DAPPER: Scaling Dynamic Author Persona Topic Model to Billion Word Corpora

Robert Giaquinto, Arindam Banerjee
International Conference on Data Mining 2018

Nov 19, 2018

Introduction

Setting. Dynamic Author Persona (DAP) topic model

- Common narratives in multi-author corpora

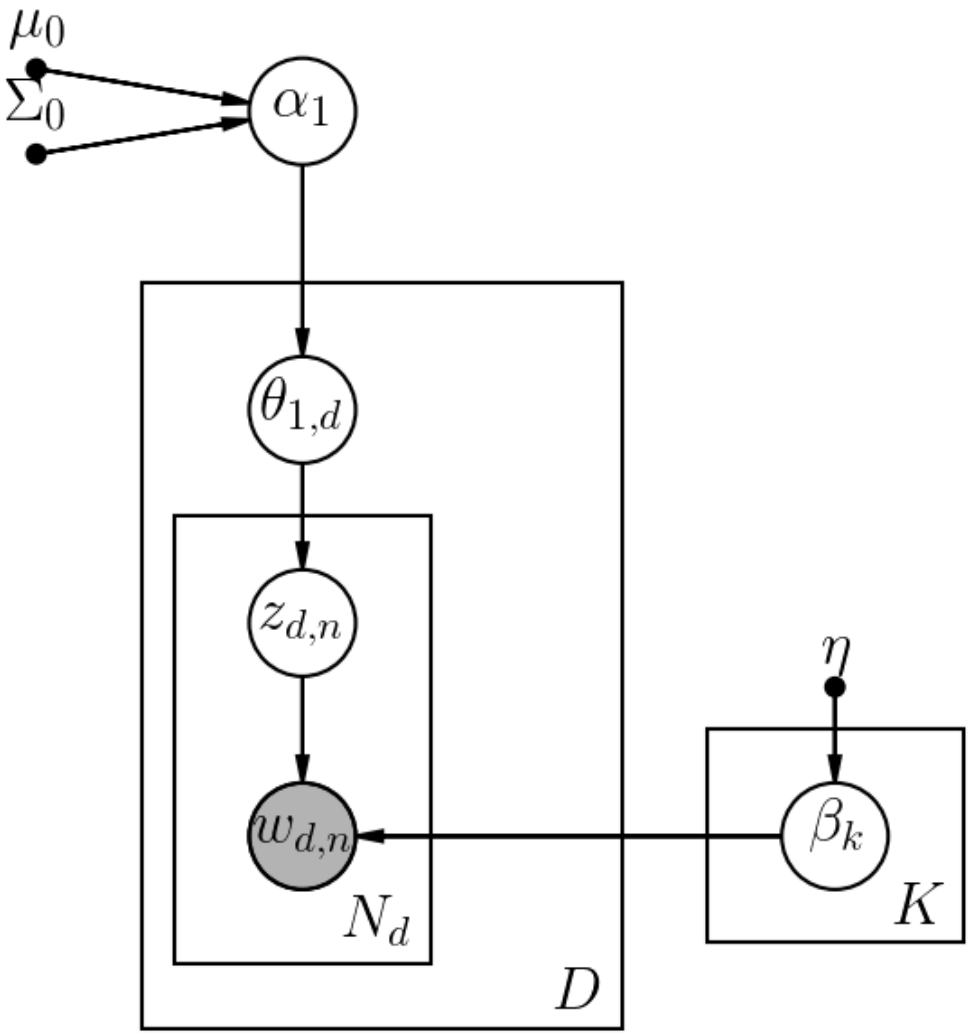
Challenge. Scalability

- Non-conjugate terms

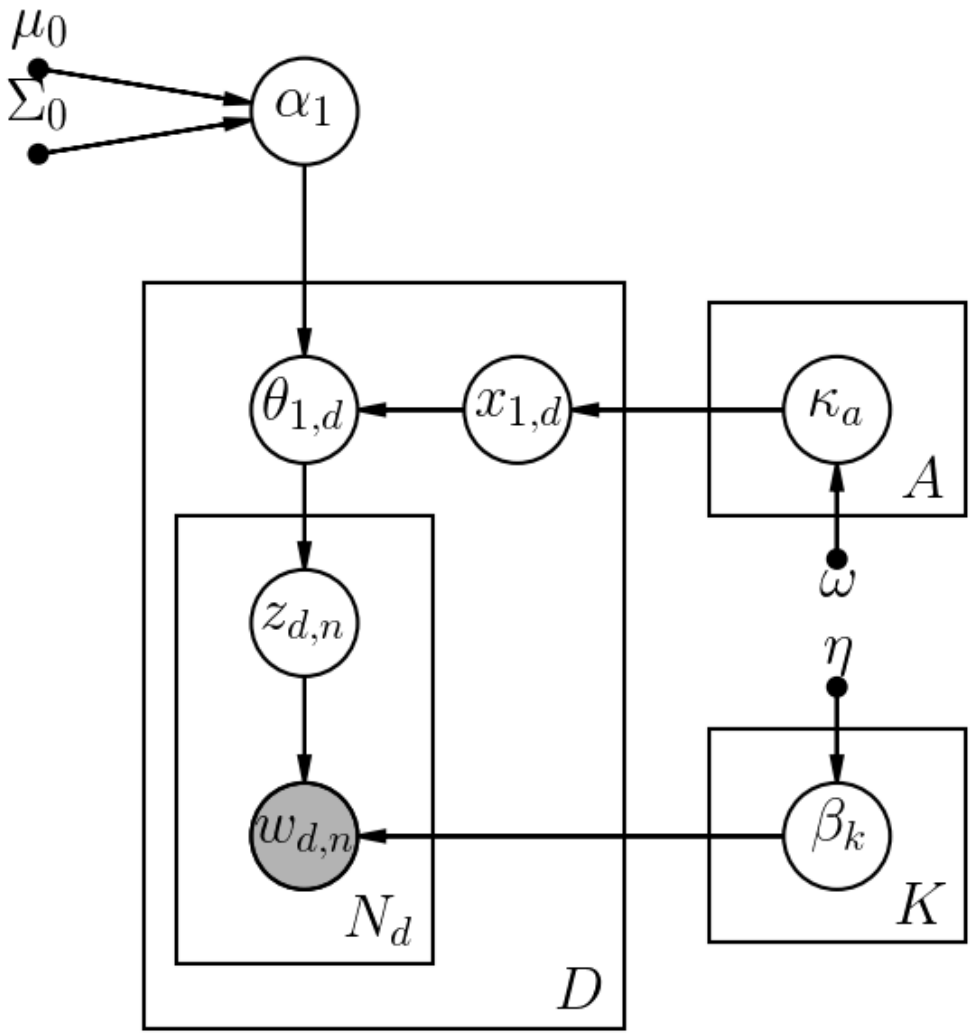
Goals. (1) DAP Performed Exceedingly Rapidly (DAPPER)

