Natural Language Processing in Economics

Distance and Dimensionality

Methods

Supervised Learning

- We observe the true value of a target variable for some subset of documents. We form predictions about other documents given the training set.

Unsupervised Learning

- We lack a target variable, and instead rely on the algorithm to discover relevant target variables (ie. topics)

- The distinction between the two approaches is not clear cut in text analysis
    - Supervised models can be used to discover unexpected themes
    - Unsupervised models can be used downstream to predict known variables
    - Semi-supervised models are useful to limit randomness of fully unsupervised models

Unsupervised handbook (Ash, E.)

What is the research question

Corpus and Data

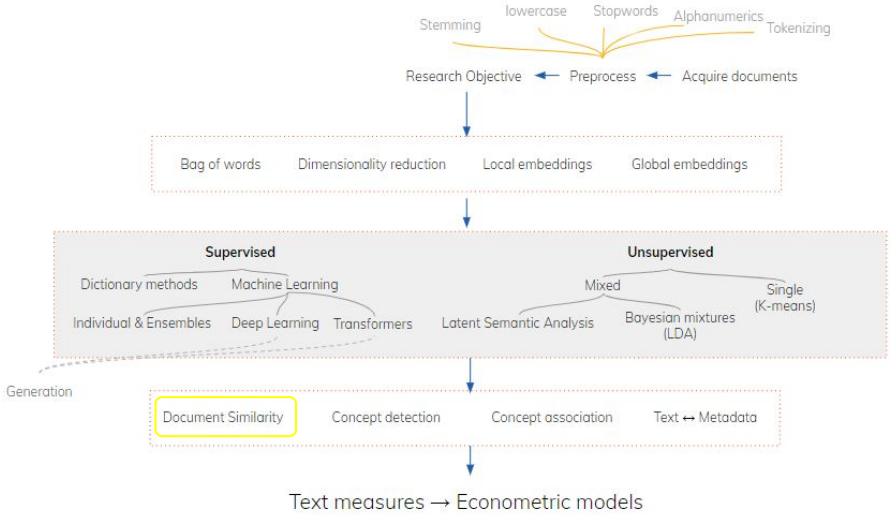
- Obtain, clean, preprocess, and link
- Produce descriptive visuals and statistics on the text and metadata

Unsupervised learning

- What are we trying to measure?
- Select a model and train it
- Probe sensitivity to hyperparameters
- Validate that the model is measuring what we want

Empirical analysis

- Produce statistics or predictions with the trained model
- Answer the research question


Stemming
lowercase
Stopwords
Alphanumerics
Tokenizing
Research Objective
Preprocess
Acquire documents
Bag of words
Dimensionality reduction
Local embeddings
Global embeddings
Supervised
Unsupervised
Dictionary methods
Machine Learning
Individual & Ensembles
Deep Learning
Transformers
Mixed
Single (K-means)
Latent Semantic Analysis
Bayesian mixtures (LDA)
Generation
Document Similarity
Concept detection
Concept association
Text ↔ Metadata
Text measures → Econometric models

Distance

Recall the document-term matrix

Input: **D**
- A sequence of tokens **W**
- A document-term matrix **C**

- A **document** $i \in \{1, ..., N\}$ becomes a finite list of features
- Each **feature** is some $j \in \{1, ..., J\}$, where J is the number of unique terms.
- The representation of your corpus D becomes a J x N matrix, where each row of **C** corresponds to the frequency distribution over words in the document corresponding to that row.

Counts and frequencies

- **Document frequency**: number of documents where a feature appears
- **Term counts**: number of total appearances of a feature in corpus D.
- **Term frequency**:

$$tf_{j,i} = \frac{c_{i,j}}{\sum_j c_{i,j}}$$

- **Inverse document frequency**:

$$idf_{i,D} = \log \frac{N}{|\{d \in D : i \in d\}|}$$

$$C = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1J} \\ c_{21} & c_{22} & \dots & c_{2J} \\ \vdots & \vdots & \vdots & \vdots \\ c_{I1} & c_{I2} & \dots & c_{IJ} \end{bmatrix}$$

Document-term matrix

A row

- Each document row c_i is a distribution over terms
- These vectors have a **spatial interpretation** → geometric distances between vectors in the **token** space reflect semantic distances between documents.

$$C = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1J} \\ c_{21} & c_{22} & \dots & c_{2J} \\ \vdots & \vdots & \vdots & \vdots \\ c_{I1} & c_{I2} & \dots & c_{IJ} \end{bmatrix}$$

A column

- Each column c□ is a distribution over documents
- These vectors also have a **spatial interpretation** → geometric distances between vectors in the **document** space reflect semantic distances between features.

$$C = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1J} \\ c_{21} & c_{22} & \dots & c_{2J} \\ \vdots & \vdots & \vdots & \vdots \\ c_{I1} & c_{I2} & \dots & c_{IJ} \end{bmatrix}$$

Cosine Similarity

$d(x,y) \geq 0 \quad d(A,B) = 0 \text{ iff } A = B \quad d(A,B) = d(B,A) \quad d(A,C) \leq d(A,B) + d(B,C)$

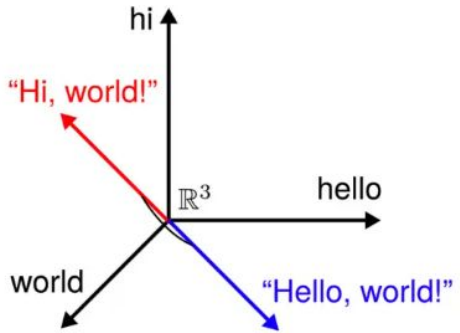
- Relatedness between individual terms or documents is commonly measured using cosine similarity.

- This is because the use of count frequency matrices or TF-IDF matrices create non-negative vectors in constant n-dimensional Euclidean spaces
  The latter implies that documents are projections, and similar documents have similar rays

- Similarity between two arbitrary document vectors **x** and **y** is then measured by the cosine of the angle between the two rays.
  - Perfectly collinear documents → cos(x,y) = 1
  - Orthogonal documents → cos(x,y) = 0

- Cosine similarity is the normalised dot product between the vectors.

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$



Paper application

Kelly et al. (2021) - Measuring technological innovation

Goal

Construct a measure of **novelty** and **impact** of innovations based on similarity and distance between patents

Methodology

- Collect 9 million patents published since 1840, from USPTO and Google Patents
- Information in a patent contains date, inventor, backward citations, and text (abstract, claims, description)
- Pre-process by removing HTML markup, punctuations, stopwords, and terms that appear in <20 patents
- Using the resulting 1.6M wide and 9M long TF matrix, compute the Backward IDF weighting of each term
- Using time-dependent similarities, identify novel and impactful technological breakthroughs

Findings

- Robust technology breakthrough metrics spanning two centuries over private and public firm activities
- Breakthrough innovations strongly correlate with long term firm profitability

Paper application

Kelly et al. (2021) - Measuring technological innovation

- Create a TF-IDF matrix $TFIDF_{pw} \equiv TF_{pw} \times IDF_{w\cdot}$, where $TF_{pw} \equiv \frac{c_{pw}}{\sum_k c_{pk}}$, and $IDF_w \equiv \log\left(\frac{\#\text{ documents in sample}}{\#\text{ documents that include term } w}\right)$
- Replace the IDF outputs by the **backward IDF**, which are defined as

$$BIDF_{wt} = \log\left(\frac{\#\text{ patents prior to } t}{1 + \#\text{ documents prior to } t \text{ that include term } w}\right)$$

- The similarities between patent pairs (i, j) are constructed as follows.
  - First, for each term **w** in **i**, we estimate the individual TF-BIDF score

$$TFBIDF_{w,i,t} = TF_{w,i} \times BIDF_{w,t}, \quad t \equiv \min(\text{filing year for i}, \text{filing year for j})$$

  - These are normalised by their Euclidean norm to have unit length, ie.

$$V_{i,t} = \frac{TFBIDF_{i,t}}{||TFBIDF_{i,t}||}.$$

  - Cosine similarities between normalised patents are estimated as $\rho_{i,j} = V_{i,t} \cdot V_{j,t}$.

Paper application

Kelly et al. (2021) - Measuring technological innovation

Novelty

Novelty is defined by the negative similarity of a patent to any patents previously granted

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

where B(j) is the set of previous patents (20Y)

Impact

Impact is defined as the similarity of subsequent patents

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

where F(j) is the set of future patents (20Y)

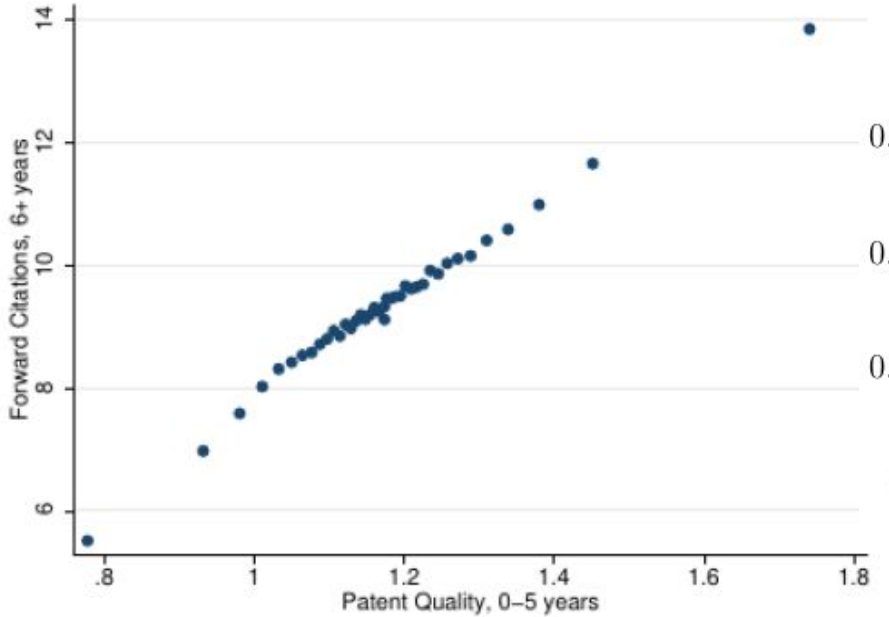
Patent Quality

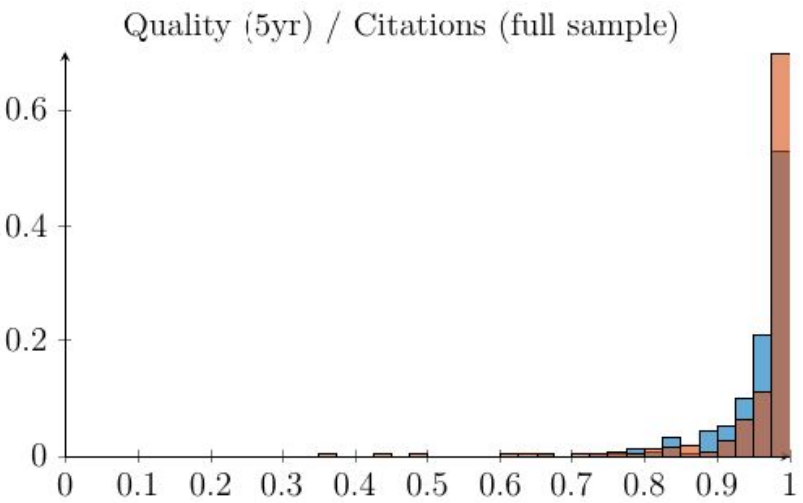
$$\text{Quality}_i = \frac{\text{Impact}_i}{-\text{Novelty}_i}$$

