ECPR Methods Summer School: Big Data Analysis in the Social Sciences

Pablo Barberá

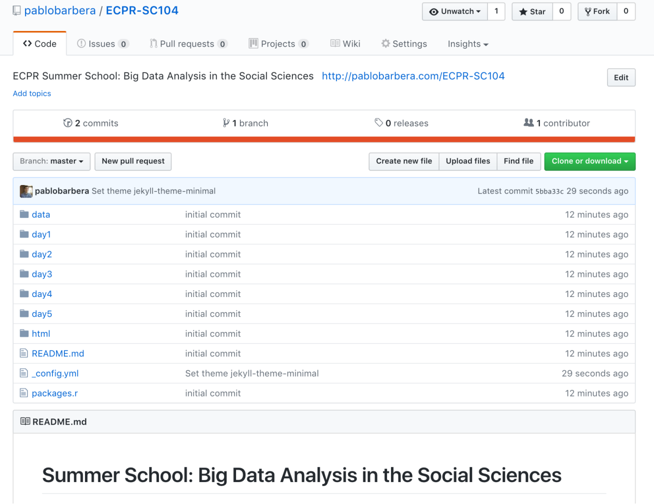
School of International Relations
University of Southern California
pablobarbera.com

Networked Democracy Lab
www.netdem.org

Course website:

github.com/pablobarbera/ECPR-SC104

Course website



ECPR Summer School: Big Data Analysis in the Social Sciences http://pablobarbera.com/ECPR-SC104

pablobarbera Set theme jekyll-theme-minimal Latest commit 5bba33c 29 seconds ago

data	initial commit	12 minutes ago
day1	initial commit	12 minutes ago
day2	initial commit	12 minutes ago
day3	initial commit	12 minutes ago
day4	initial commit	12 minutes ago
day5	initial commit	12 minutes ago
html	initial commit	12 minutes ago
README.md	initial commit	12 minutes ago
_config.yml	Set theme jekyll-theme-minimal	29 seconds ago
packages.r	initial commit	12 minutes ago

README.md

Summer School: Big Data Analysis in the Social Sciences

github.com/pablobarbera/ECPR-SC104

Social event

Save the date:

TODAY, 6pm, Ötkert

(Front door N15, 5.45pm)

Topic discovery in textual datasets

Overview of text as data methods

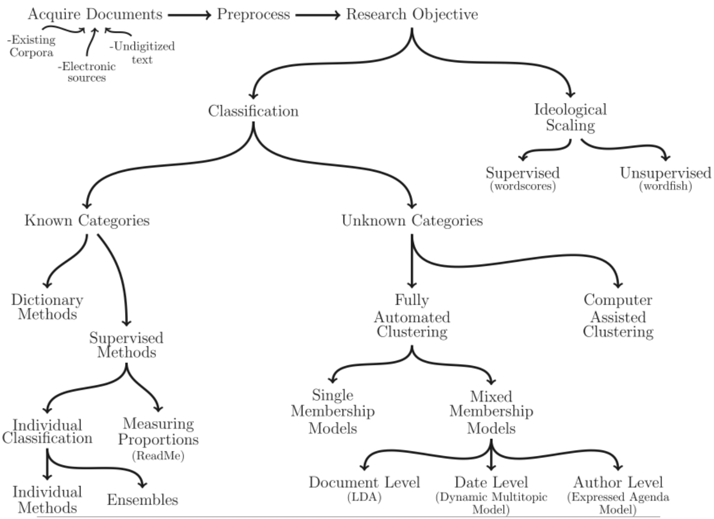
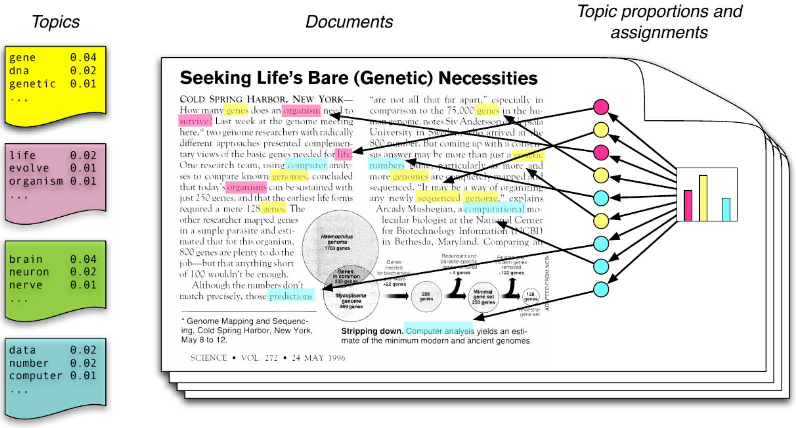


