Online Latent Dirichlet Allocation with Infinite Vocabulary

Ke Zhai and Jordan Boyd-Graber
University of Maryland

1. Online LDA: What is the vocabulary?

Latent Dirichlet allocation (LDA) reveals topics in a corpus.

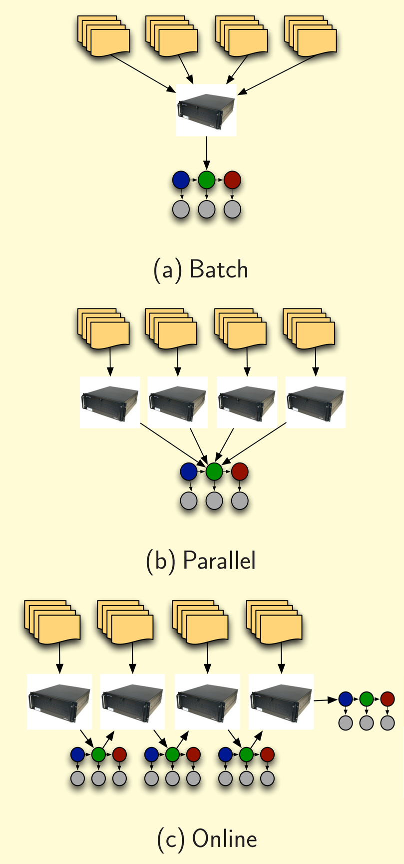
- Batch approach does not scale
- Two solutions: parallel and **online** inference
- **Online**: after observing a minibatch of documents, reestimate latent variables

Existing online approaches share same flaw: immutable vocabulary, drawn from a **fixed** Dirichlet distribution.

Cannot capture the appearance of new words Fixed vocabularies conceal when

- words are invented, e.g., "crowdsourcing"
- words cross languages, e.g.,"Gangnam"
- words cross topics, e.g., "vuvuzelas"

We replace the Dirichlet distribution over topics with a Dirichlet process (Blunsom et al., 2011). We develop new online inference techniques for "infinite vocabularies".



2. Dirichlet Process

Dirichlet Process Stick Breaking Construction Dirichlet process (DP) is a a two-parameter infinite extension to the Dirichlet distribution (scale parameter α^β and base distribution G_0). A draw G from DP(α^β, G_0) is

$$b_1, \dots, b_i, \dots \sim \text{Beta}(1, \alpha^\beta), \qquad \rho_1, \dots, \rho_i, \dots \sim G_0.$$
$$\beta_i \equiv b_i \prod_{j=1}^{i-1}(1 - b_j), \qquad G \equiv \sum_i \beta_i \delta_{\rho_i},$$

where the weights β_i give the probability of selecting any particular atom ρ_i drawn from the base distribution.

3. Base Distribution Intuition

Base Distribution: Character n-gram Model

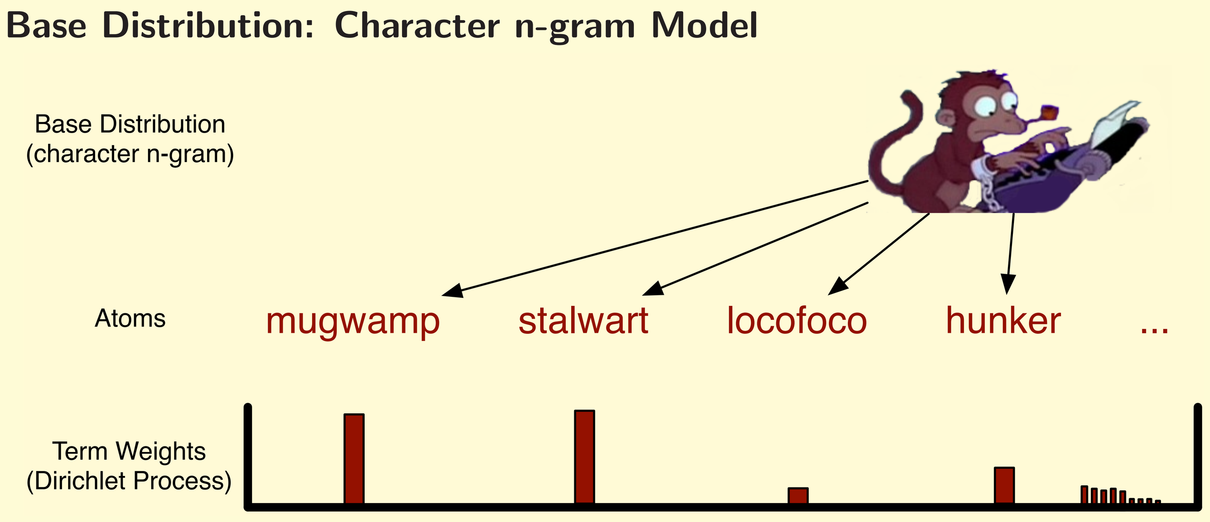


Figure: A Dirichlet process provides a distribution over an unbounded set of words