Paper application

Kelly et al. (2021) - Measuring technological innovation

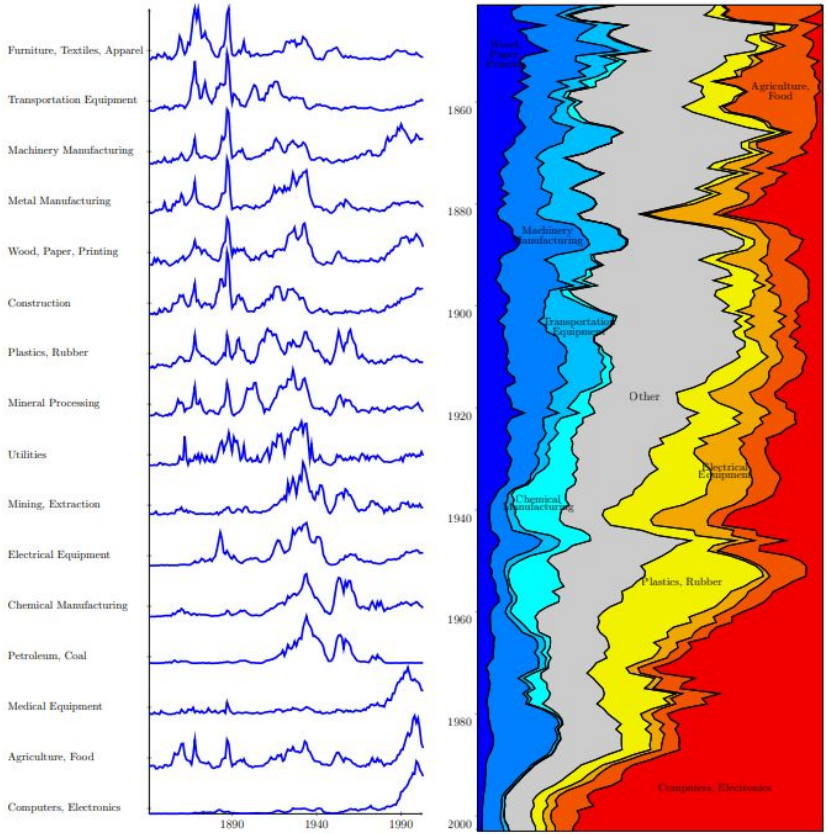
Validation

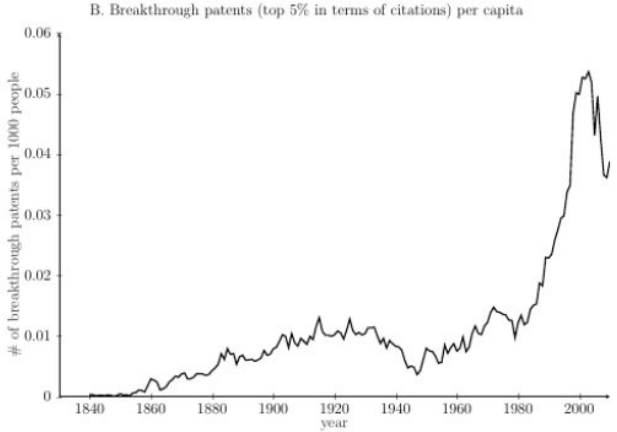


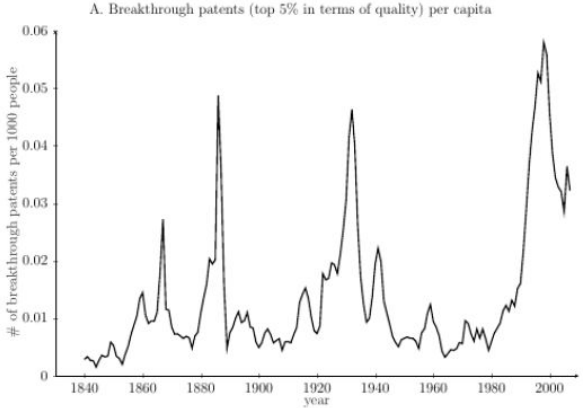


Paper application

Kelly et al. (2021) - Measuring technological innovation







Paper application

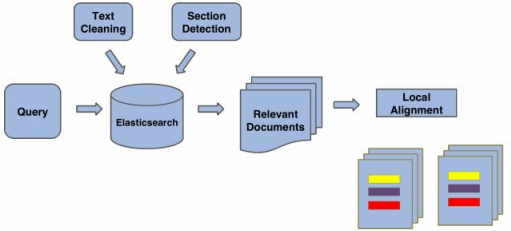
Burgess et al. (2016) - The Legislative Influence Detector

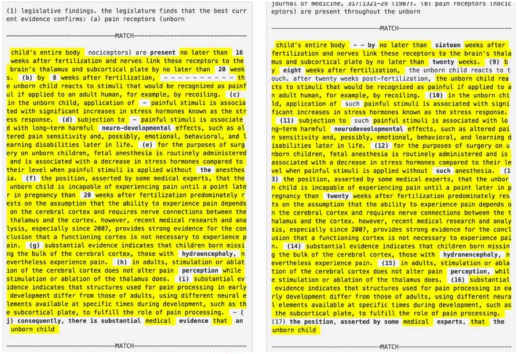
Goal

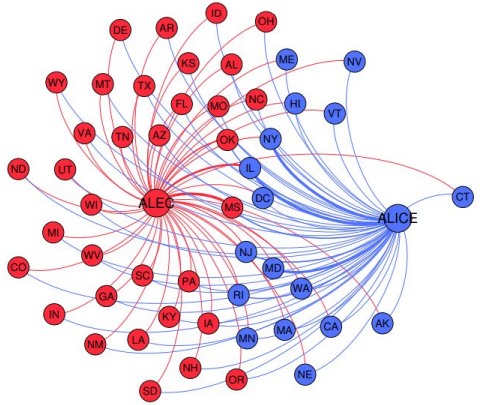
Develop efficient approaches to detect legislative text reuse and unearth partisan influential US states

Methodology

- **Search module**: use inverted index databases (elasticsearch) to identify similar bills
- **Alignment module**: use Smith-Waterman local alignment algorithm to extract parallel passages.
- **Classification module**: use a logistic regression to identify substantive text.







Paper application

Cage et al. (2019) - The Production of Information

Goal

Study how much online media outlets copy content from each other in the news production process

Methodology

- Consider all online news content produced by French media in 2013
- Identify 25,000 news events with an event-detection algorithm
- Identify first news item that breaks news about an event
- Measure “copying” by subsequent news stories related to event using a plagiarism detection algorithm
- Measure effects of copying on readership & audience

Findings

- Online copying in news production is widespread: 61.8% of content presents some form of copying
- Producing original content is rewarded with larger viewership shares

Paper application

Cage et al. (2019) - The Production of Information

Data

- **Online news content**
  - More than 2.5M French news articles published online in 2013
  - Transmedia approach: content from 86 media outlets including 1 news agency (AFP), 59 newspapers, 10 online-only media outlets, 7 radio stations, 9 TV channels
  - **Source**: French National Audiovisual Institute
- **Viewership**
  - Daily audience measures for 58 out of the 86 outlets (AFP and local newspapers missing)
  - Number of shares of the article on Facebook and Twitter. Proxy for the number of views of an article

TABLE 1
Summary statistics: articles (classified in events)

	Mean	Median	SD	Min	Max
Content					
Length (number of characters)	2,467	2,192	1,577	100	98,340
Original content (number of characters)	805	253	1,287	1	53,424
Non-original content (number of characters)	1,661	1,326	1,539	0	48,374
Originality (%)	36.5	14.5	39.8	0	100
Reactivity in hours	41.7	19.1	65.2	0	6,257
Audience					
Number of shares on Facebook	64	0	956	0	240,450
Number of shares on Facebook (winsorized)	37	0	136	0	1,017
Number of shares on Twitter	9	0	42	0	11,908
Number of shares on Twitter (winsorized)	7	0	19	0	126
Obs	851,864				

Notes: The table gives summary statistics. Year is 2013. Variables are values for the articles classified in events. The observations are at the article level. The "Number of shares on Facebook (winsorized)" variable is the version of the Facebook variable winsorized at the 99th percentile. Similarly, the "Number of shares on Twitter (winsorized)" variable is the version of the Twitter variable winsorized at the 99th percentile. Variables are described in more details in the text.

TABLE 2
Summary statistics: media outlets

	Mean	Median	SD	Min	Max
Online audience (daily)					
Number of unique visitors	248,529	107,856	384,001	3,689	2,031,580
Number of visits	340,506	156,735	543,690	4,650	2,945,172
Number of pages views	1,617,616	647,576	2,956,979	12,203	15,203,845
Audience share	1.66	0.72	2.57	0.02	13.65
Facebook (annual)					
Total number of shares	1,137,580	309,176	2,190,098	1,066	13,459,510
Twitter (annual)					
Total number of direct tweets	138,648	27,188	343,000	0	2,464,651
Total number of indirect tweets	3,627	577	8,792	0	58,507
Content (nb of characters) (annual)					
Total content not classified	32,255,744	14,999,537	114,887,872	419,234	1,065,079,616
Total content classified	19,708,659	11,580,943	23,729,089	1,114	101,246,288
Total original content	6,381,766	3,787,462	7,395,088	1,114	31,799,058
Total non-original content	13,326,893	6,860,454	19,705,976	0	76,923,528
Number of breaking news	115	54	174	0	1,011
Observations	85				

Notes: The table gives summary statistics. Year is 2013. Variables are values for media outlets (excepting the AFP and Reuters). The observations are at the media outlet/day level for the online audience statistics (first four rows) at the media outlet/year level for the total number of Facebook shares and the content data.

