A Appendix: sufficiency of l and full conditional for b

Recall that the one-step-ahead conditional probability mass function in a Pólya sequence taking values in $\mathbb{N}$ with concentration parameter α and base probability mass function $\mathbf{\Psi}$ is

$$p(z_i \mid z_{i-1}, .., z_1, \mathbf{\Psi}) = \sum_{j=1}^{i-1} \frac{1}{i-1+\alpha} \mathbb{1}_{z_j}(z_i) + \frac{\alpha}{i-1+\alpha} \Psi_{z_i}. \tag{30}$$

Introducing the random variable

$$b_i \sim \text{Ber}\left(\frac{\alpha}{i-1+\alpha}\right) \tag{31}$$

we can express the one-step-ahead conditional distribution as

$$p(z_i \mid z_{i-1}, .., z_1, b_i, \mathbf{\Psi}) = \mathbb{1}_{b_i=0} \sum_{j=1}^{i-1} \frac{1}{i-1} \mathbb{1}_{z_j}(z_i) + \mathbb{1}_{b_i=1} \Psi_{z_i}. \tag{32}$$

The joint probability mass function for $\boldsymbol{z} \mid \boldsymbol{b}, \mathbf{\Psi}$ is then

$$p(\boldsymbol{z} \mid \boldsymbol{b}, \mathbf{\Psi}) = \prod_{i=1}^{N} p(z_i \mid z_{i-1}, .., z_1, \boldsymbol{b}, \mathbf{\Psi}) = \prod_{i=1}^{N} \left[\sum_{j=1}^{i-1} \mathbb{1}_{b_i=0} \mathbb{1}_{z_j}(z_i) + \mathbb{1}_{b_i=1} \Psi_{z_i} \right]. \tag{33}$$

Note that $\mathbb{1}_{b_i=0} = 1 \iff \mathbb{1}_{b_i=1} = 0$ and vice versa. Thus each term in the product for $\boldsymbol{z} \mid \boldsymbol{b}, \mathbf{\Psi}$ only has one component, and we may express $\boldsymbol{z} \mid \boldsymbol{b}, \mathbf{\Psi}$ as

$$p(\boldsymbol{z} \mid \boldsymbol{b}, \mathbf{\Psi}) = \underbrace{\prod_{\substack{i=1 \\ b_i \neq 1}}^{N} \sum_{j=1}^{i-1} \frac{1}{i-1} \mathbb{1}_{z_j}(z_i)}_{\text{doesn't enter posterior}} \prod_{\substack{i=1 \\ b_i=1}}^{N} \prod_{k=1}^{\infty} \Psi_k^{\mathbb{1}_k(z_i)} \tag{34}$$

where we have re-expressed the probability mass function of $\mathbf{\Psi}$ in a form that emphasizes conjugacy. Thus for any prior, the posterior will only depend on the likelihood of the values of z_i for which $b_i = 1$. The sufficient statistic is

$$l_k = \sum_{\substack{i=1 \\ b_i=1}}^{N} \mathbb{1}_{z_i=k}. \tag{35}$$

Next, for a given $i' \in \{1, .., N\}$, we can calculate the posterior of a component $b_{i'}$ as

$$\mathbb{P}(b_{i'} = 1 \mid \boldsymbol{z}, \mathbf{\Psi}, \boldsymbol{b}_{-i'}) \propto \left(\frac{\alpha}{i'-1+\alpha}\right) \prod_{\substack{i=1 \\ b_i \neq 1}}^{N} \sum_{j=1}^{i-1} \frac{1}{i-1} \mathbb{1}_{z_j}(z_i) \prod_{\substack{i=1 \\ b_i=1}}^{N} \Psi_{z_i} \tag{36}$$

$$\propto \alpha \Psi_{z_{i'}} \tag{37}$$

$$\mathbb{P}(b_{i'} = 0 \mid \boldsymbol{z}, \mathbf{\Psi}, \boldsymbol{b}_{-i'}) \propto \left(\frac{i'-1}{i'-1+\alpha}\right) \prod_{\substack{i=1 \\ b_i \neq 1}}^{N} \sum_{j=1}^{i-1} \frac{1}{i-1} \mathbb{1}_{z_j}(z_i) \prod_{\substack{i=1 \\ b_i=1}}^{N} \Psi_{z_i} \tag{38}$$

$$\propto \sum_{i=1}^{i'-1} \mathbb{1}_{z_i}(z_i') \tag{39}$$

where we have divided both expressions by

$$\frac{1}{i'-1+\alpha} \prod_{\substack{i=1 \\ b_i \neq 1 \\ i \neq i'}}^{N} \sum_{j=1}^{i-1} \frac{1}{i-1} \mathbb{1}_{z_j}(z_i) \prod_{\substack{i=1 \\ b_i=1 \\ i \neq i'}}^{N} \Psi_{z_i} \tag{40}$$

which is constant with respect to $b_{i'}$. Note that full conditionally, we have $b_i \perp\!\!\!\perp b_{i'}$ for $i \neq i'$. This gives the desired expressions and concludes the derivation.