4. Base Distribution Definition

Generative process of the n-gram character model:

1: Choose a length $l \sim \text{mult}(\boldsymbol{\lambda})$.
2: Iteratively generate a word's i-th character c_i given context $c_i \sim p(c_i|\mathbf{c}_{1,\dots,i-1})$.

The probability of a word $\rho = c_1 c_2 \dots$ under base distribution G_0:

$$G_0(\rho) \equiv p(|\rho| \mid \boldsymbol{\lambda}) \prod_{i=1}^{|\rho|} p(c_i|\mathbf{c}_{i-n+1,\dots,i-1}),$$

where $|\rho|$ is word length. The multinomial distribution λ over lengths prevents bias toward short words. Parameters trained on a English dictionary.

5. Generative Model

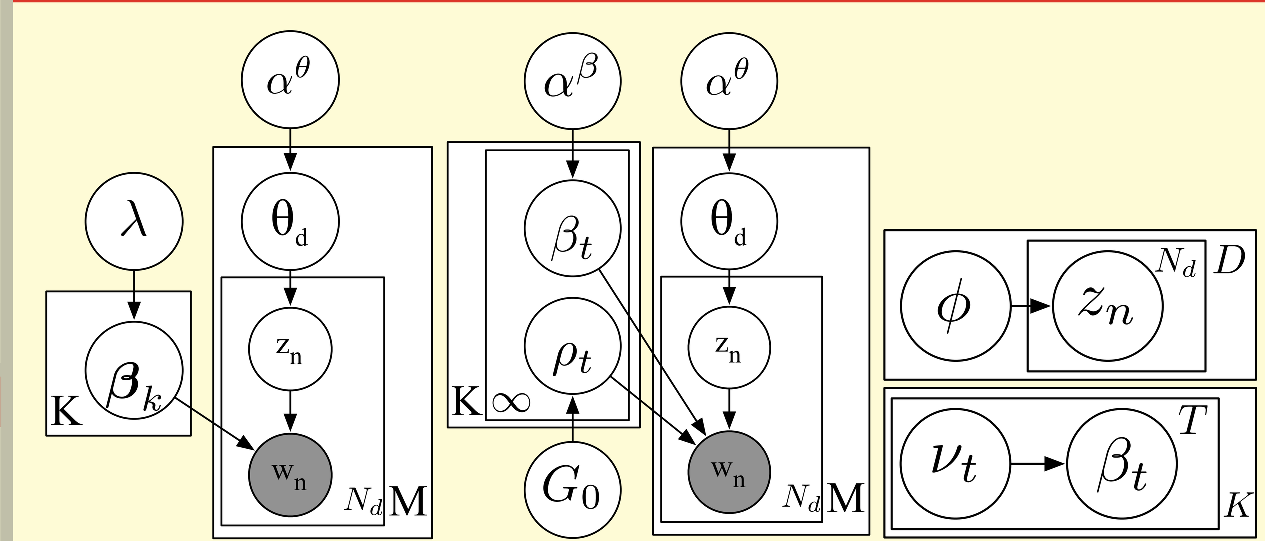


Figure: Plate representation for latent Dirichlet allocation (left), latent Dirichlet allocation with infinite vocabulary (middle) and its variational distribution (right).

Generative Process of Online LDA with Infinite Vocabulary

1: **for** each topic k **do**
2: Draw words ρ_{kt}, $(t = \{1, 2, \dots\})$ from base distribution G_0.
3: Draw $b_{kt} \sim \text{Beta}(1, \alpha^\beta)$, $(t = \{1, 2, \dots\})$.
4: Set the stick weights to be $\beta_{kt} = b_{kt} \prod_{s<t}(1 - b_{ks})$.
5: **for** each document d in a corpus D **do**
6: Draw θ_d from a Dirichlet distribution $\boldsymbol{\theta}_d \sim \text{Dir}(\alpha^\theta)$.
7: **for** each of the $n = 1, \dots, N_d$ word indexes **do**
8: Draw z_n from the topic distribution $z_n \sim \text{multi}(\boldsymbol{\theta}_d)$.
9: Draw w_n from the word distribution $p(w_n|\boldsymbol{\beta}_{z_n})$.

