Fast, Flexible Models for Discovering Topic Correlation across Weakly-Related Collections

Jingwei Zhang[1], **Aaron Gerow**[2], **Jaan Altosaar**[3], **James Evans**[2,4], **Richard Jean So**[5]

[1]Department of Computer Science, Columbia University
jz2541@columbia.edu
[2] Computation Institute, University of Chicago
{gerow,jevans}@uchicago.edu
[3] Department of Physics, Princeton University
altosaar@princeton.edu
[4] Department of Sociology, University of Chicago
[5] Department of English Language and Literature, University of Chicago
richardjeanso@uchicago.edu

Abstract

Weak topic correlation across document collections with different numbers of topics in individual collections presents challenges for existing cross-collection topic models. This paper introduces two probabilistic topic models, *Correlated LDA* (C-LDA) and *Correlated HDP* (C-HDP). These address problems that can arise when analyzing large, asymmetric, and potentially weakly-related collections. Topic correlations in weakly-related collections typically lie in the tail of the topic distribution, where they would be overlooked by models unable to fit large numbers of topics. To efficiently model this long tail for large-scale analysis, our models implement a parallel sampling algorithm based on the Metropolis-Hastings and alias methods (Yuan et al., 2015). The models are first evaluated on synthetic data, generated to simulate various collection-level asymmetries. We then present a case study of modeling over 300k documents in collections of sciences and humanities research from JSTOR.

1 Introduction

Comparing large text collections is a critical task for the curation and analysis of human cultural history. Achievements of research and scholarship are most accessible through textual artifacts, which are increasingly available in digital archives. Text-based research, often undertaken by humanists, historians, lexicographers, and corpus linguists, explores patterns of words in documents across time-periods and distinct collections of text. Here, we introduce two new topic models designed to compare large collections, Correlated LDA (C-LDA) and Correlated HDP (C-HDP), which are sensitive to document-topic asymmetry (where collections have different topic distributions) and topic-word asymmetry (where a single topic has different word distributions in each collection). These models seek to address terminological questions, such as how a topic on physics is articulated distinctively in scientific compared to humanistic research. Accommodating potential collection-level asymmetries is particularly important when researchers seek to analyze collections with little prior knowledge about shared or collection-specific topic structure. Our models extend existing cross-collection approaches to accommodate these asymmetries and implement an efficient parallel sampling algorithm enabling users to examine the long tail of topics in particularly large collections.

Using topic models for comparative text mining was introduced by Zhai et al. (2004), who developed the ccMix model which extended pLSA (Hofmann, 1999). Later work by Paul and Girju (2009) developed ccLDA, which adopted the hierarchical Bayes framework of Latent Dirichlet Allocation or LDA (Blei et al., 2003). These models account for topic-word asymmetry by assuming variation in the vocabularies of topics is due to collection-level differences. Nevertheless, they require the same topics to be present in each collection. These models are useful for comparing collections under specific assumptions, but cannot accommodate collection-topic asymmetry (which

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1554–1564,
Lisbon, Portugal, 17-21 September 2015.

arises in collections that do not share every topic or that have different numbers of topics). In situations where collections do not share all topics, the results often include junk, mixed, or sparse topics, making them difficult to interpret (Paul and Girju, 2009). Such asymmetries make it difficult to use models like ccLDA and ccMix when little is known about collections in advance. This motivates our efforts to model variation in the long tail of topic distributions, where correlations are more likely to appear when collections are weakly related.

C-LDA and C-HDP extend ccLDA (Paul and Girju, 2009) to accommodate collection-topic level asymmetries, particularly by allowing non-common topics to appear in each collection. This added flexibility allows our models to discover topic correlations across arbitrary collections with different numbers of topics, even when there are few (or unknown) numbers of common topics. To demonstrate the effectiveness of our models, we evaluate them on synthetic data and show that they outperform related models such as ccLDA and differential topic models (Chen et al., 2014). We then fit C-LDA to two large collections of humanities and sciences documents from JSTOR. Such historical analyses of text would be intractable without an efficient sampler. An optimized sampler is required in such situations because common topics in weakly-correlated collections are usually found in the tail of the document-topic distribution of a sufficiently large set of topics. To make this feasible on large datasets such as JSTOR, we employ a parallelized Metropolis-Hastings (Kronmal and Peterson Jr, 1979) and alias-table sampling framework, adapted from LightLDA (Yuan et al., 2015). These optimizations, which achieve $\mathcal{O}(1)$ amortized sampling time per token, allow our models to be fit to large corpora with up to thousands of topics in a matter of hours — an order of magnitude speed-up from ccLDA.

After reviewing work related to topic modeling across collections, section 3 describes C-LDA and C-HDP, and then details their technical relationship to existing models. Section 5 introduces the synthetic data and part of the JSTOR corpus used in our evaluations. We then compare our models' performances to other models in terms of held-out perplexity and a measure of distinguishability. The final results section exemplifies the use of C-LDA in a qualitative analysis of humanities and sciences research. We conclude with a brief discussion of the strengths of C-LDA and C-HDP, and outline directions for future work and applications.

2 Related Work

Our models seek to enable users to compare large collections that may only be weakly correlated and that may contain different numbers of topics. While topic models could be fit to separate collections to make post-hoc comparisons (Denny et al., 2014; Yang et al., 2011), our goal is to account for both document-topic asymmetry and topic-word asymmetry "in-model". In short, we seek to model the correlation between *arbitrary* collections. Prioritizing in-model solutions for document-topic asymmetry has been explored elsewhere, such as in hierarchical Dirichlet processes (HDP), which use an additional level to account for collection variations in document-topic distributions (Teh et al., 2006).

