Networks and complex data: Discussion

Edoardo Airoldi
Department of Statistics
Harvard University

Becker Friedman Institute, University of Chicago, September 24th 2016

Analysis of network data

Remarks

- What researchers might do in practice
- Complications arising from a population network
- What we know and what we do not

A field experiment in Honduras

- Education on standards of care for newborns

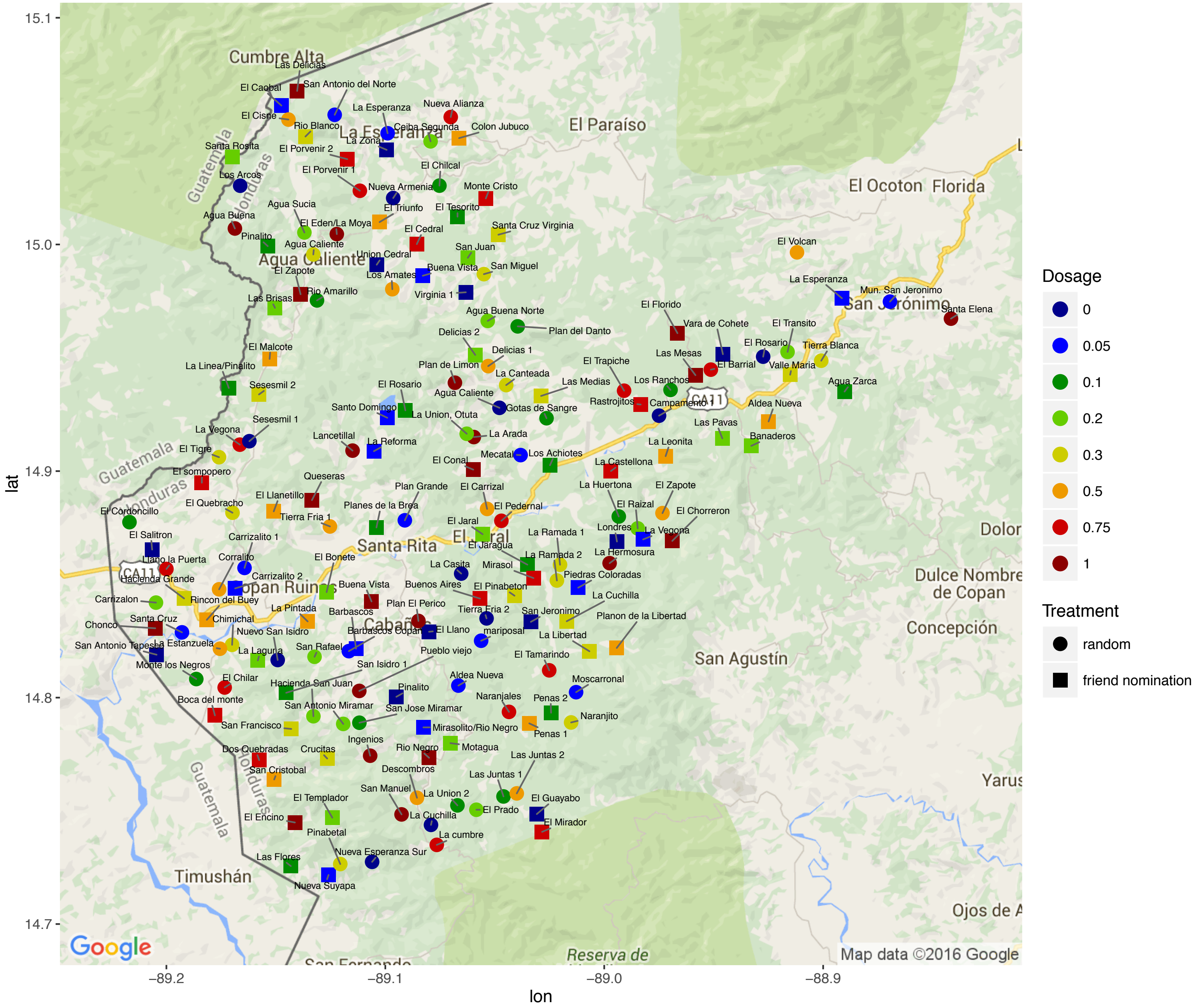
  (176 villages; 30,000 people; 3,000 target households)

- Design: 2 target nomination schemes and 8 levels of treatment

- Short- and long-term causal effects, including on social interactions 2-year post intervention

(N Christakis, J Fowler, D Spielman, R Negon, et al.)

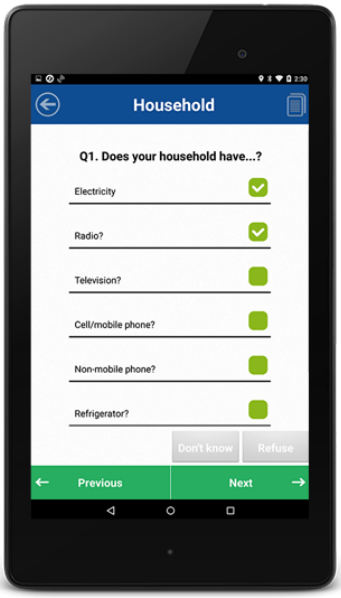
Aldeas – Copan, Honduras – Treatment Blocking



Networks mapped pre-intervention

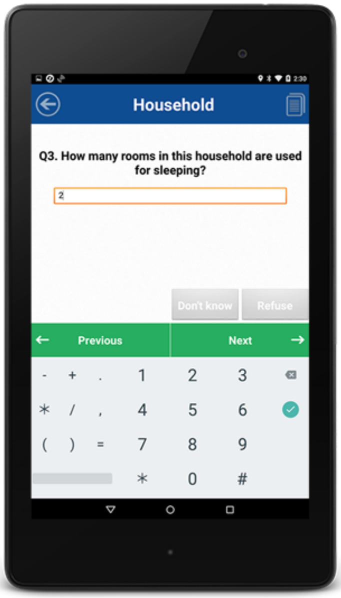
- Trellis for mapping 20+ aspects of social relations

  (Arguably no measurement error, but missing data)



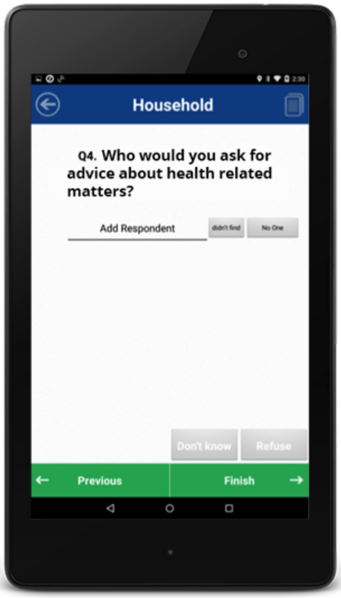
Q1. Does your household have...?

Electricity ☑
Radio? ☑
Television? ☐
Cell/mobile phone? ☐
Non-mobile phone? ☐
Refrigerator? ☐

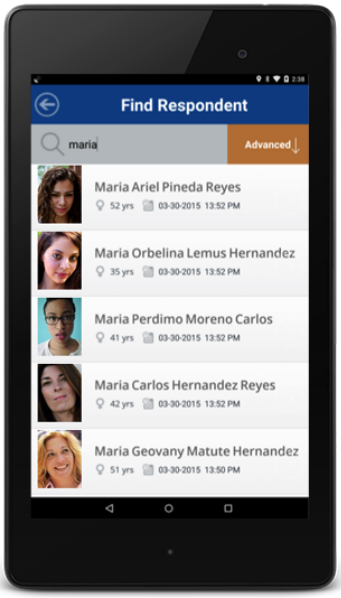


Q3. How many rooms in this household are used for sleeping?