6. Variational Distribution

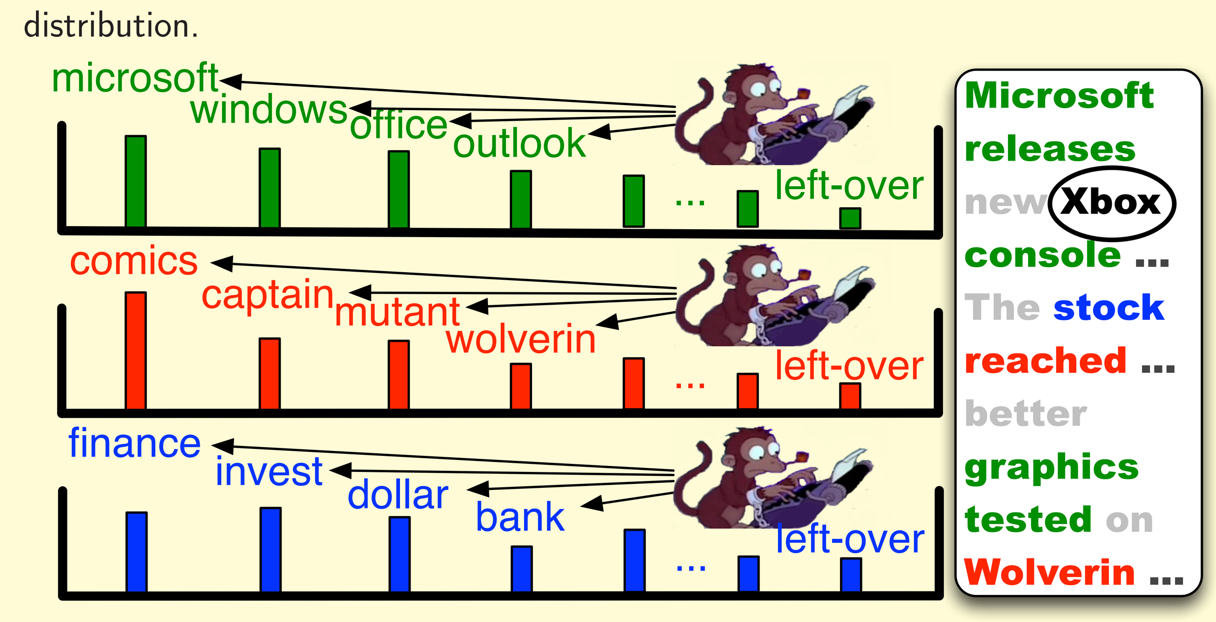
Variational distribution is $q(\mathbf{Z}) \equiv q(\beta, \mathbf{z}) = \prod_D q(\mathbf{z}_d|\eta) \prod_K q(\mathbf{b}_k|\boldsymbol{\nu}_k^1, \boldsymbol{\nu}_k^2)$.

- ν: variational parameter for stick breaking Beta distributions
- ϕ: variational parameter for topic multinomial distributions
- $\mathbf{T}_k$: Truncation Ordered Set

ν updated in online variational gradient step (Hoffman et al, 2009); ϕ by MCMC (Mimno et al, 2012).

7. Truncation Ordered Set (TOS)

We define our truncation $\mathcal{T}_k$ for topic k as an ordered set of words (atoms). This set controls the number and identity of words modeled by the variational distribution.



8. Updating the TOS

- New words are added to the TOS as they appear, appended to end of TOS
- After observing $U = 10$ minibatches, we use a heuristic inspired by Chinese restaurant process to reorder the words in the TOS according to $R(\rho_{kt}) = p(\rho_{kt}|G_0) \sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{dnk} \delta_{\omega_{dn}=\rho_{kt}}$.
- Retain only the top T terms (truncation size) according to the ranking score.
- All the previous information (e.g., rank and variational parameters) is discarded.