Fig. 1 in Grimmer and Stewart (2013)

Overview of techniques

- Descriptive analysis:
  - What are the characteristics of this corpus? How do some documents compare to others?
  - Keyness, collocation analysis, readability scores, Cosine/Jaccard similarity...
- Clustering and scaling:
  - What groups of documents are there in this corpus? Can documents be placed on a latent dimension?
  - Cluster analysis, principal component analysis, wordfish..
- Topic modeling:
  - What are the main themes in this corpus? How do different documents relate to words differently?
  - LDA, STM

Latent Dirichlet Allocation



Topics

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

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

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed – 4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

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

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

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions and assignments

Latent Dirichlet Allocation

- Document = random mixture over latent topics
- Topic = distribution over n-grams

Probabilistic model with 3 steps:

1. Choose $\theta_i \sim \mathrm{Dirichlet}(\alpha)$
2. Choose $\beta_k \sim \mathrm{Dirichlet}(\delta)$
3. For each word in document *i*:
   - Choose a topic $z_m \sim \mathrm{Multinomial}(\theta_i)$
   - Choose a word $w_{im} \sim \mathrm{Multinomial}(\beta_{i,k=z_m})$

where:

α=parameter of Dirichlet prior on distribution of topics over docs.

θ_i=topic distribution for document *i*

δ=parameter of Dirichlet prior on distribution of words over topics

β_k=word distribution for topic *k*

Latent Dirichlet Allocation

Key parameters:

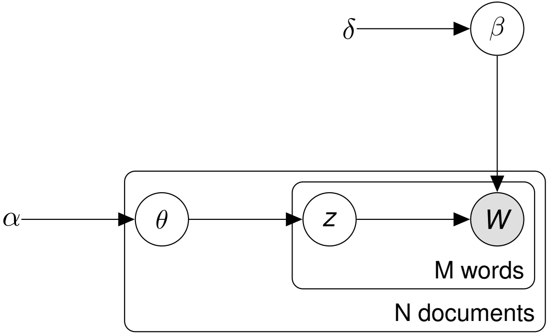
1. θ = matrix of dimensions N documents by K topics where θ_{ik} corresponds to the probability that document i belongs to topic k; i.e. assuming $K = 5$:

	T1	T2	T3	T4	T5
Document 1	0.15	0.15	0.05	0.10	0.55
Document 2	0.80	0.02	0.02	0.10	0.06
. . .					
Document N	0.01	0.01	0.96	0.01	0.01

2. β = matrix of dimensions K topics by M words where β_{km} corresponds to the probability that word m belongs to topic k; i.e. assuming $M = 6$:

	W1	W2	W3	W4	W5	W6
Topic 1	0.40	0.05	0.05	0.10	0.10	0.30
Topic 2	0.10	0.10	0.10	0.50	0.10	0.10
. . .						
Topic k	0.05	0.60	0.10	0.05	0.10	0.10

Plate notation



$\beta = M \times K$ matrix where β_{im} indicates prob(topic=k) for word m
$\theta = N \times K$ matrix where θ_{ik} indicates prob(topic=k) for document i

Validation

From Quinn et al, AJPS, 2010:

1. Semantic validity
   - Do the topics identify coherent groups of tweets that are internally homogenous, and are related to each other in a meaningful way?
2. Convergent/discriminant construct validity
   - Do the topics match existing measures where they should match?
   - Do they depart from existing measures where they should depart?
3. Predictive validity
   - Does variation in topic usage correspond with expected events?
4. Hypothesis validity
   - Can topic variation be used effectively to test substantive hypotheses?

Example: open-ended survey responses

Bauer, Barberá *et al*, *Political Behavior*, 2016.

- Data: General Social Survey (2008) in Germany
- Responses to questions: *Would you please tell me what you associate with the term "left"? and would you please tell me what you associate with the term "right"?*
- Open-ended questions minimize priming and potential interviewer effects
- Sparse Additive Generative model instead of LDA (more coherent topics for short text)
- $K = 4$ topics for each question

