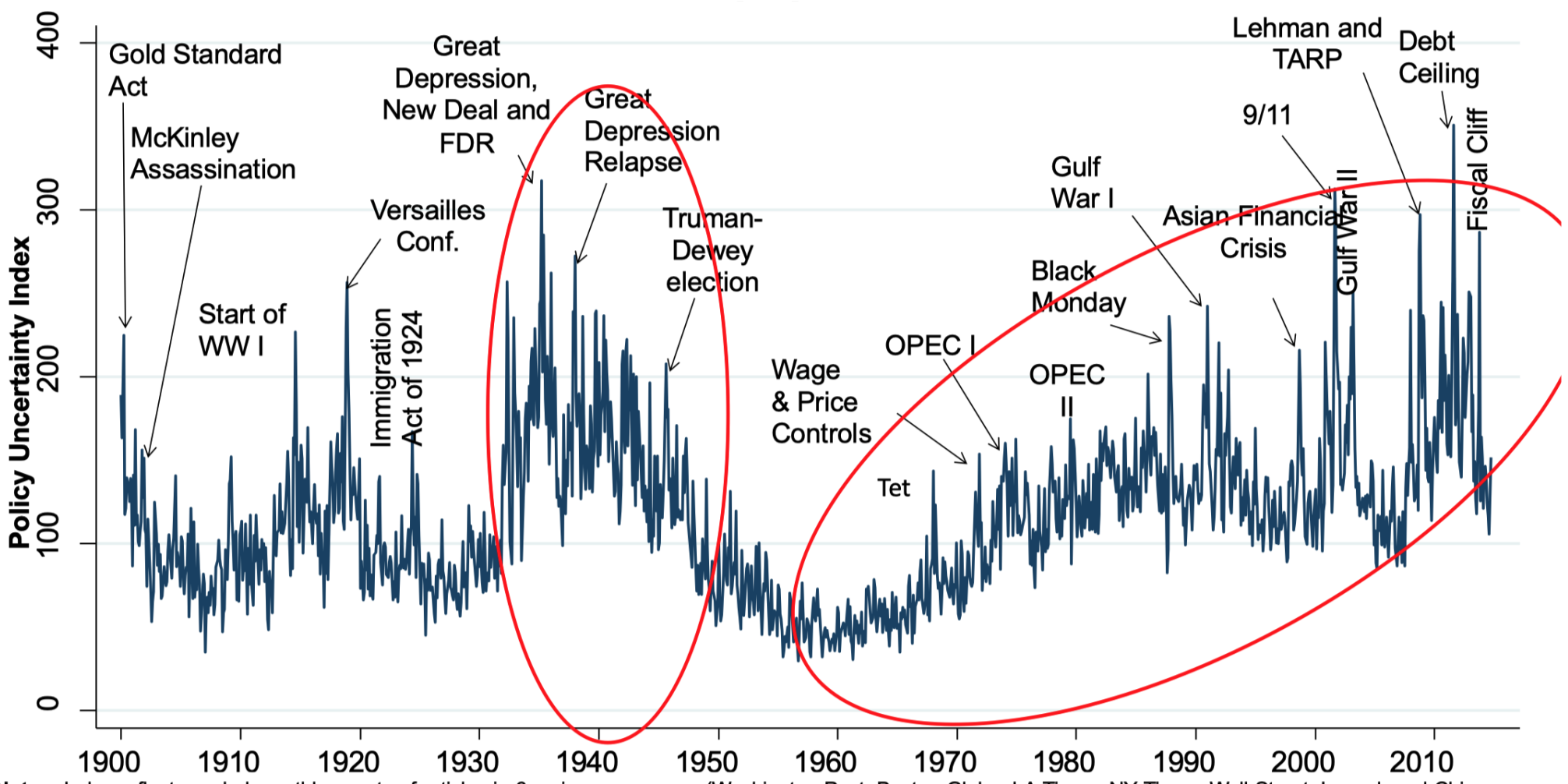
- Boston Globe
- Chicago Tribune
- Dallas Morning News
- Los Angeles Times
- Miami Herald
- New York Times
- SF Chronicle
- USA Today
- Wall Street Journal
- Washington Post

Validation: Running Detailed Human Audits

10 undergraduates read ≈ 10,000 newspaper articles to date using a 63-page audit guide to code articles if they discuss “economic uncertainty” and “economic policy uncertainty”

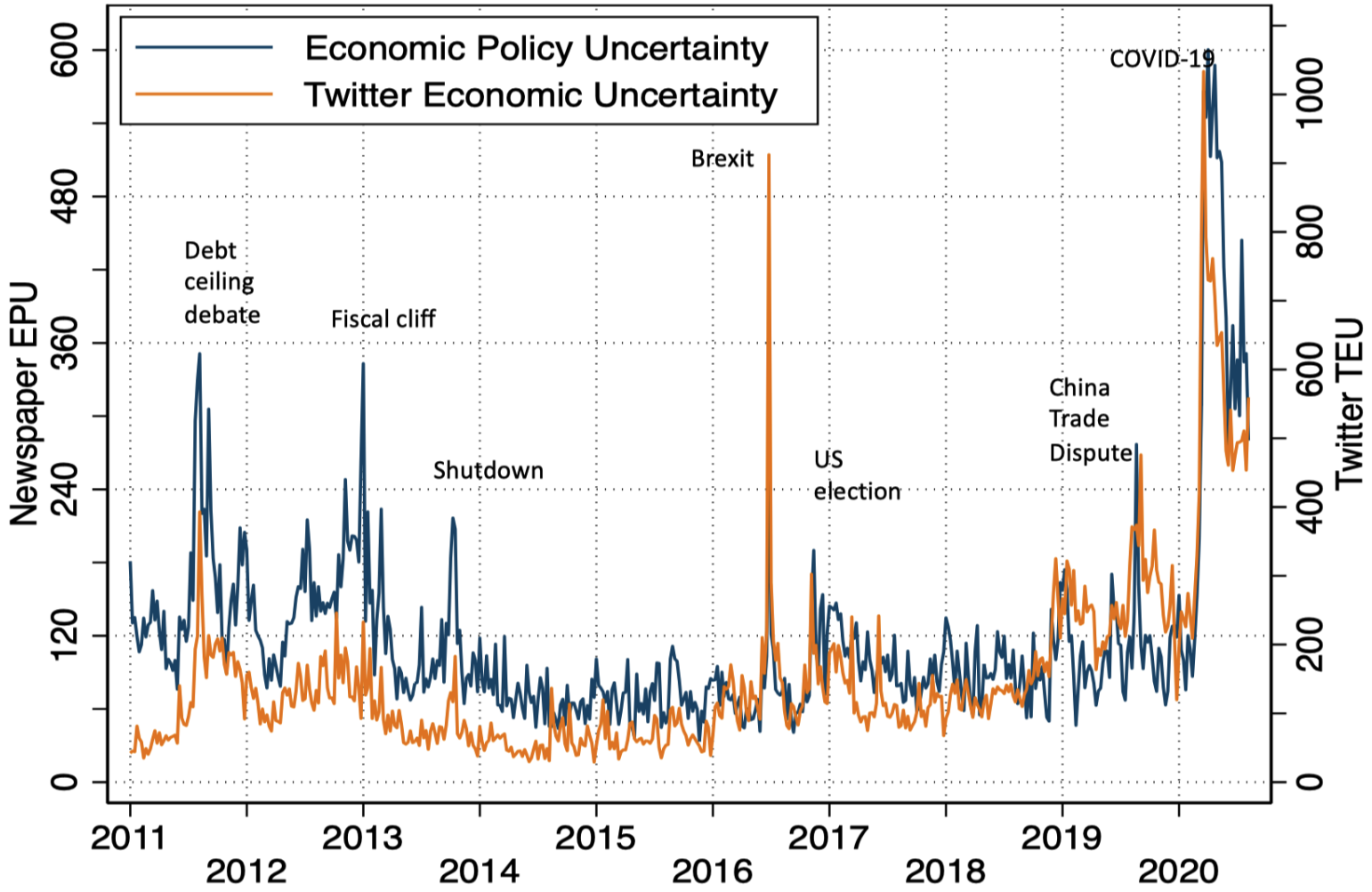


US News-based economic policy uncertainty index



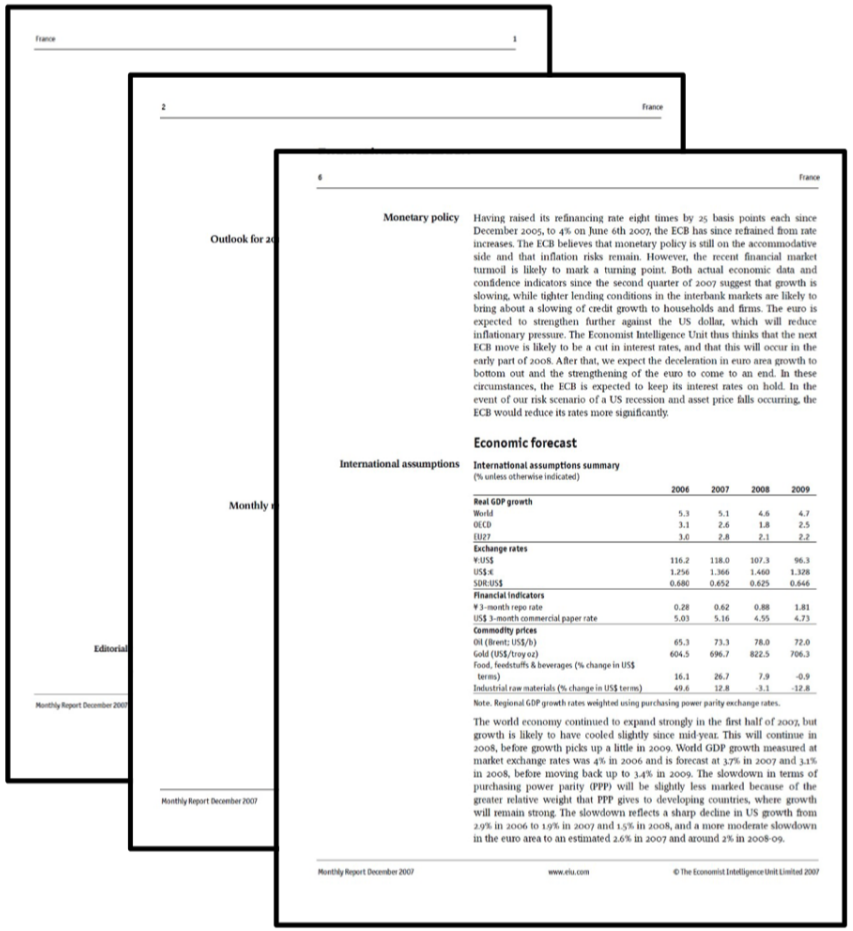
Notes: Index reflects scaled monthly counts of articles in 6 major newspapers (Washington Post, Boston Globe, LA Times, NY Times, Wall Street Journal, and Chicago Tribune) that contain the same triple as in Figure 1, except the economy term set includes "business", "commerce" and "industry" and the policy term set includes "tariffs" and "war". Data normalized to 100 from 1900-2011.