B Appendix: full conditional for Ψ

Before proceeding with the derivation, we first comment on Proposition 1 and differences between the GEM distribution and Dirichlet process, which otherwise appear superficially similar. The GEM distribution $\Psi^{\mathrm{GEM}} \sim \mathrm{GEM}(\gamma)$ is defined as

$$\Psi_k^{\mathrm{GEM}} = \varsigma_k \prod_{i=1}^{k-1}(1-\varsigma_i) \qquad\qquad \varsigma_k^{\mathrm{GEM}} \sim \mathrm{Beta}(1,\gamma). \tag{41}$$

On the other hand, a Dirichlet process $\Psi^{\mathrm{DP}} \sim \mathrm{DP}(\gamma, F)$ is defined as

$$\boldsymbol{\Psi}^{\mathrm{DP}} = \sum_{k=1}^{\infty} \pi_k \delta_{\vartheta_k} \qquad \vartheta_k \sim F \qquad \pi_k = \varsigma_k \prod_{i=1}^{k-1}(1-\varsigma_i) \qquad \varsigma_k \sim \mathrm{Beta}(1,\gamma). \tag{42}$$

From a Bayesian perspective, this extra stage—the presence of ϑ_k—prevents one from applying standard results on conjugacy of Dirichlet processes. The joint distribution of a finite set of states $(\Psi_{k_1}^{\mathrm{GEM}}, .., \Psi_{k_K}^{\mathrm{GEM}})$ does not admit a closed-form expression, so we seek to derive the posterior conditional in a different way.

Rather than proving conjugacy for $(\Psi_{k_1}^{\mathrm{GEM}}, .., \Psi_{k_K}^{\mathrm{GEM}})$ directly, we look for a larger finite-dimensional distribution within which $(\Psi_{k_1}^{\mathrm{GEM}}, .., \Psi_{k_K}^{\mathrm{GEM}})$ sits that has better conjugacy properties. The *generalized Dirichlet* distribution of Connor and Mosimann (1969) fulfills this criteria. The conjugacy relationship we seek follows from the general property that conditioning and marginalization commute. This will be shown to yield the posterior

$$\Psi_k^{\mathrm{GEM}} = \varsigma_k \prod_{i=1}^{k-1}(1-\varsigma_i) \qquad \varsigma_k \sim \mathrm{Beta}(a_k^{(\boldsymbol{\Psi})}, b_k^{(\boldsymbol{\Psi})}) \qquad a_k^{(\boldsymbol{\Psi})} = 1 + l_k \qquad b_k^{(\boldsymbol{\Psi})} = \gamma + \sum_{i=k+1}^{\infty} l_i. \tag{43}$$

For comparison, a posterior Dirichlet process is given by

$$\boldsymbol{\Psi}^{\mathrm{DP}} = \sum_{k=1}^{\infty} \pi_k \delta_{\vartheta_k} \qquad \vartheta_k \sim \frac{n}{\alpha+n}\boldsymbol{l} + \frac{\alpha}{\alpha+n}F \qquad \pi_k = \varsigma_k \prod_{i=1}^{k-1}(1-\varsigma_i) \qquad \varsigma_k \sim \mathrm{Beta}(1, \gamma+n) \tag{44}$$

which shows that this relatively mild difference in the prior yields a posterior of a rather different form.

We now proceed to formally calculate this posterior distribution, starting from a GEM prior and discrete likelihood. Since we are working in a nonparametric setting, we begin by introducing the necessary formalism. We then introduce our finite-dimensional approximating prior and compute the posterior under it. For this, we use commutativity of conditioning and marginalization to deduce the full infinite-dimensional posterior.

Definition 2 (Preliminaries)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $\mathcal{M}_s(\mathbb{N})$ be the space of signed measures, equipped with the topology of weak convergence. Let $\mathcal{M}_1(\mathbb{N}) \subset \mathcal{M}_s(\mathbb{N})$ be the space of probability measures over $\mathbb{N}$, and identify $\mathcal{M}_1(\mathbb{N})$ with the probability simplex by the homeomorphism $\mathcal{M}_1(\mathbb{N}) \cong \{\boldsymbol{x} \in \ell^1 : \forall i, x_i > 0, \sum_{i=1}^{\infty} x_i = 1\}$. Let $N \in \mathbb{N}$, let $\boldsymbol{x} \in \mathbb{N}^N$, and let $\boldsymbol{l} \in N^N$ be its empirical counts, defined by $\boldsymbol{l} = \sum_{i=1}^{N} \mathbf{1}_{x_i}$ where $\mathbf{1}_{x_i}$ is equal to 1 for coordinate x_i and 0 for all other coordinate. Let $\gamma \in \mathbb{R}^+$. Recall that $\mathbb{N}^N$ and $\mathcal{M}_1(\mathbb{N})$, endowed with the discrete topology and topology of weak convergence, respectively, are both Polish spaces—hence, the Disintegration Theorem (Ambrosio et al. (2005), Theorem 5.3.1; Bogachev (2007), Corollary 10.4.15) holds in both spaces. We associate each random variable $y : \Omega \to Y$ with its pushforward probability measure $\pi_y(A_y) = [y_* \mathbb{P}](A_y) = \mathbb{P}[y^{-1}(A_y)]$, and each conditional random variables $\theta \mid y : \Omega \times Y \to \Theta$ with its pushforward regular conditional probability measure $\pi_{y|\theta}(A_y \mid \theta) = [(y \mid \theta)_* \mathbb{P}](A_y) = \mathbb{P}[(y \mid \theta)^{-1}(A_y)]$, where the preimage is taken with respect to y.*