2

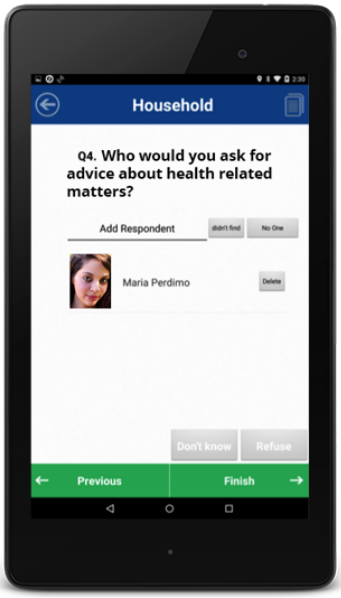


Q4. Who would you ask for advice about health related matters?



maria

Maria Ariel Pineda Reyes 52 yrs 03-30-2015 13:52 PM
Maria Orbelina Lemus Hernandez 35 yrs 03-30-2015 13:52 PM
Maria Perdimo Moreno Carlos 41 yrs 03-30-2015 13:52 PM
Maria Carlos Hernandez Reyes 42 yrs 03-30-2015 13:52 PM
Maria Geovany Matute Hernandez 51 yrs 03-30-2015 13:50 PM



Q4. Who would you ask for advice about health related matters?

Maria Perdimo

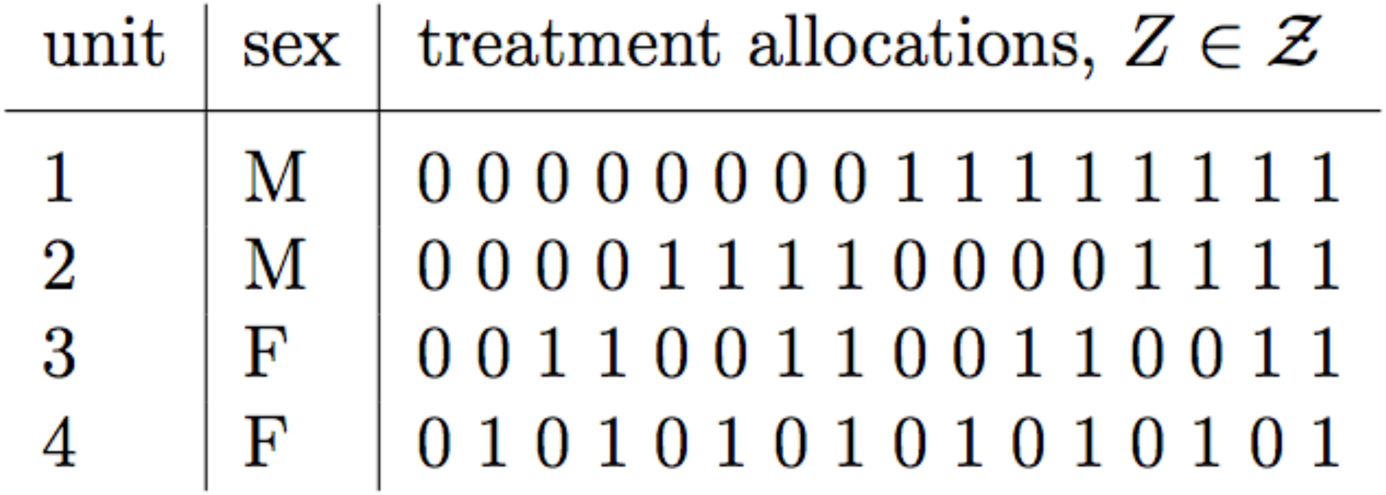
Other illustrative applications

- LinkedIn
  - Labor mobility, employment histories, recruiting
  - Network only provides limited / partial view of professional interactions
  - Connections may be interventions of interest
- Google display ads
  - Selective callouts problem (who to invite in auctions)
  - Many networks available
  - Causal mechanisms are not well developed

Remarks

- What researchers might do in practice
- Complications arising from a population network
- What we know and what we do not

Potential outcomes / table of science



unit	sex	treatment allocations, $Z \in \mathcal{Z}$															
1	M	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
2	M	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
3	F	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
4	F	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1

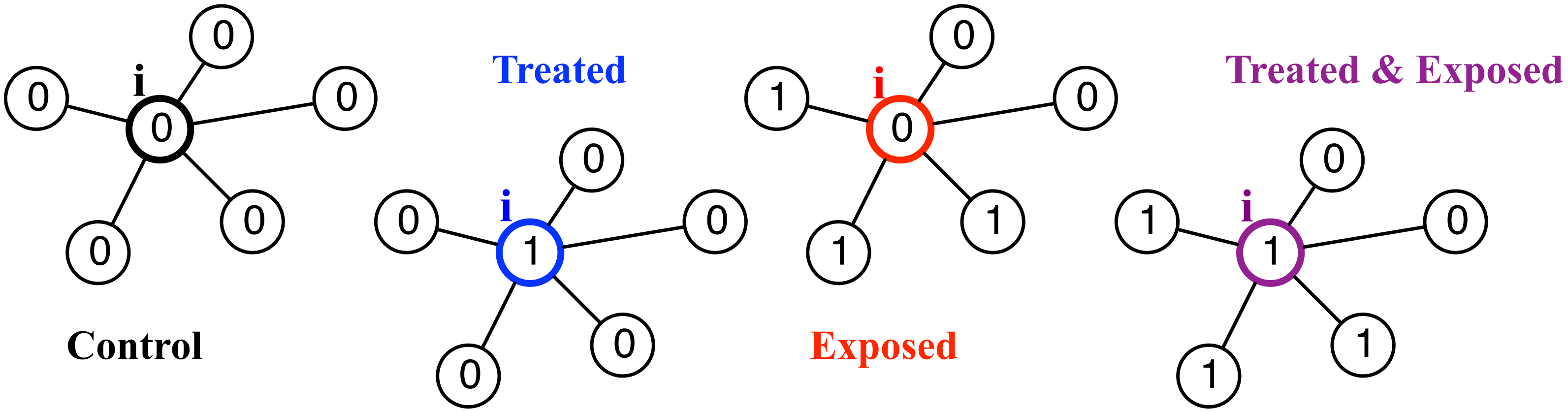
Table of science **Y** is N=4 x |$\mathcal{Z}$|=2^4, with element $Y_i(\mathbf{Z})$

Causal inferential targets are defined as a function of **Y**

Typically we assume $Y_i(Z_i)$ – **Y** becomes Nx2

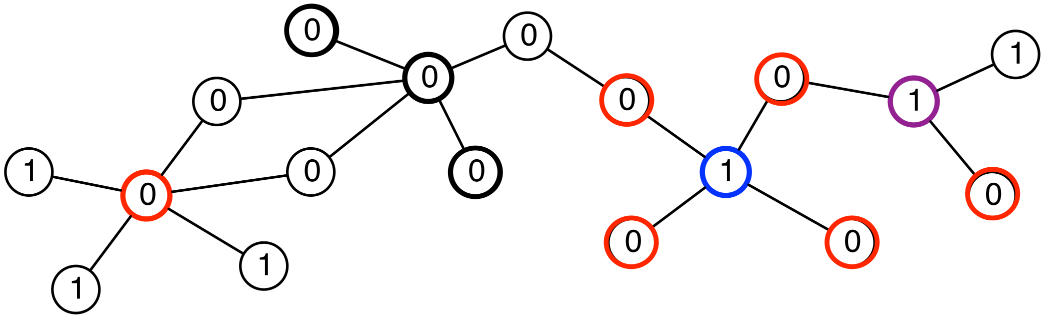
Network interference and exposure

- Given network *G*, if no interference is untenable, we must assume how $\mathbf{Z}_{\mathcal{N}_i}$ affects $Y_i(Z_i, g(\mathbf{Z}_{\mathcal{N}_i}), \text{rest})$
- Example: "**i** is exposed if it has 1+ treated neighbors" leads to 4 distinct potential outcomes



What fails?