9. Inference Algorithm

1: Randomly initialize variational parameters.
2: **repeat**
3: **for** each document d in minibatch S **do**
4: **for** every word n in document d **do**
5: Empirically sample the variational distribution $q(z_{dn}|\phi_{dn})$ according to

$$q(z_{dn} = k|\mathbf{z}_{-dn}, t = \mathcal{T}_k(w_{dn})) \propto \Big(\sum_{m=1, m\neq n}^{N_d} \mathbb{I}_{z_{dm}=k} + \alpha_k^\theta\Big) \exp\left\{\mathbb{E}_{q(\nu)}[\log \beta_{kt}]\right\}$$

6: Update variational parameters $\boldsymbol{\nu}$ using stochastic gradient descent algorithm

$$\Delta\nu_{kt}^1 = 1 + \frac{D}{|S|} \sum_{d\in S} \sum_{n=1}^{N_d} \phi_{dnk} \delta_{\omega_{dn}=\rho_{kt}} - \nu_{kt}^1$$
$$\Delta\nu_{kt}^2 = \alpha^\beta + \frac{D}{|S|} \sum_{d\in S} \sum_{n=1}^{N_d} \phi_{dnk} \delta_{\omega_{dn}>\rho_{kt}} - \nu_{kt}^2$$

7: Update the ranking score according to

$$R_{ik}(\rho) = (1 - \epsilon) \cdot R_{i-1,k}(\rho) + \epsilon \cdot R_{ik}(\rho)$$

8: Contract vocabulary for every topic if necessary.
9: **until** model convergence

10. Results: Topic Coherence

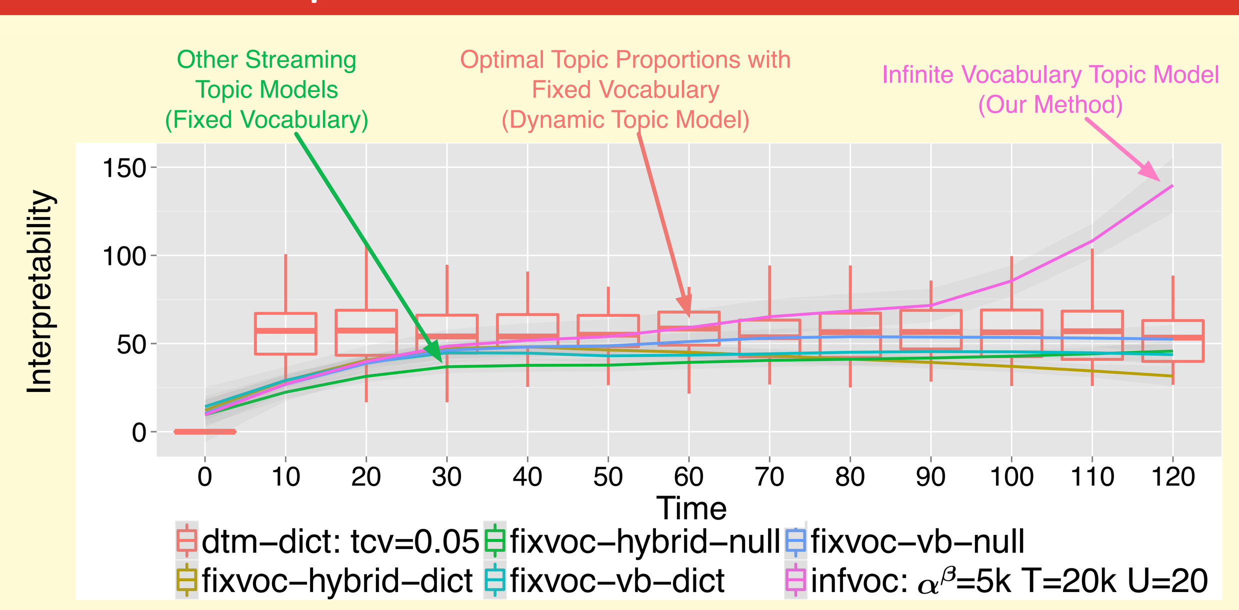


Figure: Topic interpretability score (Newman et al., 2009) on *20 newsgroups*.