One method designed to model topic-word asymmetry is ccMix (Zhai et al., 2004), which models the generative probability of a word in topic z from collection c as a mixture of shared and collection-specific distributions θ_z:

$$p(w) = \lambda_c \, p(w|\theta_z) + (1 - \lambda_c) \, p(w|\theta_{z,c})$$

where $\theta_{z,c}$ is collection-specific and λ_c controls the mixing between shared and collection-specific topics. ccLDA extends ccMix to the LDA framework and adds a beta prior over λ_c that reduces sensitivity to input parameters (Paul and Girju, 2009). Another approach, differential topic models (Chen et al., 2014), is based on hierarchical Bayesian models over topic-word distributions. This method uses the transformed Pitman-Yor process (TPYP) to model topic-word distributions in each collection, with shared common base measures. As (Paul and Girju, 2009) note, ccLDA cannot accommodate a topic if it is not common across collections — an assumption made by ccMix, ccLDA and the TPYP. In a situation where a topic is found in only one collection, it would either dominate the shared topic portion (resulting in a noisy, collection-specific portion), or it would appear as a mixed topic, revealing two sets of unrelated words (Newman et al., 2010b). C-LDA ameliorates this situation by allowing the number of common and non-common topics to be specified separately and by efficiently sampling the tail

of the document-topic distribution, allowing users to examine less prominent regions of the topic space. C-HDP also grants collections document-topic independence using a hierarchical structure to model the differences between collections.

Due to increased demand for scalable topic model implementations, there has been a proliferation of optimized methods for efficient inference, such as SparseLDA (Yao et al., 2009) and AliasLDA (Li et al., 2014). AliasLDA achieves $\mathcal{O}(K_d)$ complexity by using the Metropolis-Hastings-Walker algorithm and an alias table to sample topic-word distributions in $\mathcal{O}(1)$ time. Although this strategy introduces temporal staleness in the updates of sufficient statistics, the lag is overcome by more iterations, and converges significantly faster. A similar technique by Yuan et al. (2015), LightLDA, employs cycle-based Metropolis Hastings mixing with alias tables for both document-topic and topic-word distributions. Despite introducing lag in the sufficient statistics, this method achieves $\mathcal{O}(1)$ amortized sampling complexity and results in even faster convergence than AliasLDA. In addition to being fully parallelized, C-LDA adopts this sampling framework to make comparing large collections more tractable for large numbers of topics. Our models' efficient sampling methods allow users to fit large numbers of topics to big datasets where variation might not be observed in sub-sampled datasets or models with fewer topics.

3 The Models

3.1 Correlated LDA

In ccLDA (and ccMix), each topic has shared and collection-specific components for each collection. C-LDA extends ccLDA to make it more robust with respect to topic asymmetries between collections (Figure 1a). The crucial extension is that by allowing each collection to define a set of non-common topics in addition to common topics, the model removes an assumption imposed by ccLDA and other inter-collection models, namely that collections have the same number of topics. As a result, C-LDA is suitable for collections without a large proportion of common topics, and can also reduce noise (discussed in Section 2). To achieve this, C-LDA assumes document d in collection c has a multinomial document-topic distribution θ with an asymmetric Dirichlet prior for K_c topics, where the first $K^{\emptyset}$ are common across collections. It is also possible to introduce a tree structure into the model that uses a binomial distribution to decide whether a word was drawn from common or non-common topics. This yields collection-specific background topics by using a binomial distribution instead of a multinomial. However, we prefer the simpler, non-tree version because background topics are unnecessary when using an asymmetric α prior (Wallach et al., 2009a).

The generative process for C-LDA is as follows:

1. Sample a distribution ϕ_k (shared component) from Dir(β) and a distribution σ_k from Beta(δ_1, δ_2) for each common topic $k \in \{1, \ldots, K^{\emptyset}\}$;
2. For each collection c, sample a distribution ϕ_k^c (collection-specific component) from Dir(β) for each common topic $k \in \{1, \ldots, K^{\emptyset}\}$ and non-common topic $k \in \{K^{\emptyset} + 1, \ldots, K_c\}$;
3. For each document d in c, sample a distribution θ from Dir(α_c);
4. For each word w_i in d:
   (a) Sample a topic $z_i \in \{1, \ldots, K_c\}$ from Multi(θ);
   (b) If $z_i \leq K^{\emptyset}$, sample y_i from Binomial(σ_{z_i});
   (c) Sample w_i from Multi($\phi_{z_i}^{\xi}$), where
   $$\xi = \begin{cases} \texttt{null} & , z_i \leq K^{\emptyset} \text{ and } y_i = 0; \\ c & , \text{otherwise.} \end{cases}$$

Note that to capture common topics, $K^{\emptyset}$ should be set such that $\exists\ c$ where $K_c = K^{\emptyset}$. Otherwise, words sampled as a non-common topic will not have information about non-common topics in other collections. Then a "common-topic word" is found among non-common topics in all collections (a local minima) and it will take a long time to stabilize as a common topic. To avoid this, when determining the number of topics for sampling, the number of non-common topics for the collection with the smallest number of total topics should be zero. After inference, to distinguish common and non-common topics in this collection, we model σ independently by assuming collections have the same mixing ratio for common topics. With this reasonable assumption and an asymmetric α, common topics become sparse enough that some σ distributions reduce nearly to 0, distinguishing them as non-common topics. Although this may seem counterintuitive, it does not negatively affect results.