- ATE and TTE are no longer equal
- Allocations $\mathbf{Z}^j$ on G with the same (n_t,n_c) have different configurations of treatment, exposure and control



- Need randomization schemes that leverage G
- Need to revisit causal interpretation of parameters

Causal interpretation of parameters

Recall $Y_i^{obs} = Z_i\, Y_i(1) + (1-Z_i)\, Y_i(0)$; assume SUTVA and additivity of treatment effect $Y_i(1) = Y_i(0) + \beta$

- Slope coefficient in a regression model for Y_i^{obs} equals the ATE; thus it has a clear interpretation as causal effect defined on table of science **Y**

In the presence of network interference? The practice is

- State a model for Y_i^{obs} and argue why its parameter(s) capture (aspects of) the causal effect(s) of interest

Main assumption and model

Neighborhood interference assumption (__N__IA)

Definition 2.1. For each unit $i \in [n]$ and all treatment allocations $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}$, if for all $j \in \mathcal{N}_i \cup \{i\}$ it holds that $z_j = z'_j$ then $Y_i(\mathbf{z}) = Y_i(\mathbf{z}')$.

Convenient to define

$$\widetilde{Y}_i : \{0,1\} \times \{0,1\}^{d_i} \mapsto \mathbb{R} \quad \text{such that} \quad Y_i(\mathbf{z}) = \widetilde{Y}_i(z_i, \mathbf{z}_{\mathcal{N}_i})$$

And parameters

$$\alpha_i = \widetilde{Y}_i(0, \mathbf{0}) \qquad \beta_i = \widetilde{Y}_i(1, \mathbf{0}) - \widetilde{Y}_i(0, \mathbf{0})$$

$$\Gamma_i(\mathbf{z}_{\mathcal{N}_i}) = \widetilde{Y}_i(0, \mathbf{z}_{\mathcal{N}_i}) - \widetilde{Y}_i(0, \mathbf{0}) \qquad \Delta_i(\mathbf{z}_{\mathcal{N}_i}) = \widetilde{Y}_i(1, \mathbf{z}_{\mathcal{N}_i}) - \widetilde{Y}_i(1, \mathbf{0})$$

Additive model: $Y_i(\mathbf{Z}) = \alpha_i + \beta_i Z_i + \Gamma_i(\mathbf{Z}_{\mathcal{N}i}) + Z_i \Delta_i(\mathbf{Z}_{ni})$

Structural assumptions

Definition 2.2 (Additivity of Main Effects)**.** The potential outcomes satisfy additivity of main effects if for all treatments $\mathbf{z}$ and all units i, it holds that

$$\widetilde{Y}_i(z_i, \mathbf{z}_{\mathcal{N}_i}) = \widetilde{Y}_i(0, \mathbf{0}) + (\widetilde{Y}_i(z_i, \mathbf{0}) - \widetilde{Y}_i(0, \mathbf{0})) + (\widetilde{Y}_i(0, \mathbf{z}_{\mathcal{N}_i}) - \widetilde{Y}_i(0, \mathbf{0})). \tag*{(\_\textbf{A}N\_\_IA)}$$

Definition 2.3 (Symmetrically Received Interference Effects)**.** The potential outcomes have symmetrically received interference if for all allocations $\mathbf{z}$, units i, and permutations σ of vectors of length d_i, it holds that

$$\widetilde{Y}_i(z_i, \mathbf{z}_{\mathcal{N}_i}) = \widetilde{Y}_i(z_i, \sigma(\mathbf{z}_{\mathcal{N}_i})). \tag*{(\textbf{S}\_N\_\_IA)}$$

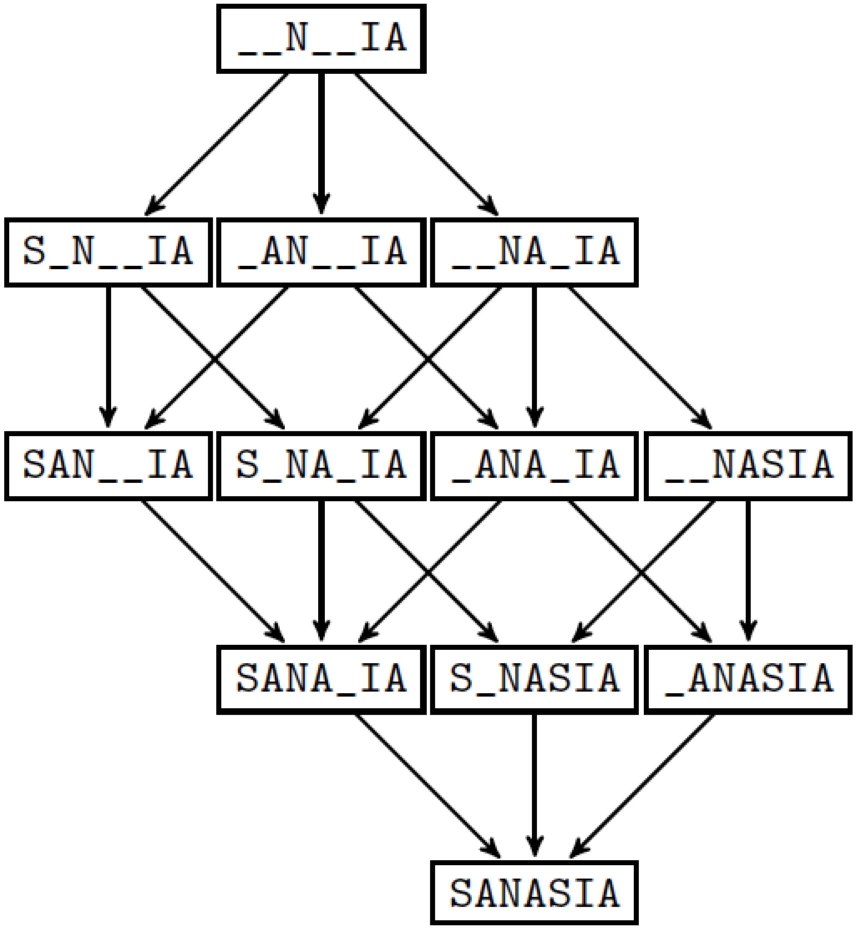
Definition 2.4 (Additivity of Interference Effects)**.** The potential outcomes satisfy additivity of interference effects if for all allocations $\mathbf{z}$ and units i

$$\widetilde{Y}_i(z_i, \mathbf{z}_{\mathcal{N}_i}) = \widetilde{Y}_i(z_i, \mathbf{0}) + \textstyle\sum_{\{j \in \mathcal{N}_i\}} (Y_i(z_j \mathbf{e}_j + z_i \mathbf{e}_i) - \widetilde{Y}_i(z_i, \mathbf{0})) \tag*{(\_\_N\textbf{A}\_IA)}$$

Definition 2.5 (Symmetrically Sent Interference Effects)**.** The potential outcomes have symmetrically sent interference if for units j and allocations $\mathbf{z}$ where $z_j = 0$, and units i, i' where $g_{ji} = g_{ji'} = 1$,

$$Y_i(\mathbf{z} + \mathbf{e}_j) - Y_i(\mathbf{z}) = Y_{i'}(\mathbf{z} + \mathbf{e}_j) - Y_{i'}(\mathbf{z}). \tag*{(\_\_N\_\textbf{S}IA)}$$

Relations among twelve unique models



__N__IA

S_N__IA _AN__IA __NA_IA

SAN__IA S_NA_IA _ANA_IA __NASIA

SANA_IA S_NASIA _ANASIA

SANASIA

_/S	Symmetrically Received Interference Effects
_/A	Additivity of Main Effects
N	**Neighborhood**
_/A	Additivity of Interference Effects
_/S	Symmetrically Sent Interference Effects
I	**Interference**
A	**Assumption**