Definition 3 (Discrete likelihood)**.** *For all $\boldsymbol{\Psi} \in \mathcal{M}_1(\mathbb{N})$, define the conditional random variable $\boldsymbol{x} \mid \boldsymbol{\Psi} : \Omega \times \mathcal{M}_1(\mathbb{N}) \to \mathbb{N}^N$ by its probability mass function*

$$p(\boldsymbol{x} \mid \boldsymbol{\Psi}) = \prod_{i=1}^{N} \prod_{k=1}^{\infty} \Psi_k^{\mathbb{1}_k(x_i)}. \tag{45}$$

We say $\boldsymbol{x} \mid \boldsymbol{\Psi} \sim \mathrm{Discrete}(\boldsymbol{\Psi})$.

Definition 4 (GEM)**.** *Let $\boldsymbol{\Psi} : \Omega \to \mathcal{M}_1(\mathbb{N})$ be a random variable defined by*

$$\Psi_k = \varsigma_k \prod_{i=1}^{k-1} (1 - \varsigma_i) \qquad\qquad \varsigma_k \sim \mathrm{Beta}(1, \gamma). \tag{46}$$

We say $\boldsymbol{\Psi} \sim \mathrm{GEM}(\gamma)$.

Definition 5 (Finite GEM)**.** *Let $\boldsymbol{\Psi} : \Omega \to \mathcal{M}_1(\mathbb{N})$ be a random variable defined by*

$$\Psi_k = \varsigma_k \prod_{i=1}^{k-1} (1 - \varsigma_i) \qquad\qquad \varsigma_k \sim \mathrm{Beta}(1, \gamma) \qquad\qquad \varsigma_K = 1. \tag{47}$$

We say $\boldsymbol{\Psi} \sim \mathrm{FGEM}(\gamma, K)$.

Definition 6 (Posterior)**.** *Let $\boldsymbol{\Psi} \mid \boldsymbol{x}$ be the unique conditional random variable given by the Disintegration Theorem, where uniqueness follows from almost sure uniqueness by virtue of the marginal measure $\pi_{\boldsymbol{x}}(\cdot) = \int_{\mathcal{M}_1(\mathbb{N})} \pi_{\boldsymbol{x}|\boldsymbol{\Psi}}(\cdot \mid \boldsymbol{\Psi}) \mathrm{d}\pi_{\Psi}$ being absolutely continuous with respect to the counting measure on $\mathbb{N}^N$, which has no non-empty null sets.*

Result 7. *Let $\boldsymbol{x} \mid \boldsymbol{\Psi} \sim \mathrm{Discrete}(\boldsymbol{\Psi})$. Let $\boldsymbol{x} \in \mathbb{N}^N$, and let $K > \sup \boldsymbol{x}$. Let $\boldsymbol{\Psi} \sim \mathrm{FGEM}(\gamma, K)$. Then for any $\boldsymbol{x}$ with empirical counts $\boldsymbol{l}$, we have that $\boldsymbol{\Psi} \mid \boldsymbol{x} : \Omega \times \mathbb{N}^N \to \mathcal{M}_1(\mathbb{N})$ is a conditional random variable defined by*

$$\Psi_k = \varsigma_k \prod_{i=1}^{k-1} (1 - \varsigma_i) \qquad\qquad \varsigma_k \sim \mathrm{Beta}(a_k^{(\boldsymbol{\Psi})}, b_k^{(\boldsymbol{\Psi})}) \qquad\qquad \varsigma_K = 1 \tag{48}$$

where

$$a_k^{(\boldsymbol{\Psi})} = 1 + l_k \qquad\qquad b_k^{(\boldsymbol{\Psi})} = \gamma + \sum_{i=k+1}^{K} l_i. \tag{49}$$

Proof. It is shown by Connor and Mosimann (1969) that $\boldsymbol{\Psi} \sim \mathrm{FGEM}(\gamma, K)$ is a special case of the *generalized Dirichlet* distribution, which admits a general stick-breaking representation. Thus, its probability density function is

$$f(\boldsymbol{\Psi}) \propto \Psi_K^{\gamma - 1} \prod_{k=1}^{K-1} \left[\sum_{k'=1}^{K} \Psi_{k'} \right]^{-1} \tag{50}$$

which we have expressed in a simplified form. By conjugacy, for a given $\boldsymbol{x}$ and associated $\boldsymbol{l}$ the posterior probability density is

$$f(\boldsymbol{\Psi} \mid \boldsymbol{x}) \propto \Psi_K^{(\gamma + l_K) - 1} \prod_{k=1}^{K-1} \left[\Psi_k^{(1 + l_k) - 1} \left[\sum_{k'=k}^{K} \Psi_{k'} \right]^{\gamma + \sum_{i=k}^{K} l_i - \left[(1 + l_k) + \gamma + \sum_{i=k+1}^{K} l_i \right]} \right] \tag{51}$$

which is again a generalized Dirichlet admitting the necessary stick-breaking representation, which we have expressed in a form that emphasizes its posterior hyperparameters. □