Three kinds of collection-level imbalance can confound inter-collection topic models: 1) in the numbers of topics between collections, 2) in the numbers of documents between collections, and 3) in the document-topic distributions. Each of

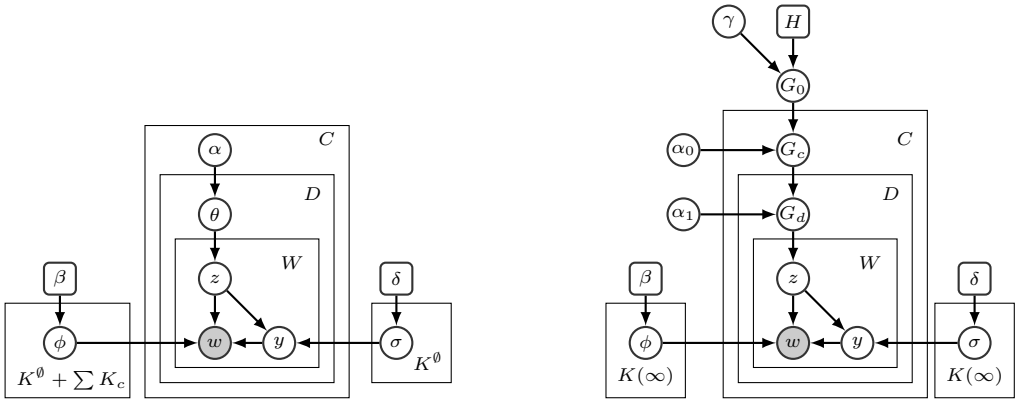

Figure 1: Graphical models of C-LDA (a; left) and C-HDP (b; right).

these can cause topics in different collections to have significantly different numbers of words assigned to the same topic. In this way, a topic can be dominated by the collection comprising most of its words. C-LDA addresses imbalances in the document-topic distributions between collections by estimating α. For imbalance in the number of topics and documents, C-LDA mimics document over-sampling in the Gibbs sampler using a different unit-value in the word count table for each collection. Specifically, a unit η_c is chosen for each collection such that the average equivalent number of assigned words per-topic ($\sum_{d \in c} \eta_c N_d / K_c$, where N_d is the length of document d) is equal. This process both increases the topic quality (in terms of collection balance) in the resulting held-out perplexity of the model.

3.2 Correlated HDP

To alleviate C-LDA's requirement that $\exists\ c$ such that $K_c = K^{\emptyset}$, we introduce a variant of the model, the correlated hierarchical Dirichlet process (C-HDP), that uses a 3-level hierarchical Dirichlet process (Teh et al., 2006). The generative process for C-HDP is the same as C-LDA shown above, except that here we assume a word's topic, z, is generated by a hierarchical Dirichlet process:

$$\begin{aligned} G_0|\gamma, H &\sim \text{DP}(\gamma, H) \\ G_c|\alpha_0, G_0 &\sim \text{DP}(\alpha_0, G_0) \\ G_d|\alpha_1, G_c &\sim \text{DP}(\alpha_1, G_c) \\ z|G_d &\sim G_d \end{aligned}$$

where G_0 is a base measure for each collection-level Dirichlet process, and G_c are base measures of document-level Dirichlet processes in each collection (Figure 1b). Thus, documents from the same collection will have similar topic distributions compared to those from other collections, and collections are allowed to have distinct sets of topics due to the use of HDP.

4 Inference

4.1 Posterior Inference in C-LDA

C-LDA can be trained using collapsed Gibbs sampling with ϕ, θ, and σ integrated out. Given the status assignments of other words, the sampling distribution for word w_i is given by:

$$\begin{aligned} & p(y_i, z_i | \boldsymbol{w}, \boldsymbol{y}_{-i}, \boldsymbol{z}_{-i}, \delta, \alpha, \beta) \\ \propto\ & \underbrace{(N(d, z_i) + \alpha_{c,z_i})}_{q_d} \\ & \times \underbrace{\begin{cases} \frac{N(y_i, z_i) + \delta_{y_i}}{N(z_i) + \sum_k \delta_k} \times \frac{N(w_i, y_i, z_i, \zeta) + \beta}{N(y_i, z_i, \zeta) + V\beta} & z_i \le K^{\emptyset} \\ \frac{N(w_i, z_i, c) + \beta}{N(z_i, c) + V\beta} & z_i > K^{\emptyset} \end{cases}}_{q_w} \end{aligned} \tag{1}$$

where $\zeta = \begin{cases} * & y_i = 0 \\ c & y_i = 1 \end{cases}$, $N(\cdots)$ is the number of status assignments for $(\cdots)$, not including w_i.