Unambiguous causal interpretation

- Can now write causal effects of interest in terms of parameters of the 12 additive models for Y_i^{obs}, e.g.,

$$
\begin{aligned}
\text{ATE} &= 1/n\, \Sigma_i\, (Y_i(\mathbf{e_i}) - Y_i(\mathbf{0})) \\
&= 1/n\, \Sigma_i\, \beta_i && \text{(under NIA)} \\
\text{TTE} &\equiv 1/n\, \Sigma_i\, (Y_i(\mathbf{1}) - Y_i(\mathbf{0})) \\
&= 1/n\, \Sigma_i\, (\beta_i + \Gamma_i(d_i) + \Delta_i(d_i)) && \text{(under NIA)} \\
&= 1/n\, \Sigma_i\, (\beta_i + \gamma\, d_i) && \text{(under SANASIA)}
\end{aligned}
$$

- We can now spell out assumptions that justify previously published estimators / estimates

Remarks

- What researchers might do in practice
- Complications arising from a population network
- What we know and what we do not

Complex landscape and set-up

- Inferential targets: ATE, TTE, AIE, …
- Assumptions about the network: fixed vs. model, observed pre-intervention or outcome; fully vs. partially observed, with and without errors, formal notion of interference
- Sampling mechanism; finite vs. infinite population inference; models for the outcomes; observational vs. experimental
- Treatment allocation strategy; estimator; complications …

What do we know? Not much …

- Network observed pre-intervention without error
  - Fisher tests for interference (beyond first order neighbors)
  - Durbin-Wu-Hausman style test SUTVA violations
  - Estimation theory (LUE / MIVE) for ATE and failures
  - New randomization and rerandomization strategies
  - Homophily vs peer influence in observational studies
- Network observed with error
  - Inference from non-ignorable network sampling designs
  - Partially revealed interference (network as outcome)
  - Condition on a model for the network

Analysis of text data

Analysis of word counts

- Statistics and machine learning^{++}
- Common elements

  Matrix of word counts **W** (n documents x v terms)

  Mixture models (k components, interpreted as ??)

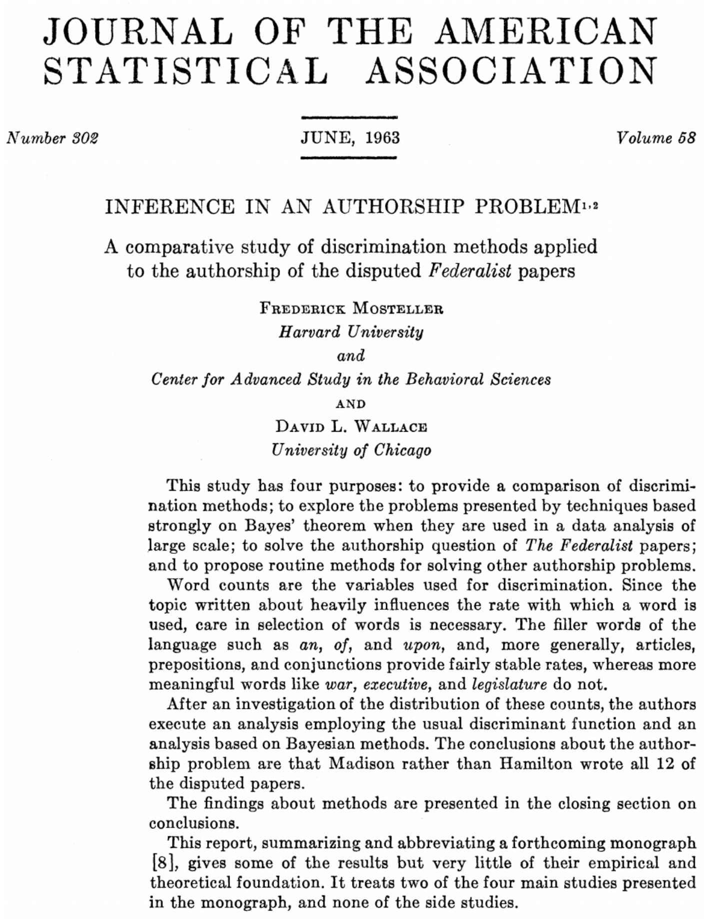
  Parameters **μ** are rates of occurrence (v terms x k components)
- Problem specific

  Document covariates **L** (n documents x …; e.g., author(s), publication year, topic annotations by professional editors)

Remarks

- Statistics and machine learning: Classic papers
- Evaluation and interpretation issues
- Bringing causality back

An authorship attribution problem



JOURNAL OF THE AMERICAN
STATISTICAL ASSOCIATION

Number 302 JUNE, 1963 *Volume 58*

INFERENCE IN AN AUTHORSHIP PROBLEM[1,2]

A comparative study of discrimination methods applied
to the authorship of the disputed *Federalist* papers

FREDERICK MOSTELLER
Harvard University
and
Center for Advanced Study in the Behavioral Sciences
AND
DAVID L. WALLACE
University of Chicago

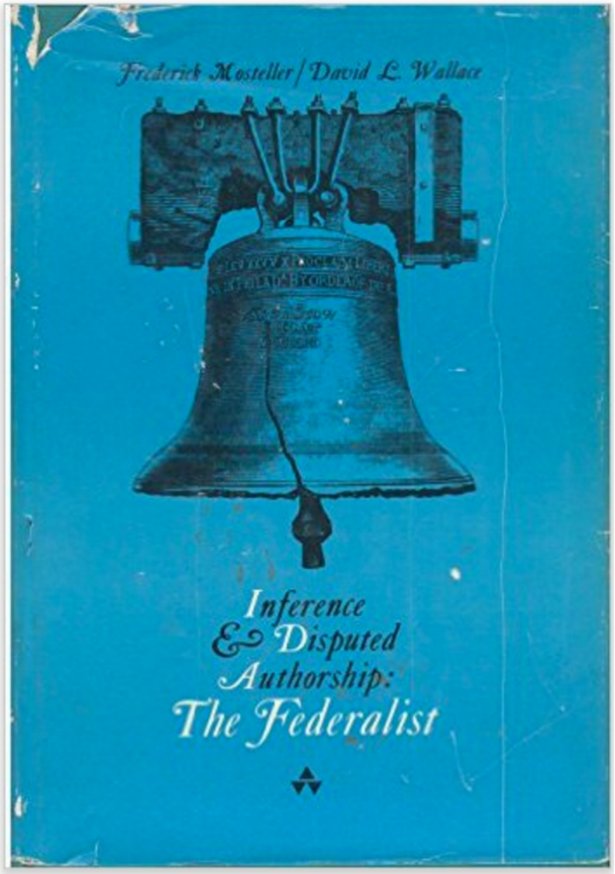
This study has four purposes: to provide a comparison of discrimination methods; to explore the problems presented by techniques based strongly on Bayes' theorem when they are used in a data analysis of large scale; to solve the authorship question of *The Federalist* papers; and to propose routine methods for solving other authorship problems.

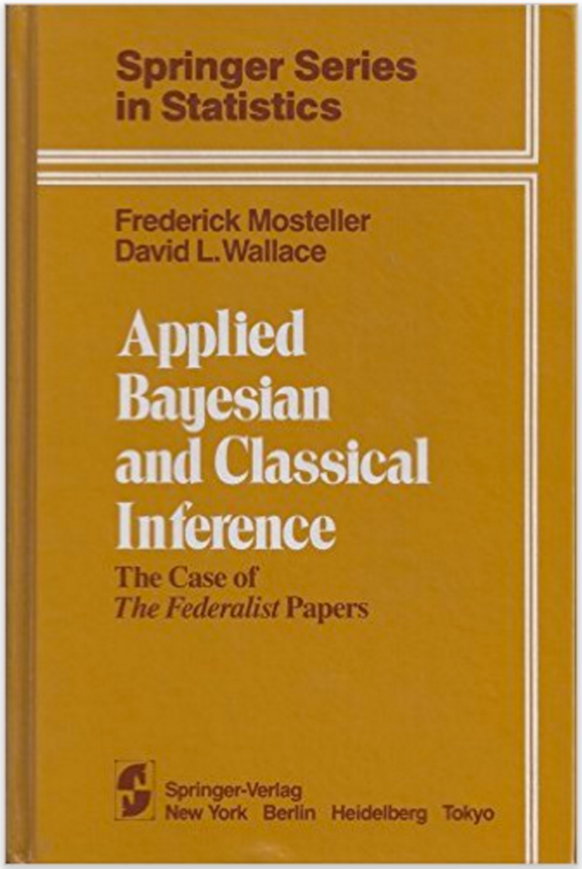
Word counts are the variables used for discrimination. Since the topic written about heavily influences the rate with which a word is used, care in selection of words is necessary. The filler words of the language such as *an*, *of*, and *upon*, and, more generally, articles, prepositions, and conjunctions provide fairly stable rates, whereas more meaningful words like *war*, *executive*, and *legislature* do not.