Remark 8. *It is now clear that the assumption $\boldsymbol{x} \mid \boldsymbol{\Psi} \sim \mathrm{Discrete}(\boldsymbol{\Psi})$ is indeed taken without loss of generality, because if we instead took $\boldsymbol{x} \mid \boldsymbol{\Psi}$ to be given by a Pólya sequence, then by sufficiency the prior-to-posterior map would be identical.*

Proposition 1. *Without loss of generality, suppose*

$$\boldsymbol{\Psi} \sim \mathrm{GEM}(\gamma) \qquad\qquad \boldsymbol{x} \mid \boldsymbol{\Psi} \sim \mathrm{Discrete}(\boldsymbol{\Psi}). \tag{52}$$

Then $\boldsymbol{\Psi} \mid \boldsymbol{x}$ is given by

$$\Psi_k = \varsigma_k \prod_{i=1}^{k-1}(1-\varsigma_i) \qquad \varsigma_k \sim \mathrm{Beta}(a_k^{(\boldsymbol{\Psi})}, b_k^{(\boldsymbol{\Psi})}) \qquad a_k^{(\boldsymbol{\Psi})} = 1 + l_k \qquad b_k^{(\boldsymbol{\Psi})} = \gamma + \sum_{i=k+1}^{\infty} l_i \tag{53}$$

where $\boldsymbol{l}$ are the empirical counts of $\boldsymbol{x}$.

Proof. Let $I \subset \mathbb{N}$ be an arbitrary finite index set, and let $\boldsymbol{\Psi}_I \mid \boldsymbol{x}$ be the finite-dimensional marginal projection of $\boldsymbol{\Psi} \mid \boldsymbol{x}$ onto the coordinates contained in I. Let $K > \sup I$, let $\boldsymbol{\Psi}^{(K)} \mid \boldsymbol{x}$ be the posterior conditional random variable under $\boldsymbol{\Psi}^{(K)} \sim \mathrm{FGEM}(\gamma, K)$, and let $\boldsymbol{\Psi}_I^{(K)} \mid \boldsymbol{x}$ be the marginal consisting of those coordinates contained in I. By construction, $\boldsymbol{\Psi}_I^{(K)} \mid \boldsymbol{x}$ equals $\boldsymbol{\Psi}_I \mid \boldsymbol{x}$ in distribution. Since by the Disintegration Theorem, conditioning and marginalization commute, the set I is arbitrary, and $\boldsymbol{\Psi} \mid \boldsymbol{x}$ is uniquely determined by its finite-dimensional marginal projections, the claim follows. □

C Appendix: quantile summary of topics for AP

Here we display a multi-quantile summary for AP, obtained by ranking all topics with at least 100 tokens by their total number of tokens, computing the $\varpi = 100\%$, 75%, 50%, 25%, and 5% quantiles. We compute the five topics closest to each quantile by number of tokens, and display their top-eight words.

AP partially collapsed $\mathbb{B} = 100\%$

k	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
$n_{k,\bullet}$	93 207	57 249	15 874	13 360	10 176
	week	people	police	percent	trial
	made	years	people	year	court
	president	year	killed	prices	charges
	officials	time	man	economic	case
	tuesday	don	officials	economy	judge
	million	back	city	rate	attorney
	thursday	day	shot	increase	prison
	national	home	authorities	report	jury

AP partially collapsed $\mathbb{B} = 75\%$

k	Topic 54	Topic 55	Topic 56	Topic 57	Topic 58
$n_{k,\bullet}$	1 055	1 032	1 025	1 014	1 013
	children	north	hostages	aids	percent
	parents	walsh	red	virus	poll
	child	reagan	release	blood	survey
	ms	iran	held	disease	points
	year	contra	hostage	drug	found
	mother	documents	anderson	infected	surveys
	boys	gesell	gunmen	immune	margin
	girl	arms	thursday	health	reported

AP partially collapsed $\mathbb{B} = 50\%$

k	Topic 108	Topic 109	Topic 110	Topic 111	Topic 112
$n_{k,\bullet}$	473	472	451	446	436
	abortion	women	solidarity	waste	train
	souter	club	walesa	garbage	railroad
	anti	members	poland	recycling	cars
	state	men	polish	city	trains
	women	male	government	ash	transportatio
	abortions	membership	mazowiecki	trash	skinner
	rights	female	jaruzelski	state	transit
	hampshire	black	talks	dump	policy