11. Results: Classification Accuracy

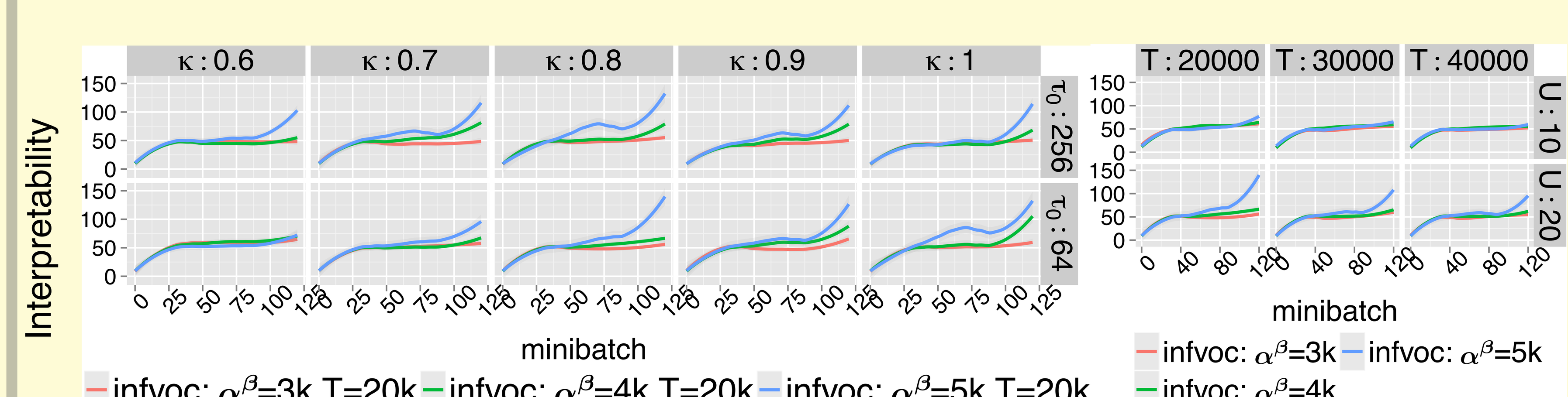
To test the quality of the model, we fit a topic model with 50 topics to the 20-newsgroups dataset. We train a classifier on training fold and report accuracy on the test fold. Messages are ordered by the date they were posted.

		model settings		accuracy %
$S = 155$	$\tau_0 = 64$ $\kappa = 0.6$	*infvoc*	$\alpha^\beta = 3k$ $T = 40k$ $U = 10$	52.683
		fixvoc	vb-dict	45.514
		fixvoc	hybrid-dict	46.720
		fixvoc	vb-null	49.390
		fixvoc	hybrid-null	50.474
		fixvoc	vb-dict hash	52.525
		fixvoc	hybrid-dict hash	50.948
		fixvoc	vb-full hash $T = 30k$	51.653
		fixvoc	hybrid-full hash $T = 30k$	50.948
		dtm-dict	$tcv = 0.001$	**62.845**

		model settings		accuracy %
$S = 310$	$\tau_0 = 64$ $\kappa = 0.6$	*infvoc*	$\alpha^\beta = 3k$ $T = 40k$ $U = 20$	52.317
		fixvoc	vb-dict	44.701
		fixvoc	hybrid-dict	46.368
		fixvoc	vb-null	51.815
		fixvoc	hybrid-null	50.569
		fixvoc	vb-dict hash	48.130
		fixvoc	hybrid-dict hash	51.558
		fixvoc	vb-full hash $T = 30k$	47.276
		fixvoc	hybrid-full hash $T = 30k$	43.008
		dtm-dict	$tcv = 0.001$	**64.186**

Our infinite vocabulary topic model performs as well as hash-based topic models while remaining interpretable. Dynamic topic models, which view all data at once, perform better.

12. Results: Parameter Sensitivity



PMI score on *20 newsgroups* against different settings of DP scale parameter α^β, decay factor κ and τ_0 (left), and against different settings of DP scale parameter α^β, truncation level T and reordering delay U (right).