After an investigation of the distribution of these counts, the authors execute an analysis employing the usual discriminant function and an analysis based on Bayesian methods. The conclusions about the authorship problem are that Madison rather than Hamilton wrote all 12 of the disputed papers.

The findings about methods are presented in the closing section on conclusions.

This report, summarizing and abbreviating a forthcoming monograph [8], gives some of the results but very little of their empirical and theoretical foundation. It treats two of the four main studies presented in the monograph, and none of the side studies.


Frederick Mosteller / David L. Wallace
Inference & Disputed Authorship: The Federalist


Springer Series in Statistics
Frederick Mosteller
David L. Wallace
Applied Bayesian and Classical Inference
The Case of The Federalist Papers
Springer-Verlag
New York Berlin Heidelberg Tokyo

Parameterization and priors

- Counts for term v are Poisson with rates (μ_v^H, μ_v^M)
- Re-parameterize with total and differential rates

$$\sigma_v = \mu_v^H + \mu_v^M$$

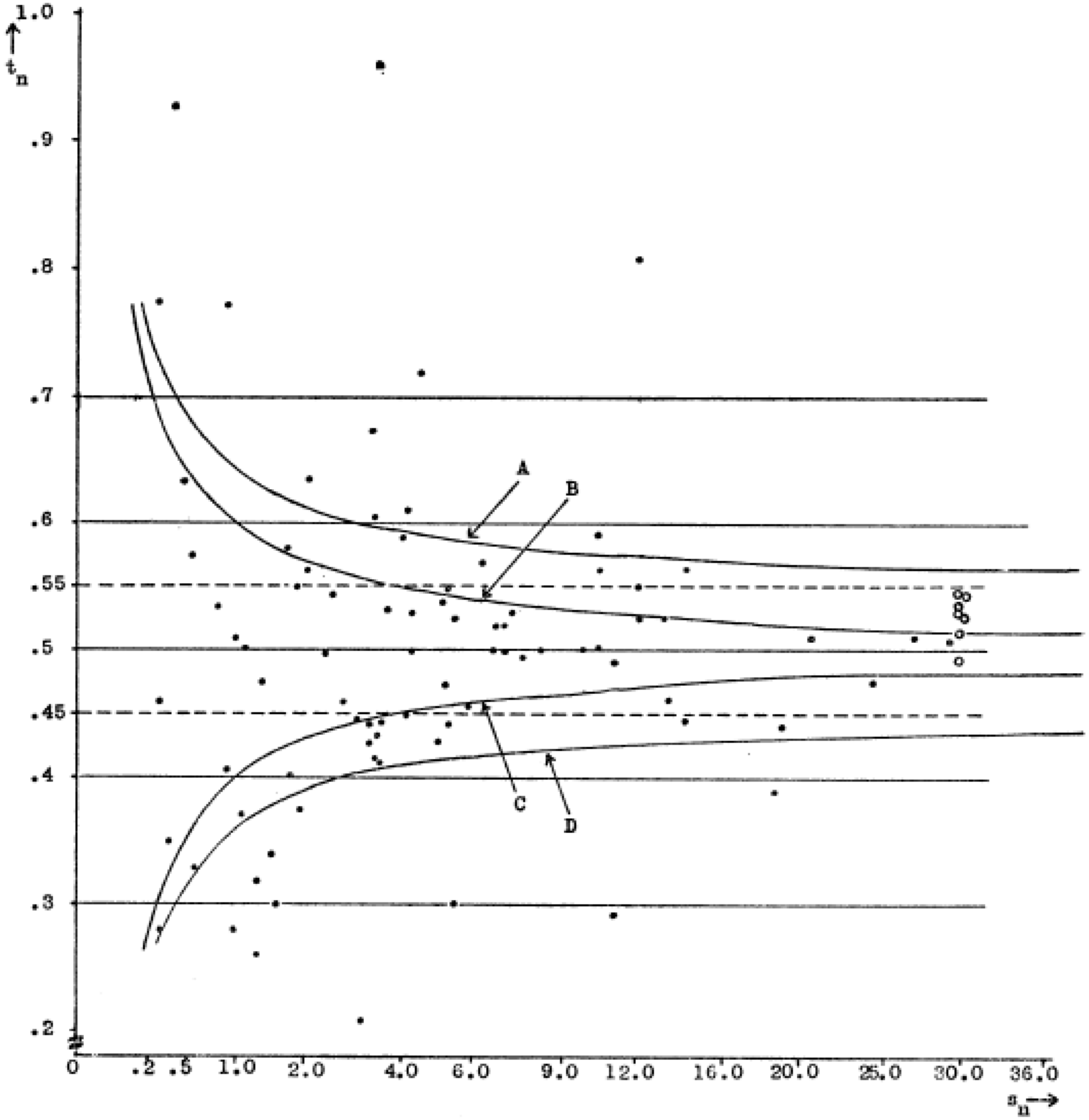
$$\tau_v = \mu_v^H / (\mu_v^H + \mu_v^M)$$

- Priors

$$\sigma_v \propto \text{constant}$$

$$\tau_v = \text{symmetric beta } (\alpha_1 + \alpha_2\, \sigma_v)$$

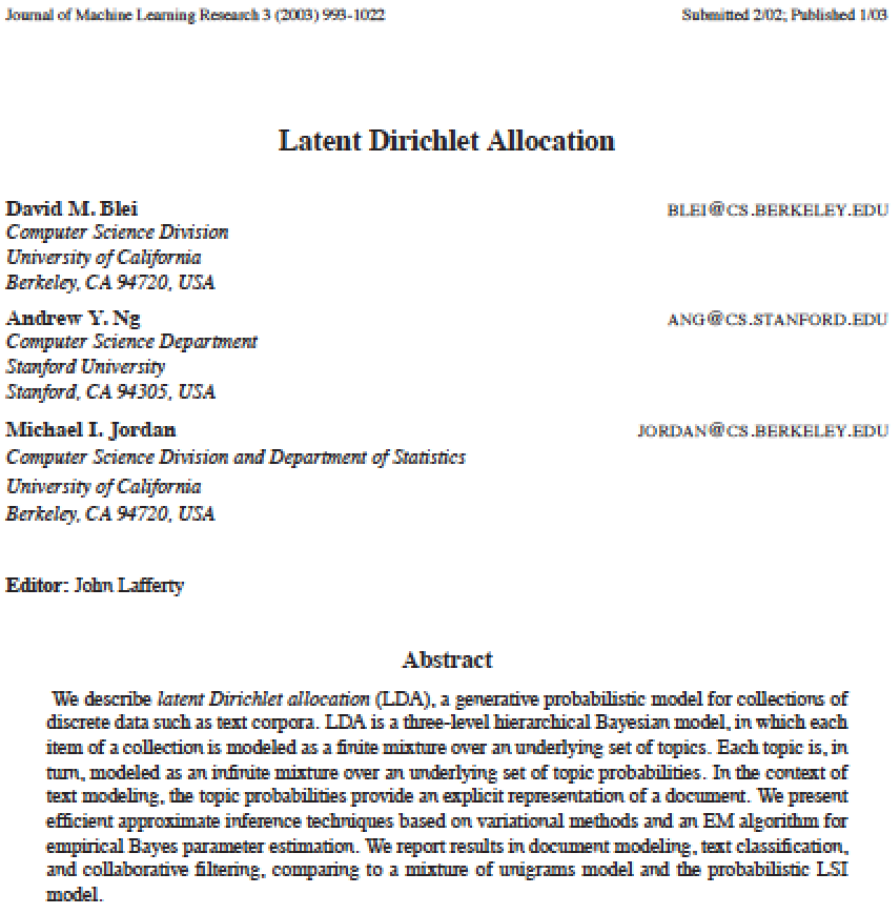
Also see Negative-Binomial (MW) and COM Poisson distributions (Kadane, Shmueli, and co-authors)