AP partially collapsed $\mathbb{B} = 25\%$

k	Topic 162	Topic 163	Topic 164	Topic 165	Topic 166
$n_{k,\bullet}$	193	187	185	184	184
	health	wine	miners	dixon	barry
	care	warmus	mine	yates	moore
	spe	solomon	coal	count	jackson
	bc	california	mines	tosh	mayor
	weight	bar	hull	rogers	statehood
	american	gallo	pittston	rig	mr
	diet	test	benefits	russell	gregory
	cholesterol	questions	platform	cookies	room

AP partially collapsed $\mathbb{B} = 5\%$

k	Topic 206	Topic 207	Topic 208	Topic 209	Topic 210
$n_{k,\bullet}$	117	115	112	111	111
	pageant	mall	roberts	stuart	gold
	miss	malls	shell	lawn	polaroid
	cereal	pinochet	boigny	dea	shamrock
	boxes	shopping	houphouet	boston	fields
	contestants	downtown	travelers	ruth	consolidated
	box	park	leonard	yankees	suit
	america	oak	arsenal	foundation	proposals
	bruce	usa	oil	richman	mining

AP direct assignment $\mathbb{B} = 100\%$

k	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
$n_{k,\bullet}$	90 497	18 626	10 832	9 923	9 430
	year	years	police	dollar	percent
	people	year	people	market	year
	time	people	killed	stock	rose
	president	time	government	yen	sales
	years	don	reported	index	million
	made	home	today	late	billion
	state	day	capital	trading	month
	week	back	violence	exchange	reported

AP direct assignment $\mathbb{B} = 75\%$

k	Topic 93	Topic 94	Topic 95	Topic 96	Topic 97
$n_{k,\bullet}$	784	757	753	745	738
	keating	bus	eastern	united	smoking
	deconcini	driver	pilots	states	cigarettes
	lincoln	train	airline	nations	farmers
	senators	greyhound	orion	resolution	tobacco
	regulators	accident	air	international	ban
	meeting	passengers	union	plo	insurance
	committee	railroad	airlines	mission	batus
	gray	passenger	service	assembly	smokers

AP direct assignment $\mathbb{B} = 50\%$

k	Topic 186	Topic 187	Topic 188	Topic 189	Topic 190
$n_{k,\bullet}$	346	338	338	338	334
	power	cable	conservatives	water	dental
	franc	television	flag	dam	funds
	jersey	nbc	conservative	river	claims
	bradley	tempo	amendment	area	plough
	utility	hsn	speaker	reservoir	oral
	wppss	industry	darman	savannah	counter
	utilities	subscribers	kemp	corps	embassy
	west	tv	republicans	canyon	mid

AP direct assignment $\mathbb{B} = 25\%$

k	Topic 279	Topic 280	Topic 281	Topic 282	Topic 283
$n_{k,\bullet}$	220	219	219	219	218
	fernandez	water	bloom	canadian	election
	fdic	lake	minnick	lee	grenada
	republicbank	mussels	walters	ritalin	boigny
	weicker	neill	lawyer	murphy	houphouet
	virginia	erie	athletes	domestic	gairy
	ruth	problem	college	security	coast
	robinson	plant	suspect	woods	nov
	station	north	signing	radio	failed

AP direct assignment $\mathbb{B} = 5\%$

k	Topic 354	Topic 355	Topic 356	Topic 357	Topic 358
$n_{k,\bullet}$	133	133	133	132	132
	machine	young	count	reynolds	turkey
	stop	johnston	forman	premier	department
	reed	golf	festival	bond	bird
	gun	notes	rig	release	cooking
	chief	bodies	arts	news	wash
	sununu	homes	hughes	regulated	bacteria
	geneva	call	lights	address	stuffed
	formal	shortage	staged	petition	adams

D Appendix: quantile summary of topics for CGCBIB

Here we display a multi-quantile summary for CGCBIB, obtained by ranking all topics with at least 100 tokens by their total number of tokens, computing the $\varpi = 100\%$, 75%, 50%, 25%, and 5% quantiles. We compute the five topics closest to each quantile by number of tokens, and display their top-eight words.

CGCBIB partially collapsed $\mathbb{B} = 100\%$

k	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
$n_{k,\bullet}$	110 702	58 811	27 084	21 215	19 832
	elegans	elegans	elegans	gene	mutations
	caenorhabditi	protein	genetic	elegans	gene
	nematode	caenorhabditi	molecular	sequence	mutants
	results	gene	development	protein	genes
	found	function	caenorhabditi	caenorhabditi	mutant
	show	proteins	nematode	amino	elegans
	observed	required	studies	cdna	caenorhabditi
	specific	show	model	acid	alleles

CGCBIB partially collapsed $\mathbb{B} = 75\%$

k	Topic 54	Topic 55	Topic 56	Topic 57	Topic 58
$n_{k,\bullet}$	2 166	2 061	2 048	2 040	2 025
	germ	egl	emb	spe	wnt
	germline	egg	temperature	sperm	mom
	early	laying	mutants	fer	pop
	granules	serotonin	sensitive	spermatozoa	signaling
	cells	neurons	zyg	membrane	bar
	embryos	cat	maternal	spermatids	pathway
	somatic	dopamine	expression	spermatogenes	lin
	line	mutants	embryonic	pseudopod	wrm