Inference in C-LDA employs two optimizations: a parallelized sampler and an efficient sampling algorithm (Algorithm 1). We use the parallel schema in (Smola and Narayanamurthy, 2010; Lu et al., 2013) which applies atomic updates to the sufficient statistics to avoid race conditions. The key idea behind the optimized sampler is the combination of alias tables and the Metropolis-Hastings method (MH), adapted from (Yuan et al., 2015; Li et al., 2014). Metropolis-Hastings is a Markov chain Monte Carlo method that uses a proposal distribution to approximate the true distribu-

Algorithm 1 Sampling in C-LDA

repeat
  for all documents $\{d\}$ in parallel **do**
    for words $\{w\}$ in d **do**
      $z \leftarrow \text{CYCLEMH}(p, q_w, q_d, z)$
      sample y given z
    Atomic update sufficient statics
  Estimate α
until convergence

procedure $\text{CYCLEMH}(p, q_w, q_d, z)$
  for $i = 1$ **to** N **do**
    if i is even **then**
      proposal $q \leftarrow q_w$
    else
      proposal $q \leftarrow q_d$
    sample $z' \sim \text{ALIASTABLE}(q)$
    if RandUnif(1) $< \min(1, \frac{p(z')q(z)}{p(z)q(z')})$ **then**
      $z \leftarrow z'$
  return z

tion when exact sampling is difficult. In a complimentary way, Walker's alias method (2004) allows one to effectively sample from a discrete distribution by using an alias table, constructed in $\mathcal{O}(K)$ time, from which we can sample in $\mathcal{O}(1)$. Thus, reusing the sampler K times as the proposal distribution for Metropolis-Hastings yields $\mathcal{O}(1)$ amortized sampling time per-token.

Notice that in Eq. 1, the sampling distribution is the product of a single document-dependent term q_d and a single word-dependent term q_w. After burn-in, both terms will be sparse (without the smoothing factor). It is therefore reasonable to use q_d and q_w as cycle proposals (Yuan et al., 2015), alternating them in each Metropolis-Hastings step. Our experiments show that the primary drawback of this method — stale sufficient statistics — does not empirically affect convergence. Our implementation uses proposal distributions q_w and q_d, with y marginalized out. After the Metropolis-Hastings steps, y is sampled to update z, to reduce the size of the alias tables, yielding even faster convergence.

Lastly, the use of an asymmetric α allows C-LDA to discover correlations between less dominant topics across collections (Wallach et al., 2009a). We use Minka's fixed-point method, with a gamma hyper-prior to optimize α_c for each collection separately (Wallach, 2008). All other hyperparameters were fixed during inference.

4.2 Posterior Inference in C-HDP

C-HDP uses the block sampling algorithm described in (Chen et al., 2011), which is based on the Chinese restaurant process metaphor. Here, rather than tracking all assignments (as the samplers given in (Teh et al., 2006)), table indicators are used to track only the start of new tables, which allows us to adopt the same sampling framework as C-LDA. In the Chinese restaurant process, each Dirichlet process in the hierarchical structure is represented as a restaurant with an infinite number of tables, each serving the same dish. New customers can either join a table with existing customers, or start a new table. If a new table is chosen, a proxy customer will be sent to the parent restaurant to determine the dish served to that table.

In the block sampler, indicators are used to denote a customer creating a table (or tables) up to level u (0 as the root, 1 for collection level, and 2 for the document level), and $u = \emptyset$ indicates no table has been created. For example, when a customer creates a table at the collection level, and the proxy customer in the collection level creates a table at the root level, u is 0. With this metaphor, let n_{lz} be the number of customers (including their proxies) served dish z at restaurant l, and let t_{lz} be the number of tables serving dish z at restaurant l ($l = 0$ for root, $l = c$ for collection level or $l = d$ for document level), with $N_0 = \sum_z n_{0z}$ and $N_c = \sum_z n_{cz}$. By the chain rule, the conditional probability of the state assignments for w_i, given all others, is

$$
\begin{aligned}
&p(y_i, z_i, u_i | \boldsymbol{w}, \boldsymbol{y}_{-i}, \boldsymbol{z}_{-i}, \boldsymbol{u}_{-i}, \ldots) \\
\propto &\frac{N(y,z) + \delta_y}{N(z) + \sum_k \delta_k} \times \frac{N(w,y,z,\zeta) + \beta}{N(y,z,\zeta) + V\beta} \\
\times &\begin{cases}
\frac{\gamma\alpha_0}{\gamma + N_0} & u = 0 \\
\frac{\alpha_0}{\gamma+N_0} \frac{S^{n_{cz}+1}_{t_{cz}+1}}{S^{n_{cz}}_{t_{cz}}} \frac{S^{n_{dz}+1}_{t_{dz}+1}}{S^{n_{dz}}_{t_{dz}}} \frac{n^2_{0z}(t_{cz}+1)(t_{dz}+1)}{(n_{0z}+1)(n_{cz}+1)(n_{dz}+1)} & u = 1 \\
\frac{S^{n_{cz}+1}_{t_{cz}}}{S^{n_{cz}}_{t_{cz}}} \frac{S^{n_{dz}+1}_{t_{dz}+1}}{S^{n_{dz}}_{t_{dz}}} \frac{(t_{dz}+1)(n_{cz}-t_{cz}+1)}{(n_{cz}+1)(n_{dz}+1)} & u = 2 \\
\frac{\alpha_0+N_1}{\alpha_1} \frac{S^{n_{dz}+1}_{t_{dz}}}{S^{n_{dz}}_{t_{dz}}} \frac{n_{dz}-t_{dz}+1}{n_{dz}+1} & u = \emptyset
\end{cases}
\end{aligned}
$$

Here, S^n_t is the Stirling number, the ratios of which can be efficiently precomputed (Buntine and Hutter, 2010). The concentration parameters γ, α_0, and α_1 can be sampled using the auxiliary variable method (Teh et al., 2006).