Example: open-ended survey responses

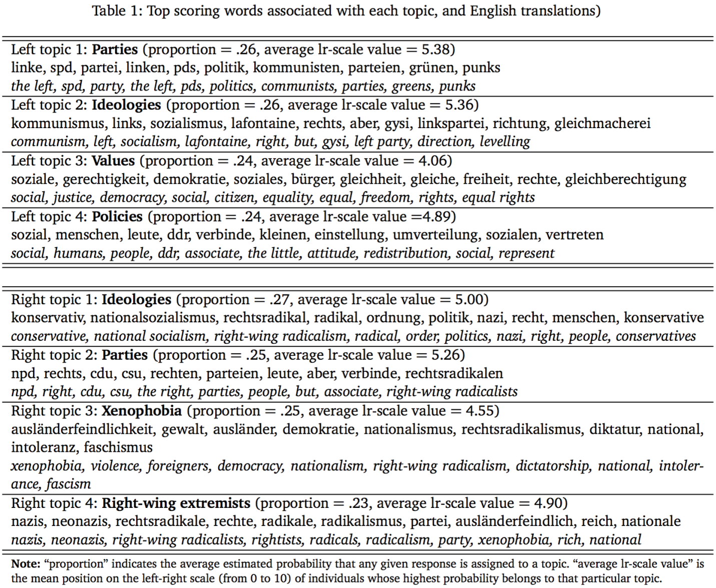


Table 1: Top scoring words associated with each topic, and English translations)

Left topic 1: **Parties** (proportion = .26, average lr-scale value = 5.38) linke, spd, partei, linken, pds, politik, kommunisten, parteien, grünen, punks *the left, spd, party, the left, pds, politics, communists, parties, greens, punks*
Left topic 2: **Ideologies** (proportion = .26, average lr-scale value = 5.36) kommunismus, links, sozialismus, lafontaine, rechts, aber, gysi, linkspartei, richtung, gleichmacherei *communism, left, socialism, lafontaine, right, but, gysi, left party, direction, levelling*
Left topic 3: **Values** (proportion = .24, average lr-scale value = 4.06) soziale, gerechtigkeit, demokratie, soziales, bürger, gleichheit, gleiche, freiheit, rechte, gleichberechtigung *social, justice, democracy, social, citizen, equality, equal, freedom, rights, equal rights*
Left topic 4: **Policies** (proportion = .24, average lr-scale value =4.89) sozial, menschen, leute, ddr, verbinde, kleinen, einstellung, umverteilung, sozialen, vertreten *social, humans, people, ddr, associate, the little, attitude, redistribution, social, represent*

Right topic 1: **Ideologies** (proportion = .27, average lr-scale value = 5.00) konservativ, nationalsozialismus, rechtsradikal, radikal, ordnung, politik, nazi, recht, menschen, konservative *conservative, national socialism, right-wing radicalism, radical, order, politics, nazi, right, people, conservatives*
Right topic 2: **Parties** (proportion = .25, average lr-scale value = 5.26) npd, rechts, cdu, csu, rechten, parteien, leute, aber, verbinde, rechtsradikalen *npd, right, cdu, csu, the right, parties, people, but, associate, right-wing radicalists*
Right topic 3: **Xenophobia** (proportion = .25, average lr-scale value = 4.55) ausländerfeindlichkeit, gewalt, ausländer, demokratie, nationalismus, rechtsradikalismus, diktatur, national, intoleranz, faschismus *xenophobia, violence, foreigners, democracy, nationalism, right-wing radicalism, dictatorship, national, intolerance, fascism*
Right topic 4: **Right-wing extremists** (proportion = .23, average lr-scale value = 4.90) nazis, neonazis, rechtsradikale, rechte, radikale, radikalismus, partei, ausländerfeindlich, reich, nationale *nazis, neonazis, right-wing radicalists, rightists, radicals, radicalism, party, xenophobia, rich, national*

Note: "proportion" indicates the average estimated probability that any given response is assigned to a topic. "average lr-scale value" is the mean position on the left-right scale (from 0 to 10) of individuals whose highest probability belongs to that particular topic.

Bauer, Barberá *et al*, *Political Behavior*, 2016.

Example: open-ended survey responses

Fig. 6: Left-right scale means for different subsamples of associations with **left** (dashed = sample mean, bars = 95% Cis)

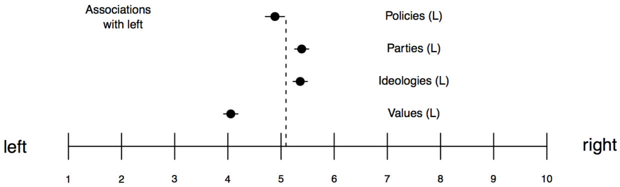
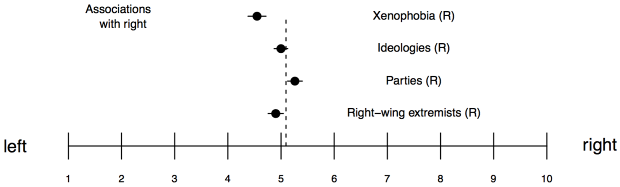