Paper application

Cage et al. (2019) - The Production of Information

Methodology Step 1 - Event detection

- Consider headline and text of each article and compute its TF-IDF vector representation.
- Compute the cosine similarity of each article-pair
- Iteratively aggregate articles into event-clusters if the cosine similarity is above a certain threshold (ad hoc)
- “Close” an event if no article is aggregated to it within a 24-hour window
- Drop events that contain less than 2 distinct media outlets and less than 10 articles
- Results in 25,200 news events, each lasting about 41 hours
- 33.4% of the 2.5M articles are classified into an event
- Large clusters are mostly uninformative

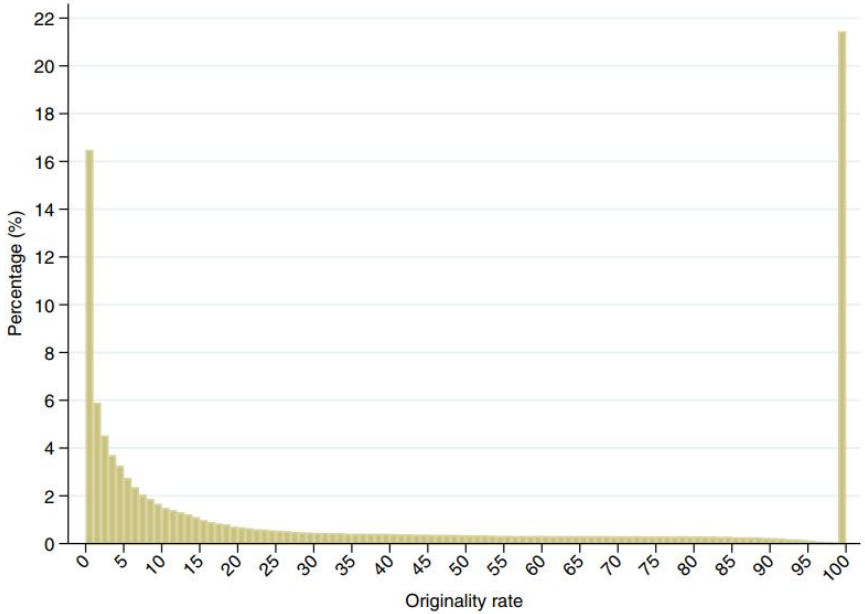
Methodology Step 2 - Plagiarism det.

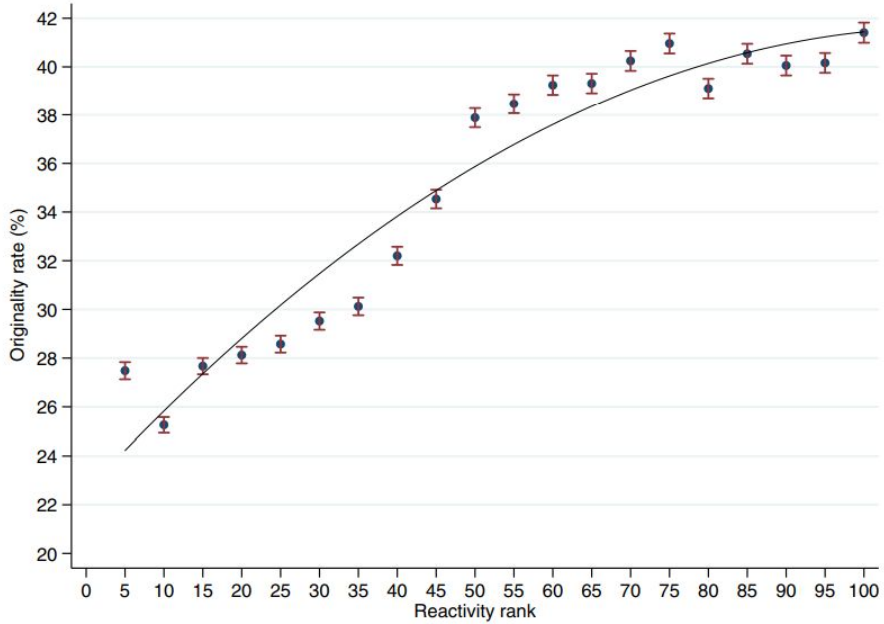
- Consider all articles within an event-cluster and order them by time. The first one is the **news-breaker**.
- Define the **reaction time** of an article as the time elapsed since publication of the news-breaker.
- Compare the text of each article to that of all the preceding articles in the event-cluster.
- If a portion of text of at least 100 hashed characters in the article is identical to any previous portion that is already published, then that portion is a **copy**.
- The **originality rate** is computed as the portion of hashes original to the event-cluster.

$$\text{originality rate} = \frac{\#\text{ original characters}}{\#\text{ total characters}}$$

Paper application

Cage et al. (2019) - The Production of Information



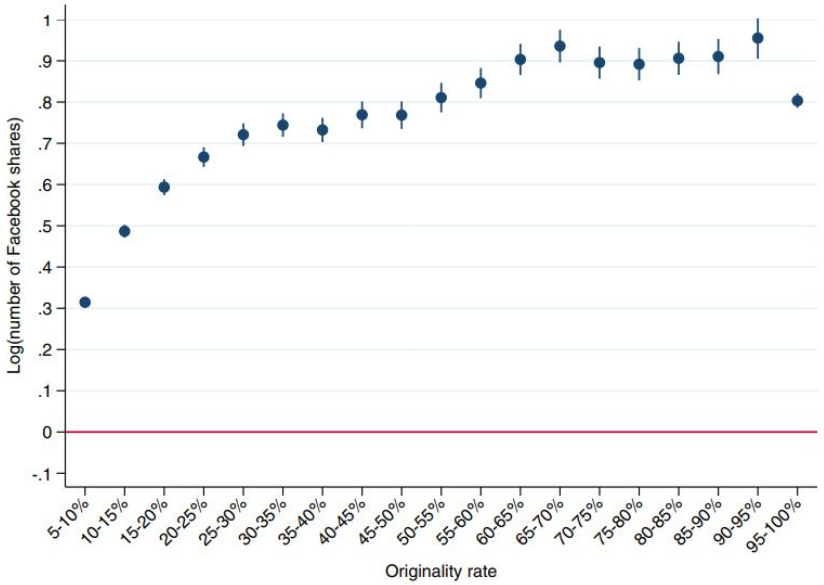


Paper application

Cage et al. (2019) - The Production of Information

Methodology Step 3 - Originality ↔ Views

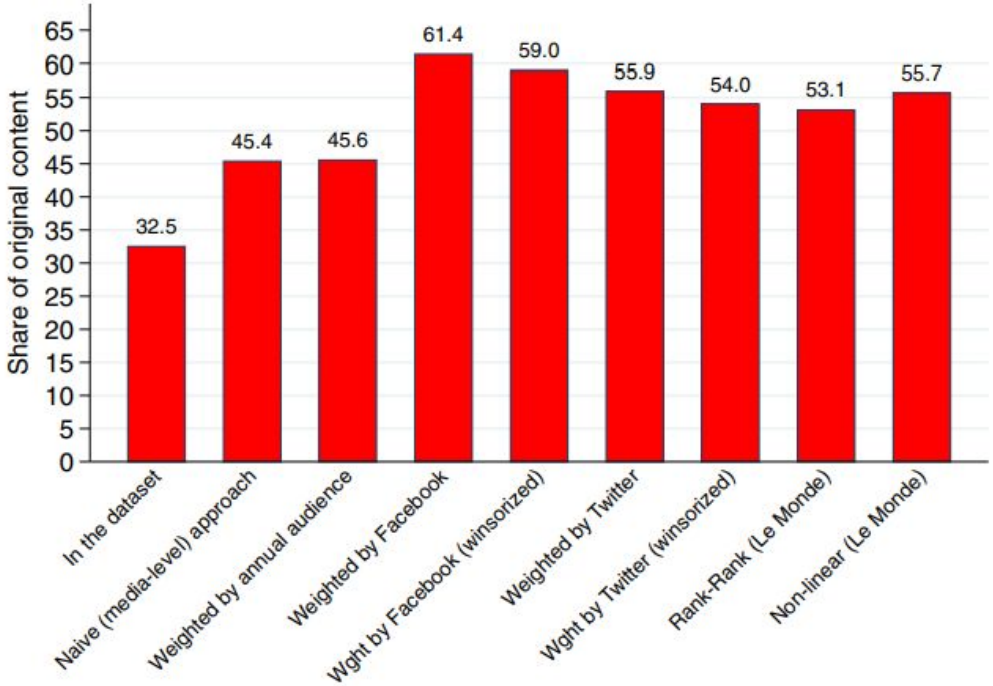
- **Naive approach**
  - Assume all articles are equally popular in a site on a given day
  - Number of views as number of total page views over the number of articles
- **Linear approach**
  - Number of article views is proportional to the relative number of shares of the article
- **Social media approach**
  - Collect data from *Le Monde* on article views from April to August 2017
  - Link the URL of the online article to SM data in order to map article views to shares
  - Use relationship to infer the number of views for other outlets

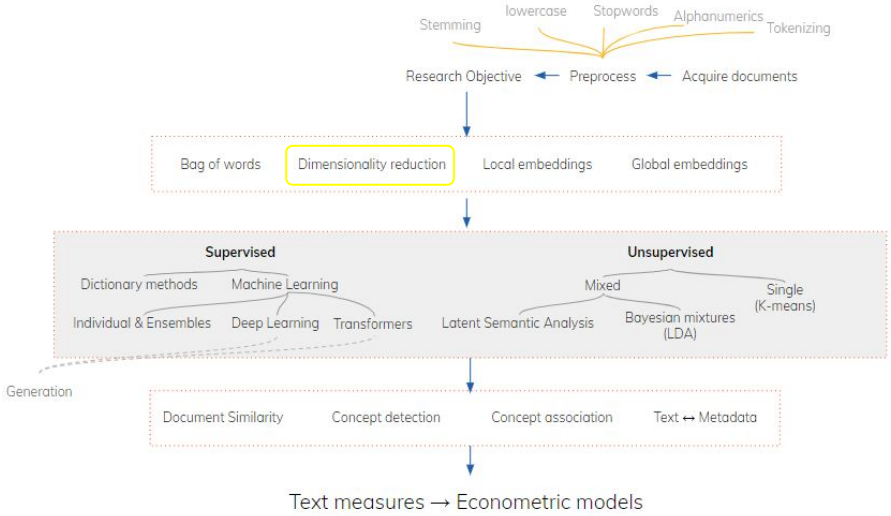


Paper application

Cage et al. (2019) - The Production of Information

$$\frac{\sum_a \text{original content}_a * \text{number of views}_a}{\sum_a \text{original content}_a * \text{number of views}_a + \sum_a \text{non-original content}_a * \text{number of views}_a},$$