Note that because conditional probability has the same separability as C-LDA (to give term q_w and q_d), the same sampling framework can be used with two alterations: 1) when a new topic is created or removed at the root, collection, or document level, the related alias tables must be reset, which makes the sampling slightly slower

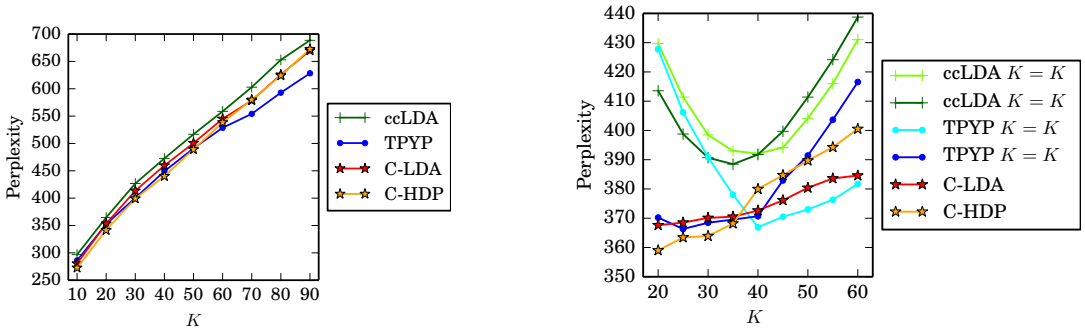

Figure 2: Held-out perplexity of C-LDA, C-HDP, ccLDA and TPYP fit to synthetic data, where $K_1 = K_2 = K$ (a; left) and data with an asymmetric number of topics (b; right).

than $\mathcal{O}(1)$, and 2) while the document alias table samples z and u simultaneously, after sampling z from the word alias table u must be sampled using t_{lc}/n_{lz} (Chen et al., 2011). Parallelizing C-HDP requires an additional empirical method of merging new topics between threads (Newman et al., 2009), which is outside of the scope of this work. Our implementation of both models, C-LDA and C-HDP, are open-sourced online [1].

5 Experiments

5.1 Model Comparison

We use perplexity on held-out documents to evaluate the performance of C-LDA and C-HDP. In all experiments, the gamma prior for α in C-LDA was set to $(1, 1)$, and $(5, 0.1)$, $(5, 0.1)$, $(0.1, 0.1)$ for γ, α_0, α_1 respectively in C-HDP. In the hold-out procedure, 20% of documents were randomly selected as test data. LDA, C-LDA and ccLDA were run for 1,000 iterations and C-HDP and the TII-variant of TPYP for 1,500 iterations (unless otherwise noted), all of which converged to a state where change in perplexity was less than 1% for ten consecutive iterations.

Perplexity was calculated from the marginal likelihood of a held-out document $p(\mathbf{w}|\mathbf{\Phi}, \alpha)$, estimated using the "left-to-right" method (Wallach et al., 2009b). Because it is difficult to validate real-world data that exhibits different kinds of asymmetry, we use synthetic data generated specifically for our evaluation tasks (AlSumait et al., 2009; Wallach et al., 2009b; Kucukelbir and Blei, 2014).

5.1.1 Topic Correlation

C-LDA is unique in the amount of freedom it allows when setting the number of topics for collections. To assess the models' performances with various topic correlations in a fair setting, we generated two collections of synthetic data by following the generative process (varying the number of topics) and measured the models' perplexities against the ground truth parameters. In each experiment, two collections were generated, each with 1,000 documents containing 50 words each, over a vocabulary of 3,000. β and δ were fixed at 0.01 and 1.0 respectively, and α was asymmetrically defined as $1/(i + \sqrt{K_c})$ for $i \in [0, K_c - 1]$.

Completely shared topics The assumptions imposed by ccLDA and TPYP effectively make them a special case of our model where $K^{\emptyset} = K_1 = K_2 = \ldots$. To compare results, data was generated such that all numbers of topics were equal to $K \in [10, 90]$. Additionally, all models were configured to use this ground truth parameter when training. Not surprisingly, ccLDA, C-LDA, and C-HDP have almost the same perplexity with respect to K because their structure is the same when all topics are shared (Figure 2a).

Asymmetric numbers of topics To explore the effect of asymmetry in the number of topics, data was generated such that one collection had $K_1 \in [20, 60]$ topics while a second had a fixed $K_2 = 40$ topics. The number of shared topics was set to $K^{\emptyset} = 20$. The parameters for C-LDA and C-HDP (initial values) were set to ground truths, and, to retain a fair comparison, versions of ccLDA and TPYP were fit with both $K = K_1$ and $K = K_2$.

We find that ccLDA performs nearly as well as C-LDA and C-HDP when there is more symmetry between collection, namely when $K_1 \approx K_2$ (Figure 2b). TPYP, on the other hand, performs well with more topics ($2 \times \max(K_1, K_2)$ where the ground truth is K_1 & K_2). In contrast, C-LDA

[1] https://github.com/iceboal/correlated-lda

and C-HDP perform more consistently than other models across varying degrees of asymmetry.

Partially-shared topics When collections have the same number of topics, C-LDA, C-HDP and ccLDA exhibit adequate flexibility, resulting in similar perplexities. When collections have increasingly few common topics, however, common and non-common topics from ccLDA are considerably less distinguishable than those from C-LDA. To evaluate the models' abilities in such situations, data was generated for two collections having $K_1 = K_2 = 50$ topics, but with the shared number of topics $K^{\emptyset} \in [5, 45]$. We also set $\delta^{(0)} = \delta^{(1)} = 5$, and for comparison to ccLDA we used $K = 50$.