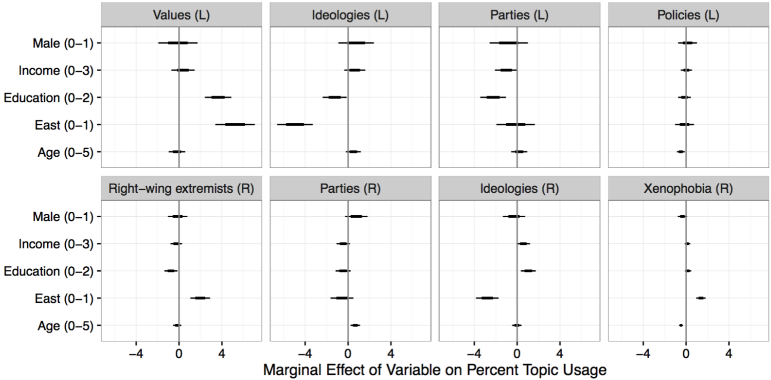
Fig. 7: Left-right scale means for different subsamples of associations with **right** (dashed = sample mean, bars = 95% Cis)



Bauer, Barberá *et al*, *Political Behavior*, 2016.

Example: open-ended survey responses

Fig. 9: Systematic relationship between associations with "left" and "right" and characteristics of respondents



Note: Each line indicates a 95% confidence interval (and 66% confidence interval in darker color) for the coefficient of eight different regressions of topic usage (in a scale from 0 to 100) at the respondent level on seven individual-level characteristics. The line on the bottom right corner (second row, second plot), for example, shows that individual a one-category change in age is associated with around one percentage point increase in the probability that the individual associated "right" with political parties.

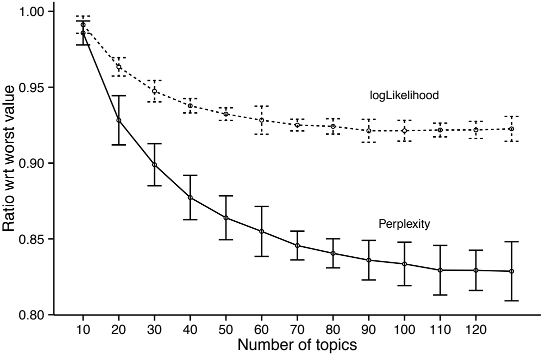
Bauer, Barberá *et al*, *Political Behavior*, 2016.

Example: topics in US legislators' tweets

- Data: 651,116 tweets sent by US legislators from January 2013 to December 2014.
- 2,920 documents = 730 days × 2 chambers × 2 parties
- Why aggregating? Applications that aggregate by author or day outperform tweet-level analyses (Hong and Davidson, 2010)
- $K = 100$ topics (more on this later)
- Validation: http://j.mp/lda-congress-demo

Choosing the number of topics

- Choosing K is “one of the most difficult questions in unsupervised learning” (Grimmer and Stewart, 2013, p.19)
- We chose $K = 100$ based on cross-validated model fit.



- **BUT**: “there is often a negative relationship between the best-fitting model and the substantive information provided”.
- GS propose to choose K based on “substantive fit.”

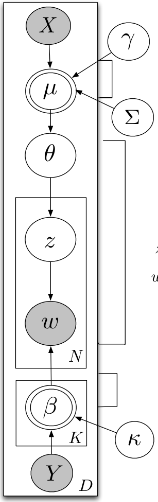
Extensions of LDA

1. Structural topic model (Roberts et al, 2014, AJPS)
2. Dynamic topic model (Blei and Lafferty, 2006, ICML; Quinn et al, 2010, AJPS)
3. Hierarchical topic model (Griffiths and Tenembaun, 2004, NIPS; Grimmer, 2010, PA)

Why?

- Substantive reasons: incorporate specific elements of DGP into estimation
- Statistical reasons: structure can lead to better topics.