(2) Scale to 9M CaringBridge health journals

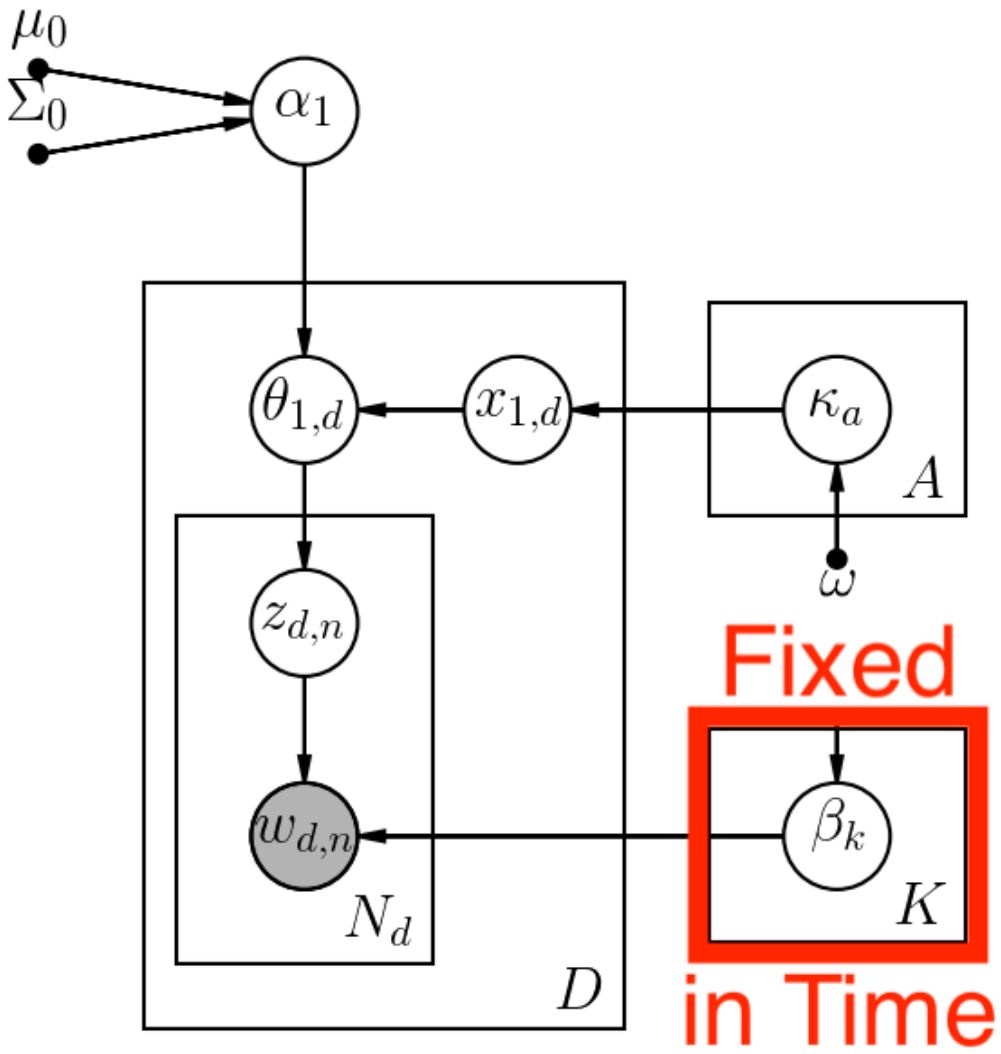
DAP: Similar Parameterization to CTM



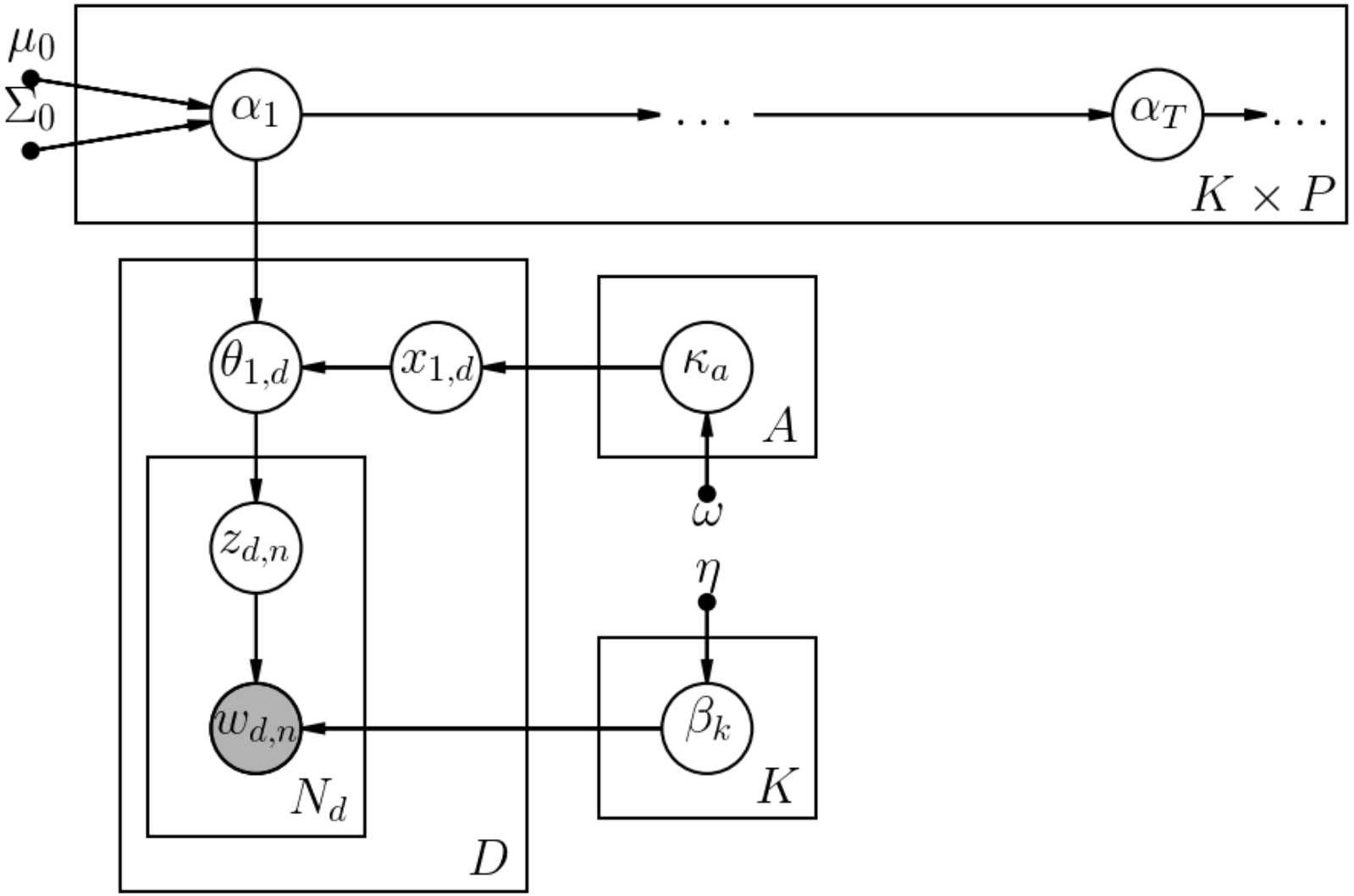
DAP: Assigns Persona to Each Author



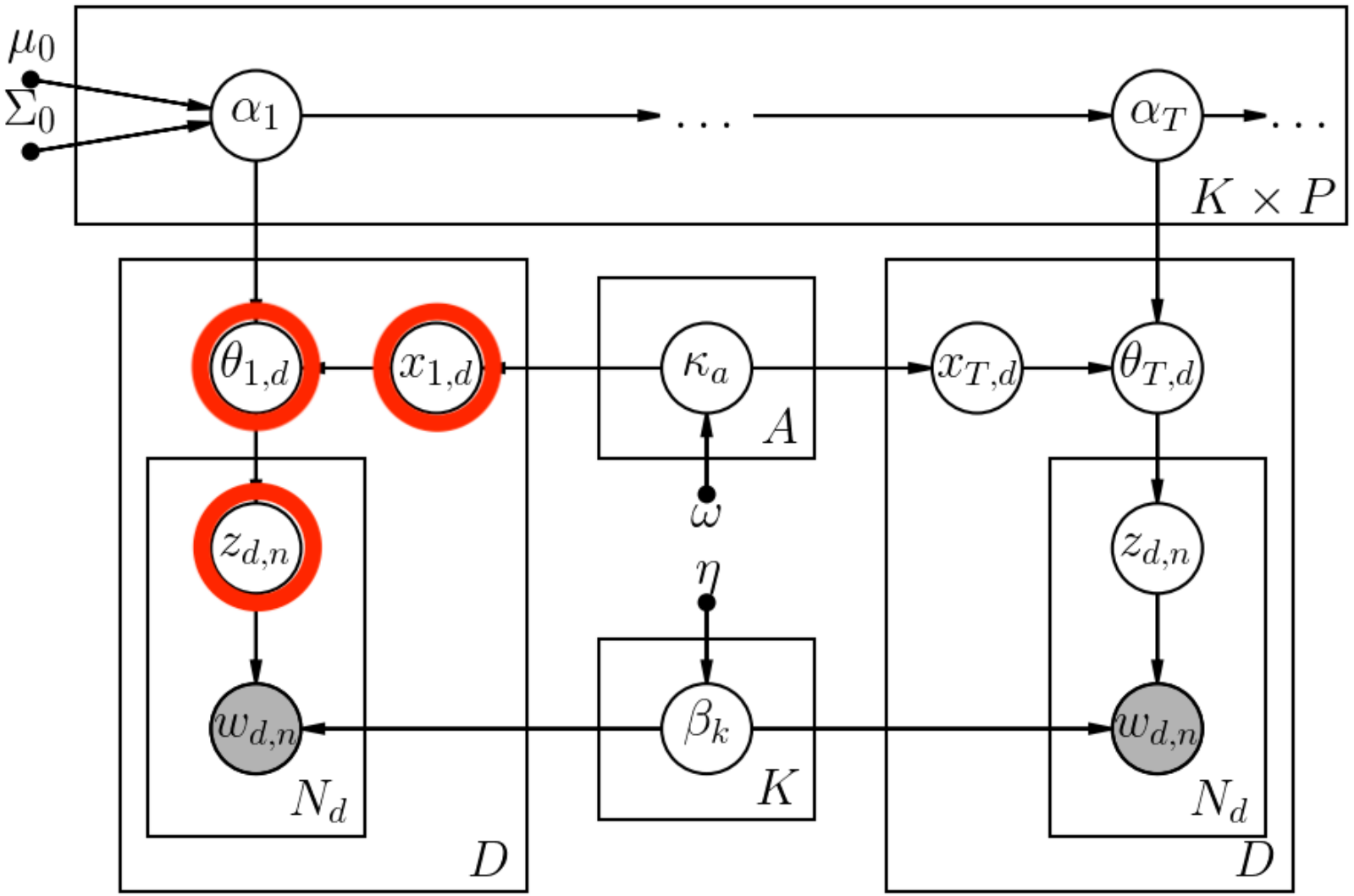
DAP: Words in Topic are Fixed in Time