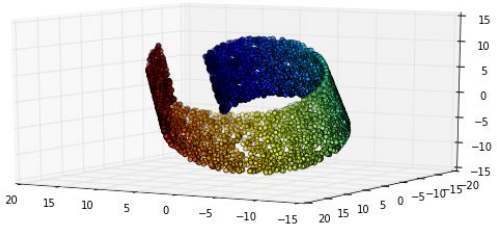




Dimensionality reduction

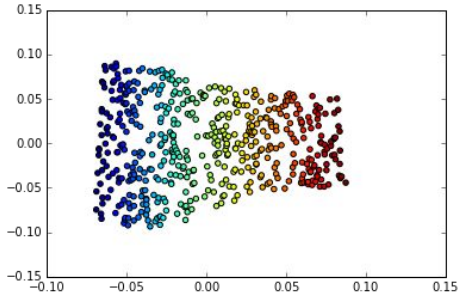
Latent spaces 🔗

- A dataset is not distributed uniformly across the feature (input) space
- They typically have a lower-dimensional latent structure that can be learned - a **manifold**



Why

- **More interpretable**: high-dimensional plots are unintuitive
- **More efficient**: data becomes more tractable
- **Better performing**: by removing statistical noise, results may improve



Dimension feature reduction

Document-term matrix

- Each **feature** column $j \in \{1, ..., J\}$ represents a word, and each row $i \in \{1, ..., N\}$ represents a document
- The resulting matrix is J x N dimensional, and this is compounded by the inclusion of n-grams (collocations)

Disjoint prior approaches

- **Pointwise mutual information**: order words by how often they collocate relative to how often they occur in isolation

$$\text{PMI}(w_1, w_2) = \log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)}$$

- **Constituency parsing**: n-grams with specific Part of Speech patterns

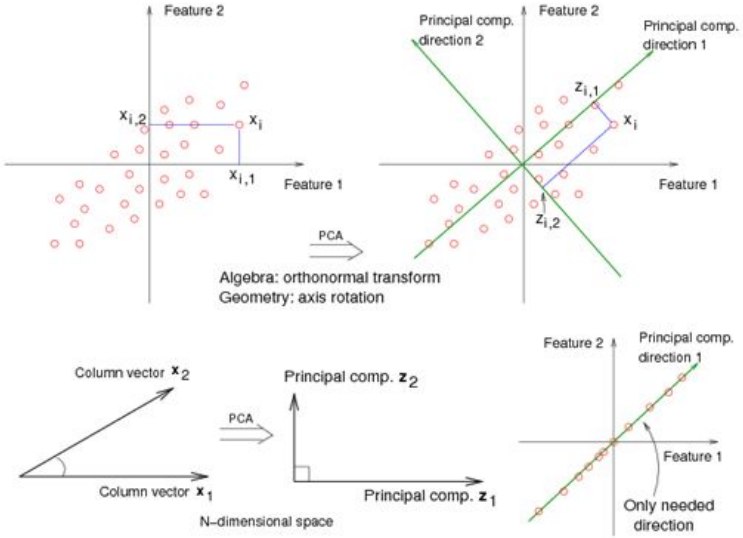
Baseline posterior approaches

- **Principal Component Analysis**: map the input space into some lower-dimensional representation while preserving high-informative dimensions.
- **K-means clustering**: use document representation of features to cluster these given their geometric similarity.

Baseline posterior approaches

Principal Component Analysis

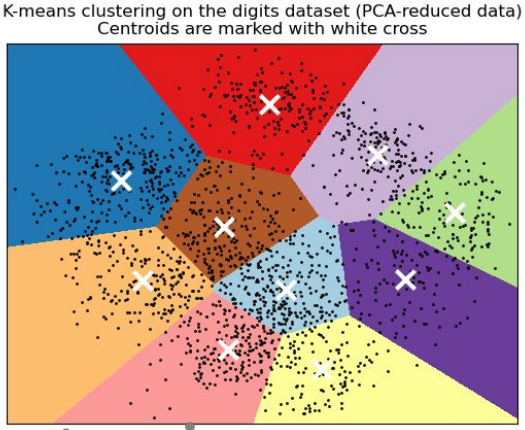
- PCA enables dimensionality reduction by projecting data onto the first principal components, reducing the number of dimensions while preserving important information.
- Distance metrics are largely preserved in the lower-dimensional space.
- The reduced matrix can serve as predictors, but performance may be subpar.
- PCA dimensions are not directly interpretable.
    - If interpretability matters, Non-negative Matrix Factorisation is an alternative to PCA.

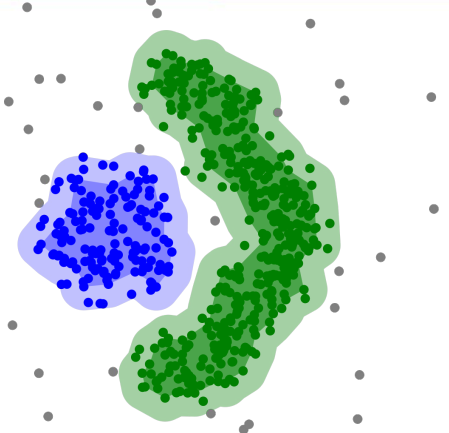


Disjoint posterior approaches

Clustering approaches

- **K-means**: initialise K cluster centroids randomly, iteratively re-assign to minimise the sum of within-cluster squared Euclidean distances. *(Elbow or Gap statistic)*
- **K-medoids**: clustering using L1 distance rather than L2 (Euclidean) distance, produces median vectors.
- **Community detection**: borrowed from network analysis, implement CD algorithms on term-document collocations
- **DBSCAN**: surrogate continuous space of feature clusters, prioritises regions of high density and discards outliers
- **Agglomerative clustering**: creates a hierarchy of clusters, a dendrogram.

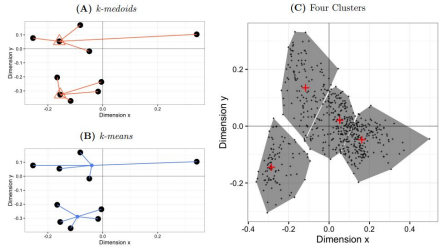




Paper application

Ganglmair, Wardlaw (2017) - Complexity, Standardisation and Loan Agreements

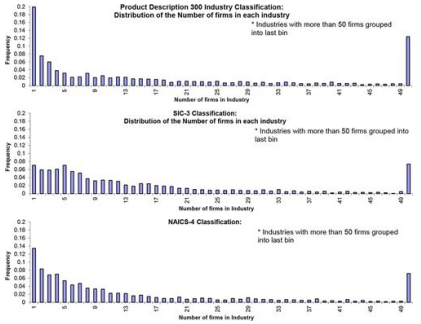
- Use k-medoids to identify different types of debt contracts, and analyse customisation
- Descriptive analysis → larger deals have more customisation and ad hoc content



Paper application

Hoberg, Phillips (2016) - Text based network industries and endog. product differentiation

- Business descriptions from annual regulatory filings
- Description vectors: per-word booleans
- Clusters of these vectors are industries - sets of firms with similar lists of nouns in their descriptions



Mixture-based dimension reduction

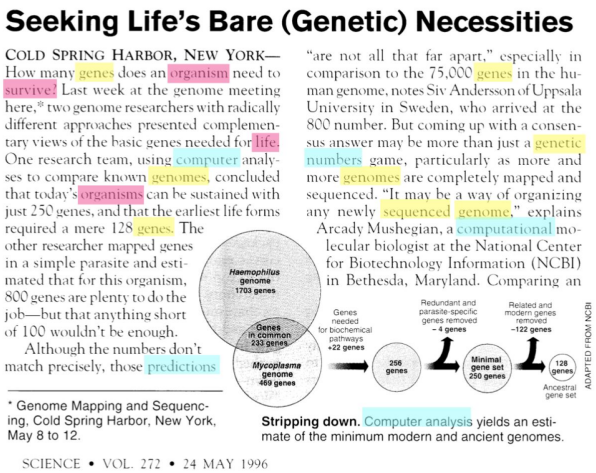
Topic models

- These models are based on the idea that documents are **mixtures of topics**, where a topic is a probability distribution over words. They act like statistical highlighters of topic associations to tokens in a text.
- A topic model is a generative model for documents: it specifies a probabilistic process by which documents are generated and populated with features.
- For many years, topic modelling has been the workhorse of advanced economics NLP applications.
    - What people talk about on social networks?
    - What themes are prevalent in the news today, and how does this predict conflict?
    - What products share similar descriptions on Amazon or EBay?
    - What are economists studying?
- Importantly, topic models are more interpretable than other dimension reduction methods, such as PCA.

LDA modelling

Latent Dirichlet allocation

- The model treats each document as a **mixture of topics**, and each topic as a **mixture of words**.
- This allows documents to overlap with each other in terms of content, rather than being separated into discrete groups, in a way that mirrors typical use of natural language.



Seeking Life's Bare (Genetic) Necessities

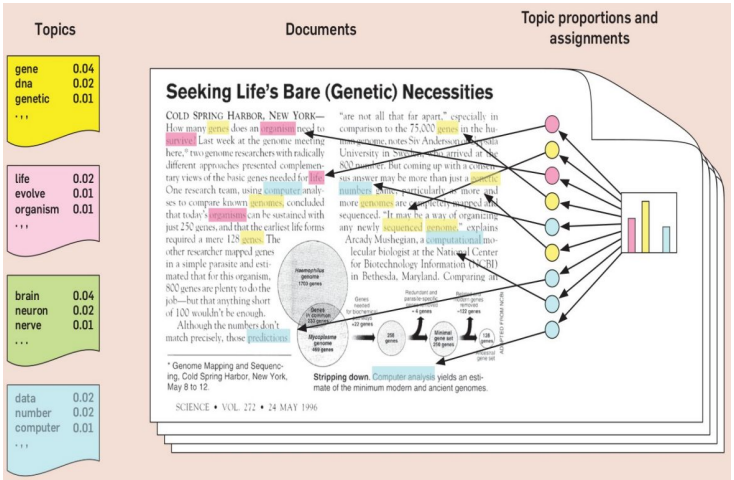
COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996



