Claremont GRADUATE UNIVERSITY

Topic Modeling / LDA

Topic Modeling / LDA

David Blei, one of the originators, has expanded to L-LDA, hLDA, …

What are topic models?

- Most popular = LDA (Latent Dirichlet Allocation)
    - Bayesian / Inferential method – *backing into* a generative model
    - Assumption: authors sample from a set of discourse-specific topics
- Vector-based model
    - Bag-of-words approach
    - Words are only "*visible*" (!= *latent*) feature
    - Treated as "random" variable – independent of sequence, linguistic meaning
- Mixed-membership assumption
    - Words can appear in >1 topic (approximates meaning / nuance)
    - Each *article* is a (vector-based) probability / likelihood distribution over *topics*
    - Each *topic* is a (vector-based) probability / likelihood distribution over *words*

Topic Modeling / LDA

Special Issue of journal *Poetics* – December 2013

- **Editors' Introduction: "Topic models: What they are and why they matter."** John Mohr (Sociology, UCSB) and Petko Bogdanov (Computer Science, UCSB)

- **Paper #1: "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of Government Arts Funding in the U.S."** Paul DiMaggio (Sociology, Princeton University), Manish Nag (Sociology, Princeton University), and David Blei (Computer Science, Princeton University).

- **Paper #2: "Differentiating Language-Usage Through Topic Models."** Daniel A. McFarland (Education, Stanford), Daniel Ramage, Jason Chuang, Jeff Heer, Christopher D. Manning (Computer Science, Stanford) and Daniel Jurafsky (Linguistics, Stanford)

Topic Modeling / LDA

LDA Application 1: *Regulatory Debates + Stakeholder Positions*

- We use online comment data and topic modeling strategies to investigate 25 years of regulatory debates around the use of electronic monitoring systems in the U.S. long-haul trucking industry. (source: regulations.gov)
- Electronic monitoring is hugely contentious within the trucking community and has engendered vigorous debate among stakeholders around issues like safety and privacy.
- We use topic models to uncover thematic patterns in public comments on the proposed regulations.
- In addition, by supplementing the model with covariates labeling commenter identity, we identify systematic differences among the interests and evaluative principles that different groups of stakeholders emphasize

Levy and Franklin, "Driving Regulation: Using Topic Models to Examine Political Contention in the U.S. Trucking Industry" (Social Science Computer Review, 2014)