DAP: Model Topics Over Time



DAP: Non-conjugate Terms

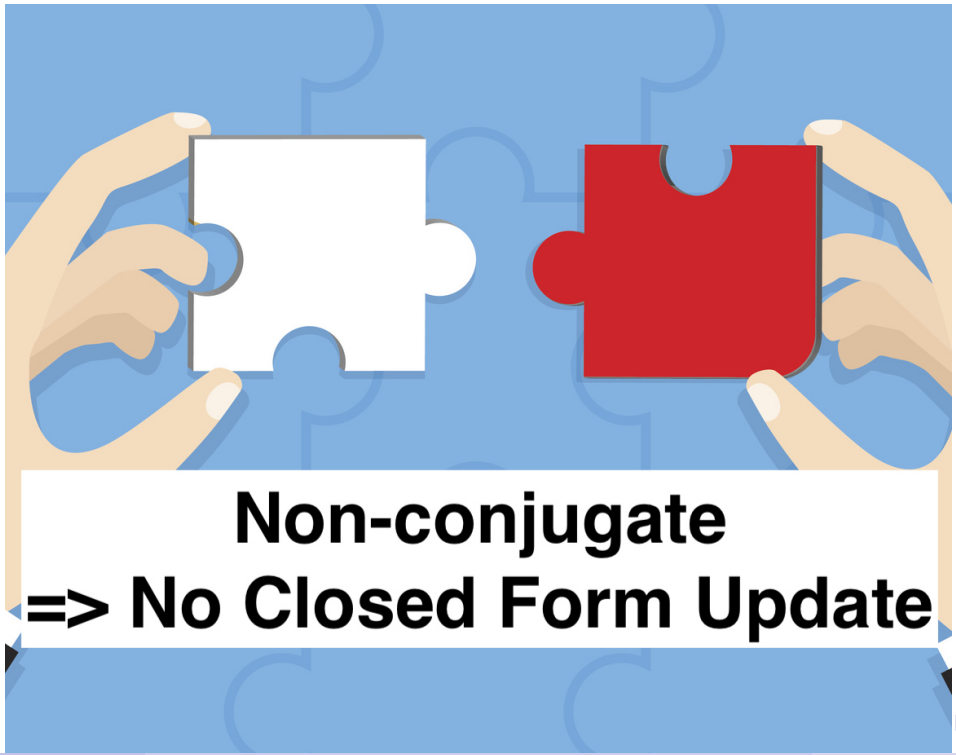


DAP Challenges #1: Inference

"Variational inference is that thing you implement while waiting for your Gibbs sampler to converge."

— David Blei