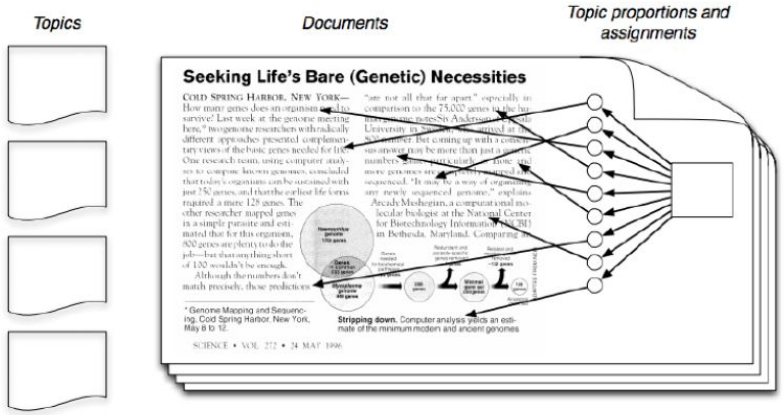
LDA modelling

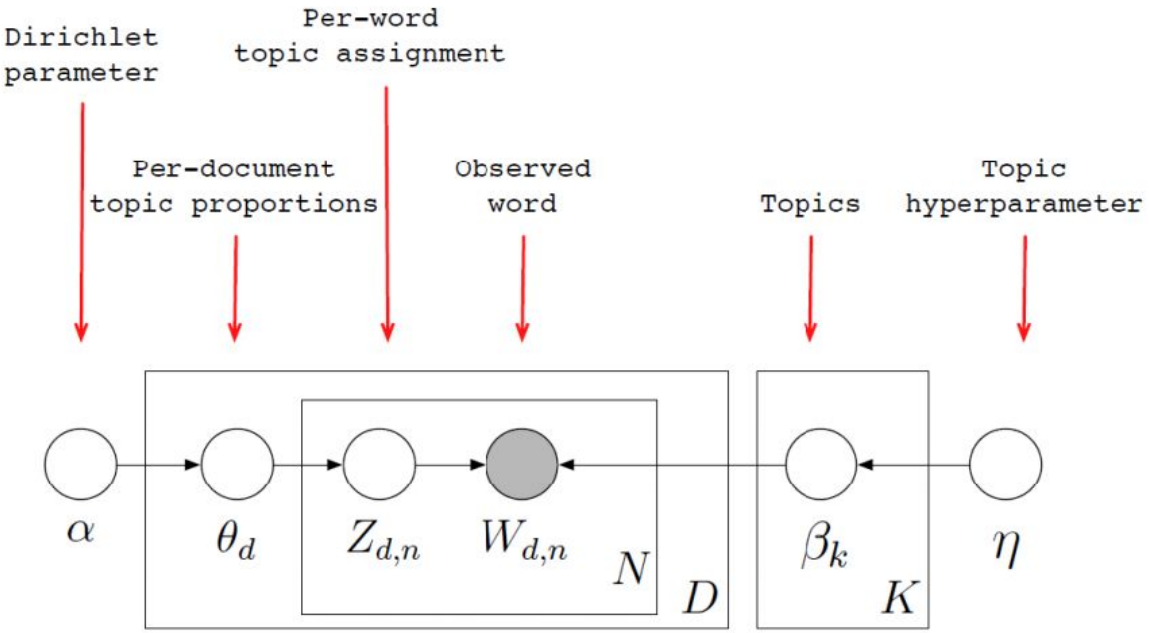
Latent Dirichlet allocation

- We do not observe the parameters of the model: we only observe N documents and J word count vectors
- Conditional on the observed documents or word counts, we want to infer the parameters of

$$\mathbf{x}_i \sim Multinomial(\beta_1\theta_{i1} + ... + \beta_k\theta ik)$$

where θ_{ik} is the weight on *k*-th **topic** for document *i* and β_k is a (1 x J) vector of **word probabilities** for k





1. For each topic $k = 1, \ldots, K$
   (a) Draw a distribution of words $\boldsymbol{\beta}_k \sim \text{Dir}_V(\boldsymbol{\eta})$
2. For each document $i = 1, \ldots, n$
   (a) Draw a random vector of topic proportions $\boldsymbol{\theta}_i \sim \text{Dir}_K(\boldsymbol{\alpha})$
   (b) For each word $j = 1, \ldots, d_i$
      (i) Draw a topic assignment $z_{ij} \sim \text{Multinomial}(\boldsymbol{\theta}_i)$, $z_{ij} \in \{1, \ldots, K\}$
      (ii) Draw a word $x_{ij} \sim \text{Multinomial}(\boldsymbol{\beta}_{z_{ij}})$, $x_{ij} \in \{1, \ldots, V\}$

LDA modelling

Latent Dirichlet allocation

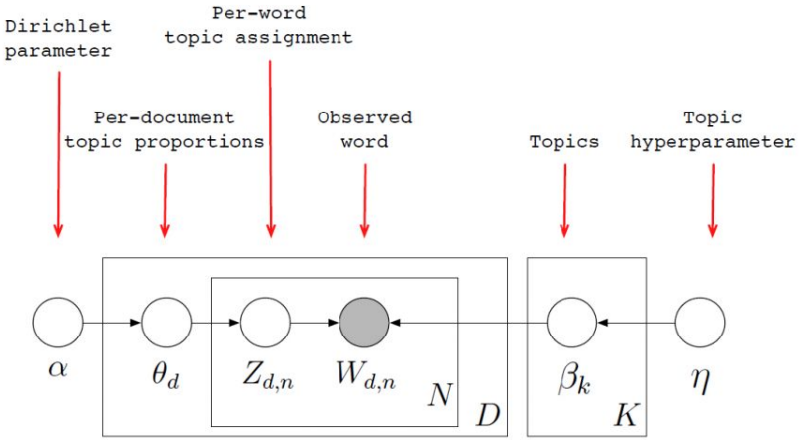
- α: document-topic density
  - higher α means documents contain more topics
- η: topic-word density
  - higher η means topics have more words
- For each word in a document-specific **N** plate, we have a variable Z which gives the topic assignment for the n-th word **placeholder** from the θ distribution.

Why Dirichlet distributed

- It is an exponential family distribution over the simplex of positive numbers that sum to one.
- A generalised version of the Beta distribution, assigns prior probabilities to all multinomial parameter vectors

$$p(\mathbf{p} \mid \alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} p_i^{\alpha_i - 1}$$

where $\sum p_i = 1$ and $p_i \geq 0$



1. For each topic $k = 1, \ldots, K$
   (a) Draw a distribution of words $\boldsymbol{\beta}_k \sim \text{Dir}_V(\boldsymbol{\eta})$
2. For each document $i = 1, \ldots, n$
   (a) Draw a random vector of topic proportions $\boldsymbol{\theta}_i \sim \text{Dir}_K(\boldsymbol{\alpha})$
   (b) For each word $j = 1, \ldots, d_i$
      (i) Draw a topic assignment $z_{ij} \sim \text{Multinomial}(\boldsymbol{\theta}_i)$, $z_{ij} \in \{1, \ldots, K\}$
      (ii) Draw a word $x_{ij} \sim \text{Multinomial}(\boldsymbol{\beta}_{z_{ij}})$, $x_{ij} \in \{1, \ldots, V\}$

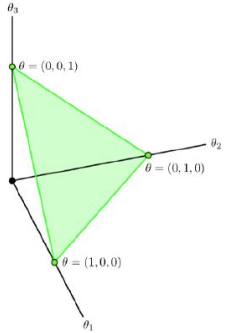
LDA modelling

Why Dirichlet distributed

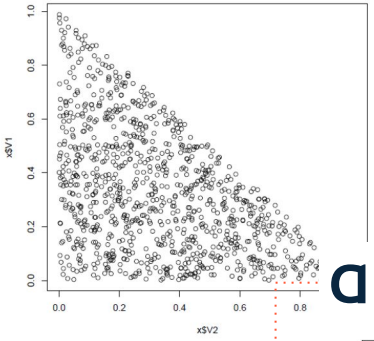
- It is an exponential family distribution over the simplex of positive numbers that sum to one.
- A generalised version of the Beta distribution, assigns prior probabilities to all multinomial parameter vectors

$$p(\mathbf{p} \mid \alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} p_i^{\alpha_i - 1}$$

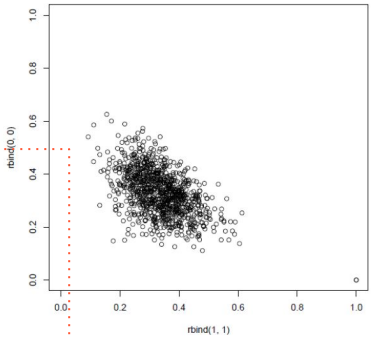
where $\sum p_i = 1$ and $p_i \geq 0$



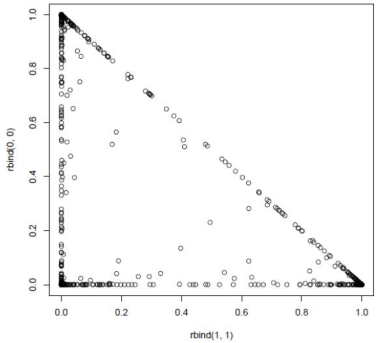
$\alpha = 1$



$\alpha = 10$



$\alpha = 0.05$



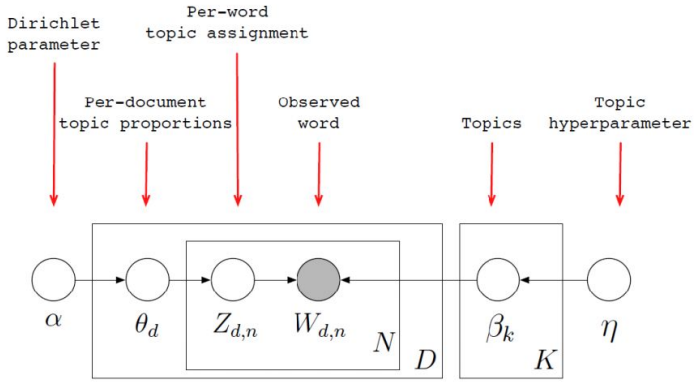
LDA modelling

Baseline LDA model

- Documents share the same topics, but each document exhibits these topics in different proportions.
- The goal of the model is to automatically discover the topics from a collection of documents given collocations
- This is possible because of the how distributions behave
  - A topic should contain as few words as possible
  - A document should contain as few topics a.p.
- The generative process corresponds to the following joint distribution of the hidden and observed variables

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{x}) = \prod_{k=1}^{K} p(\boldsymbol{\beta}_k) \prod_{i=1}^{n} p(\boldsymbol{\theta}_i) \left[\prod_{j=1}^{d_i} p(z_{ij}|\boldsymbol{\theta}_i) p(x_{ij}|z_{ij}, \boldsymbol{\beta}) \right]$$

Gibbs sampling Variational inference



Limitations & Extensions

- Probabilities are time-agnostic
- Topic probabilities are orthogonal
- Probabilities omit relevant metadata
- Word-topic assignment is fully unsupervised
- ...