Twitter text uncertainty measures



Notes: Weekly values for Economic Policy Uncertainty (EPU) index and Twitter Economic Uncertainty (TEU) index from www.policyuncertainty.com. See Baker, Bloom and Davis (2016) for details of EPU index construction and Baker, Bloom, Davis and Renault (2020) for details of the TEU index construction, with data at http://www.policyuncertainty.com. We plot data from 1 January 2011 to 12 August

“world uncertainty index” covering 143 countries from Economist Intelligence Unit text



Monetary policy

Having raised its refinancing rate eight times by 25 basis points each since December 2005, to 4% on June 6th 2007, the ECB has since refrained from rate increases. The ECB believes that monetary policy is still on the accommodative side and that inflation risks remain. However, the recent financial market turmoil is likely to mark a turning point. Both actual economic data and confidence indicators since the second quarter of 2007 suggest that growth is slowing, while tighter lending conditions in the interbank markets are likely to bring about a slowing of credit growth to households and firms. The euro is expected to strengthen further against the US dollar, which will reduce inflationary pressure. The Economist Intelligence Unit thus thinks that the next ECB move is likely to be a cut in interest rates, and that this will occur in the early part of 2008. After that, we expect the deceleration in euro area growth to bottom out and the strengthening of the euro to come to an end. In these circumstances, the ECB is expected to keep its interest rates on hold. In the event of our risk scenario of a US recession and asset price falls occurring, the ECB would reduce its rates more significantly.

Economic forecast

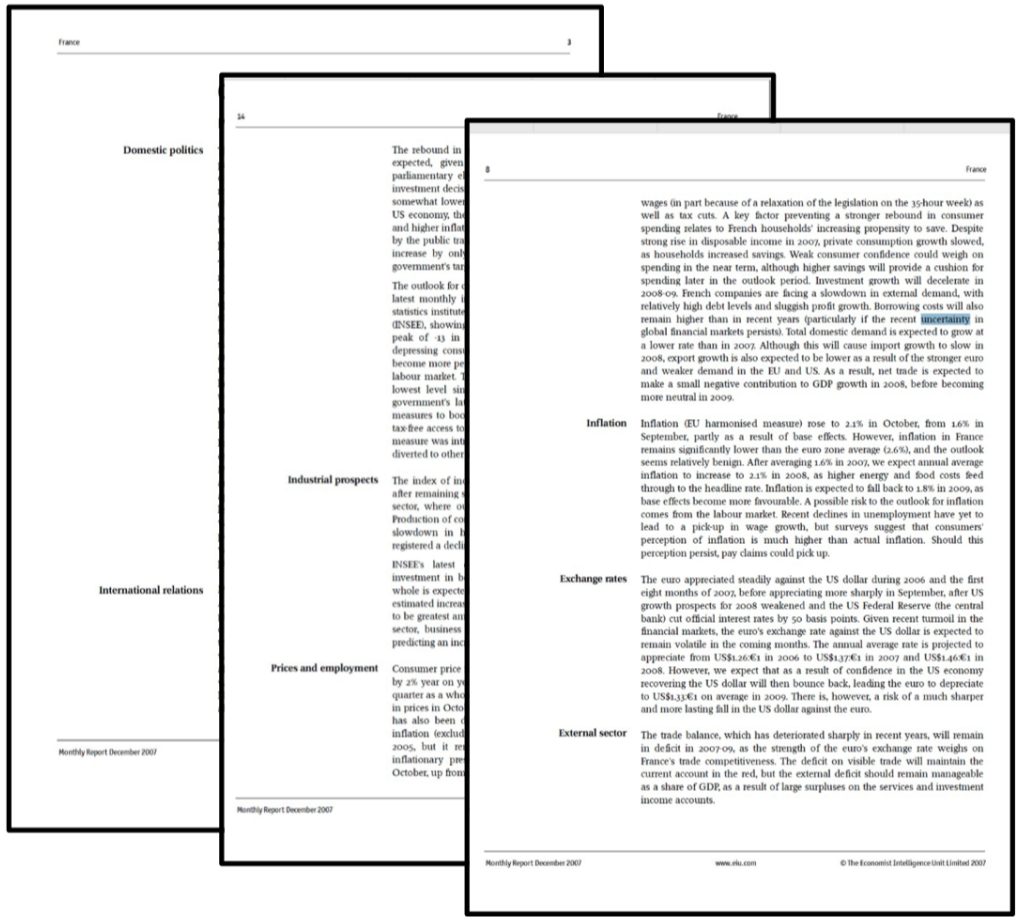
International assumptions

International assumptions summary
(% unless otherwise indicated)

	2006	2007	2008	2009
Real GDP growth				
World	5.3	5.1	4.6	4.7
OECD	3.1	2.6	1.8	2.5
EU27	3.0	2.8	2.1	2.2
Exchange rates				
¥:US$	116.2	118.0	107.3	96.3
US$:€	1.256	1.366	1.460	1.328
SDR:US$	0.680	0.652	0.625	0.646
Financial indicators				
¥ 3-month repo rate	0.28	0.62	0.88	1.81
US$ 3-month commercial paper rate	5.03	5.16	4.55	4.73
Commodity prices				
Oil ($/brent; US$/b)	65.3	73.3	78.0	72.0
Gold (US$/troy oz)	604.5	696.7	822.5	706.3
Food, feedstuffs & beverages (% change in US$ terms)	16.1	26.7	7.9	-0.9
Industrial raw materials (% change in US$ terms)	49.6	12.8	-3.1	-12.8

Note. Regional GDP growth rates weighted using purchasing power parity exchange rates.

The world economy continued to expand strongly in the first half of 2007, but growth is likely to have cooled slightly since mid-year. This will continue in 2008, before growth picks up a little in 2009. World GDP growth measured at market exchange rates was 4% in 2006 and is forecast at 3.7% in 2007 and 3.1% in 2008, before moving back up to 3.4% in 2009. The slowdown in terms of purchasing power parity (PPP) will be slightly less marked because of the greater relative weight that PPP gives to developing countries, where growth will remain strong. The slowdown reflects a sharp decline in US growth from 2.9% in 2006 to 1.9% in 2007 and 1.5% in 2008, and a more moderate slowdown in the euro area to an estimated 2.6% in 2007 and around 2% in 2008-09.



Domestic politics

International relations

Industrial prospects

Prices and employment

wages (in part because of a relaxation of the legislation on the 35-hour week) as well as tax cuts. A key factor preventing a stronger rebound in consumer spending relates to French households' increasing propensity to save. Despite strong rise in disposable income in 2007, private consumption growth slowed, as households increased savings. Weak consumer confidence could weigh on spending in the near term, although higher savings will provide a cushion for spending later in the outlook period. Investment growth will decelerate in 2008-09. French companies are facing a slowdown in external demand, with relatively high debt levels and sluggish profit growth. Borrowing costs will also remain higher than in recent years (particularly if the recent uncertainty in global financial markets persists). Total domestic demand is expected to grow at a lower rate than in 2007. Although this will cause import growth to slow in 2008, export growth is also expected to be lower as a result of the stronger euro and weaker demand in the EU and US. As a result, net trade is expected to make a small negative contribution to GDP growth in 2008, before becoming more neutral in 2009.

Inflation

Inflation (EU harmonised measure) rose to 2.1% in October, from 1.6% in September, partly as a result of base effects. However, inflation in France remains significantly lower than the euro zone average (2.6%), and the outlook seems relatively benign. After averaging 1.6% in 2007, we expect annual average inflation to increase to 2.1% in 2008, as higher energy and food costs feed through to the headline rate. Inflation is expected to fall back to 1.8% in 2009, as base effects become more favourable. A possible risk to the outlook for inflation comes from the labour market. Recent declines in unemployment have yet to lead to a pick-up in wage growth, but surveys suggest that consumers' perception of inflation is much higher than actual inflation. Should this perception persist, pay claims could pick up.