To measure this distinguishability, we examine the inferred σ. Recall that σ indicates what percentage of a common topic is shared. When a topic is actually non-common, the value of σ should be small. We sort σ_k for $k \in [1, K]$ in reverse and use

$$\begin{aligned} \bar{\sigma}_{\text{common}} &= \frac{1}{K^{\emptyset}} \sum_{k=1}^{K^{\emptyset}} \sigma_k \\ \bar{\sigma}_{\text{non-common}} &= \frac{1}{K - K^{\emptyset}} \sum_{k=K^{\emptyset}+1}^{K} \sigma_k \end{aligned} \tag{2}$$

as measures of how well common and non-common topics were learned[2]. $\bar{\sigma}_{\text{common}}$ is the average of the $K^{\emptyset}$ largest σ values, and $\bar{\sigma}_{\text{non-common}}$ is the average of the rest. When $\delta^{(0)} = \delta^{(1)}$ in the synthetic data, σ in the common portion should be 0.5, whereas it should be 0 in the non-common part. Figure 3 shows that C-LDA better distinguishes between common and non-common topics, especially when $K^{\emptyset}$ is small. This allows non-common topics to be separated from the results by examining the value of σ. C-HDP has similar performance but larger σ values. In ccLDA, all topics are shared between collections which means that common and non-common topics are mixed. As expected, ccLDA performs similarly when all topics are common across collections.

5.2 Semantic Coherence

Semantic coherence is a corpus-based metric of the quality of a topic, defined as the average pairwise similarity of the top n words (Newman et al., 2010a; Mimno et al., 2011). A PMI-based form of coherence, which has been found to be the best

[2]TPYP is not comparable using this metric, but its hierarchical structure will cause topics to mix naturally.

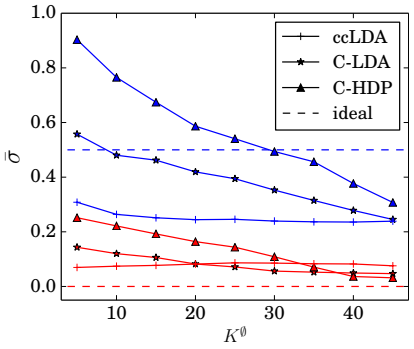

Figure 3: Distinguishability (Eq. 2) of topics fit with C-LDA,C-HDP and ccLDA. Blues lines denote $\bar{\sigma}_{\text{common}}$ and red denote $\bar{\sigma}_{\text{non-common}}$.

proxy human judgements of topic quality, is defined for a topic k as:

$$\text{C}(k) = \frac{2}{n(n-1)} \sum_{\substack{(w_i, w_j) \in k \\ i<j}}^{n} \log \frac{D(w_i, w_j) + 1}{D(w_i) D(w_j)}$$

where $D(\cdot)$ computes the document co-occurrence. To accommodate coherence with common topics in C-LDA that have shared and collection-specific components we define *mutual coherence*, $\text{MC}(k)$, as

$$\text{MC}(k) = \frac{1}{n^2} \sum_{\substack{w_i \in \text{shared}, \\ w_j \in \text{collection-specific}}}^{n} \log \frac{D(w_i, w_j) + 1}{D(w_i) D(w_j)}$$

so that for each collection, $\text{C}(k)$ ($2n$ words) is equal to $\text{C}(k, \text{shared})$ + $\text{C}(k, \text{collection-specific})$ + $\text{MC}(k)$. Table 1 shows the semantic coherence of topics fit with ccLDA and C-LDA. We used a 10% sample of JSTOR due to the limited speed of ccLDA, using 50 (common) topics for ccLDA / C-LDA, and 250 non-common humanities topics for C-LDA. Although these settings are different for the models, the science topics are still comparable because they both have 50 topics. We found that C-LDA provides improved coherence in nearly all situations.

5.2.1 Inference Efficiency

To compare the model efficiency, we timed runs on a sample of 5,036 documents from JSTOR (introduced in the next section) with a 20% held-out and set $K = K_1 = K_2 = 200$ run on a commodity computer with four cores and 16GB of memory. Figure 4a shows the perplexity over

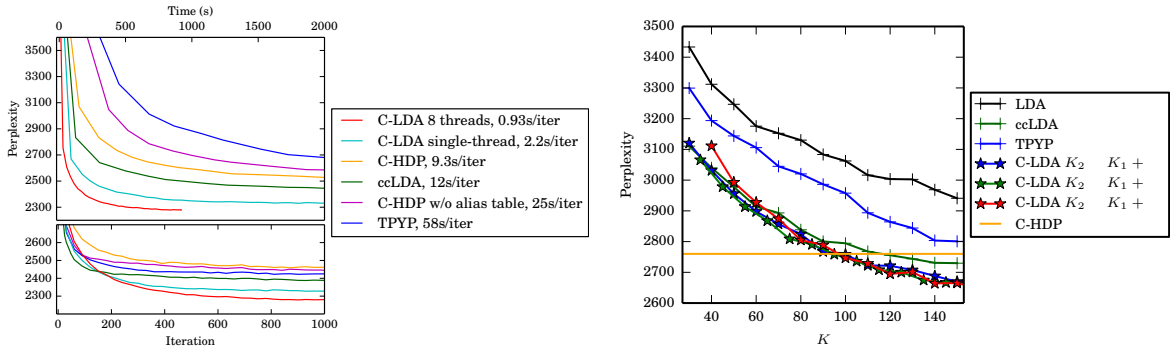

Figure 4: Using JSTOR: perplexity vs. runtime and iterations (a; left) and perplexity vs. K (b; right).