*assuming the model doesn't have difficult **non-conjugate** terms.



DAP Challenges #2: Regularized Variation Inference

RVI [1] := encourage *certain* solutions

- In DAP, encourage distinct personas

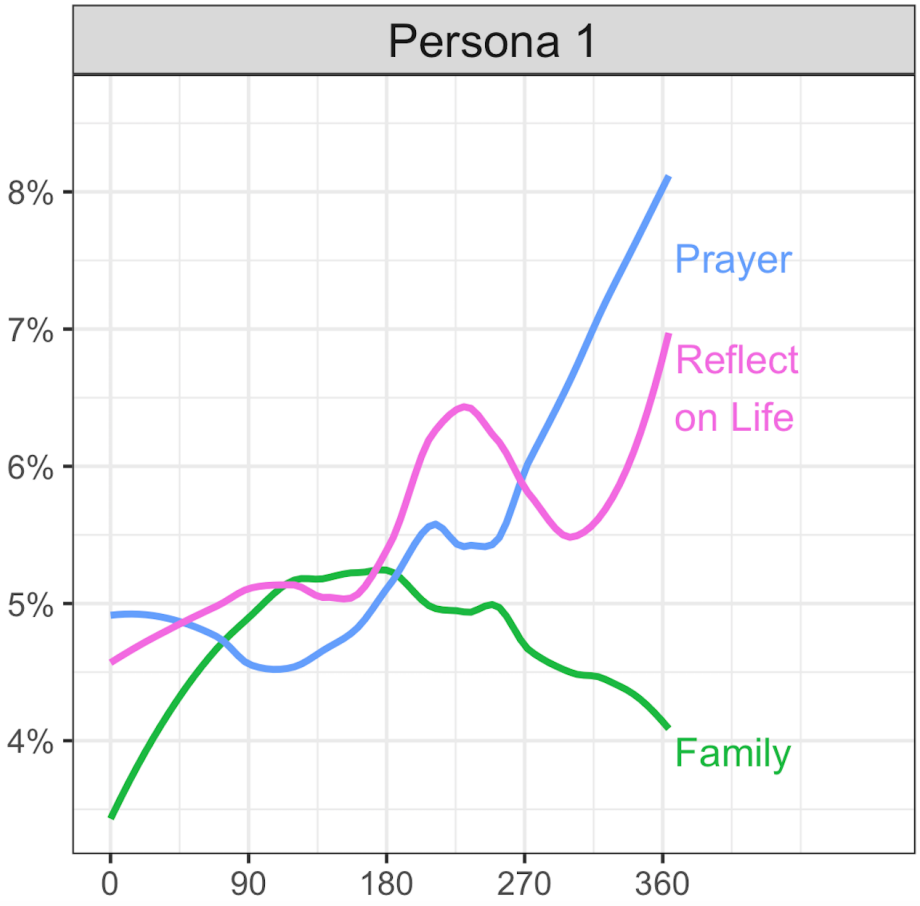


Figure: One persona found by unregularized DAP. Seems fine, right?