1.0
t_n
.9
.8
.7
.6
.55
.5
.45
.4
.3
.2
A
B
C
D
0
.2
.5
1.0
2.0
4.0
6.0
9.0
12.0
16.0
20.0
25.0
30.0
36.0
s_n

Remarks

- Authorship vector **L** is largely observed
- Clear interpretation of the 2 mixture components
- Evaluation
  - In-sample using agreement between posterior odds of authorship for undisputed papers and $\mathbf{L}^{obs}$
  - Out-of-sample predictions for $\mathbf{L}^{mis}$

Characterizing topics



Journal of Machine Learning Research 3 (2003) 993-1022 Submitted 2/02; Published 1/03

Latent Dirichlet Allocation

David M. Blei BLEI@CS.BERKELEY.EDU
Computer Science Division
University of California
Berkeley, CA 94720, USA

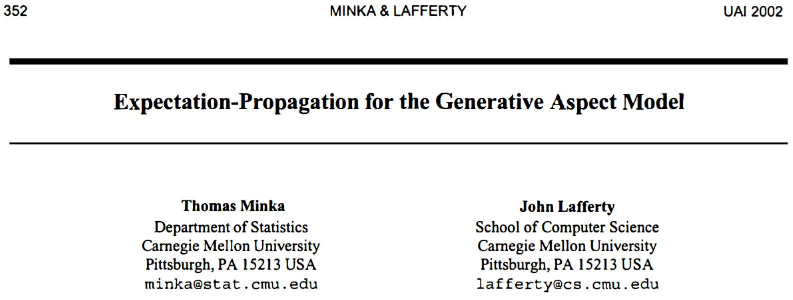
Andrew Y. Ng ANG@CS.STANFORD.EDU
Computer Science Department
Stanford University
Stanford, CA 94305, USA

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
Computer Science Division and Department of Statistics
University of California
Berkeley, CA 94720, USA

Editor: John Lafferty

Abstract

We describe *latent Dirichlet allocation* (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document. We present efficient approximate inference techniques based on variational methods and an EM algorithm for empirical Bayes parameter estimation. We report results in document modeling, text classification, and collaborative filtering, comparing to a mixture of unigrams model and the probabilistic LSI model.

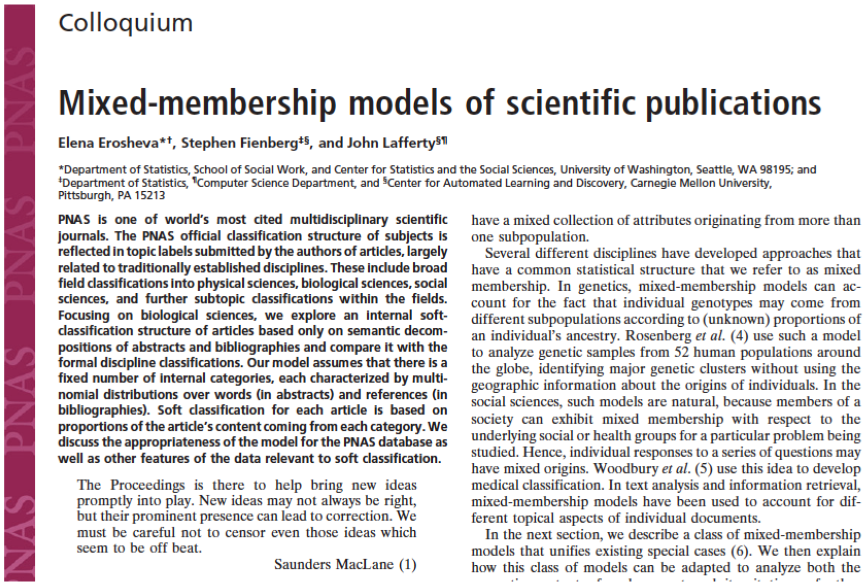


352 MINKA & LAFFERTY UAI 2002

Expectation-Propagation for the Generative Aspect Model

Thomas Minka
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213 USA
`minka@stat.cmu.edu`

John Lafferty
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213 USA
`lafferty@cs.cmu.edu`



Colloquium

Mixed-membership models of scientific publications

Elena Erosheva[*][†], Stephen Fienberg[‡][§], and John Lafferty[§][¶]

[*]Department of Statistics, School of Social Work, and Center for Statistics and the Social Sciences, University of Washington, Seattle, WA 98195; and [‡]Department of Statistics, [¶]Computer Science Department, and [§]Center for Automated Learning and Discovery, Carnegie Mellon University, Pittsburgh, PA 15213

PNAS is one of world's most cited multidisciplinary scientific journals. The PNAS official classification structure of subjects is reflected in topic labels submitted by the authors of articles, largely related to traditionally established disciplines. These include broad field classifications into physical sciences, biological sciences, social sciences, and further subtopic classifications within the fields. Focusing on biological sciences, we explore an internal soft-classification structure of articles based only on semantic decompositions of abstracts and bibliographies and compare it with the formal discipline classifications. Our model assumes that there is a fixed number of internal categories, each characterized by multinomial distributions over words (in abstracts) and references (in bibliographies). Soft classification for each article is based on proportions of the article's content coming from each category. We discuss the appropriateness of the model for the PNAS database as well as other features of the data relevant to soft classification.

> The Proceedings is there to help bring new ideas promptly into play. New ideas may not always be right, but their prominent presence can lead to correction. We must be careful not to censor even those ideas which seem to be off beat.
>
> Saunders MacLane (1)

have a mixed collection of attributes originating from more than one subpopulation.

Several different disciplines have developed approaches that have a common statistical structure that we refer to as mixed membership. In genetics, mixed-membership models can account for the fact that individual genotypes may come from different subpopulations according to (unknown) proportions of an individual's ancestry. Rosenberg *et al.* (4) use such a model to analyze genetic samples from 52 human populations around the globe, identifying major genetic clusters without using the geographic information about the origins of individuals. In the social sciences, such models are natural, because members of a society can exhibit mixed membership with respect to the underlying social or health groups for a particular problem being studied. Hence, individual responses to a series of questions may have mixed origins. Woodbury *et al.* (5) use this idea to develop medical classification. In text analysis and information retrieval, mixed-membership models have been used to account for different topical aspects of individual documents.