	Coherence					**Mutual Coherence**	
	shared component			collection-specific		shared & collection-specific	
	all documents	science	humanities	science	humanities	science	humanities
C-LDA	-8.83	-7.73	-8.04	-8.38	-8.14	-8.54	-8.37
ccLDA	-9.04	-8.22	-8.27	-8.38	-8.15	-8.69	-8.40
C-LDA	-7.22	-3.68	-6.11	-8.25	-8.09	-7.75	-7.97
ccLDA	-8.11	-5.68	-7.12	-8.24	-7.88	-8.22	-7.95

Table 1: Average semantic coherence of the 50 common topics from JSTOR (top) and the average of the 10 best common topics judged by the mean value of different types of coherence (bottom).

time and iterations. The inference algorithm introduces some staleness, which yields slower convergence in the first 200 iterations. This effect, however, is outweighed in both C-LDA and C-HDP by the increased sampling speed. With 8 threads, C-LDA not only converges faster, but yields lower perplexity, likely due to threads introducing additional stochasticity.

5.3 Performance on JSTOR

To compare our models against slower models, we sampled 2,465 documents from JSTOR, withholding 20% as testing set. We fit a model with 100 common and 50 non-common initial topics using C-HDP, which produced 272 root topics after 2,000 iterations.The perplexity scores are roughly the same when C-LDA uses the same average number of topics per collection (Figure 4b), except when numbers of topics are very asymmetric. Our model begins to outperform ccLDA after 80 topics. C-HDP did not, however, out-perform C-LDA despite the original HDP outperforming LDA. This could be do to the fact that the hierarchical structure of C-HDP is considerably different than the typical 2-level HDP. Held-out perplexity on real data provides a quantitative evaluation of our models' performance in a real-world setting. However, the goal of our models is to enable a deeper analysis of large, weakly-related corpora, which we next discuss.

5.4 Qualitative Analysis

Our models are designed to enable researchers to compare collections of text in a way that is scalable and sensitive to collection-level asymmetries. To demonstrate that C-LDA can fill this role, we fit a model to the entire JSTOR sciences and humanities collections with 100 science topics and 1000 humanities topics (to reveal the less popular science-related topics in the humanities), and $\beta = 0.01, \delta = 1.0$. JSTOR includes books and journal publications in over 9 million documents across nearly 3 thousand journals. We used the journal *Science* to represent a collection of scientific research and 76 humanist journals to represent humanities research[3]. Words were lemmatized, and the most and least frequent words discarded. The final humanities collection contained 149,734 documents and the sciences collection had 160,680 documents, with a combined vocabulary of 21,513 unique words. Together, these collections typify a real-world situation where there is likely some, but not overwhelming correlation.

The results indicate that the sciences and humanities share several topics. Both exhibit an interest in a "non-human" theme (common topic #2; Table 2). This topic is quite similar in both collections (*pig* and *monkey* for science documents; *bird* and *gorilla* for humanities documents), while their shared component forms a cohesive topic (*animal*,

[3]The list is available at http://j.mp/humanities-txt.

Topic 2			Topic 21			Topic 23		
shared	**science**	**humanities**	**shared**	**science**	**humanities**	**shared**	**science**	**humanities**
animal	pig	beast	economic	cost	rural	particle	energy	universe
specie	fly	creature	government	industry	local	physic	electron	quantum
dog	monkey	nonhuman	economy	company	community	physicist	ray	physic
wild	guinea	natural	trade	price	village	energy	ion	technical
wolf	primate	humanity	major	market	region	experiment	atom	scientific
monkey	worm	bird	growth	product	urban	event	particle	relativity
horse	dog	living	capital	income	country	measurement	mass	physical
sheep	cat	gorilla	industry	industrial	area	atom	neutron	mechanic
lion	mammal	brute	institution	business	regional	interaction	proton	law
cat	cattle	ape	support	private	population	atomic	nucleus	reality

Table 2: Three topics from the JSTOR collections with their top words in shared and specific components. Complete results available at http://j.mp/jstor-html.

specie, and *monkey*). This kind of correlation is also evident in topic #23, about physics. While the science documents clearly represent research in particle physics, it is interesting to find the topic is also represented by humanist research focused on cultural representations of science. This reflects a growing interest in science and technology studies that has gained recent traction in the humanities. Despite their differences, both collections engage with a similar theme, seen in the shared component with words like particle, *energy* and *atom*.

The results also indicate that while sciences and humanities documents can share themes, they often diverge in how they are discussed. For example, common topic #21 could be identified as *economic* or *capitalist*, but in the collection-specific components, the two disciplines differ in their articulatation. Science uses terms like *price* and *market*, indicating an acceptance of free-market capitalism (especially as it affects the practice of science), while the humanities, which has long been critical of free-market capitalism, uses terms like *rural* and *community*, highlighting cultural facets of modern economics. These results provide evidence about how ideas move between the sciences and humanities — a phenomenon that constitutes a growing area of research for historians (Galison, 2003; Canales, 2015). C-LDA provides empirical, measurable, and reproducible evidence of the shared research between these disciplines, as well as how concepts are articulated.