Structural topic model



Topic Prevalence:

$$\mu_{d,k} = X_d \gamma_k$$
$$\gamma_k \sim \mathcal{N}(0, \sigma_k^2)$$
$$\sigma_k^2 \sim \text{Gamma}(s^\gamma, r^\gamma)$$

Language Model:

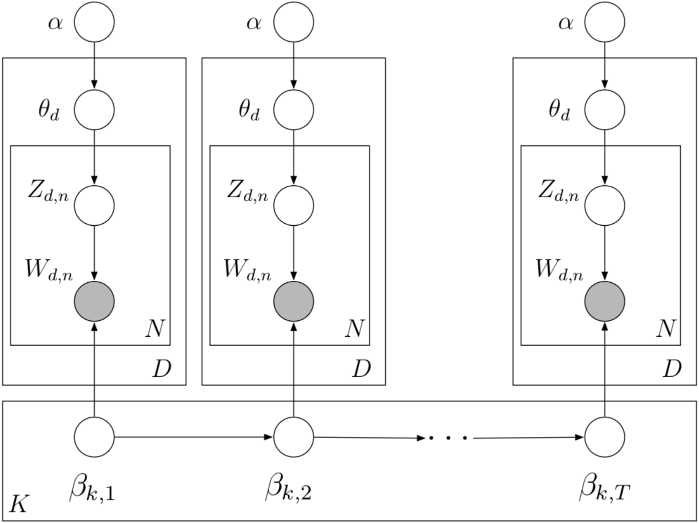
$$\theta_d \sim \text{LogisticNormal}(\mu_d, \Sigma)$$
$$z_{d,n} \sim \text{Mult}(\theta_d)$$
$$w_{d,n} \sim \text{Mult}(\beta_d^{k=z_{d,n}})$$

Topical Content:

$$\beta_{d,v}^k \propto \exp(m_v + \kappa_v^{\cdot,k} + \kappa_v^{y,\cdot} + \kappa_v^{y,k})$$
$$\kappa_v^{y,k} \sim \text{Laplace}(0, \tau_v^{y,k})$$
$$\tau_v^{y,k} \sim \text{Gamma}(s^\kappa, r^\kappa)$$

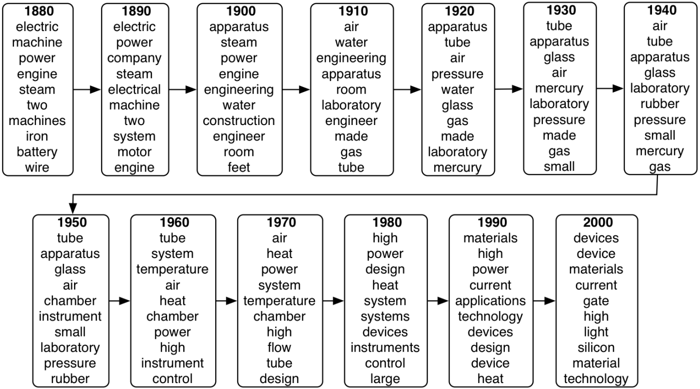
- **Prevalence**: Prior on the mixture over topics is now document-specific, and can be a function of covariates (documents with similar covariates will tend to be about the same topics)
- **Content**: distribution over words is now document-specific and can be a function of covariates (documents with similar covariates will tend to use similar words to refer to the same topic)

Dynamic topic model



Source: Blei, “Modeling Science”

Dynamic topic model



1880
electric
machine
power
engine
steam
two
machines
iron
battery
wire

1890
electric
power
company
steam
electrical
machine
two
system
motor
engine

1900
apparatus
steam
power
engine
engineering
water
construction
engineer
room
feet

1910
air
water
engineering
apparatus
room
laboratory
engineer
made
gas
tube

1920
apparatus
tube
air
pressure
water
glass
gas
made
laboratory
mercury

1930
tube
apparatus
glass
air
mercury
laboratory
pressure
made
gas
small

1940
air
tube
apparatus
glass
laboratory
rubber
pressure
small
mercury
gas

1950
tube
apparatus
glass
air
chamber
instrument
small
laboratory
pressure
rubber

1960
tube
system
temperature
air
heat
chamber
power
high
instrument
control

1970
air
heat
power
system
temperature
chamber
high
flow
tube
design

1980
high
power
design
heat
system
systems
devices
instruments
control
large

1990
materials
high
power
current
applications
technology
devices
design
device
heat

2000
devices
device
materials
current
gate
high
light
silicon
material
technology

Source: Blei, “Modeling Science”

Wordfish (Slapin and Proksch, 2008, AJPS)

- Goal: unsupervised scaling of ideological positions
- Ideology of politician *i*, θ_i is a position in a latent scale.
- Word usage is drawn from a Poisson-IRT model:

$$W_{im} \sim \mathrm{Poisson}(\lambda_{im})$$
$$\lambda_{im} = exp(\alpha_i + \psi_m + \beta_m \times \theta_i)$$

- where:

  α_i is "loquaciousness" of politician *i*
  ψ_m is frequency of word *m*
  β_m is discrimination parameter of word *m*

- Estimation using EM algorithm.
- Identification:
  - Unit variance restriction for θ_i
  - Choose *a* and *b* such that $\theta_a > \theta_b$