Exchange rates

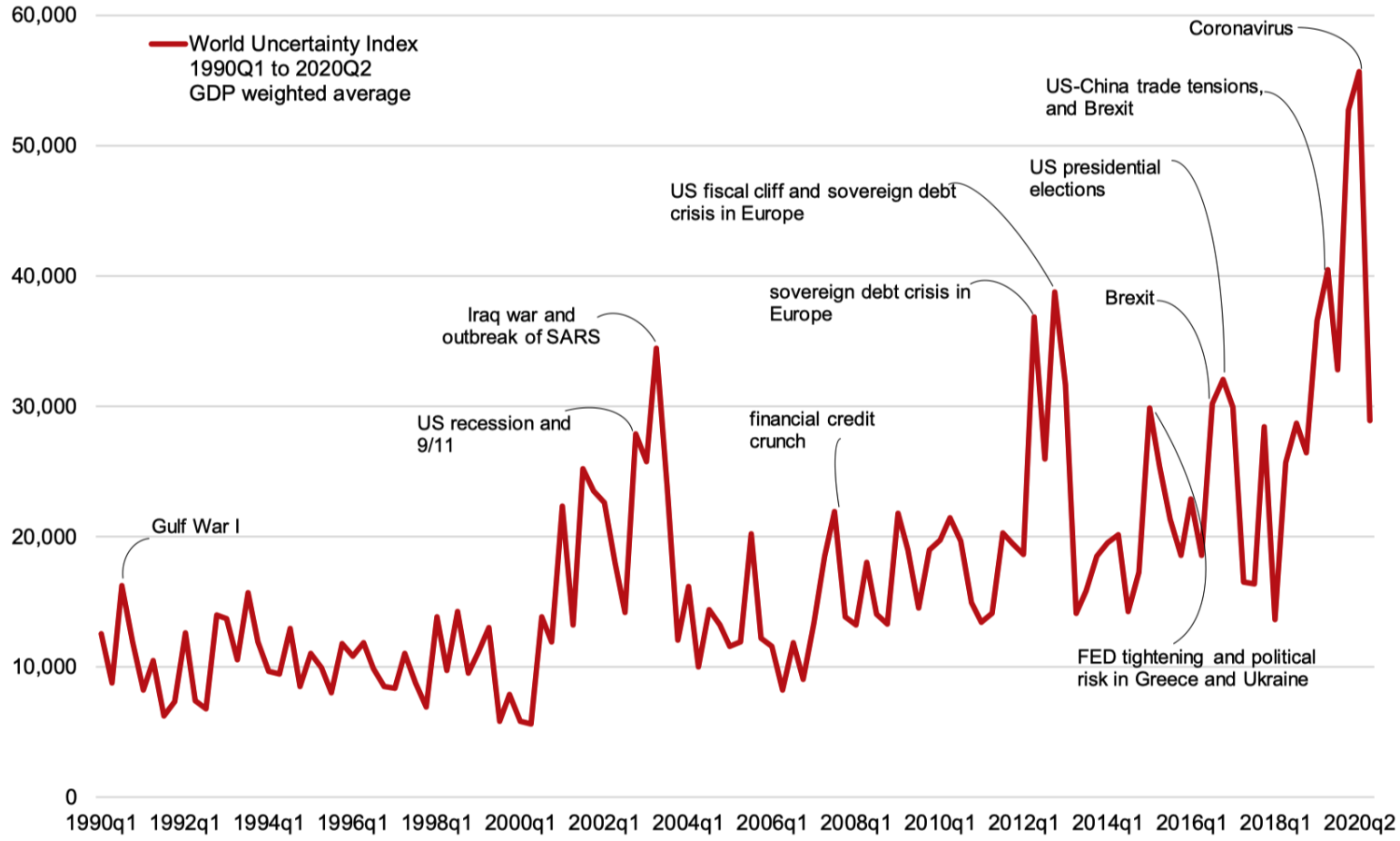
The euro appreciated steadily against the US dollar during 2006 and the first eight months of 2007, before appreciating more sharply in September, after US growth prospects for 2008 weakened and the US Federal Reserve (the central bank) cut official interest rates by 50 basis points. Given recent turmoil in the financial markets, the euro's exchange rate against the US dollar is expected to remain volatile in the coming months. The annual average rate is projected to appreciate from US$1.26:€1 in 2006 to US$1.37:€1 in 2007 and US$1.46:€1 in 2008. However, we expect that as a result of confidence in the US economy recovering the US dollar will then bounce back, leading the euro to depreciate to US$1.33:€1 on average in 2009. There is, however, a risk of a much sharper and more lasting fall in the US dollar against the euro.

External sector

The trade balance, which has deteriorated sharply in recent years, will remain in deficit in 2007-09, as the strength of the euro's exchange rate weighs on France's trade competitiveness. The deficit on visible trade will maintain the current account in the red, but the external deficit should remain manageable as a share of GDP, as a result of large surpluses on the services and investment income accounts.

EUI quarterly reports standard format, mean (and median) of 29 pages.

Global average of all 143 countries



Source: https://worlduncertaintyindex.com/

The Diffusion of Disruptive Technologies

Bloom, Kalyani, Lerner, and Tahoun (2021), The Diffusion of Disruptive Technologies, mimeo Stanford U, Boston U, and HBS

- Construct text-based measures of exposure to 20 different technologies at the firm, patent, and job-level, 2002-19.
- Use these novel data to study the spread of new technologies across firms, regions, occupations, and skill-levels.

Five Stylized Facts on Disruptive Technologies

1. Development & initial employment in disruptive technologies is geographically highly concentrated.
2. Over time, hiring associated with new technologies gradually spreads: “region broadening.”
3. Over time, skill level in tech jobs declines sharply: “skill broadening.”
4. Low-skill jobs associated with a given technology spread out significantly faster than high-skill jobs.
5. Pioneer locations retain long-lasting advantage in high-skilled jobs.

Data Sources

1. Full text of USPTO patents (1976-2016)
    - Typically follow a research paper format – invention title, abstract, claim, description.
2. Transcripts of Earnings Conference Calls (2002-19)
    - Discussions of 300k+ quarterly earnings by 12k publicly listed firms.
    - Typically contains management presentation followed by analyst Q & A.
3. Full text of 200 M+ online job postings from BG (2007, 2010-19).
    - Scraped from job forums (e.g., Glassdoor.com) and employer websites.
    - Geo-coded and assigned to SOC Codes

Step 1: Identify Technical Bigrams from Patents

Identify two-word combinations (bigrams) that are indicative of discussion of novel technologies.

1. Extract all (17 mil+) bigrams US patents (1976-2016)
2. Remove any bigrams that were commonly in use prior to 1970 (Corpus of Historical American English)
3. Keep bigrams which account for at least 1000 citations.

> List of 35,063 ‘technical bigrams’ associated with influential inventions.

Top Bigrams in Patents

tech	Citations (std)	tech	Citations (std)	tech	Citations (std)
readable medium	204379	computer implemented	71392	acceptable carrier	57709
user interface	197444	acid sequence	69521	disposed adjacent	55760
readable storage	131350	pharmaceutical compositions	69082	computing system	55655
fluid communication	117896	positioned adjacent	69018	optical fiber	54593
storage media	97478	pharmaceutical composition	68895	disk drive	54282
electrically conductive	96715	data structures	68401	plan view	53676
transitory computer	85426	service provider	67424	digital data	52337
readable media	82035	output signals	66494	acceptable salts	52291
conductive material	80232	data structure	62563	graphical user	52256
machine readable	80079	vapor deposition	61255	electrically coupled	51310
user input	76641	image data	59499	dielectric layer	50642
polyethylene glycol	75889	fiber optic	59395	temperature sensor	50612
data stored	72884	personal computer	59216	polymeric material	50360
proc natl	71562	volatile memory	58589	acceptable salt	48690
natl acad	71523	computer executable	58501	data stream	48320

Step 2: Identify Disruptive Technologies from Earnings Calls

Identify technical bigrams that are discussed in EC with increasing frequency (keep those at <10% of max in first year) – Total 305.

bigram	# ECs
mobile devices	6597
machine learning	2860
cloud computing	2781
cloud services	2450
quality metrics	2029
flow profile	1966
smart phones	1957
mobile platform	1605
public cloud	1569
social networking	1548
smart grid	1441
cloud service	1393
connected devices	1304
cloud infrastructure	1136

bigram	# ECs
nand flash	1002
virtual reality	903
digital channel	896
delivery network	887
social networks	883
autonomous driving	839
smart devices	765
active user	735
augmented reality	730
mobile payment	717
cloud environment	668
production site	664
ethanol production	662
power outage	643

bigram	# ECs
autonomous vehicles	586
global warming	565
cloud based	545
hydraulic fracturing	506
optimization process	505
software defined	482
wifi network	474
results page	454
user behavior	441
additive manufacturing	438
millimeter wave	426
identity theft	424
relevant content	423
local search	420

Technical vs non-technical bigrams

Non technical bigrams = bigrams in earnings calls and NOT in patents

Statistic	Supervised bigrams	Non-technical bigrams (top 221)	Technical bigrams (Unsupervised)	Non-technical bigrams* (top 305)	Non-technical bigrams (ext)* (top 4000)
# bigrams	221	221	305	305	4000
Avg. postings/bigram	59,013	142	49,677	157	474
Bigrams w/ more than 100 postings	88.3%	10.0%	92.4%	9.2%	8.1%

Top Technical bigrams

bigram	# earnings	# job postings
mobile devices	6597	1078049
machine learning	2860	525286
cloud computing	2781	485333
cloud services	2450	380980
quality metrics	2029	196497

Top Non-technical Bigrams

bigram	# earnings	# job postings
bofa merrill	34490	221
stifel nicolaus	28877	256
division associate	12472	4237
keefe bruyette	11682	16
bruyette woods	11498	14

Step 3: Bigrams to Technologies

Two alternative approaches

- `“Supervised”`: Group bigrams with similar meaning to measure the spread of 29 specific technologies, add `synonyms’ and manually audit each bigram. (Main specification)
    - Smart Devices - mobile devices; smartphone tablet; android phones; smart phones ...
    - 3d printing - 3d printer; 3d printing; additive manufacturing; d printed
- `“Unsupervised”`: Treat each tech bigram as a separate technology without any further intervention. (Robustness check)

Technology Exposure

1. Measure technology exposure at the patent, earnings call, and job level as

$$exposure_{i,\tau,t} = 1\{b_\tau \epsilon\, D_{i,t}\},$$

where $D_{i,t}$ is the set of bigrams contained in a job posting/earnings call posted at time t and b_τ is a bigram associated with technology τ.