Must Keep RVI's Closed Form Update!

Figure: Unregularized DAP: **boring**! Personas are **homogenous**.

Background

- Topic models with non-conjugate terms: Correlated, Dynamic, Continuous Dynamic Topic Models [2, 3, 4, 1]
- Stochastic VI: Conjugate models only [5, 6]
  - Natural gradients needed! [7]
- Black-Box VI: Solves non-conjugate problem [8]
  - Ignores existing closed-form updates :(
  - Like SGVB [9], BBVI builds on [10]
- Conjugate-computation VI: Best of SVI and BBVI [11]
  - Preserves existing closed-form updates
  - Stochastic gradients elsewhere

CVI-based Inference

- [11] develops CVI for regression, Gaussian process, and matrix factorization
- Replace non-conjugates with exponential family approximation (EFA)
- Approximate posterior:

$$q(\mathbf{z}; \lambda_{i+1}) \propto \text{EFA}\langle \text{suff-stats}(\mathbf{z}), \widetilde{\lambda}_i \rangle \cdot p_c(\mathbf{w}, \mathbf{z})$$

  - $p_c(\mathbf{w}, \mathbf{z})$ joint distribution of conjugate terms
- EFA's parameter $\widetilde{\lambda}_i$ updated in closed-form

DAPPER: DAP + CVI

DAP Performed Exceedingly Rapidly (DAPPER)

Fast Inference: Replaces difficult non-conjugates

Distinct Personas: Preserves existing closed-form RVI-based updates

New inference algorithm:

- Maximize Evidence Lower BOund (ELBO) [12, 13, 14]
  - Stochastic mirror descent updates
- Variational EM algorithm:

  E-step: CVI-based updates to local parameters

  M-step: RVI derived updates to global parameters

Better Likelihoods

TABLE II

OVERALL COMPARISON OF MODELS AFTER A MAXIMUM OF 24 HOURS OF TRAINING ON CB-SUBSET AND SM-BLOGS CORPORA. PER-WORD LOG-LIKELIHOODS ARE REPORTED FOR THE TEST CORPUS.

Model	CB-subset	SM-blogs
DAPPER (full batch)	-6.73	-5.76
DAPPER (batch size = 512)	-8.19	-6.31
DAP	-8.84	-7.50
CDTM	-8.81	-8.24
DTM	-9.59	-7.93
LDA	-9.23	-7.79

Figure: DAPPER achieves better likelihoods than similar topic models.