LDA modelling

Correlated LDA model

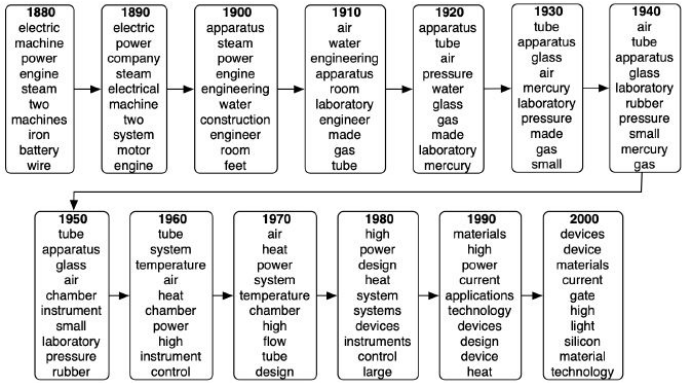
1. For each topic $k = 1, \ldots, K$
   (a) Draw a distribution of words $\boldsymbol{\beta}_k \sim \text{Dir}_V(\boldsymbol{\eta})$
2. For each document $i = 1, \ldots, n$
   (a) Draw a random vector of topic proportions $\boldsymbol{\xi}_i | \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0 \sim \mathcal{N}_K(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$
   (b) For each word $j = 1, \ldots, d_i$
      (i) Draw a topic assignment $z_{ij} \sim \text{Multinomial}(f(\boldsymbol{\xi}_i))$, $z_{ij} \in \{1, \ldots, K\}$
      (ii) Draw a word $x_{ij} \sim \text{Multinomial}(\boldsymbol{\beta}_{z_{ij}})$, $x_{ij} \in \{1, \ldots, V\}$

where

$$f(\xi_{ik}) = \frac{\exp(\xi_{ik})}{\sum_{k=1}^{K} \exp(\xi_{ik})} \qquad k = 1, \ldots, K \ ,$$

Dynamic LDA model

1. For each topic $k = 1, \ldots, K$
   (a) Draw a distribution $\boldsymbol{\psi}_{t,k} | \boldsymbol{\psi}_{t-1,k} \sim N(\boldsymbol{\psi}_{t-1,k}, \sigma I_V)$
2. For each document $i = 1, \ldots, n$
   (a) Draw a random vector of topic proportions $\boldsymbol{\theta}_i \sim \text{Dir}_K(\boldsymbol{\alpha})$
   (b) For each word $j = 1, \ldots, d_i$
      (i) Draw a topic assignment $z_{ij} \sim \text{Multinomial}(\boldsymbol{\theta}_i)$, $z_{ij} \in \{1, \ldots, K\}$
      (ii) Draw a word $x_{ij} \sim \text{Multinomial}(f(\boldsymbol{\psi}_{t,z_{ij}}))$, $x_{ij} \in \{1, \ldots, V\}$



Paper application

Hansen, McMahon, Prat (2018) - FOMC transparency

Goal

Analyse how making the FOMC meetings more **transparent** affects policymakers' incentives & their discussion

Motivation - Transparency

- Can incentivise effort and relevant contributions, a **discipline** effect
- Can discourage broad and creative debate, a **conformity** effect

Idea

- Use changes in the Federal Reserve's disclosure policy and transcripts
- Exploit dynamics before and after the change and identify variation in communication patterns

Results

- **Economic situation discussion**: inexperienced members show increased **diversity** and **quantitative** focus
- **Monetary policy strategy discussion:** inexperienced members engage in **conformity**. The former effect dom.

Paper application

Hansen, McMahon, Prat (2018) - FOMC transparency

Context

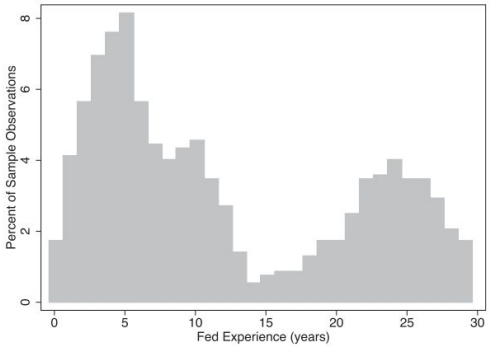
- The FOMC meets eight times a year to draft monetary policy and lay out the Federal Reserve policies
- Authors exploit as a **natural experiment** a change in the Fed's FOMC meetings disclosure policy
  - **1970-1993**: members thought that their debates were not recorded. Greenspan reveals in 1993 that meetings had been transcribed and stored
  - **1993**: following the revelation, the Federal Reserve agrees to publish all past transcripts and to release any transcripts henceforth with a five-year lag

Empirical design

Compare experts of varying experience pre- & post-1993

Hypothesis

Rookie FOMC members react stronger due to **career concerns**



Paper application

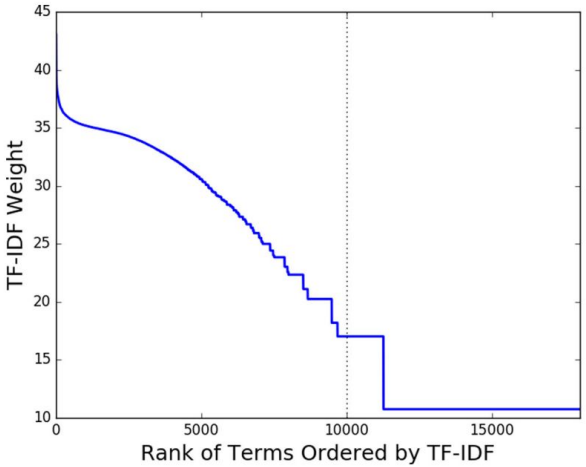
Hansen, McMahon, Prat (2018) - FOMC transparency

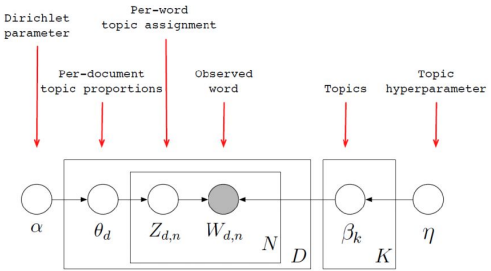
Data

- 26,645 recorded member statements
- 24,314 unique vocabulary words
- 8,206 unique terms after pre-process

LDA model

- Set **K=40**, each with a probability vector β over all 8,206 tokens
- A document **d** has its **own distribution** over topics given by θ, which is **K** dimensional
- Using Dirichlet priors on β and θ, estimate the posterior distributions with Gibbs sampling
  (Griffiths, Steyvers, 2004)





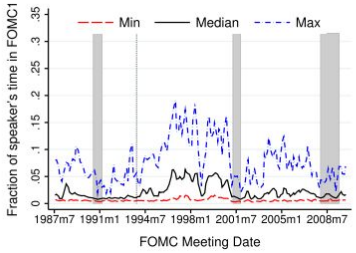
Paper application

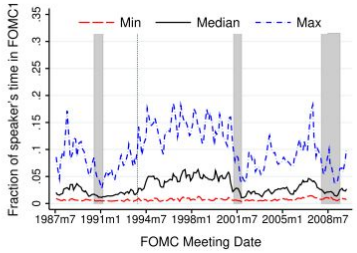
Hansen, McMahon, Prat (2018) - FOMC transparency

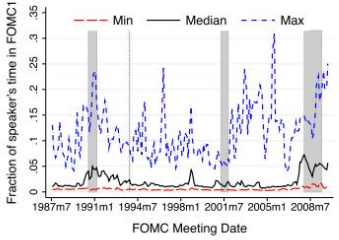
													Pro-cyclicality
Topic0[1]	product	increas	wage	price	cost	labor	rise	acceler	inflat	pressur	trend	compens	0.024
Topic1[1,2]	growth	slow	economi	continu	expans	strong	trend	inflat	will	recent	slowdown	moder	0.023
Topic2[2]	inflat	expect	core	measur	higher	path	slack	gradual	continu	remain	view	suggest	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

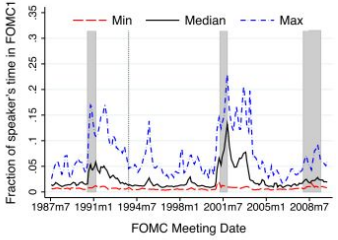
Paper application

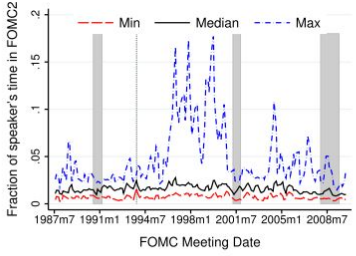
Hansen, McMahon, Prat (2018) - FOMC transparency



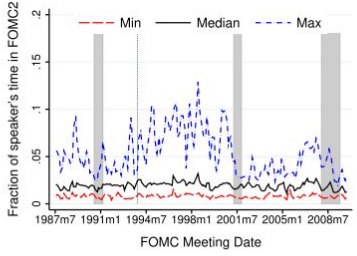




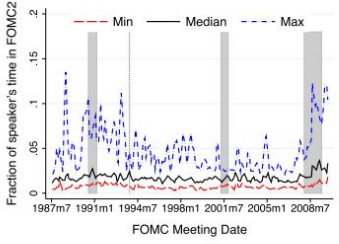




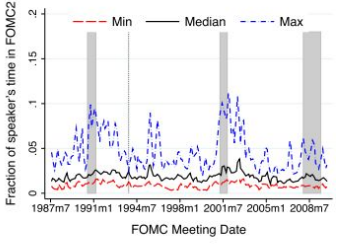
(A) Topic 0 'Productivity'



(B) Topic 1 'Growth'



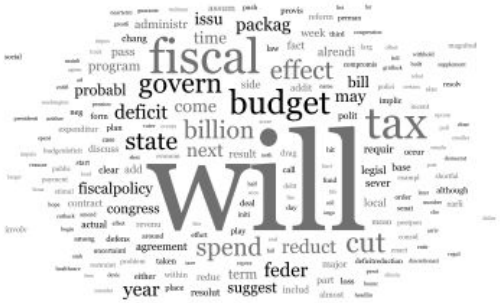
(A) Topic 38 'Financial sector'



(B) Topic 39 'Economic weakness'

Paper application

Hansen, McMahon, Prat (2018) - FOMC transparency



(A) Topic 35 'Fiscal issues'

(B) Topic 12 'Risk'

Paper application

Hansen, McMahon, Prat (2018) - FOMC 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)

Paper application

$$y_{it} = \alpha_i + \gamma D(Trans)_t + \lambda X_t + \varepsilon_{it},$$

Hansen, McMahon, Prat (2018) - FOMC transparency

DIFFERENCE RESULTS FOR ECONOMIC SITUATION DISCUSSION (FOMC1): COUNT MEASURES

Main regressors	Words (1)	Statements (2)	Questions (3)	Numbers (4)
D(Trans)	56.7*	−0.52	−0.039	3.71***
	[.076]	[.162]	[.659]	[.003]
D(Recession)	−1.95	−0.69	−0.19	−0.71
	[.952]	[.159]	[.314]	[.488]
EPU index	0.30	−0.00094	0.00088	0.0040
	[.186]	[.876]	[.586]	[.520]
D(2 day)	27.1	1.36*	0.56*	1.28
	[.256]	[.085]	[.051]	[.188]
# of PhDs	6.68	−0.45***	−0.11***	0.51
	[.561]	[.005]	[.009]	[.109]
Constant	528***	10.0***	2.44***	1.50
	[.002]	[.000]	[.000]	[.740]
Unique members	19	19	19	19
Observations	903	903	903	903
Member FE	Yes	Yes	Yes	Yes
Time FE	No	No	No	No
Meeting section	FOMC1	FOMC1	FOMC1	FOMC1
Transparency effect	9.5*	−10	−2.5	53.2***

Notes. This table reports the results of estimating (DIFF) on FOMC member statements from the economic situation discussion. Dependent variable definitions are in Table IV. Coefficients are labeled according to significance (***$p < .01$, **$p < .05$, *$p < .1$) while brackets below coefficients report p-values calculated using Driscoll-Kraay standard errors. The transparency effect reports the estimated coefficient on $D(Trans)$ as a percentage of the average value of the dependent variable before November 1993. These effects carry the same star labels as the corresponding estimated coefficient on $D(Trans)$.