Example Jobs Exposed to Smart Devices

Systems Engineer – Produce

We are looking for a senior systems engineer with years of experience board design skills for embedded processing and mixed signal applications based on a good understanding of signal integrity familiarity with digital analog and video interfaces experience with high speed digital interfaces good problem solving and trouble shooting skills …

…*as a member of the digital entertainment business unit you will play a key role in the development testing and validation of new chips in the growing* ***smart tv*** *market.* You will be involved in a wide range of activities that cover all phases of the chip development cycle from prototyping support to final silicon validation…

Sales Representative – Use

We are in search of an outgoing driven and reliable individual who is looking for a part time or full time opportunity to become a brand rep for a regulated product in convenience store locations in your area you will be a key asset to the program…

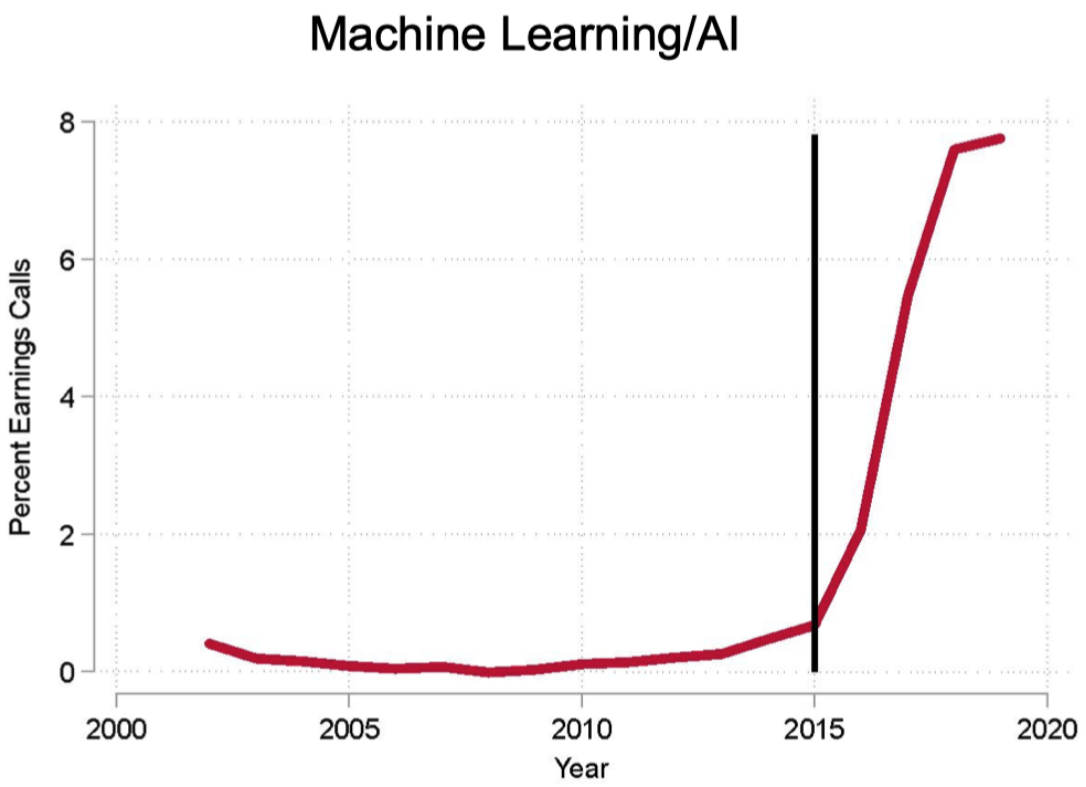
Responsibilities and requirements:

- work on product displays pull product out of back stock and merchandise replenish displays as needed.
- *use third channel technology on a* ***smart device*** *to collect crucial data engage with consumers and provide sales support/brand education to retail associates*
- *reliable transportation a* ***smart phone*** *with internet access.*

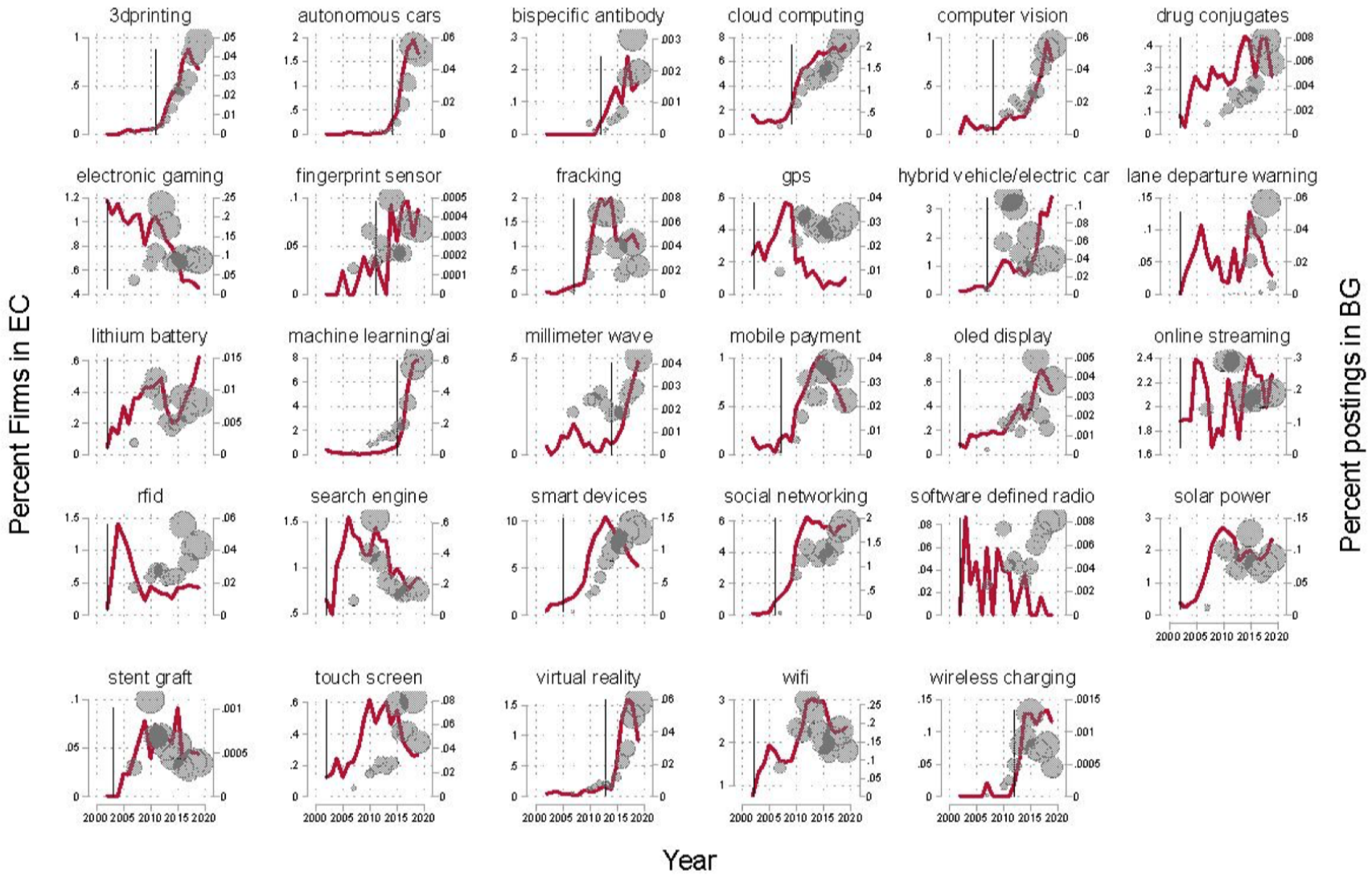
- On average, each technical bigram appears in 59,013 job postings. Compare to 157 average mentions of top non-technical bigrams from earnings calls.

Define an Emergence Year for each Technology

1. Measure the share of earnings calls mentioning a technology
2. Define a “technology year of emergence” as year in earnings calls when `the time series first attains at least 10% of its maximum`.