In the next section, we describe a class of mixed-membership models that unifies existing special cases (6). We then explain how this class of models can be adapted to analyze both the

Basic model

Data: count matrix $\mathbf{W}$, document lengths $\mathbf{N}$

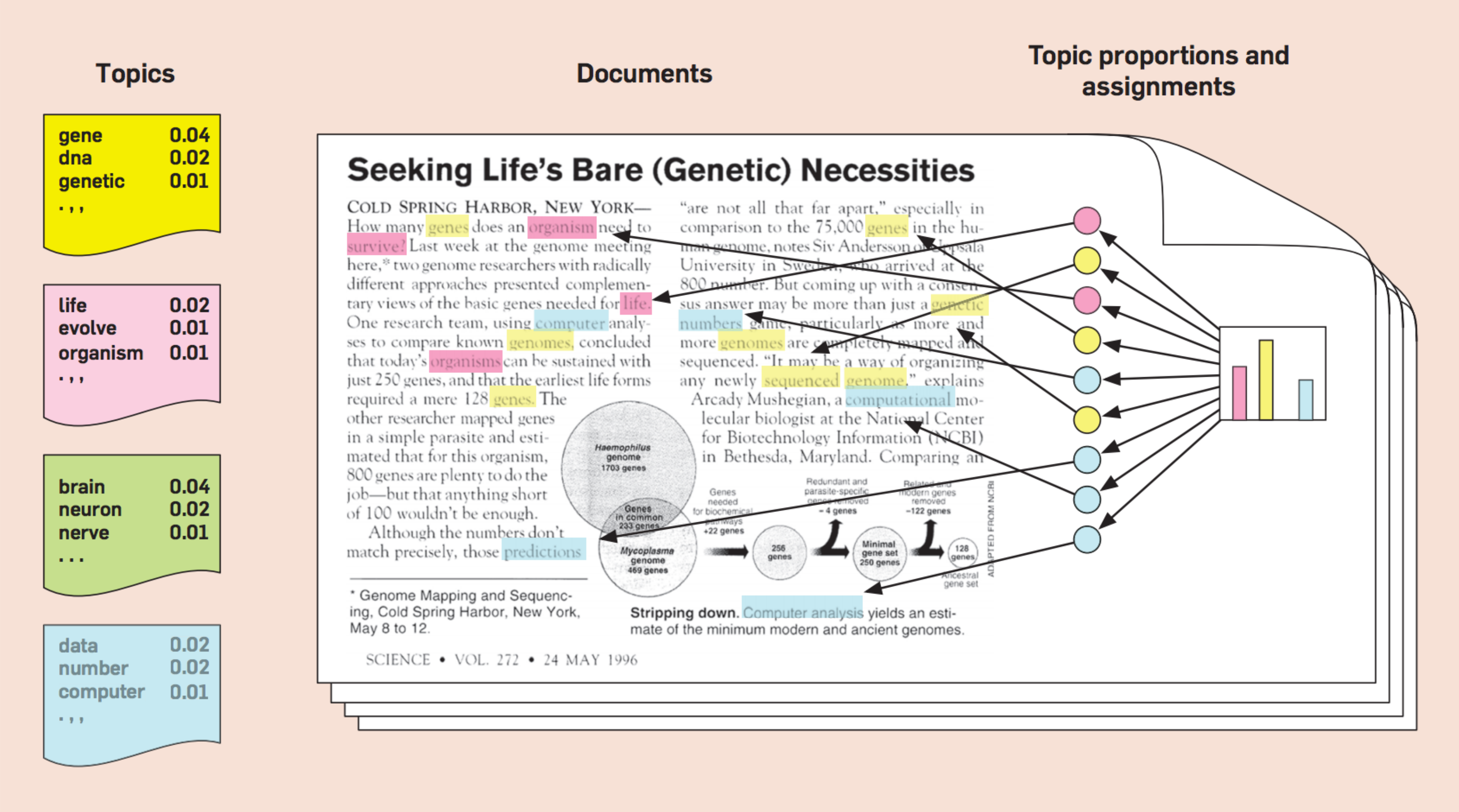
Re-parameterize rate matrix $\boldsymbol{\beta}$ where $\beta_{vk} = \mu_{vk} / \Sigma_v \mu_{vk}$

For document d

- $\boldsymbol{\theta}_d \sim$ Dirichlet ($\boldsymbol{\alpha}$) where $\boldsymbol{\theta}_d$ is a k x 1 vector
- $\mathbf{z}_d \sim$ Multinomial ($\boldsymbol{\theta}_d$,N_d) where $\mathbf{z}_d$ is a k x 1 vector
- $\mathbf{w'}_{dk} \sim$ Multinomial ($\boldsymbol{\beta}_{\cdot k}$, z_{dk})
- $W_{dv} = \Sigma_k w'_{dkv}$

- Place symmetric Dirichlet prior on columns $\boldsymbol{\beta}_{\cdot k}$

Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.


Topics
Documents
Topic proportions and assignments
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
. . .
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

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

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

(Source: Blei, 2012)

Remarks

- Topic vector **L** is entirely unobserved
- Often unclear interpretation of many of the k mixture components
- Evaluation
  - Lists of most frequent words
  - Predictions for $\mathbf{L}^{mis}$ using cross-validation, held-out log-lik

Remarks

- Statistics and machine learning: Classic papers
- Evaluation and interpretation issues
- Bringing causality back

Issues with evaluation standards

- Original papers
  - Interpret most frequent words for most frequent topics
  - Supplemental websites to explore the entire model output
- Follow-up papers
  - Almost exclusively focus on frequency
  - Qualitative and anecdotal evaluations
- We need exhaustive and quantitative evaluations
  - How to quantify topic diversity and coherence?
  - How to maximize interpretability of components/topics?

Hypotheses and MTurk experiments

1. Topic summaries based on frequency and exclusivity are more interpretable than frequency alone
2. Regularizing rates by word yields better estimates of FREX scores than regularizing rates by topic

- However, interpretability is hard to quantify
- We carry out two experiments on Amazon MTurk that enlist human evaluators to execute a comparative analysis of the interpretability of topic summaries

Design of experiments

- Three strategies to generate topic summaries

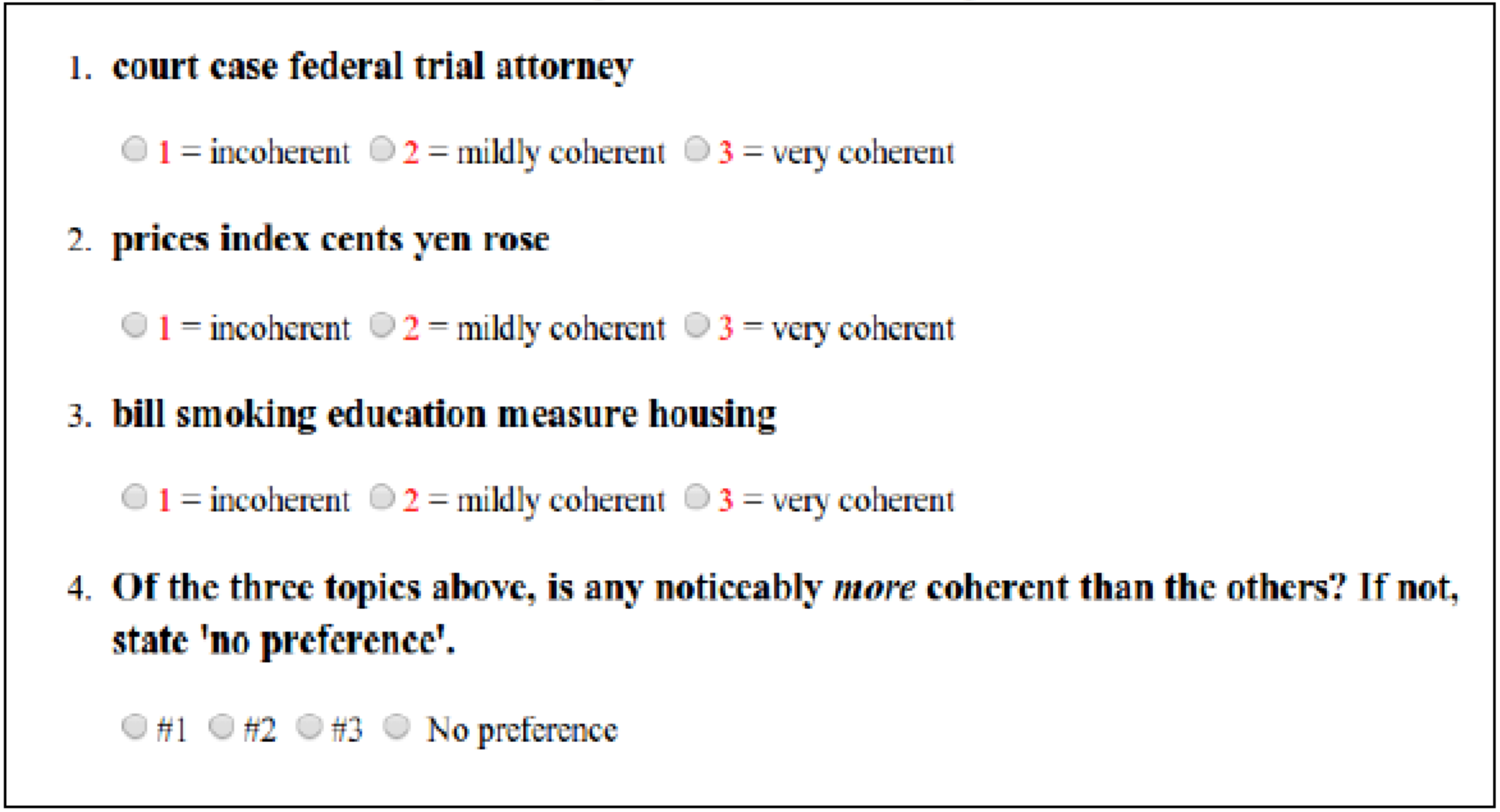
  PCM FREX: Poisson, regularize by word, max FREX