Faster Inference

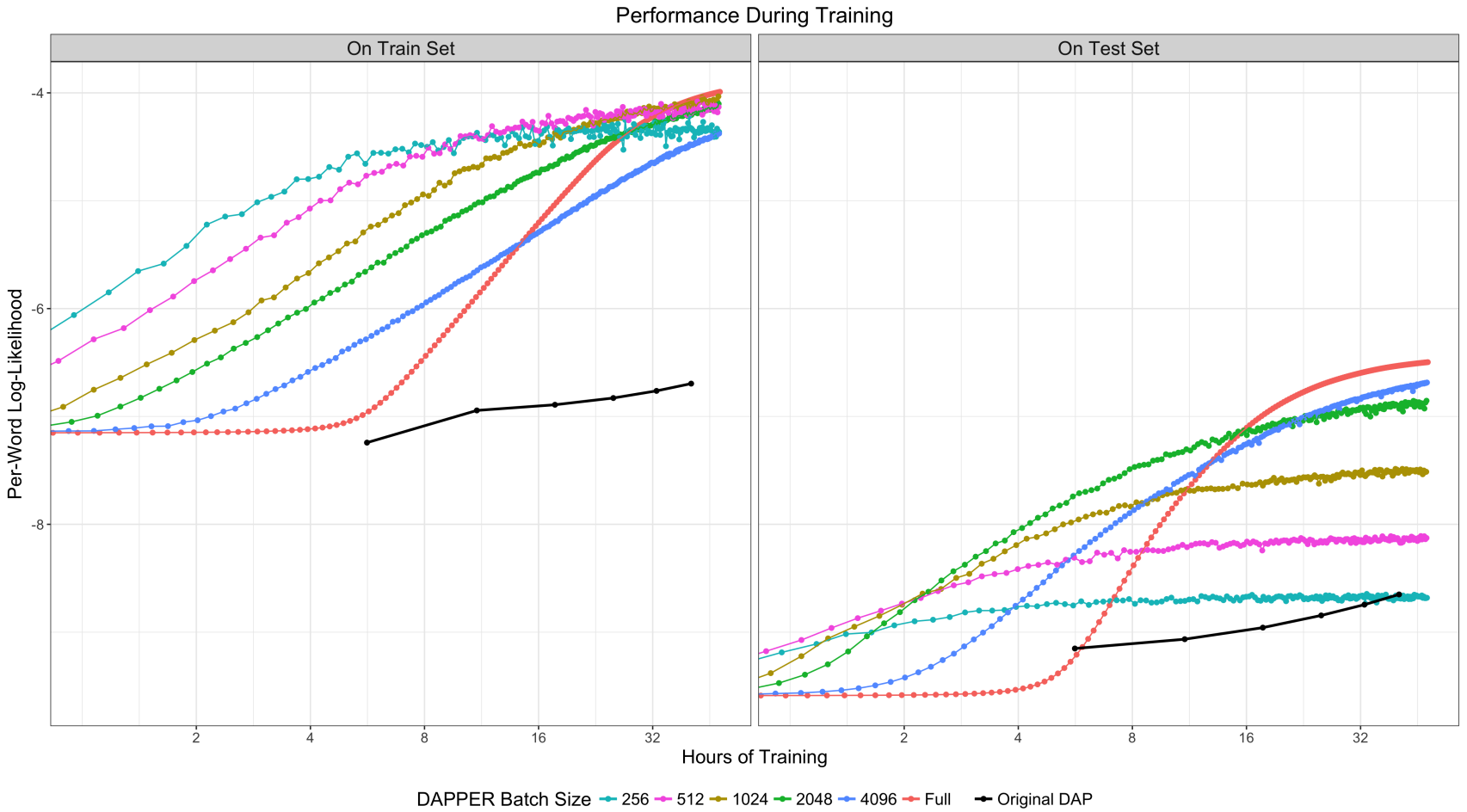


Figure: DAPPER is faster than DAP.

Compelling Shared Narratives

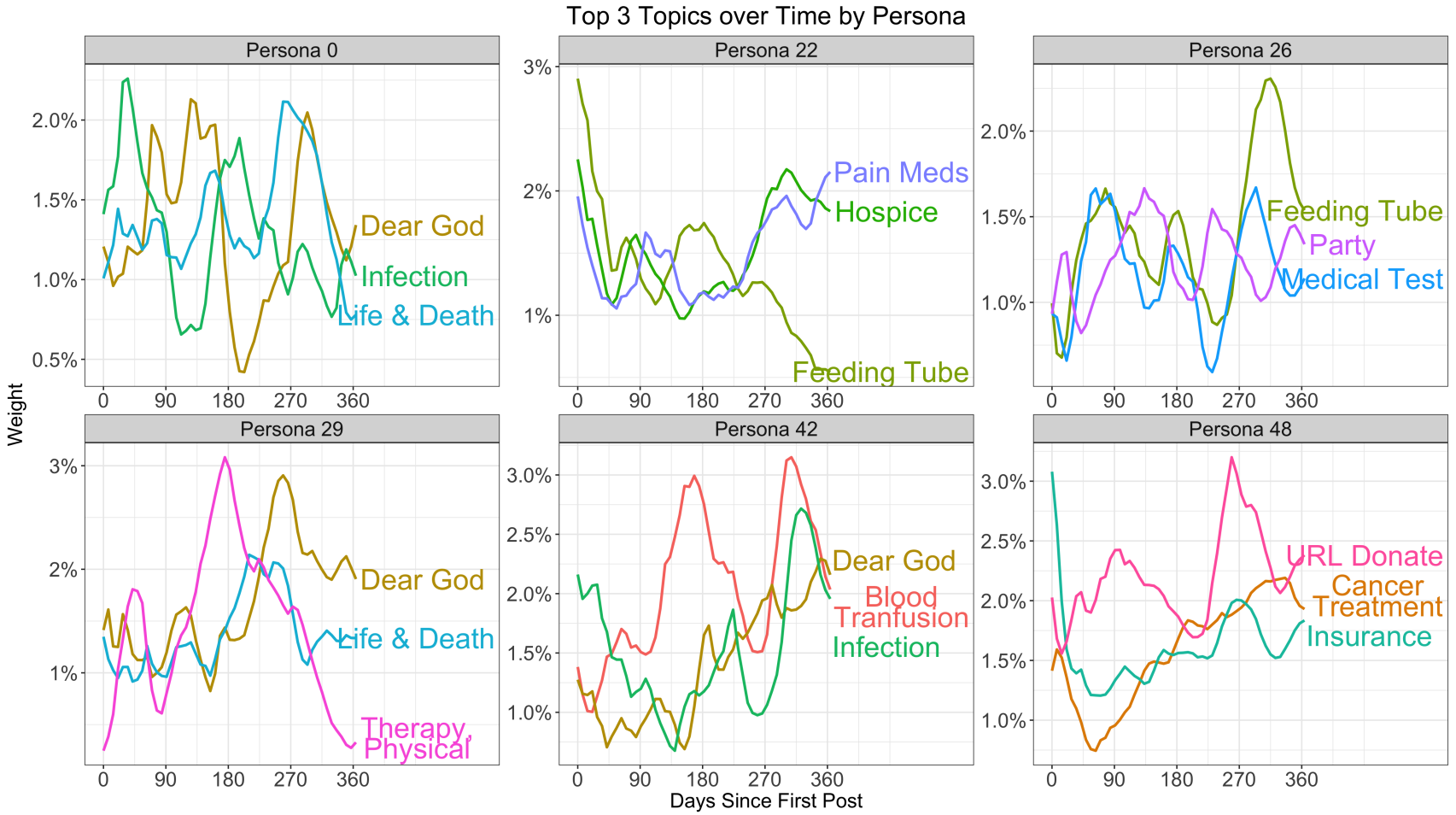


Figure: DAPPER finds personas that reflect common "health journeys" shared by CaringBridge users. Each plot shows how the 3 topics most associated with a persona evolve over time. Topic labels are hand-selected based on top words in each topic.

Conclusion

- New inference algorithm for DAPPER
  - Better models, faster
- CVI-based inference for topic models
- Integrates cleanly with RVI
- Stochastic VI despite non-conjugacy
- Demonstrated scalability

Questions?

Thank you!!!