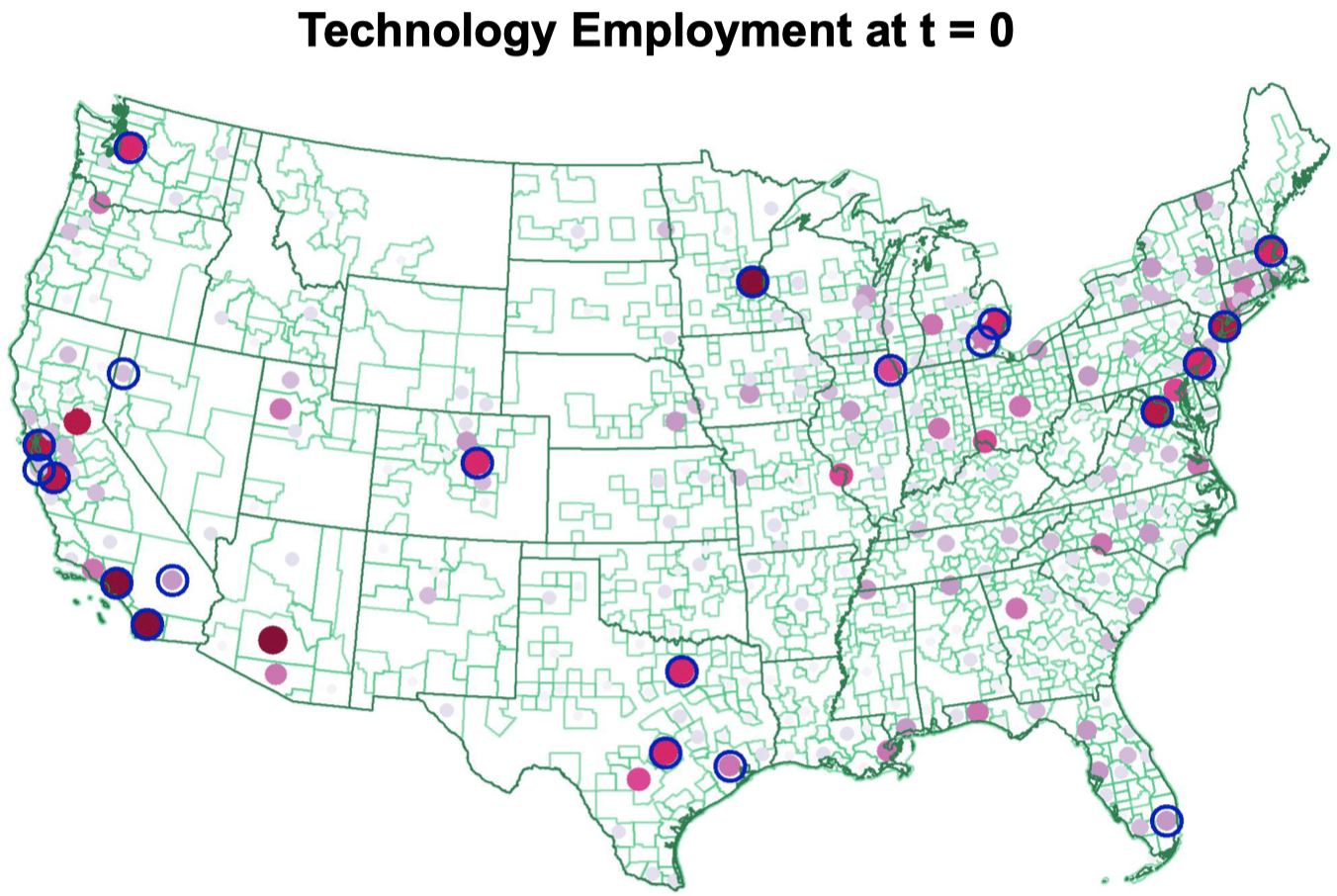


Share Exposed Firms and Job Postings – Corr. 80%

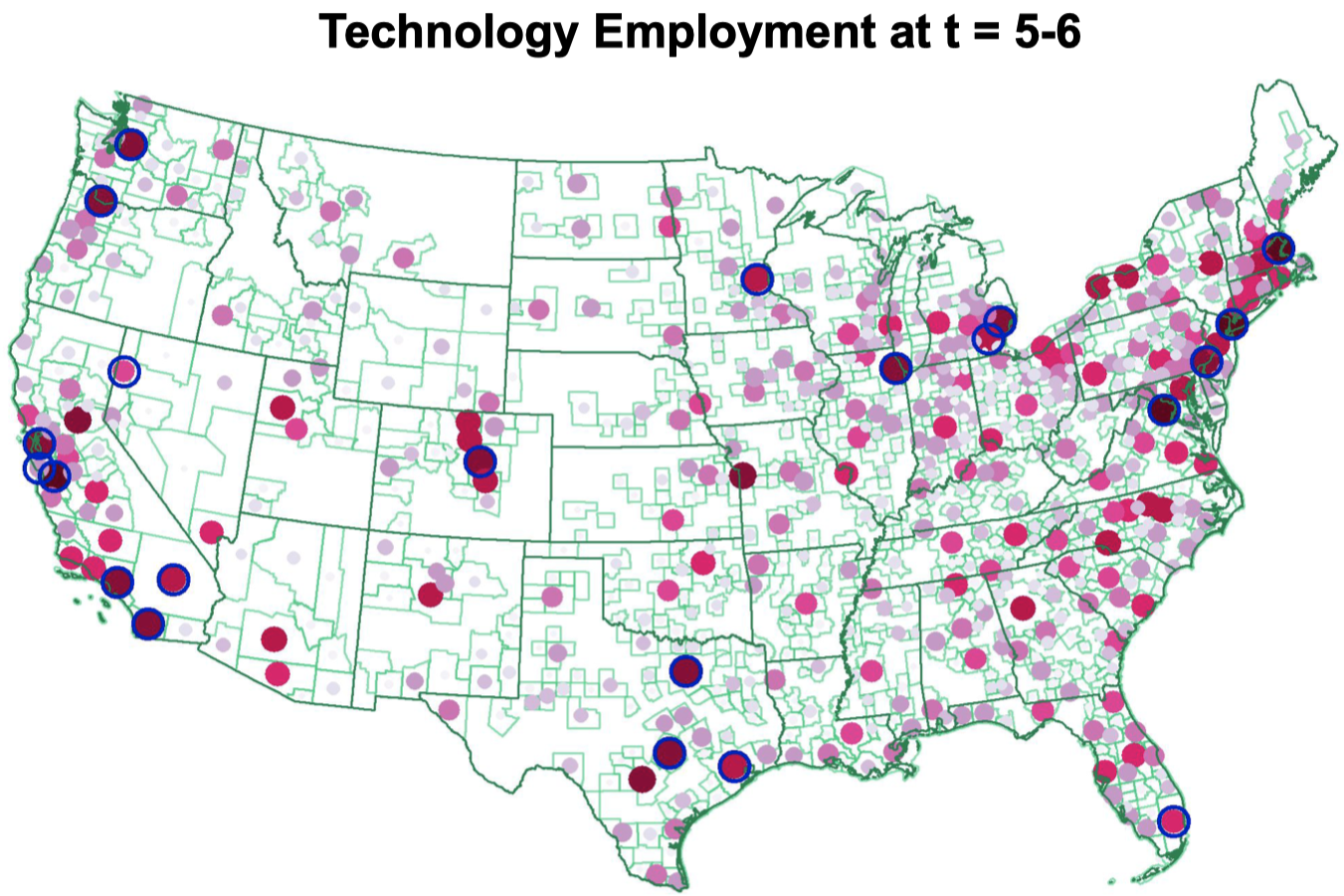


Pioneer Locations

Define pioneer locations as ones which account for 50% of technology patents 10 years before emergence year.



Broadening over Time and Pioneer Locations



Parts of Speech Predict Loan Repayment

Netzer, Lemaire, and Herzenstein (2019), "When Words Sweat"

Imagine you consider lending $2,000 to one of two borrowers on a crowdfunding website. The borrowers are identical in terms of demographic and financial characteristics. However, the text they provided when applying for a loan differs:

```
Borrower #1:

"I am a hard working person, married for 25 years, and have two wonderful boys.
Please let me explain why I need help.
I would use the $2,000 loan to fix our roof.
Thank you, god bless you, and I promise to pay you back."

Borrower #2:

"While the past year in our new place has been more than great,
the roof is now leaking and I need to borrow $2,000 to cover the cost of the repair.
I pay all bills (e.g., car loans, cable, utilities) on time."
```

Parts of Speech Predict Loan Repayment

Which borrower is more likely to default?

“Loan requests written by defaulting borrowers are more likely to include words (or themes) related to the borrower’s family, financial and general hardship, mentions of god, and the near future, as well as pleading lenders for help, and using verbs in present and future tenses.”

Loan Application Words Predicting Repayment (Netzer, Lemaire, and Herzenstein 2019)

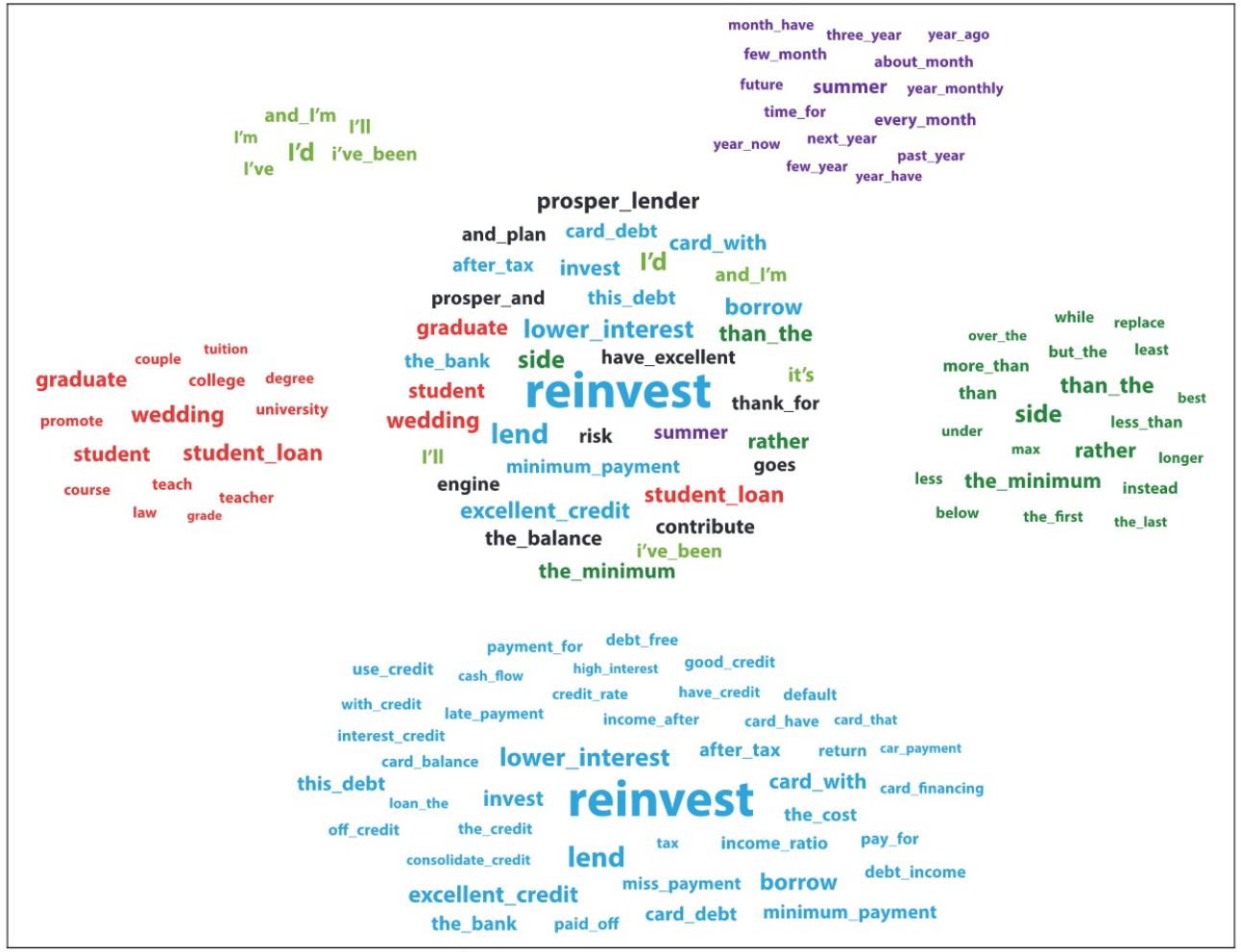


Figure 2. Words indicative of loan repayment.

Notes: The most common words appear in the middle cloud (cutoff = 1:1.5) and are then organized by themes. Starting on the right and moving clockwise: relative words, financial literacy words, words related to a brighter financial future, "I" words, and time-related words.