CGCBIB partially collapsed $\mathbb{B} = 50\%$

k	Topic 109	Topic 110	Topic 111	Topic 112	Topic 113
$n_{k,\bullet}$	930	916	915	900	893
	vit	binding	kinesin	growth	eat
	yolk	affinity	klp	survival	pharyngeal
	vitellogenin	site	transport	mortality	pharynx
	genes	activity	motor	population	pumping
	yp	sites	ift	rate	inx
	proteins	avermectin	cilia	populations	gap
	vpe	elegans	dynein	parameter	feeding
	lrp	membrane	movement	size	junctions

CGCBIB partially collapsed $\mathbb{B} = 25\%$

k	Topic 164	Topic 165	Topic 166	Topic 167	Topic 168
$n_{k,\bullet}$	386	369	368	364	360
	mlc	dom	innate	vha	ife
	mel	effects	immune	atpase	cap
	myosin	humic	immunity	subunit	eife
	nmy	pyrene	abf	genes	capping
	chain	effect	lys	vacuolar	cel
	elongation	bioconcentrat	toll	subunits	gtp
	rho	dissolved	antimicrobial	atpases	isoforms
	phosphatase	substances	pathway	type	rna

CGCBIB partially collapsed $\mathbb{B} = 5\%$

k	Topic 208	Topic 209	Topic 210	Topic 211	Topic 212
$n_{k,\bullet}$	141	141	140	140	136
	ubq	asp	da	ion	hcf
	gc	salmonella	cl	diet	cehcf
	tbp	poona	fli	relative	vp
	footprints	enterica	gs	xpa	ldb
	oscillin	clp	db	groups	cell
	tlf	serotype	glu	carbon	mammalian
	ubiquitin	necrotic	phospholipid	characteristi	phosphorylati
	tata	mug	tg	atoms	neural

CGCBIB direct assignment $\varpi = 100\%$

k	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
$n_{k,\bullet}$	65 059	41 005	33 714	27 221	22 813
	elegans	elegans	elegans	mutations	elegans
	caenorhabditi	genetic	caenorhabditi	elegans	gene
	protein	caenorhabditi	nematode	gene	sequence
	gene	nematode	results	mutants	caenorhabditi
	function	molecular	observed	genes	protein
	proteins	development	high	caenorhabditi	amino
	required	studies	type	mutant	cdna
	show	model	effect	function	acid

CGCBIB direct assignment $\varpi = 75\%$

k	Topic 68	Topic 69	Topic 70	Topic 71	Topic 72
$n_{k,\bullet}$	1 921	1 894	1 836	1 828	1 776
	loci	worm	cell	alpha	unc
	genetic	elegans	epithelial	gpa	gaba
	strains	research	junctions	egl	receptors
	lines	caenorhabditi	membrane	signaling	receptor
	life	brenner	cells	protein	resistance
	mutations	years	dlg	goa	lev
	mutation	nematode	hmp	rgs	levamisole
	inbred	biology	exc	proteins	cholinergic

CGCBIB direct assignment $\varpi = 50\%$

k	Topic 137	Topic 138	Topic 139	Topic 140	Topic 141
$n_{k,\bullet}$	782	779	779	763	756
	cell	hsp	survival	acid	yeast
	dimensional	heat	mortality	amino	cerevisiae
	microscopy	shock	model	acids	saccharomyces
	embryo	chaperone	data	nematode	pombe
	analysis	small	gompertz	glycine	cell
	system	proteins	parameter	briggsae	budding
	computer	crystallin	population	cytochrome	schizosacchar
	time	hsps	rate	multiple	cycle

CGCBIB direct assignment $\varpi = 25\%$

k	Topic 206	Topic 207	Topic 208	Topic 209	Topic 210
$n_{k,\bullet}$	364	359	358	343	343
	activity	pgp	telomere	gcy	mediator
	lh	mrp	telomeres	guanylyl	med
	activities	aat	ceh	cyclase	sop
	juvenile	cells	yeast	wee	transcription
	nematodes	glycoprotein	nematode	ase	development
	antiallatal	mammalian	mrt	receptor	pvl
	hormone	resistance	telomerase	cyclases	transcription
	insect	glycoproteins	telomeric	gfp	dhp

CGCBIB direct assignment $\varpi = 5\%$

k	Topic 261	Topic 262	Topic 263	Topic 264	Topic 265
$n_{k,\bullet}$	164	164	159	157	156
	cog	atp	calcineurin	selection	srl
	wd	structures	egg	flow	rol
	repeat	oligomerizati	bovine	separation	threshold
	connection	family	laying	redundancy	ra
	native	binding	hg	flows	energy
	response	members	white	directional	free
	worm	stability	haemin	solution	external
	nr	mechanism	phosphatase	period	experimental

E Appendix: quantile summary of topics for NEURIPS

Here we display a multi-quantile summary for NeurIPS, obtained by ranking all topics with at least 100 tokens by their total number of tokens, computing the $\varpi = 100\%$, 75%, 50%, 25%, and 5% quantiles. We compute the five topics closest to each quantile by number of tokens, and display their top-eight words.