DIFFERENCE RESULTS FOR ECONOMIC SITUATION DISCUSSION (FOMC1): TOPIC MEASURES

Main regressors	Concentration (1)	Quant (2)	Avg Sim (B) (3)	Avg Sim (D) (4)	Avg Sim (KL) (5)
D(Trans)	0.0041	−0.00027	0.0082***	0.0012	0.032***
	[.205]	[.831]	[.001]	[.692]	[.000]
D(Recession)	0.0061**	−0.000056	0.0020	0.015***	−0.0017
	[.028]	[.968]	[.385]	[.000]	[.758]
EPU index	3.7e-06	−9.6e-06	0.000050*	0.000029	0.00015
	[.890]	[.541]	[.077]	[.300]	[.109]
D(2 day)	−0.0040*	0.0042**	0.00044	−0.0037***	0.00051
	[.093]	[.024]	[.802]	[.001]	[.914]
# of PhDs	0.0017	−0.00063	0.000097	0.00079	0.00018
	[.255]	[.292]	[.885]	[.671]	[.928]
# Stems	0.000075***	8.8e-06**	−3.5e-06	0.000030***	0.000049
	[.000]	[.049]	[.837]	[.001]	[.284]
Constant	0.13***	0.037***	0.89***	0.084***	0.62***
	[.000]	[.000]	[.000]	[.001]	[.000]
Unique members	19	19	19	19	19
Observation	903	903	903	903	903
Member FE	Yes	Yes	Yes	Yes	Yes
Time FE	No	No	No	No	No
Meeting section	FOMC1	FOMC1	FOMC1	FOMC1	FOMC1
Topics	P1	T4 & T23	P1	P1	P1
Similarity measure	—	—	Bhatta-charyya	Dot product	Kullback-Leibler
Transparency effect	2.5	−0.7	0.9***	1.1	4.9***

Notes. This table reports the results of estimating (DIFF) on FOMC member statements from the economic situation discussion. Dependent variable definitions are in Table IV. Coefficients are labeled according to significance (***$p < .01$, **$p < .05$, *$p < .1$) while brackets below coefficients report p-values calculated using Driscoll-Kraay standard errors. The transparency effect reports the estimated coefficient on $D(Trans)$ as a percentage of the average value of the dependent variable before November 1993. These effects carry the same star labels as the corresponding estimated coefficient on $D(Trans)$.

Paper application

$$y_{it} = \alpha_i + \gamma D(Trans)_t + \lambda X_t + \varepsilon_{it},$$

Hansen, McMahon, Prat (2018) - FOMC transparency

DIFFERENCE RESULTS FOR POLICY STRATEGY DISCUSSION (FOMC2): COUNT MEASURES

Main regressors	Words (1)	Statements (2)	Questions (3)	Numbers (4)
D(Trans)	92.1**	−0.99***	−0.41**	1.86***
	[.019]	[.007]	[.012]	[.000]
D(Recession)	23.4	1.58***	0.17	−0.34
	[.560]	[.004]	[.356]	[.692]
EPU index	0.34	−0.0025	−0.0027***	0.0031
	[.134]	[.341]	[.004]	[.468]
D(2 day)	48.9	0.45	0.19	0.92
	[.226]	[.251]	[.133]	[.153]
# of PhDs	7.26	0.16	0.039	−0.37
	[.766]	[.560]	[.587]	[.489]
Constant	143	2.76	0.81	5.78
	[.638]	[.416]	[.312]	[.376]
Unique members	19	19	19	19
Observation	895	895	895	895
Member FE	Yes	Yes	Yes	Yes
Time FE	No	No	No	No
Meeting section	FOMC2	FOMC2	FOMC2	FOMC2
Transparency effect	29.9**	−15.7***	−29.4**	44.6***

Notes. This table reports the results of estimating (DIFF) on FOMC member statements from the monetary policy strategy discussion. Dependent variable definitions are in Table IV. Coefficients are labeled according to significance (***$p < .01$, **$p < .05$, *$p < .1$) while brackets below coefficients report p-values calculated using Driscoll-Kraay standard errors. The transparency effect reports the estimated coefficient on $D(Trans)$ as a percentage of the average value of the dependent variable before November 1993. These effects carry the same star labels as the corresponding estimated coefficient on $D(Trans)$.

TABLE VIII

DIFFERENCE RESULTS FOR POLICY STRATEGY DISCUSSION (FOMC2): TOPIC MEASURES

Main regressors	Concentration (1)	Quant (2)	Avg Sim (B) (3)	Avg Sim (D) (4)	Avg Sim (KL) (5)	Pr (No dissent) (6)
D(Trans)	0.0048*	−0.00045	−0.00079	−0.013***	0.0074	−0.010
	[.097]	[.681]	[.805]	[.000]	[.473]	[.613]
D(Recession)	−0.0055*	0.00016	0.0022	−0.0080**	0.0032	−0.0028
	[.090]	[.908]	[.323]	[.049]	[.636]	[.750]
EPU index	0.000068	−0.000033**	0.000018	−0.000015	0.000097	0.00026**
	[.107]	[.016]	[.605]	[.741]	[.371]	[.012]
D(2 day)	0.0083**	0.00031	−0.0013	0.0017	−0.0032	0.0025
	[.016]	[.701]	[.690]	[.721]	[.786]	[.742]
# of PhDs	−0.0042**	0.0013***	−0.0017	−0.0054***	−0.0058	0.00044
	[.022]	[.007]	[.127]	[.000]	[.113]	[.896]
# Stems	0.000058***	3.3e-06	0.000028**	8.6e-06	0.00012***	−0.00015***
	[.000]	[.805]	[.013]	[.335]	[.001]	[.000]
Constant	0.21***	0.028***	0.94***	0.21***	0.77***	0.82***
	[.000]	[.000]	[.000]	[.000]	[.000]	[.000]
Unique members	19	19	19	19	19	19
Observation	893	893	893	893	893	893
Member FE	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	No	No	No	No	No	No
Meeting section	FOMC2	FOMC2	FOMC2	FOMC2	FOMC2	FOMC2
Topics	P2	T4 & T23	P2	P2	P2	P2
Similarity measure	—	—	Bhatta-charyya	Dot product	Kullback-Leibler	—
Transparency effect	2.6*	−1.2	−0.1	−8.8***	1	−1.3

Notes. This table reports the results of estimating (DIFF) on FOMC member statements from the monetary policy strategy discussion. Dependent variable definitions are in Table IV. Coefficients are labeled according to significance (***$p < .01$, **$p < .05$, *$p < .1$) while brackets below coefficients report p-values calculated using Driscoll-Kraay standard errors. The transparency effect reports the estimated coefficient on $D(Trans)$ as a percentage of the average value of the dependent variable before November 1993. These effects carry the same star labels as the corresponding estimated coefficient on $D(Trans)$.

$$(DinD)\; y_{it} = \alpha_i + \delta_t + \eta FedExp_{it} + \phi D(Trans)_t \times FedExp_{it} + \epsilon_{it},$$

Paper application

Hansen, McMahon, Prat (2018) - FOMC transparency

DIFFERENCE-IN-DIFFERENCES RESULTS FOR ECONOMIC SITUATION DISCUSSION (FOMC1): COUNT MEASURES

Main regressors	Words (1)	Statements (2)	Questions (3)	Numbers (4)
D(Trans) × Fed experience	−0.18	0.015	0.0023	−0.21***
	[.912]	[.586]	[.863]	[.000]
Fed experience	1,492***	4.52*	2.29	29.2***
	[.000]	[.069]	[.344]	[.001]
Observations	920	920	920	920
Unique members	19	19	19	19
Member FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes
Meeting section	FOMC1	FOMC1	FOMC1	FOMC1
Rookie effect	0.5	−6.4	−3.3	48.1***

Notes. This table reports the results of estimating (DinD) on FOMC member statements from the economic situation discussion. Dependent variable definitions are in Table IV. Coefficients are labeled according to significance (***$p < .01$, **$p < .05$, *$p < .1$) while brackets below coefficients report p-values calculated using Driscoll-Kraay standard errors. The rookie effect reports the estimated coefficient on $D(Trans)_t \times FedExp_{it}$ multiplied by 20 (approximate difference in experience between the two modes in Figure VI) as a percentage of the average value of the dependent variable before November 1993. These effects carry the same star labels as the corresponding estimated coefficient on $D(Trans)_t \times FedExp_{it}$.

DIFFERENCE-IN-DIFFERENCES RESULTS FOR POLICY STRATEGY DISCUSSION (FOMC2): TOPIC MEASURES

Main regressors	Concentration (1)	Quant (2)	Avg Sim (B) (3)	Avg Sim (D) (4)	Avg Sim (KL) (5)	Pr (No Dissent) (6)
D(Trans) × Fed experience	−0.00077**	−0.00011	−0.00019	−0.00041***	−0.00040	−0.0015**
	[.014]	[.323]	[.222]	[.006]	[.377]	[.025]
Fed experience	−0.21***	−0.0035	−0.057	−0.11***	−0.22**	−0.41**
	[.000]	[.911]	[.140]	[.006]	[.045]	[.031]
# Stems	0.000023**	0.000018	0.000015**	0.000017***	0.000070***	−0.00011***
	[.048]	[.127]	[.030]	[.000]	[.001]	[.000]
Observations	910	910	910	910	910	910
Unique members	19	19	19	19	19	19
Member FE	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes
Meeting section	FOMC2	FOMC2	FOMC2	FOMC2	FOMC2	FOMC2
Topics	P2	T4 & T23	P2	P2	P2	P2
Similarity measure	—	—	Bhatta-charyya	Dot product	Kullback-Leibler	—
Rookie effect	8.9**	5.6	0.4	5.5***	1.1	3.5**

Notes. This table reports the results of estimating (DinD) on FOMC member statements from the monetary policy strategy discussion. Dependent variable definitions are in Table IV. Coefficients are labeled according to significance (***$p < .01$, **$p < .05$, *$p < .1$) while brackets below coefficients report p-values calculated using Driscoll-Kraay standard errors. The rookie effect reports the estimated coefficient on $D(Trans)_t \times FedExp_{it}$ multiplied by 20 (approximate difference in experience between the two modes in Figure VI) as a percentage of the average value of the dependent variable before November 1993. These effects carry the same star labels as the corresponding estimated coefficient on $D(Trans)_t \times FedExp_{it}$.