Loan Application Words Predicting Default (Netzer, Lemaire, and Herzenstein 2019)

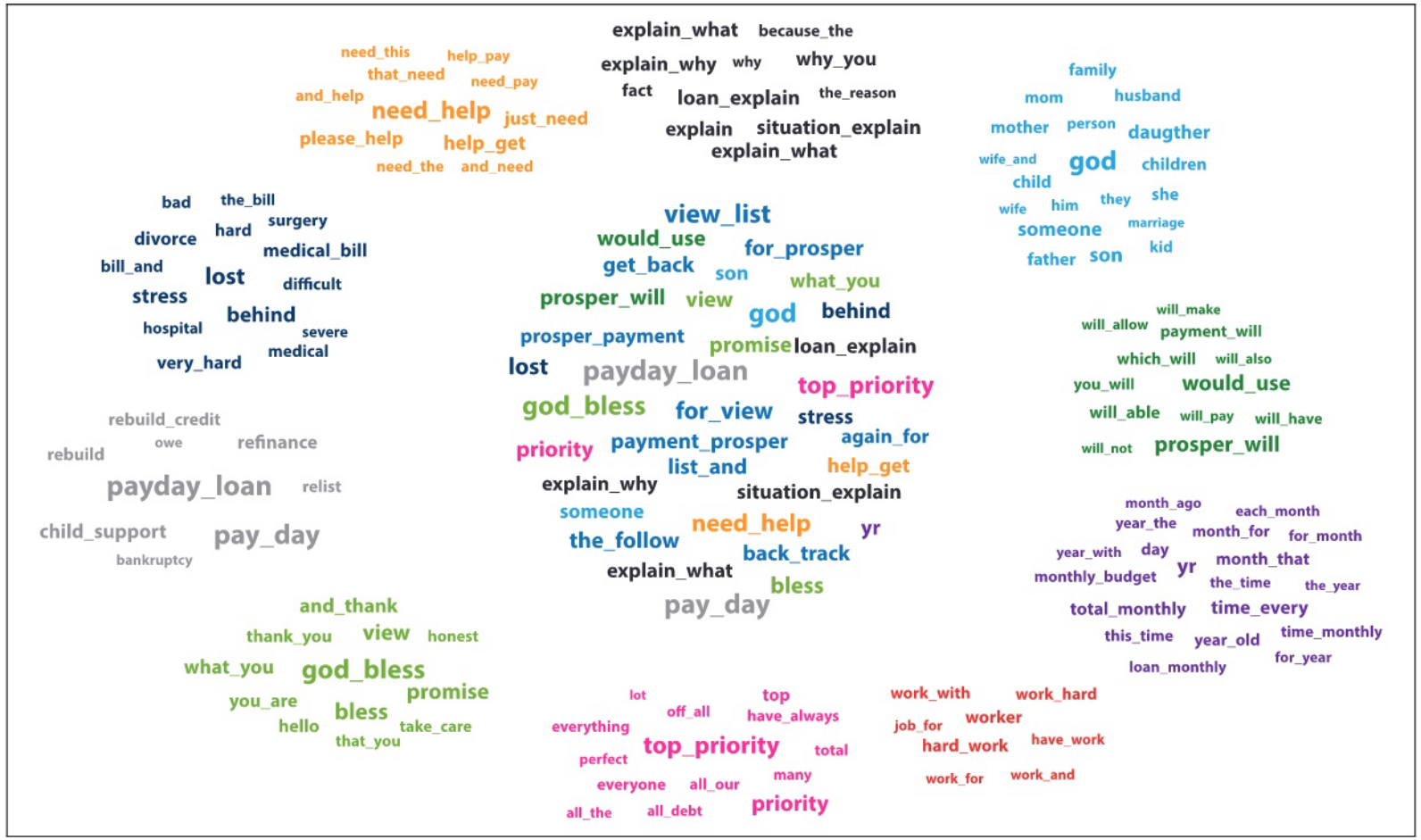


Figure 3. Words indicative of loan default.

Notes: The most common words appear in the middle cloud (cutoff = 1:1.5) and are then organized by themes. Starting on the top and moving clockwise: words related to explanations, external influence words and others, future-tense words, time-related words, work-related words, extremity words, words appealing to lenders, words relating to financial hardship, words relating to general hardship, and desperation/plea words.

Burgess et al, “Legislative Influence Detectors”

The two largest interest group associations: ALEC (on the conservative side) and ALICE (on the liberal side)

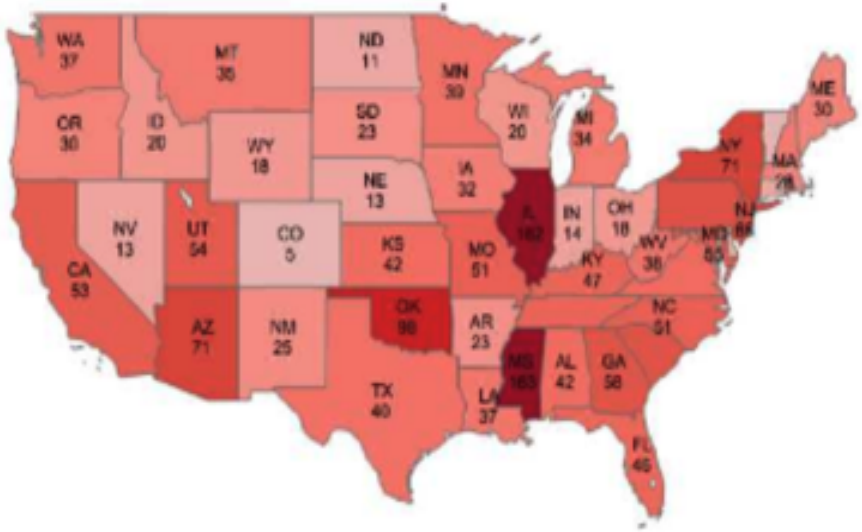


Figure 7: Introduced bills by state from ALEC model legislation

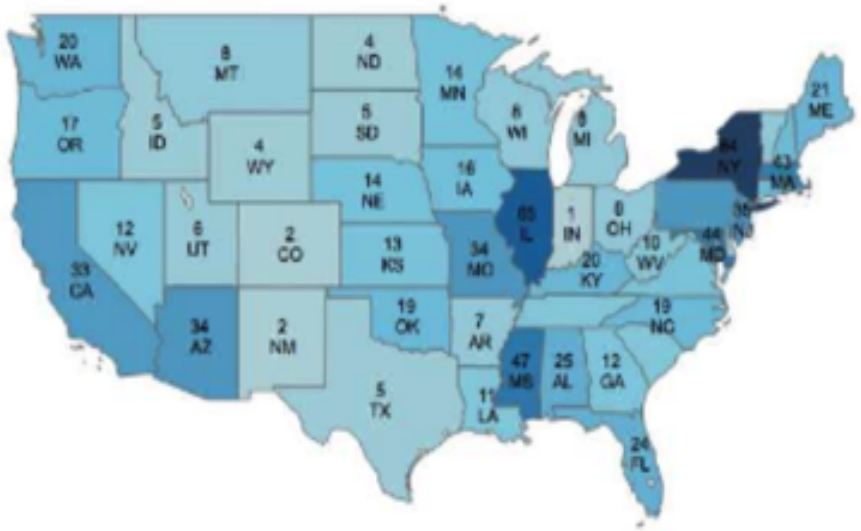
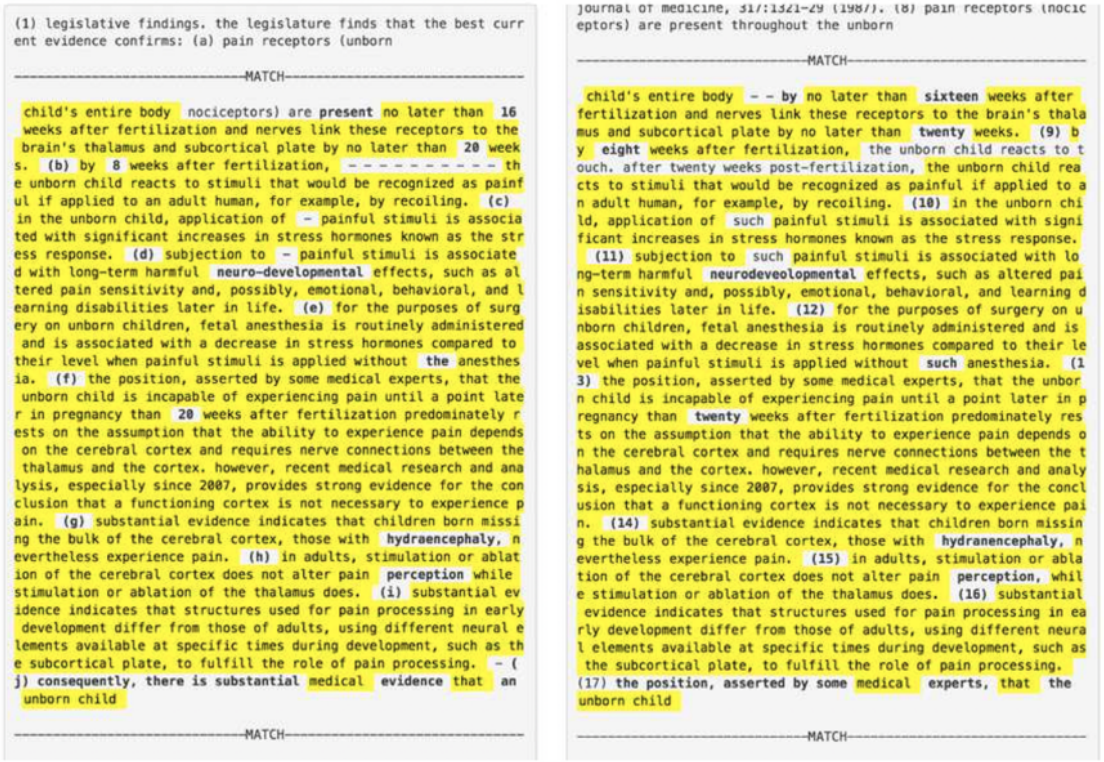


Figure 8: Introduced bills by state from ALICE model legislation

Burgess et al, “Legislative Influence Detectors”



(1) legislative findings. the legislature finds that the best curr
ent evidence confirms: (a) pain receptors (unborn