  LDA FREQ: Binomial, regularize by topic, most frequent

  LDA FREX: Binomial, regularize by topic, max FREX (exclusivity estimated by renormalizing rates post inference)

- Two tasks: (i) word intrusion, (ii) topic coherence
- Top-5 words from models with 10, 25, 50, 100 topics
- 400 turkers for each model size; 2 replicates

(b) Topic coherence example



1. **court case federal trial attorney**

   ○ 1 = incoherent ○ 2 = mildly coherent ○ 3 = very coherent

2. **prices index cents yen rose**

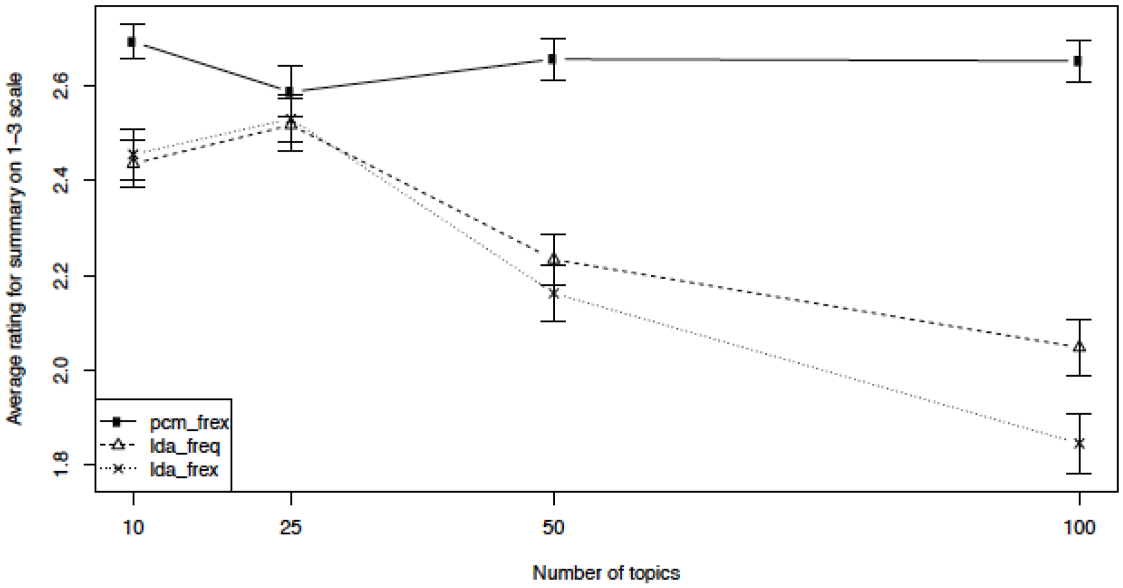
   ○ 1 = incoherent ○ 2 = mildly coherent ○ 3 = very coherent

3. **bill smoking education measure housing**

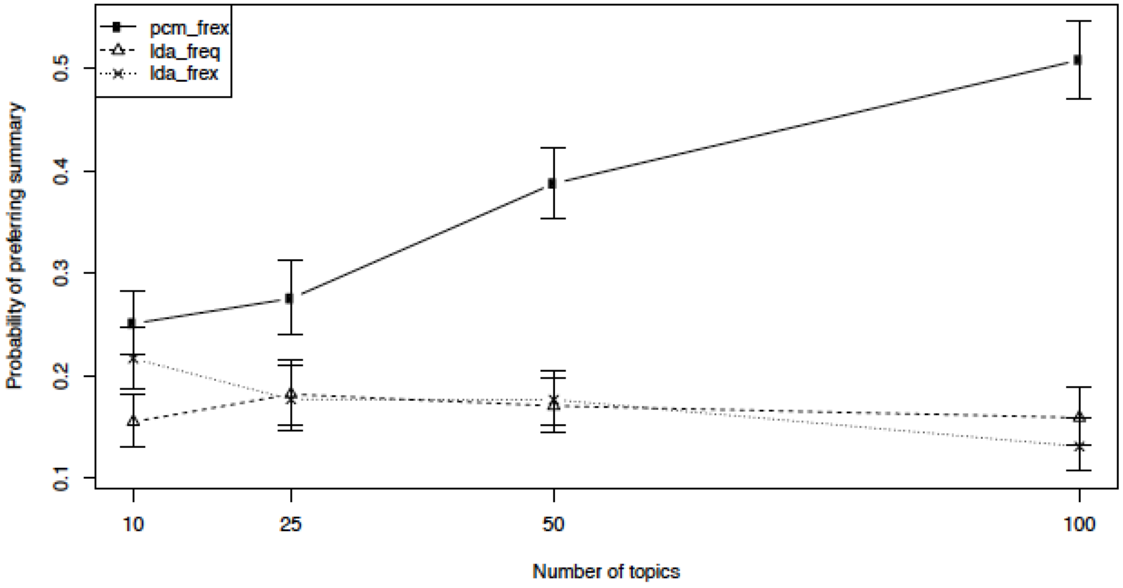
   ○ 1 = incoherent ○ 2 = mildly coherent ○ 3 = very coherent

4. **Of the three topics above, is any noticeably *more* coherent than the others? If not, state 'no preference'.**

   ○ #1 ○ #2 ○ #3 ○ No preference



(a) Average ratings for individual summaries



(b) Relative preference across summary methods

Remarks

- Statistics and machine learning: Classic papers
- Evaluation and interpretation issues
- Bringing causality back

A simple idea

- Make the topic proportions and the rates a function of document level covariates
- JASA paper with Molly Roberts, Brandon Stewart
- R package **stm**, by Molly, Brandon and Dustin Tingley
- What causal questions can we answer leveraging text data? Text as outcome or covariates.

Acknowledgements and pointers

Rubin, **Basse**, Karwa, **Sussman**, Toulis (Harvard), Aral (MIT), Imbens (Stanford), **Christakis** (Yale), **Manski** (Northwestern), Azari, Lambert (Google), **Agarwal**, Xu, Ghosh (LinkedIn).

Some fundamental ideas for causal inference on large networks.
(Donald B Rubin) Soon on PNAS.

Optimal design of experiments in the presence of network-correlated outcomes.
(Guillaume Basse) arXiv paper no. 1507.00803

Optimal design of experiments in the presence of network interference.
(with discussion) Soon on EJS.

Papers on text are on the JASA's "latest articles" page