Paper application

Bandiera et al. (2020) - CEO Behavior & Firm Performance

Goal

Estimate differences in CEO behavior using high-frequency, high-dimensional diary data. Link to performance.

Idea

- Generate a one-dimensional behavior index that represents each CEO as a combination of underlying behavioral features.
- Merge the CEO behavior index with firm balance-sheet data to study the correlation between CEO traits and performance
- A statistical model of CEO-firm assignment is used to gauge pair fit and identify mismatches, as well as alleviate causality reversal concerns

Results

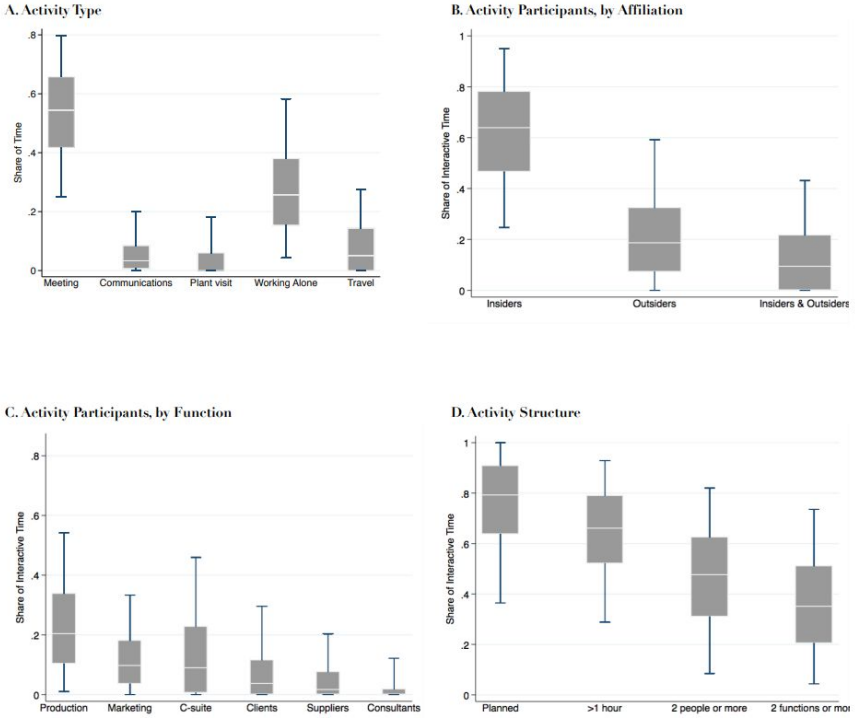
- Two flavors of CEOs: “**managers**”, and “**leaders**”.
- 17% of firms end up with the wrong type of CEO, particularly in lower-income countries.

Paper application

Bandiera et al. (2020) - CEO Behavior & Firm Performance

Survey data

- Time allocation survey data on all activities performed in 15-minute time blocks for one week by 1,1114 CEOs. Activity types are classified along five different dimensions.
  - type (meeting, public event, …)
  - duration (15 min, 30 min, …)
  - planning (planned or unplanned)
  - number of participants (1, +1)
  - functions of participants (insiders, outsiders)
- Data consists of 43,233 separate activities and 4,253 unique type combinations. A 1,114 x 4,253 matrix.

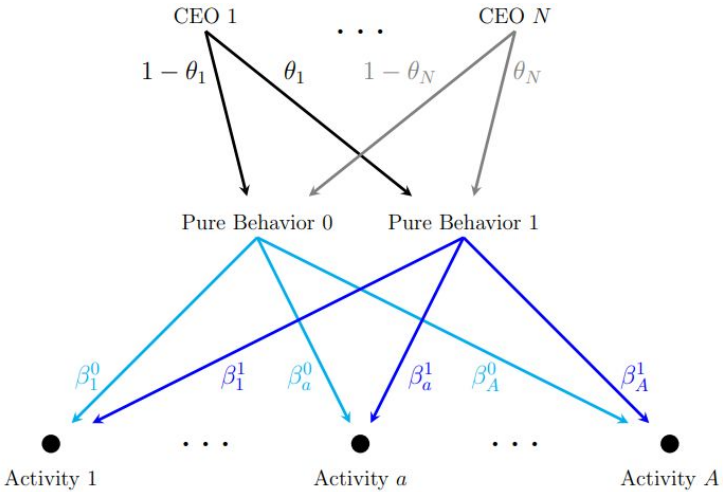


Paper application

Bandiera et al. (2020) - CEO Behavior & Firm Performance

LDA application

- Exploit the idea that the high-dimensional raw activity data is generated by a low-dimensional set of latent managerial behaviors
- Suppose that managers have **A** possible ways of organizing each 15-minute block, with x_a a particular activity
- If **X** = {x_1, x_2, x_3, ..., x_a} is the set of activities, a **pure behavior k** is the probability distribution β_a over **X** that is common to all CEOs
- Set **K = 2**, so that only two pure behaviors are possible: **managers** and **leaders**. The behavior of CEO **i** is given by a mixture of these two behaviors, according to weights $\theta_i \in [0,1]$
- The probability of CEO **i** generating activity **a** can lie anywhere between β_{a0} and β_{a1}.
- Actions are features, behaviors are topics, CEOs are documents, 15-minute blocks are text placeholders.

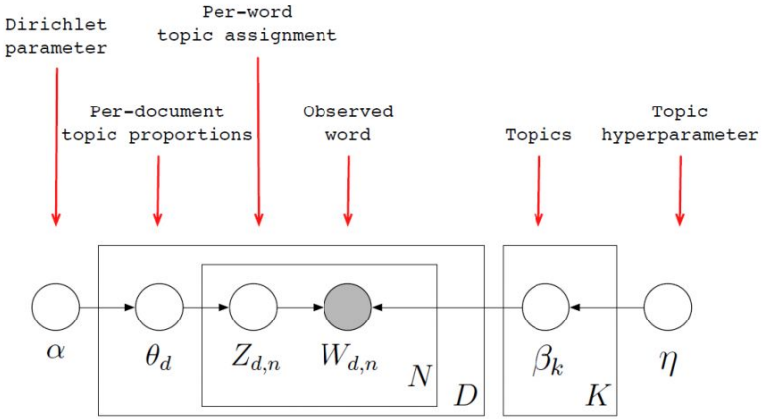


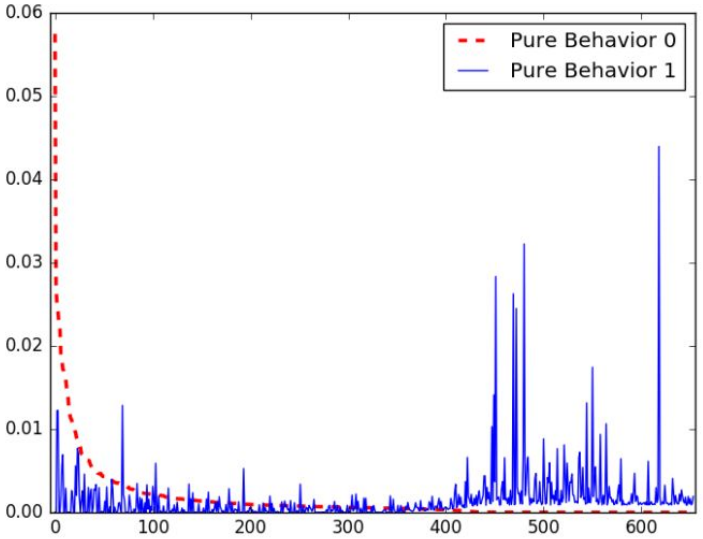
Paper application

Bandiera et al. (2020) - CEO Behavior & Firm Performance

LDA application

- Use Dirichlet priors for both β and θ, estimate posterior distributions of these using Gibbs sampling
- The process churns out
  - Two estimated pure behaviors β_0 and β_1
  - The estimated behavior indices θ_i for CEO i = 1, ..., N

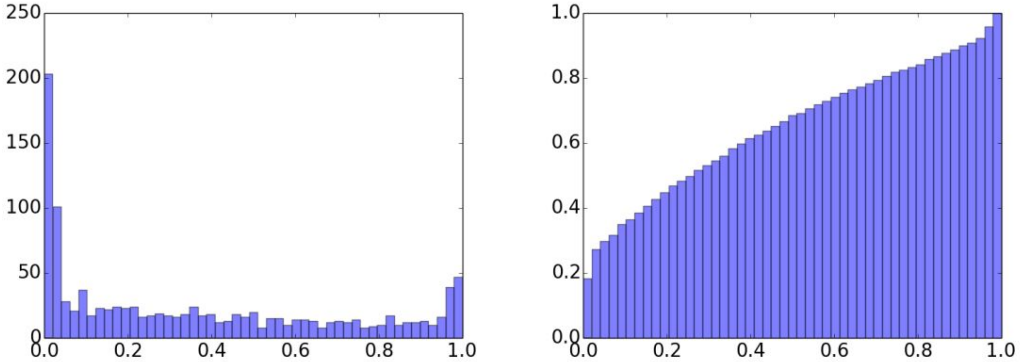




Probabilities of activities in estimated pure behaviors. 654 activities ordered on Beh0

Paper application

Bandiera et al. (2020) - CEO Behavior & Firm Performance



(a) Frequency Distribution (b) Cumulative Distribution

CEO behavior and index distribution

	X times less likely in Behavior 1		X times more likely in Behavior 1
Feature		**Feature**	
Plant Visits	0.11	Communications	1.90
Just Outsiders	0.58	Outsiders + Insiders	1.90
Production	0.46	C-suite	33.90
Suppliers	0.32	Multifunction	1.49

Paper application

Bandiera et al. (2020) - CEO Behavior & Firm Performance

Dependent Variable	(1)	(2)	(3)	(4)	(5)	(6)
	Log(sales)					Profits/Emp
CEO behavior index	0.343***	0.227**	0.322***	0.641**	0.506**	10.029***
	(0.108)	(0.111)	(0.121)	(0.278)	(0.236)	(3.456)
log(employment)	0.889***	0.555***	0.346***	0.339**	0.784***	-0.284
	(0.040)	(0.066)	(0.099)	(0.152)	(0.090)	(0.734)
log(capital)		0.387***	0.188***	0.194*		
		(0.042)	(0.056)	(0.098)		
log(materials)			0.447***	0.421***		
			(0.073)	(0.109)		
Management					0.179**	
					(0.072)	
Number of observations (firms)	920	618	448	243	156	386
Observations used to compute means	2,202	1,519	1,054	604	383	1,028
Sample	all	with k	with k & m	with k & m, listed	with management score	with profits, listed