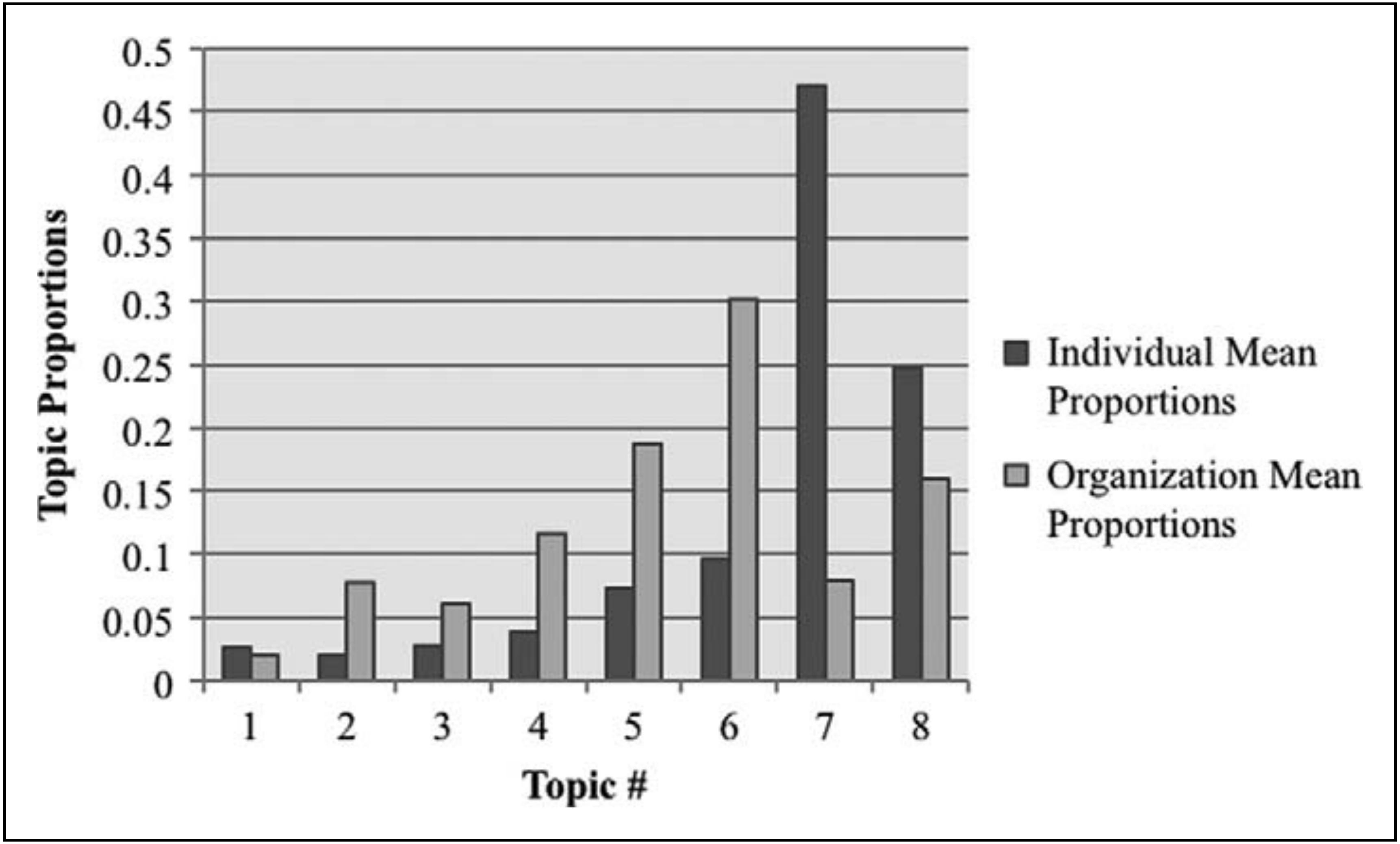
Topic Modeling / LDA

Table 1. Unsupervised Eight-Topic Model. Table 1 displays the 40 highest-ranked words for each topic. Words were “stemmed” in the model (e.g., *propose*, *proposes*, and *proposal* are treated as the same word, *propos*, for analysis) but have been rewritten as full words here for clarity when applicable. α for this model was set to .01.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
sleep	utility	fatigue	eobr	propose	propose	electronic	propose
work	work	duty	carrier	cost	construction	company	day
shift	regulate	study	require	carrier	industry	propose	work
fatigue	operate	safety	system	operate	duty	eobr	rest
day	propose	period	data	safety	safety	address	make
perform	vehicle	vehicle	motor	regulate	period	make	park
schedule	safety	crash	vehicle	require	day	safety	road
night	exempt	accident	compliance	industry	attach	work	home
study	emergency	carrier	hos	increase	work	log	company
operate	power	rest	duty	addition	transport	industry	load
effect	electric	motor	operate	motor	concrete	request	year
period	employee	fhwa	device	dot	limit	support	stop
de	day	report	cost	transport	delivery	pay	week
circadian	require	research	electronic	duty	maximum	september	duty
report	company	highway	safety	fatigue	company	law	attach
test	line	day	status	agency	product	problem	problem
worker	duty	data	log	company	washington	road	sleep
alert	state	operate	technology	benefit	clerk	load	regulate
safety	period	sleep	enforce	impact	road	owner	force
fhwa	worker	work	support	type	december	regulate	run