NeurIPS partially collapsed $\varpi = 100\%$

k	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
$n_{k,\bullet}$	182 743	162 355	129 745	52 356	44 155
	system	function	number	model	training
	information	case	result	neural	set
	approach	result	small	result	data
	set	term	values	system	test
	problem	parameter	order	activity	performance
	research	neural	large	input	number
	computer	form	effect	pattern	result
	single	defined	high	function	error

NeurIPS partially collapsed $\varpi = 75\%$

k	Topic 148	Topic 149	Topic 150	Topic 151	Topic 152
$n_{k,\bullet}$	2 585	2 585	2 574	2 559	2 549
	genetic	delay	bengio	fig	matching
	algorithm	bifurcation	output	properties	model
	population	oscillation	dependencies	proc	point
	fitness	point	input	step	correspondenc
	string	stability	experiment	range	match
	generation	fixed	frasconi	structure	problem
	bit	limit	term	calculation	set
	function	hopf	information	illinois	object

NeurIPS partially collapsed $\varpi = 50\%$

k	Topic 297	Topic 298	Topic 299	Topic 300	Topic 301
$n_{k,\bullet}$	1 310	1 309	1 309	1 297	1 295
	vor	routing	speaker	delay	memory
	storage	load	recognition	input	action
	anastasio	network	normalization	transition	states
	responses	path	male	window	agent
	velocity	packet	feature	width	sensing
	pan	traffic	female	connection	loop
	rotation	shortest	mntn	information	history
	vestibular	policy	ntn	temporal	mdp

NeurIPS partially collapsed $\varpi = 25\%$

k	Topic 446	Topic 447	Topic 448	Topic 449	Topic 450
$n_{k,\bullet}$	748	748	746	739	735
	composite	psom	limited	tau	cmm
	mdp	robot	interconnect	hypothesis	speed
	action	camera	fan	mansour	particle
	elemental	set	shunting	growth	particles
	optimal	pointing	modularity	coefficient	pattern
	payoff	coordinates	collective	function	presence
	solution	basis	linear	stem	method
	mdt	ritter	unit	large	card

NeurIPS partially collapsed $\varpi = 5\%$

k	Topic 566	Topic 567	Topic 568	Topic 569	Topic 570
$n_{k,\bullet}$	396	385	383	379	372
	morph	minimal	visualization	periodic	machine
	kernel	root	high	period	capacity
	parent	biases	low	coefficient	path
	human	attribute	diagram	primitive	trouble
	busey	remove	visualizing	homogeneous	high
	similar	rumelhart	graphic	tst	task
	exemplar	row	fund	mhaskar	increasing
	distinctivene	exponential	window	chain	measures

NeurIPS subcluster split-merge $\mathbb{B} = 100\%$

k	Topic 6	Topic 2	Topic 1	Topic 13	Topic 62
$n_{k\bullet}$	473 770	93 435	52 418	50 965	41 565
	network	network	model	model	function
	model	unit	neuron	data	network
	learning	input	input	parameter	bound
	function	learning	network	network	dimension
	input	training	cell	algorithm	learning
	neural	weight	system	mixture	result
	algorithm	neural	unit	function	number
	set	output	visual	gaussian	set

NeurIPS subcluster split-merge $\mathbb{B} = 75\%$

k	Topic 440	Topic 170	Topic 334	Topic 418	Topic 312
$n_{k\bullet}$	2 678	2 657	2 643	2 636	2 622
	learning	movement	motion	learning	cell
	critic	visual	unit	algorithm	correlation
	function	vector	direction	action	neuron
	actor	image	model	advantage	model
	algorithm	model	stage	system	unit
	system	location	input	function	interaction
	control	eye	network	policy	firing
	model	map	cell	control	set

NeurIPS subcluster split-merge $\mathbb{B} = 50\%$

k	Topic 378	Topic 322	Topic 82	Topic 344	Topic 414
$n_{k\bullet}$	1 032	1 028	1 013	1 009	1 006
	iiii	cell	model	form	component
	cell	spike	response	word	algorithm
	network	unit	neural	phone	sources
	neural	function	escape	input	analysis
	response	firing	interneuron	network	data
	model	result	cockroach	system	noise
	point	transfer	leg	training	orientation
	fixed	sorting	input	meaning	spatial

NeurIPS subcluster split-merge $\mathbb{B} = 25\%$

k	Topic 220	Topic 341	Topic 441	Topic 447	Topic 308
$n_{k\bullet}$	728	723	723	722	721
	aspect	element	network	input	traffic
	object	pairing	neural	unit	waiting
	view	grouping	constraint	spike	elevator
	node	group	match	layer	appeared
	learning	saliency	learn	learning	application
	network	contour	problem	model	compared
	weight	computation	initial	predict	department
	equation	optimal	row	prediction	found