----------MATCH----------

child's entire body nociceptors) are present no later than 16
weeks after fertilization and nerves link these receptors to the
brain's thalamus and subcortical plate by no later than 20 week
s. (b) by 8 weeks after fertilization, - - - - - - - - - - th
e unborn child reacts to stimuli that would be recognized as painf
ul if applied to an adult human, for example, by recoiling. (c)
in the unborn child, application of - painful stimuli is associa
ted with significant increases in stress hormones known as the str
ess response. (d) subjection to - painful stimuli is associate
d with long-term harmful neuro-developmental effects, such as al
tered pain sensitivity and, possibly, emotional, behavioral, and l
earning disabilities later in life. (e) for the purposes of surg
ery on unborn children, fetal anesthesia is routinely administered
and is associated with a decrease in stress hormones compared to
their level when painful stimuli is applied without the anesthes
ia. (f) the position, asserted by some medical experts, that the
unborn child is incapable of experiencing pain until a point late
r in pregnancy than 20 weeks after fertilization predominately r
ests on the assumption that the ability to experience pain depends
on the cerebral cortex and requires nerve connections between the
thalamus and the cortex. however, recent medical research and ana
lysis, especially since 2007, provides strong evidence for the con
clusion that a functioning cortex is not necessary to experience p
ain. (g) substantial evidence indicates that children born missi
ng the bulk of the cerebral cortex, those with hydraencephaly, n
evertheless experience pain. (h) in adults, stimulation or ablat
ion of the cerebral cortex does not alter pain perception while
stimulation or ablation of the thalamus does. (i) substantial ev
idence indicates that structures used for pain processing in early
development differ from those of adults, using different neural e
lements available at specific times during development, such as th
e subcortical plate, to fulfill the role of pain processing. - (
j) consequently, there is substantial medical evidence that an
unborn child

----------MATCH----------

journal of medicine, 317:1321-29 (1987). (8) pain receptors (nocic
eptors) are present throughout the unborn

----------MATCH----------

child's entire body - - by no later than sixteen weeks after
fertilization and nerves link these receptors to the brain's thala
mus and subcortical plate by no later than twenty weeks. (9) b
y eight weeks after fertilization, the unborn child reacts to t
ouch. after twenty weeks post-fertilization, the unborn child rea
cts to stimuli that would be recognized as painful if applied to a
n adult human, for example, by recoiling. (10) in the unborn chi
ld, application of such painful stimuli is associated with signi
ficant increases in stress hormones known as the stress response.
(11) subjection to such painful stimuli is associated with lo
ng-term harmful neurodeveolopmental effects, such as altered pai
n sensitivity and, possibly, emotional, behavioral, and learning d
isabilities later in life. (12) for the purposes of surgery on u
nborn children, fetal anesthesia is routinely administered and is
associated with a decrease in stress hormones compared to their le
vel when painful stimuli is applied without such anesthesia. (1
3) the position, asserted by some medical experts, that the unbor
n child is incapable of experiencing pain until a point later in p
regnancy than twenty weeks after fertilization predominately res
ts on the assumption that the ability to experience pain depends o
n the cerebral cortex and requires nerve connections between the t
halamus and the cortex. however, recent medical research and analy
sis, especially since 2007, provides strong evidence for the concl
usion that a functioning cortex is not necessary to experience pai
n. (14) substantial evidence indicates that children born missin
g the bulk of the cerebral cortex, those with hydranencephaly, n
evertheless experience pain. (15) in adults, stimulation or abla
tion of the cerebral cortex does not alter pain perception, whil
e stimulation or ablation of the thalamus does. (16) substantial
evidence indicates that structures used for pain processing in ea
rly development differ from those of adults, using different neura
l elements available at specific times during development, such as
the subcortical plate, to fulfill the role of pain processing.
(17) the position, asserted by some medical experts, that the
unborn child

----------MATCH----------

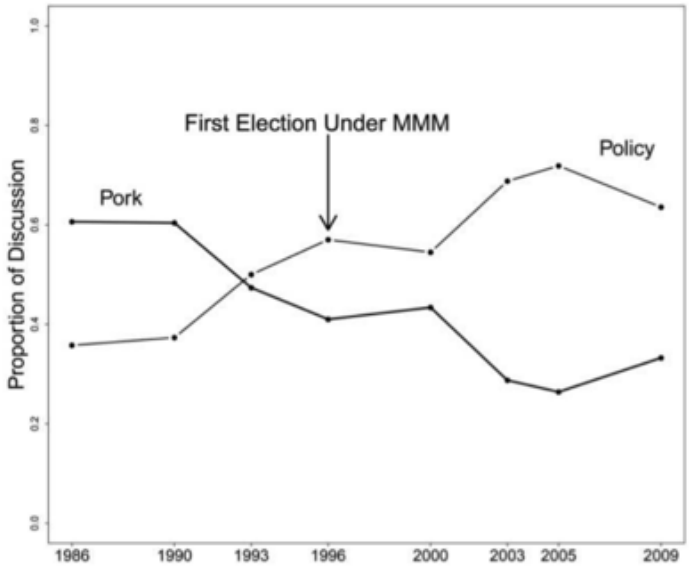
Figure 10: Match between Scott Walker’s bill and a highly similar bill from Louisiana. For a detailed view, please visit http://dssg.uchicago.edu/lid/.

Burgess et al, “Legislative Influence Detectors”

Compare bill texts across states in two-step process:

- find candidates using elasticsearch (tf-idf similarlity);
- compare candidates using text reuse score.

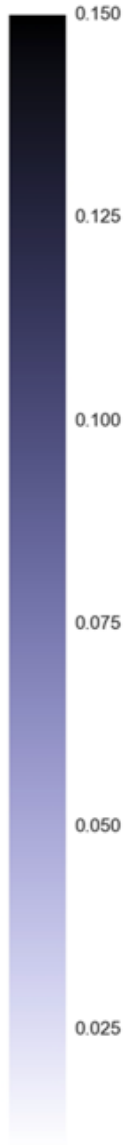
From Pork to Policy



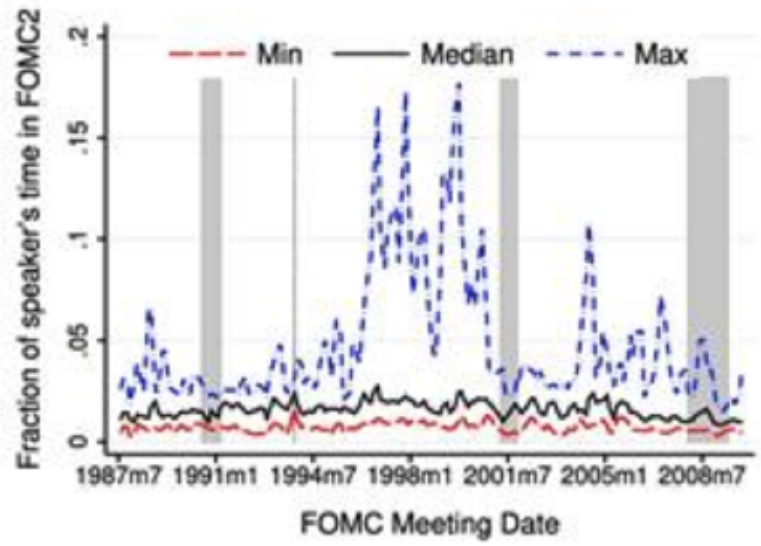
Topic modeling Federal Reserve Bank transcripts

- Analyze speech transcripts from FOMC (Federal Open Market Committee).
  - private discussions among committee members at Federal Reserve (U.S. Central Bank)
  - 150 meetings, 20 years, 26,000 speeches, 24,000 unique words.
- Pre-processing:
  - drop stopwords, stems; vocab = 10,000 words
- LDA:
  - K = 40 topics selected for interpretability / topic coherence.