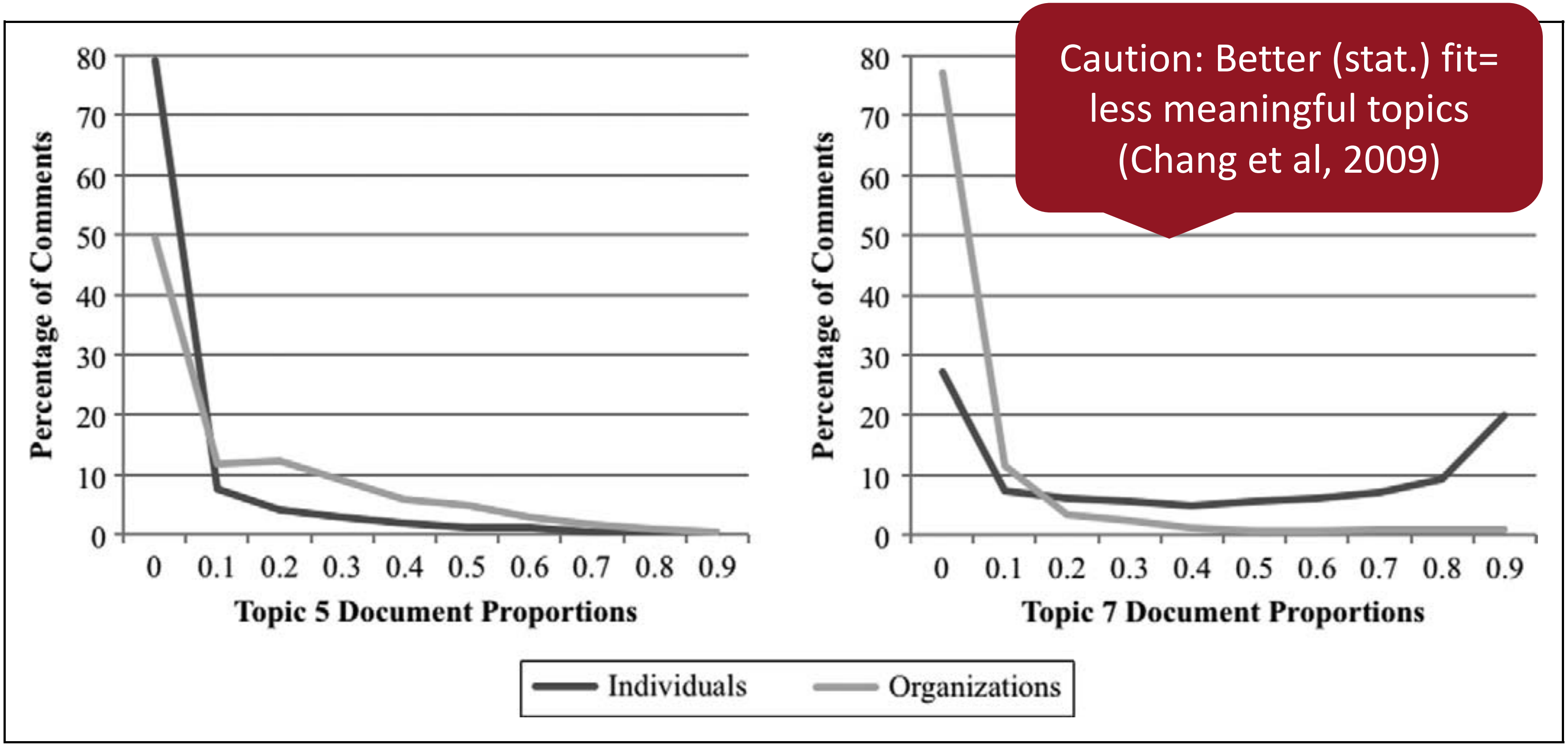
• • •

Topic Modeling / LDA



Topic Modeling / LDA

Scholars can add confidence checks (qual; parameters; data cuts)



Caution: Better (stat.) fit= less meaningful topics (Chang et al, 2009)

Figure 2. Distribution of Topic 5 proportions across comments (*N* = 3,531).
Figure 3. Distribution of Topic 7 proportions across comments (*N* = 3,531).