NeurIPS subcluster split-merge $\mathbb{B} = 5\%$

k	Topic 259	Topic 246	Topic 195	Topic 245	Topic 293
$n_{k\bullet}$	509	507	506	503	503
	input	network	network	network	network
	output	neural	symbol	equation	function
	activation	task	vtp	neuron	adaptation
	data	link	learning	moment	algorithm
	encoded	food	phrases	neural	prediction
	function	nodes	sentences	approximation	projection
	hidden	output	vpp	ohira	neural
	model	recurrent	classificatio	stochastic	training

F Appendix: topics produced by Algorithm 2 on PUBMED

Here we show top eight words for each topic together with total number of tokens assigned, which is shown at the top of each table. We display all topics containing at least eight unique word tokens.

k	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
$n_{k,\bullet}$	47 322 709	40 229 486	34 685 122	30 795 166	30 707 144
PubMed	care	age	model	cell	gene
	health	risk	data	expression	protein
	patient	children	system	growth	dna
	medical	year	time	protein	expression
	research	women	analysis	factor	sequence
	clinical	patient	effect	receptor	genes
	system	factor	test	kinase	rna
	cost	population	field	beta	region

k	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
$n_{k,\bullet}$	28 510 997	27 277 306	26 709 116	26 408 263	25 200 662
PubMed	cell	cancer	patient	rat	cell
	il	tumor	treatment	receptor	electron
	cd	patient	mg	effect	muscle
	mice	carcinoma	drug	neuron	tissue
	antigen	cell	effect	brain	fiber
	human	breast	therapy	activity	rat
	lymphocytes	survival	dose	stimulation	development
	immune	tumour	day	induced	microscopy

k	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15
$n_{k,\bullet}$	24 856 624	24 750 437	24 607 618	24 482 090	22 956 810
PubMed	patient	patient	blood	patient	infection
	surgery	artery	pressure	disease	virus
	complication	heart	flow	clinical	hiv
	surgical	coronary	min	diagnosis	strain
	treatment	ventricular	effect	lesion	infected
	year	myocardial	exercise	brain	patient
	postoperative	cardiac	arterial	syndrome	positive
	operation	left	heart	imaging	viral

k	Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
$n_{k,\bullet}$	22 095 623	21 838 239	21 363 408	20 887 061	20 828 980
PubMed	ca	structure	concentration	pregnancy	protein
	effect	binding	degrees	level	binding
	receptor	protein	samples	women	human
	channel	reaction	liquid	hormone	antibodies
	cell	acid	solution	day	acid
	calcium	interaction	assay	fetal	alpha
	concentration	compound	detection	infant	antibody
	na	site	system	concentration	gel

k	Topic 21	Topic 22	Topic 23	Topic 24	Topic 25
$n_{k,\bullet}$	20 106 260	19 788 488	18 675 096	17 163 327	16 440 018
PubMed	rat	bone	patient	gene	activity
	cell	patient	renal	mutation	acid
	effect	joint	liver	genetic	enzyme
	liver	muscle	transplantati	chromosome	liver
	mice	fractures	blood	analysis	concentration
	dose	hip	disease	genes	rat
	drug	year	acute	dna	enzymes
	mg	implant	chronic	polymorphism	synthesis

	k	Topic 26	Topic 27	Topic 28	Topic 29	Topic 30
	$n_{k,\bullet}$	16 136 164	14 201 063	13 706 016	13 191 158	13 105 245
PubMed		effect	diet	patient	strain	protein
		platelet	weight	disease	plant	membrane
		induced	intake	gastric	growth	cell
		oxide	food	asthma	acid	domain
		rat	body	test	bacteria	binding
		cell	effect	pylori	activity	receptor
		endothelial	acid	arthritis	cell	lipid
		activity	vitamin	chronic	species	membranes

	k	Topic 31	Topic 32	Topic 33	Topic 34	Topic 35
	$n_{k,\bullet}$	12 705 261	12 624 252	10 422 885	9 850 167	7 027 660
PubMed		insulin	species	exposure	skin	level
		glucose	population	concentration	patient	patient
		diabetes	infection	iron	eyes	ml
		cholesterol	animal	level	eye	control
		level	egg	water	retinal	serum
		diabetic	host	effect	laser	plasma
		plasma	parasite	exposed	visual	factor
		lipoprotein	malaria	lead	corneal	concentration

	k	Topic 36	Topic 37	Topic 38	Topic 39	Topic 40
	$n_{k,\bullet}$	6 130 945	644 182	2 264	1 325	104
PubMed		dental	sleep	ppr	pac	feather
		oral	caffeine	csc	foal	tieg
		teeth	tea	stretch	cpr	sorghum
		tooth	effect	pthrp	pacap	coii
		periodontal	theophylline	response	edm	phycocyanin
		treatment	night	br	speck	vanx
		salivary	coffee	gei	branchial	midrib
		gland	green	pth	lth	ifi

	k	Topic 41
	$n_{k,\bullet}$	104
PubMed		steer
		mca
		persistency
		buckwheat
		dnak
		eset
		branding
		akr