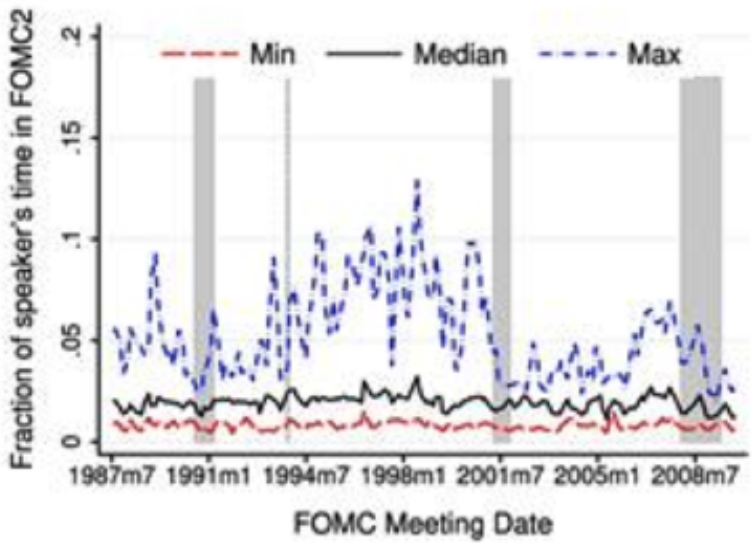
													Pro-cyclicality
Topic0[1]	product	increas	wage	price	cost	labor	rise	acceler	inflat	pressur	trend	compens	0.024
Topic1[1,2]	growth	slow	economi	continu	expans	strong	trend	inflat	will	recent	slowdown	moder	0.023
Topic2[2]	inflat	expect	core	measur	higher	path	slack	gradual	continu	remain	view	suggest	0.017
Topic3[1]	percent	year	quarter	growth	month	rate	last	next	state	averag	california	employ	0.007
Topic4	number	data	look	chang	measur	use	point	show	revis	estim	gdp	actual	0.007
Topic5[1,2]	polici	inflat	monetarpol	need	time	can	monetari	move	tighten	view	action	believ	0.005
Topic6[2]	rate	term	expect	real	lower	increas	rise	level	declin	short	nomin	year	0.005
Topic7	statement	word	chang	meet	languag	discuss	issu	want	read	sentenc	view	use	0.005
Topic8[2]	chairman	support	mr	direct	recommend	agre	asymmetr	prefer	symmetr	move	toward	favor	0.004
Topic9[1]	employ	continu	growth	job	nation	region	seem	state	manufactur	greenbook	busi	bit	0.004
Topic10	dollar	unitedstates	export	countri	import	foreign	japan	growth	abroad	trade	develop	currenc	0.003
Topic11	model	use	simul	shock	effect	scenario	nairu	differ	rule	chang	baselin	altern	0.003
Topic12[2]	risk	may	balanc	seem	side	uncertainti	possibl	economi	probabl	reason	upsid	much	0.003
Topic13	forecast	greenbook	staff	project	differ	assumpt	littl	assum	somewhat	lower	end	period	0.002
Topic14	period	committe	consist	econom	run	maintain	futur	read	slightli	stabil	expect	develop	0.002
Topic15	invest	incom	spend	capit	household	consum	busi	hous	consumpt	sector	stock	stockmarket	0.002
Topic16[1]	month	report	increas	survey	expect	indic	remain	continu	last	recent	data	activ	0.002
Topic17[1]	project	forecast	year	quarter	expect	will	percent	revis	anticip	growth	next	recent	0.002
Topic18	question	ask	issu	let	want	answer	rais	discuss	don	start	without	okay	0.001
Topic19	peopl	talk	lot	much	comment	around	differ	number	realli	look	thing	hear	0.001
Topic20	presid	ye	governor	parri	stern	vice	hoenig	minehan	kelley	jordan	moskow	mcteer	0.001
Topic21	move	can	evid	signific	stage	inde	will	issu	economi	may	quit	clearli	0.001
Topic22[2]	chairman	thank	mr	time	meet	laughter	comment	let	will	point	call	may	0.0
Topic23[1]	year	panel	line	shown	right	chart	expect	project	percent	middl	left	next	0.0
Topic24	district	nation	area	continu	sector	construct	manufactur	report	activ	region	economi	remain	0.0
Topic25	know	someth	happen	right	thing	want	look	sure	can	realli	anyth	els	0.0
Topic26[1,2]	polici	might	committe	market	may	tighten	eas	risk	action	staff	possibl	potenti	-0.001
Topic27	year	continu	product	price	level	industri	will	sale	increas	auto	last	district	-0.001
Topic28[1]	inventori	product	sale	level	order	will	sector	come	good	quarter	much	adjust	-0.001
Topic29	price	oil	increas	energi	effect	import	suppli	product	demand	will	market	oilprices	-0.002
Topic30	term	might	point	can	sens	run	short	probabl	time	longer	tri	someth	-0.002
Topic31	seem	may	time	certainli	bit	littl	quit	much	far	perhap	better	might	-0.003
Topic32	money	aggreg	borrow	seem	rang	reserv	rate	target	time	altern	suggest	million	-0.003
Topic33[2]	move	market	point	will	fundsrate	rate	basispoints	need	fed	today	basi	time	-0.004
Topic34[1]	report	busi	compani	year	contact	firm	sale	worker	expect	plan	director	industri	-0.004
Topic35	will	fiscal	ta	budget	cut	govern	effect	billion	state	spend	deficit	year	-0.005
Topic36	will	economi	world	rather	problem	believ	can	situat	much	seem	view	good	-0.008
Topic37	realli	look	side	thing	lot	problem	concern	littl	pretti	situat	kind	much	-0.012
Topic38	bank	credit	market	loan	financi	debt	lend	fund	concern	financ	problem	spread	-0.018
Topic39[1,2]	economi	weak	recoveri	recess	confid	eas	neg	econom	will	turn	declin	period	-0.059



Pro-Cyclical Topics

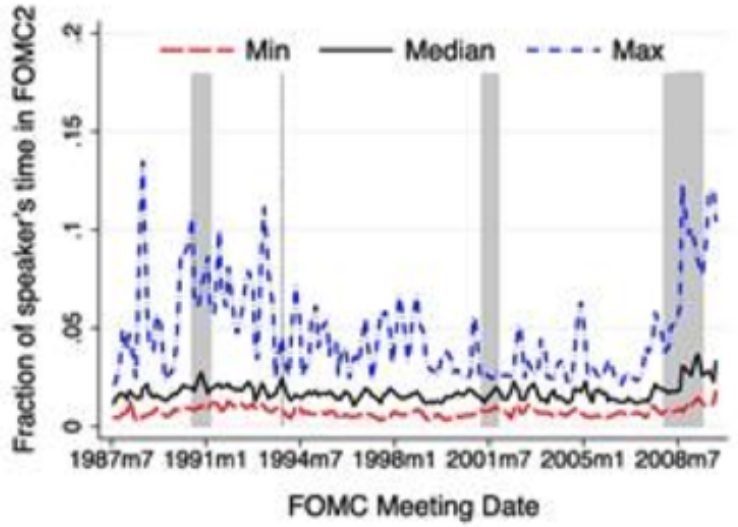


(A) Topic 0 'Productivity'

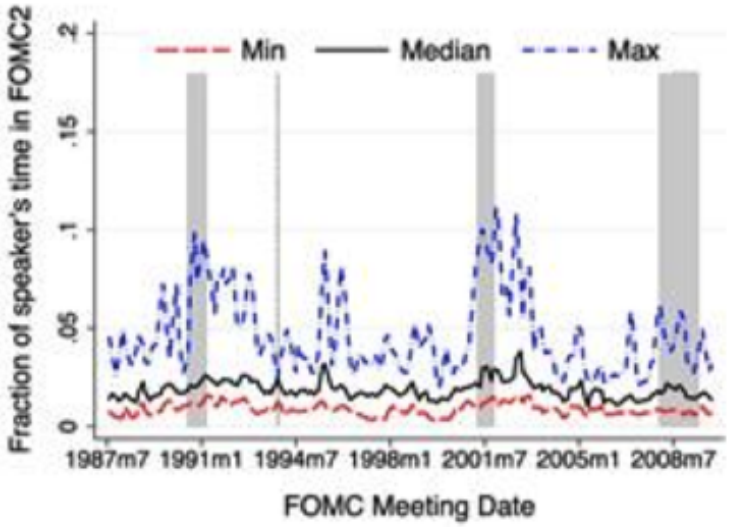


(B) Topic 1 'Growth'

Counter-Cyclical Topics



(A) Topic 38 'Financial sector'



(B) Topic 39 'Economic weakness'

Effect of Transparency

TABLE IV

SUMMARY OF COMMUNICATION MEASURES (MEETING-SECTION-SPEAKER LEVEL)

Count measures		Topic measures	
Name	Description	Name	Description
Words	The count of words spoken	Concentration	The Herfindahl index applied to distribution over policy topics
Statements	The count of statements made	Quant	Percentage of time on data topics
Questions	The count of questions asked	Avg Sim (X) $X \in \{B, D, KL\}$ B = Bhattacharyya D = dot product KL = Kullback − Leibler	The similarity between a speaker's distribution over policy topics and the FOMC average, computed using metric X
Numbers	The count of numbers spoken	Pr (no dissent)	The fitted value for no voiced dissent from the LASSO for policy topic selection (only FOMC2)

Effect of Transparency

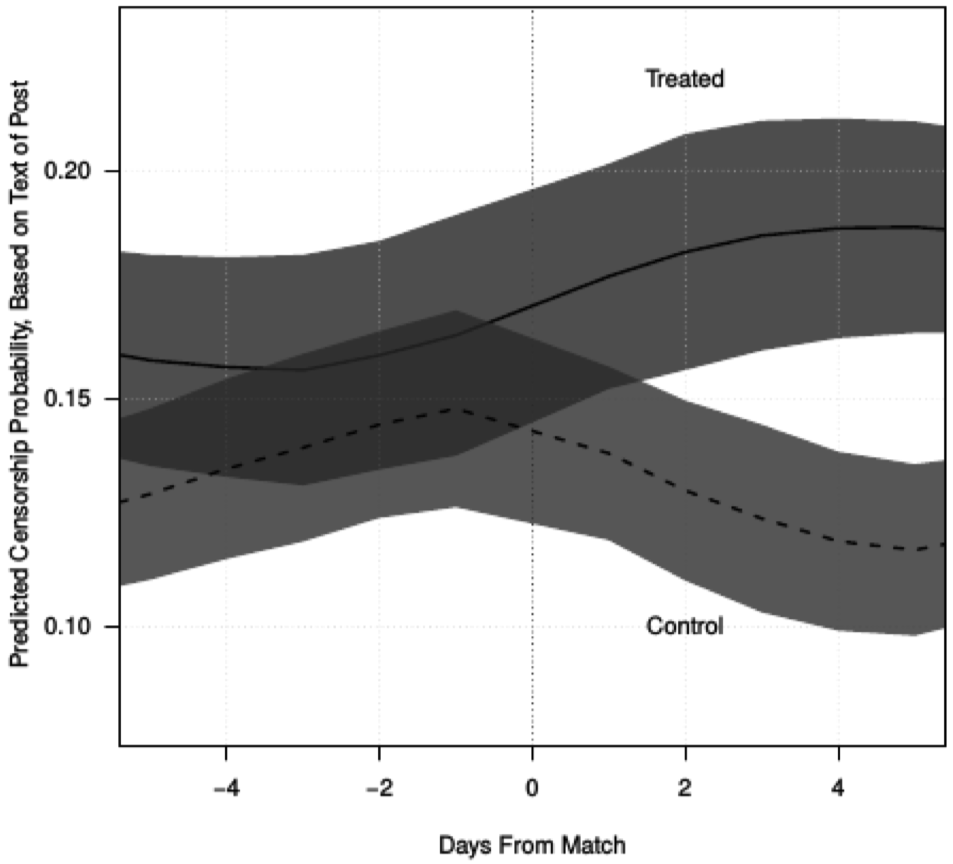
- In 1993, there was an unexpected transparency shock where transcripts became public.
- Increasing transparency results in:
  - higher discipline / technocratic language (probably beneficial)
  - higher conformity (probably costly)
- Highlights tradeoffs from transparency in bureaucratic organizations.

Text matching for causal inference: Application to online censorship in China

Roberts, Stewart, and Nielsen (2018)

- Construct a corpus of chinese social media posts, some of which are censored.
    - 593 bloggers, 150,000 posts, 6 months
- They use a variation of propensity score matching to identify almost identical posts, some of which were censored, and some of which were not.
- Outcome:
    - Using text of subsequent posts, measure how likely they are to be censored (how censorable)
    - Can see whether censorship has a deterrence or backlash effect.

Censorship has a backlash effect



- Bloggers who are censored respond with more censorable content.