13. Results: Incorporating New Words

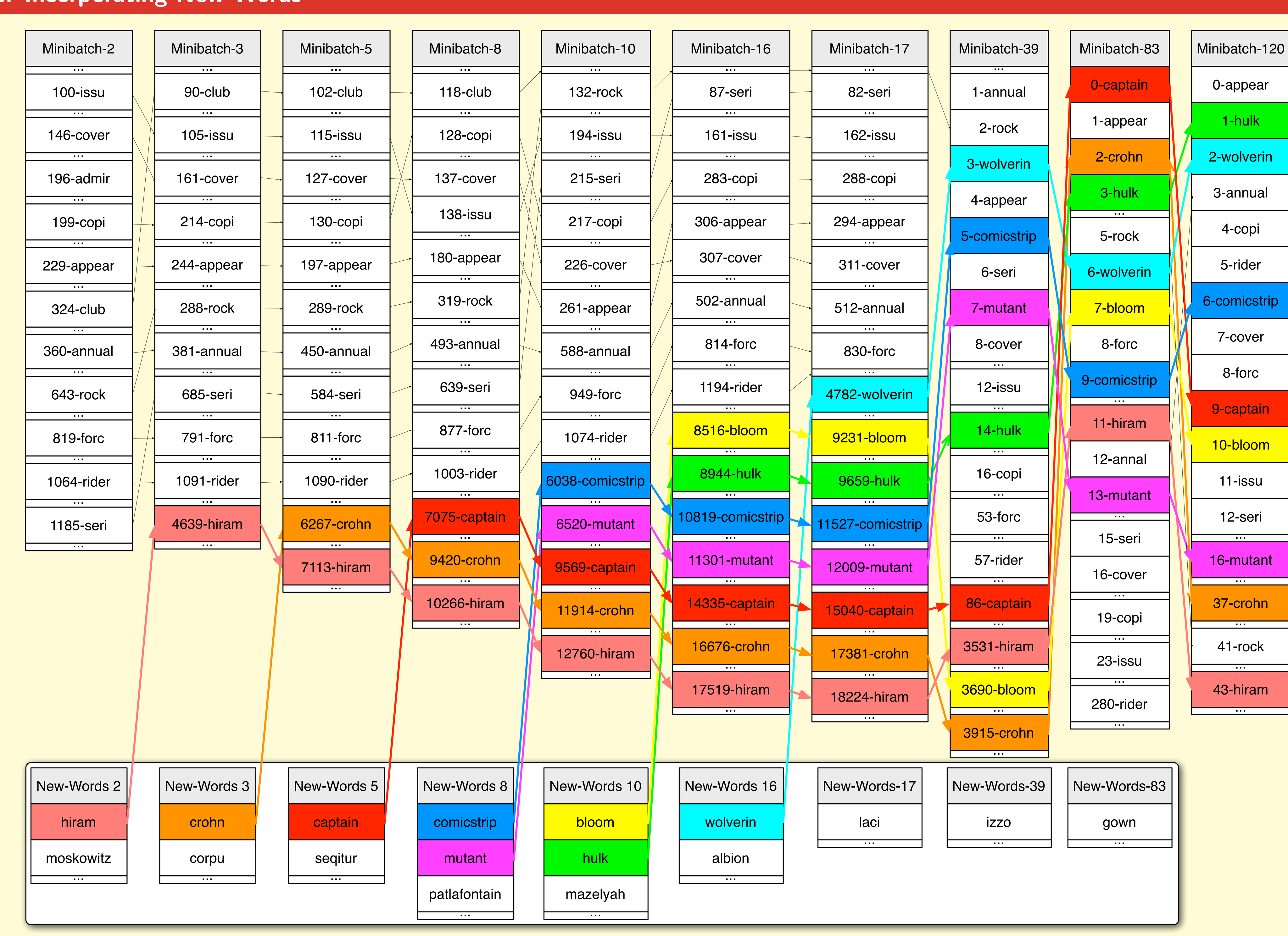


Figure: The evolution of single topic about comic books from the *20 newsgroups* corpus as new words are discovered.

14. Conclusion

- Extend LDA by drawing topics from a Dirichlet process whose base distribution is over all strings rather than from a finite Dirichlet.
- We develop inference using online variational inference and propose heuristics to dynamically order, expand, and contract our vocabulary.

References

Blunsom, Phil and Cohn, Trevor. A hierarchical Pitman-Yor process HMM for unsupervised part of speech induction. In *Proceedings of the Association for Computational Linguistics*, 2011.

Hoffman, Matthew, Blei, David M., and Bach, Francis. Online learning for latent Dirichlet allocation. In *NIPS*, 2010.

Newman, David, Karimi, Sarvnaz, and Cavedon, Lawrence. External evaluation of topic models. In *Proceedings of the Aurstralasian Document Computing Symposium*, 2009.

Mimno, David, Hoffman, Matthew, and Blei, David. Sparse stochastic inference for latent Dirichlet allocation. In *Proceedings of the International Conference of Machine Learning*, 2012.

Email: {zhaike,jbg}@umiacs.umd.edu