6 Discussion

Our models provide a robust way to explore large and potentially weakly-related text collections without imposing assumptions about the data. Like ccLDA and TPYP, our models account for topic-word variation at the collection level. The models accommodate asymmetry in the numbers of topics (set in C-LDA, fit in C-HDP) and provide an efficient inference method which allows them to fit data with large values for K, which can help find correlations in less prevalent topics. Our primary contribution is our models' ability to accommodate asymmetries between arbitrary collections. JSTOR, the world's largest digital collection of humanities research, was an ideal application setting given the size, asymmetry, and comprehensiveness of the humanities collection. As we show, humanities and science research exhibit asymmetries with regard to vocabulary and topic structure — asymmetries that would be systematically overlooked using existing models. By characterizing common topics as mixtures of shared and collection-specific components, we can capture a kind of topic-level homophily, where similar themes are articulated in different ways due to word-, document-, and collection-level variation. Future work on these models could explore methods to fit non-common topics for both collections. In general, C-LDA and C-HDP can be used whenever documents are sampled from ostensibly different populations, where the nature of the difference is unknown.

Acknowledgements

Thanks to David Blei for advice on applications of the model. This work contains analysis of private, or otherwise restricted data, made available to James Evans and Eamon Duede by ITHAKA (JSTOR), the opinions of whom are not represented in this paper. Jaan Altosaar acknowledges support from the Natural Sciences and Engineering Research Council of Canada. This work was supported by a grant from the Templeton Foundation to the Metaknowledge Research Network and by grant #1158803 from the National Science Foundation.

References

Loulwah AlSumait, Daniel Barbará, James Gentle, and Carlotta Domeniconi. 2009. Topic significance ranking of LDA generative models. In *Machine Learning and Knowledge Discovery in Databases*, pages 67–82. Springer.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Wray Buntine and Marcus Hutter. 2010. A Bayesian view of the Poisson-Dirichlet process. *arXiv preprint arXiv:1007.0296*.

Jimena Canales. 2015. *The Physicist and the Philosopher: Einstein, Bergson, and the Debate that Changed our Understanding of Time*. Princeton University Press, Princeton, NJ.

Changyou Chen, Lan Du, and Wray Buntine. 2011. Sampling table configurations for the hierarchical Poisson-Dirichlet process. In *Machine Learning and Knowledge Discovery in Databases*, pages 296–311. Springer.

Changyou Chen, Wray Buntine, Nan Ding, Lexing Xie, and Lan Du. 2014. Differential topic models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

M. Denny, J. ben Aaron, H. Wallach, and B. Desmarais. 2014. Modeling email network content and structure. In *the 72nd Annual Midwest Political Science Association Conference, 2014; the Northeast Political Methodology Meeting, 2014; the 7th Annual Political Networks Conference, the Society for Political Methodology 31st Annual Summer Meeting*.

Peter Galison. 2003. *Poincar's Maps*. Norton, New York, NY.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57.

Richard A Kronmal and Arthur V Peterson Jr. 1979. On the alias method for generating random variables from a discrete distribution. *The American Statistician*, 33(4):214–218.

Alp Kucukelbir and David M Blei. 2014. Profile predictive inference. *arXiv preprint arXiv:1411.0292*.

Aaron Q Li, Amr Ahmed, Sujith Ravi, and Alexander J Smola. 2014. Reducing the sampling complexity of topic models. In *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 891–900.

Mian Lu, Ge Bai, Qiong Luo, Jie Tang, and Jiuxin Zhao. 2013. Accelerating topic model training on a single machine. In *Web Technologies and Applications*, pages 184–195. Springer.

George Marsaglia, Wai Wan Tsang, and Jingbo Wang. 2004. Fast generation of discrete random variables. *Journal of Statistical Software*, 11:1–8.

David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272.

David Newman, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2009. Distributed algorithms for topic models. *The Journal of Machine Learning Research*, 10:1801–1828.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010a. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108.

David Newman, Youn Noh, Edmund Talley, Sarvnaz Karimi, and Timothy Baldwin. 2010b. Evaluating topic models for digital libraries. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, JCDL '10, pages 215–224. ACM.

Michael Paul and Roxana Girju. 2009. Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3*, pages 1408–1417.

Alexander Smola and Shravan Narayanamurthy. 2010. An architecture for parallel topic models. *Proceedings of the VLDB Endowment*, 3(1-2):703–710.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2006. Hierarchical Dirichlet processes. *Journal of the american statistical association*, 101(476).

Hanna M Wallach, David Mimno, and Andrew Mccallum. 2009a. Rethinking LDA: Why priors matter. In *Advances in Neural Information Processing Systems*, pages 1973–1981.

Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009b. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1105–1112, New York, NY, USA. ACM.

Hanna M Wallach. 2008. Structured topic models for language. *Unpublished doctoral dissertation, Univ. of Cambridge*.

Tze-I Yang, Andrew J Torget, and Rada Mihalcea. 2011. Topic modeling on historical newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 96–104.

Limin Yao, David Mimno, and Andrew McCallum. 2009. Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 937–946.

Jinhui Yuan, Fei Gao, Qirong Ho, Wei Dai, Jinliang Wei, Xun Zheng, Eric Po Xing, Tie-Yan Liu, and Wei-Ying Ma. 2015. Lightlda: Big topic models on modest computer clusters. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1351–1361. International World Wide Web Conferences Steering Committee.

ChengXiang Zhai, Atulya Velivelli, and Bei Yu. 2004. A cross-collection mixture model for comparative text mining. In *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining*, pages 743–748.