Topic Modeling / LDA

LDA Application 2: *Arts Funding + Heteroglossia (DiMaggio et al)*

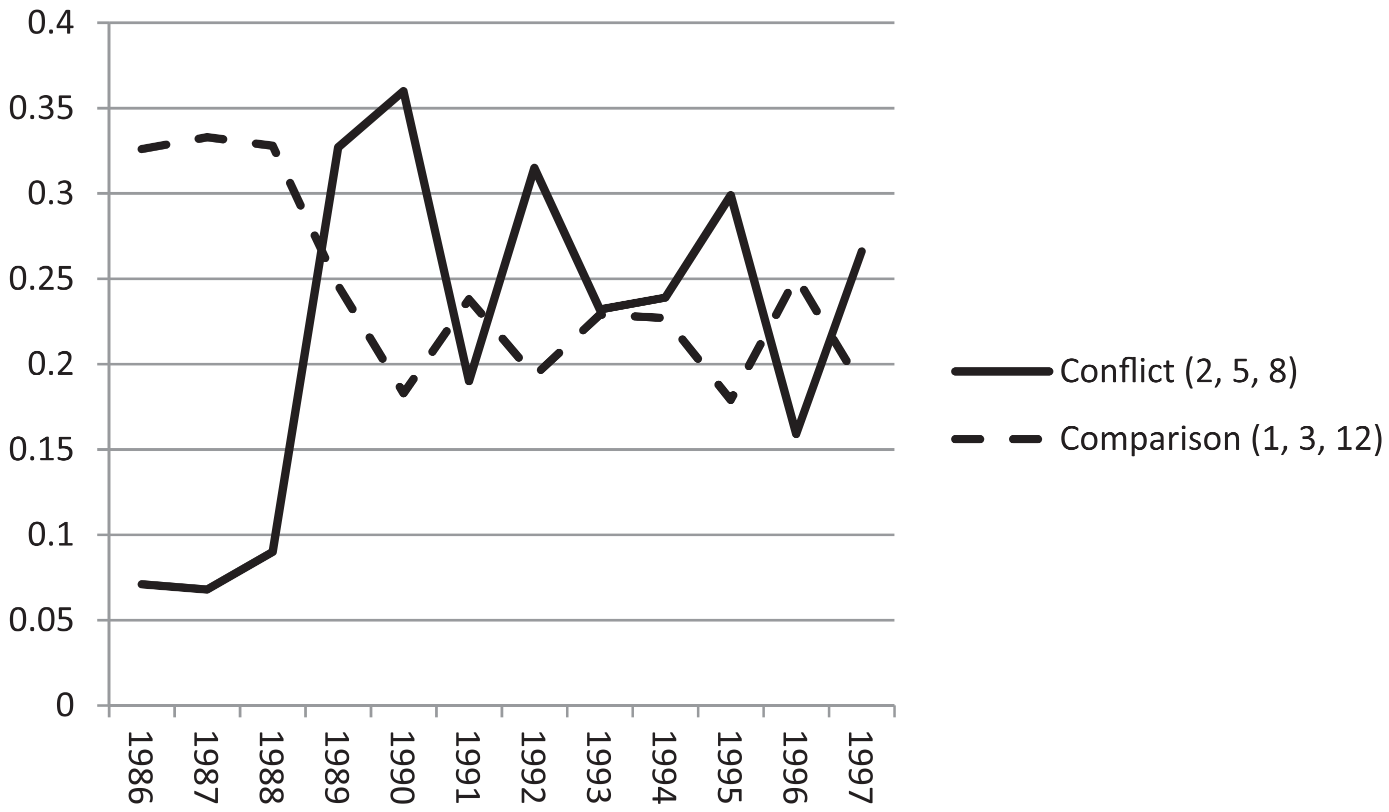


Fig. 4. Percentage of words assigned to conflict frames vs. comparison frames, 1986–1997.