Code + Slides:
`github.com/robert-giaquinto/dapper`

Appendix: Words in CaringBridge Topics

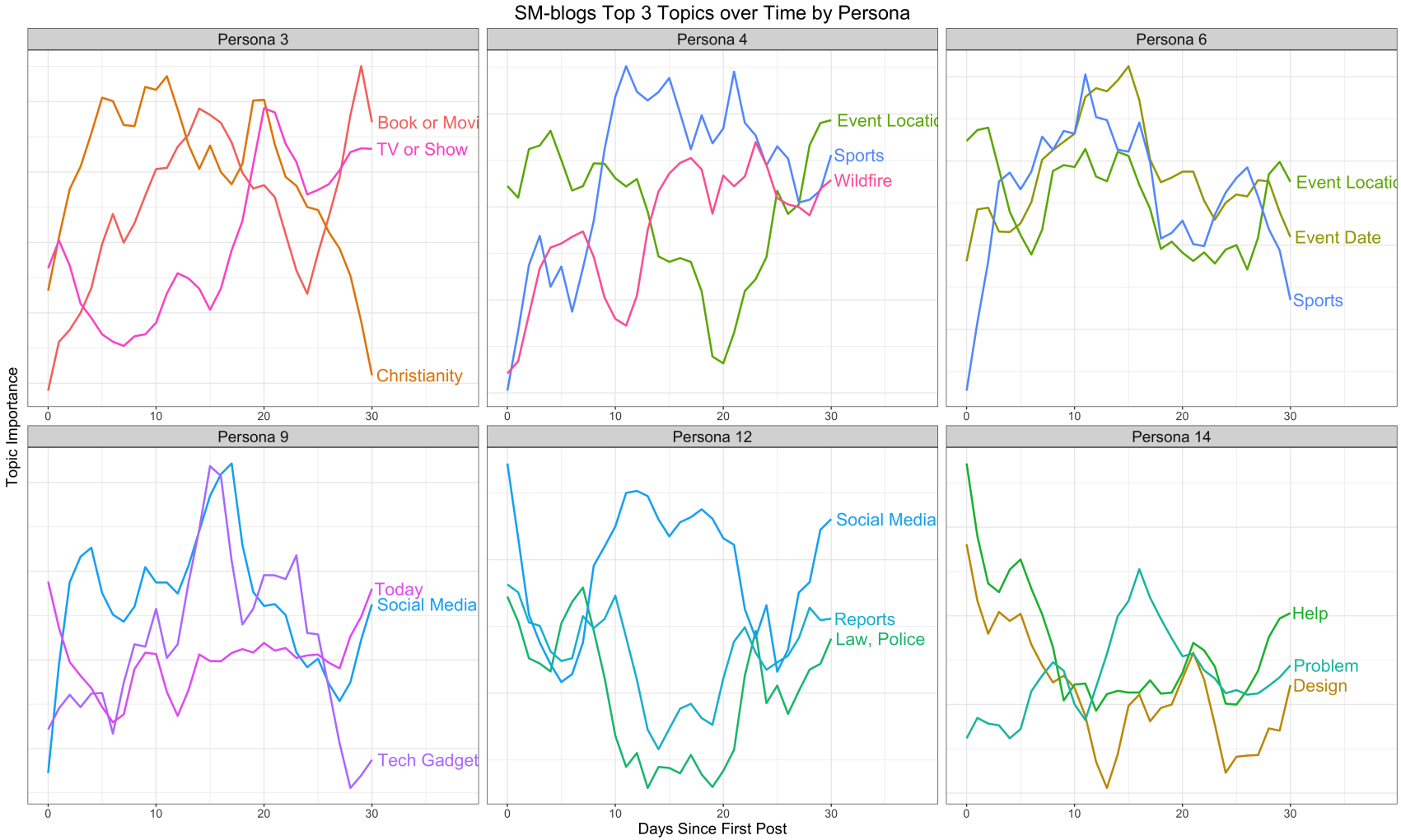
TABLE V
TOP EIGHT WORDS ASSOCIATED WITH THE MOST PREVALENT TOPICS FOUND BY THE DAPPER MODEL TRAINED ON THE FULL CARINGBRIDGE DATASET. TOPIC LABELS ARE SELECTED MANUALLY IN ORDER TO AID REFERENCE WITH FIGURE 3. THE WORDS _DOLLARS_, _NAME_, AND _URL_ REFER TO THE RESULT OF TEXT PRE-PROCESSING STEPS FOR CAPTURING COMMON PATTERNS LIKE THE DOLLAR AMOUNTS, ANONYMIZED NAMES, AND WEBSITE URLS, RESPECTIVELY.

Infection	Life & Death	Dear God	Pain Meds	Friend, Memories	Feeding Tube	Party	Medical Test
infection	life	god	cause	beautiful	tube	school	dr
fluid	live	lord	pain	friend	feed	birthday	test
lung	child	praise	medication	celebrate	breathe	fun	scan
remove	world	peace	brain	_name_	weight	_name_	result
procedure	others	pray	dose	card	oxygen	party	drug
chest	moment	trust	increase	memory	gain	aunt	mri
pressure	fear	father	level	flower	rate	game	ct
antibiotic	choose	joy	steroid	dance	ventilator	kid	liver

Therapy, Physical	Blood Tranfusion	Child	Hospice	Cancer Treatment	URL Donate	ICU	Insurance
therapy	blood	play	mom	cancer	_url_	icu	provide
physical	count	daddy	dad	radiation	_dollars_	brain	medical
leg	cell	mommy	visit	tumor	donate	monitor	information
therapist	low	girl	visitor	oncologist	money	stable	disease
arm	bone	boy	hospice	surgeon	benefit	wound	insurance
rehab	transplant	_name_	nursing	chemotherapy	en	neck	condition
foot	white	little	facility	breast	donation	doctor	regard
pt	marrow	cute	phone	biopsy	ha	unit	decision

Figure: Words in each topic.

Selected Personas of Signal Media Corpus



Selected Topics of Signal Media Corpus

TABLE VI
TOP WORDS ASSOCIATED WITH THE MOST PREVALENT TOPICS FOUND BY THE DAPPER MODEL TRAINED ON THE FULL SM-BLOGS CORPUS. TOPIC LABELS ARE SELECTED MANUALLY IN ORDER TO AID REFERENCE WITH FIGURE 4.

Christianity	Tech Gadgets	Sports	TV or Show	Social Media	Wildfire	Book or Movie	Problem
life	apple	game	show	post	area	story	may
god	feature	season	live	share	water	book	change
word	phone	play	night	free	fire	movie	number
heart	device	team	star	comment	north	film	deal
church	user	against	news	video	land	full	result
pope	plus	player	series	photo	west	character	allow
father	update	football	special	click	near	author	note
christian	version	yard	tv	link	south	title	problem
son	iphone	coach	award	facebook	california	director	within
lord	app	ball	fan	twitter	local	writer	step

Law, Police	Today	Event Date	Design	Reports	Event Location	Help	Systems & Security
case	new	_year_	include	report	city	use	service
law	best	september	add	plan	event	need	system
police	today	th	design	issue	center	help	data
court	next	watch	create	member	st	different	technology
claim	open	online	large	public	street	small	customer
charge	month	date	image	accord	art	easy	network
officer	late	october	view	continue	park	save	access
against	hour	august	base	national	friday	important	solution
act	york	episode	space	action	sept	type	security
judge	check	july	form	official	monday	choose	provide

Figure: Words corresponding to each topic from results on the Signal Media Blogs.

Effect of Hyperparameters

Number of Topics	Personas	Batch Size: 256	512	1024	2048	Full Batch
25	15	-6.46	-6.26	-6.17	-6.16	-5.65
25	25	-6.47	-6.26	-6.21	-6.23	-5.68
25	50	-6.55	-6.35	-6.34	-6.35	-5.71
50	15	-6.47	-6.30	-6.11	-6.16	**-4.97**
50	25	-6.50	-6.31	-6.30	-6.39	**-5.07**
50	50	-6.61	-6.38	-6.38	-6.52	-5.47
75	15	-6.87	-6.59	-6.46	-6.53	**-5.08**
75	25	-6.85	-6.61	-6.55	-6.61	-5.42
75	50	-6.96	-6.80	-6.79	-6.95	-6.00
100	15	-7.14	-6.93	-6.90	-6.98	-5.67
100	25	-7.07	-6.90	-6.91	-7.02	-5.99
100	50	-7.33	-7.17	-7.16	-7.31	-6.72

Figure: Shows per-word log-likelihoods for DAPPER trained on the Signal Media Blogs data for varying hyperparameters.

Description of DAPPER's Parameters

Parameter	Variational	Description
$\mathbf{w}_{t,d}$		Words in document d_t
$\mathbf{z}_n$	$\boldsymbol{\phi}_n$	Assigns word n to a topic
$\boldsymbol{\theta}_{t,d}$	$\boldsymbol{\gamma}_{t,d}$	Topic distribution for document d_t
$\mathbf{v}_{t,d}$	$\hat{\mathbf{v}}_{t,d}$	Covariance between topics for d_t
μ_0		Prior for mean of $\boldsymbol{\alpha}_0$
Σ_0		Prior for covariance of $\boldsymbol{\alpha}_0$
$\boldsymbol{\alpha}_{t,p}$	$\hat{\boldsymbol{\alpha}}_{t,p}$	Persona p's topic distribution
$\boldsymbol{\Sigma}_t$	$\hat{\boldsymbol{\Sigma}}_t$	Covariance in topic distributions
ω		Prior for $\boldsymbol{\kappa}_a$
$\boldsymbol{\kappa}_a$	$\boldsymbol{\delta}_a$	Author a's personas distribution
$\mathbf{x}_{d,t}$	$\boldsymbol{\tau}_{t,d}$	Assigns author of d_t to a persona
η		Prior parameter for β_k
$\boldsymbol{\beta}_k$	$\boldsymbol{\lambda}_k$	$\forall k$ distribution over words


Robert Giaquinto and Arindam Banerjee.
Topic Modeling on Health Journals with Regularized Variational Inference.
AAAI, pages 3021–3028, 2018.


John D. Lafferty and David M. Blei.
Correlated Topic Models.
Advances in Neural Information Processing Systems 18, pages 147–154, 2006.


David M Blei and John D Lafferty.
Dynamic Topic Models.
International Conference on Machine Learning, pages 113–120, 2006.


Chong Wang, David Blei, and David Heckerman.
Continuous Time Dynamic Topic Models.
Proc of UAI, pages 579–586, 2008.


Matthew D Hoffman, David M Blei, and Francis Bach.
Online Learning for Latent Dirichlet Allocation.
Advances in Neural Information Processing Systems, 23:1–9, 2010.


Matt Hoffman, David M. Blei, Chong Wang, and John Paisley.
Stochastic Variational Inference.
Journal of Machine Learning Research, 14:1303–1347, 2012.


Shun-Ichi Amari.
Natural gradient works efficiently in learning.
Neural computation, 10(2):251–276, 1998.


Rajesh Ranganath, Sean Gerrish, and David M Blei.
Black Box Variational Inference.
Aistats, 33, 2013.


Diederik P Kingma and Max Welling.
Auto-Encoding Variational Bayes.
Proceedings of the 2nd International Conference on Learning Representations (ICLR), pages 1–14, December 2014.


John Paisley, David Blei, and Michael Jordan.
Variational Bayesian Inference with Stochastic Search.
Icml, (2000):1367—-1374, 2012.


Mohammad Emtiyaz Khan and Wu Lin.
Conjugate-Computation Variational Inference : Converting Variational Inference in Non-Conjugate Models to Inferences in Conjugate Models.
Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 54:878–887, 2017.


Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul.
Introduction to variational methods for graphical models.
Machine Learning, 37(2):183–233, 1999.


Martin J. Wainwright and Michael I. Jordan.
Graphical Models, Exponential Families, and Variational Inference.
Foundations and Trends® in Machine Learning, 1(1–2):1–305, 2007.


David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe.
Variational Inference: A Review for Statisticians.
Journal of the American Statistical Association, 112(518):859–877, 2017.