Topic Modeling / LDA

LDA Application 3: *Mapping Cross-Conversations as Hegemony*

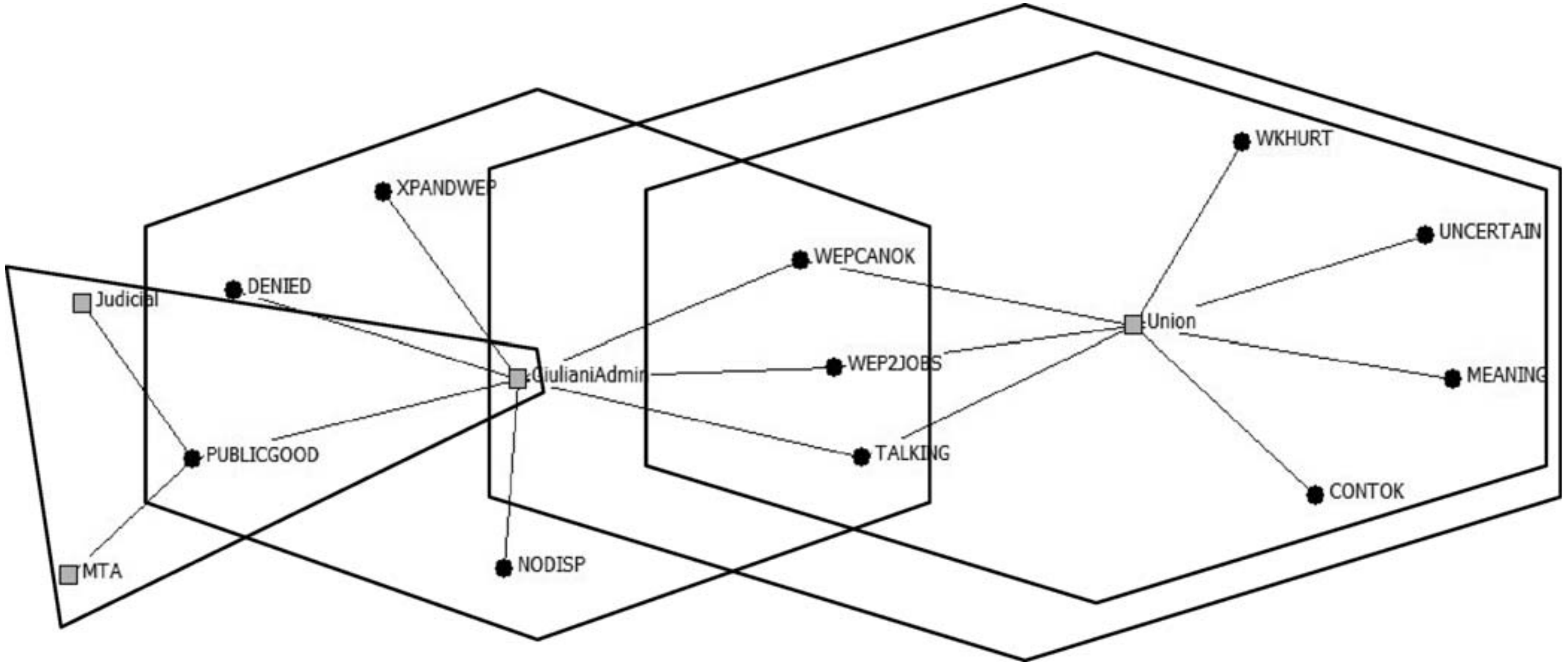


Fig. 2. Bicliques. Actors are represented by squares, claims by circles. The four shapes enclose the four bicliques in the graph. Network graph produced in Netdraw (Borgatti, 2002)

Krinsky, "Dynamics of Hegemony: Mapping Mechanisms of Cultural and Political Power in the Debates over Workfare in New York City, 1993-1999" (Poetics, 2010)

Questions?