## Title
Local and global topics in text modeling of web pages nested in web sites



## Authors
Wang, Jason
Weiss, Robert E

## Publication Date
2022-09-01

## DOI
10.1016/j.csda.2022.107518

Peer reviewed

# Local and Global Topics in Text Modeling of Web Pages Nested in Web Sites

Jason Wang[a,1,*], Robert E. Weiss[a]

[a]*Department of Biostatistics, Fielding School of Public Health, University of California at Los Angeles.*
*650 Charles E Young Dr S. 51-254 CHS*
*Los Angeles, CA 90095-1772*
*United States of America*

**Abstract**

Topic models assert that documents are distributions over latent topics and latent topics are distributions over words. A nested document collection has documents nested inside a higher order structure such as articles nested in journals, podcasts within authors, or web pages nested in web sites. In a single collection of documents, topics are global or shared across all documents. For web pages nested in web sites, topic frequencies likely vary across web sites and within a web site, topic frequencies almost certainly vary from web page to web page. A hierarchical prior for topic frequencies models this hierarchical structure with a global topic distribution, web site topic distributions varying around the global topic distribution, and web page topic distributions varying around the web site topic distribution.

Web pages in one United States local health department web site often contain local geographic and news topics not found on web pages of other local health department web sites. For web pages nested in web sites, some topics are likely *local topics* and unique to an individual web site. Regular topic models ignore the nesting structure and may identify local topics but cannot label those topics as local nor identify the corresponding web site owner. Explicitly modeling local topics identifies the owning web site and identifies the topic as local. In US health web site data, topic coverage is defined at the web site level after removing local topic words from pages. Hierarchical local topic models can be used to study how well health topics are covered.

*Keywords:* hierarchical models, text analysis, Bayesian models, public health, internet

## 1. Introduction

Topic models have been used to abstract topical information from collections of text documents such as journal abstracts, tweets, and blogs (Griffiths and Steyvers, 2004; Liu et al., 2009; Paul and Dredze, 2014; Boyd-Graber et al., 2017). Topic models are hierarchical models that define documents as distributions over latent topics and topics as distributions over words. In topic models, each topic is characterized by a vector of word probabilities and each document is characterized by a vector of topic probabilities. Topic-word distributions and document-topic distributions describe the prevalence of words in a topic and topics in a document, respectively. Topics are generally assumed global or shared across all documents (Blei et al., 2003; Rosen-Zvi et al., 2004; Blei and Lafferty, 2005; Chang and Blei, 2009; Roberts et al., 2013). However, this may not be the case for a nested document collection, where documents are nested inside a higher structure. Examples of nested document collections include articles nested within journals, blog posts nested within authors, episodes nested within television show, and web pages nested within web sites. In a nested document collection, some topics may be unique to a group of documents, and we refer to these topics as local topics.

We collected text from web pages nested in web sites of local health departments in the United States. We wish to abstract topics from the text and study if and how health topics are covered across web sites. Each web site contains many web pages. Thus, we have a collection of web pages nested within web sites. These web sites have local words and phrases such as geographical names and places that are common within a web site, but are rarely seen on other web sites. Other local words and phrases can be found in local events and local news. The content of local topics, how frequent local topic words occur and where local topics are found on a page vary substantially across web sites and web pages. Thus it is difficult to identify local topics a priori and instead we take a probabilistic approach.

We propose local topic extensions to topic models to accommodate and identify local topics. Local topics can be extensive on individual web pages and can

---

⋆Web appendix, data, and code for HALT-LDA are in supplemental materials.

*Corresponding author

*Email address:* jbwang321@gmail.com (Jason Wang)

[1]*Present address:* 650 Charles E Young Dr S. 51-254 CHS
UCLA Fielding School of Public Health
Los Angeles, California 90095-1772
United States of America

comprise substantial portions of a web site. Local topics do not contribute to our desired inferences and explicitly identifying local topics facilitates drawing conclusions. Effectively, local topics are removed from web pages before we make further inference. We apply our extensions to latent Dirichlet allocation (LDA) models which place Dirichlet priors on topic-word and document-topic distributions (Blei et al., 2003).

In a collection of documents, an asymmetric prior on document-topic distributions has been recommended for improved performance over symmetric priors, although symmetric priors remain common and default in applications (Wallach et al., 2009a; Grün and Hornik, 2011). We expect that a hierarchical asymmetric prior would then fit better for a nested collection of documents. A fully hierarchical prior for topic frequencies for nested documents includes a global topic distribution, web site topic distributions varying around the global topic distribution, and web page topic distributions varying around their web site topic distribution.

We consider four models applied to web pages as documents with all four models indexed by the number of global topics. The first model is a traditional LDA model with an asymmetric prior on document-topic distributions. The second model places a hierarchical asymmetric (HA-LDA) prior on document-topic distributions of the web pages. An asymmetric prior on document-topic distributions accommodates the belief that some topics are more common than others across all web pages and web sites. A hierarchical asymmetric prior further adds that which topics are more common varies from web site to web site and also within web site from web page to web page.

Our third (LT-LDA) and fourth models (HALT-LDA) add local topics, one unique local topic per web site, into the LDA and HA-LDA models. All four models have a fixed maximum number $K$ of global topics. We consider a wide range of values for $K$.

Documents with relationships more complex than a single collection of exchangeable documents have been studied in diverse data settings (Chang and Blei, 2009; Rosen-Zvi et al., 2004; Qiang et al., 2017; Chemudugunta et al., 2006; Hua et al., 2020). Under the author model of Rosen-Zvi et al. (2004), documents have one or several known authors, authors have distributions over topics, topics are distributions over words. Authors contribute to one or several documents in different mixtures, which means documents have commonalities defined by shared authorship. The author list is a categorical covariate associated with a document, and so each document has different distributions of topics. However, two documents with the same author list would have the same dis-

tribution over topics. Dynamic topic models are used to model collections of documents whose topic distributions evolve over time Blei and Lafferty (2006); Wang et al. (2008); Zhang et al. (2010); Hong et al. (2011). Dynamic topic models would require a sequential relationship along time or other dimension for web pages that does not exist in our data.

Nested document collections can be thought of as a special case of document networks (Chang and Blei, 2009; Guo et al., 2015; Terragni et al., 2020) where links are known. Relational topic models (Chang and Blei, 2009) model the links between any two documents and can be used to predict links to a newly published document (Liu et al., 2009; Chen et al., 2013; Chen et al., 2015). Web page links are equal within a web site, non-existent between web pages from two different web sites. Neither modeling the links in web pages nor predicting currently unknown links is of interest.

One type of document nesting involves secondary documents nested within a primary document (Yang et al., 2016), such as comments nested within blog posts. This data structure has two distinct document types, whereas web pages are our only document type and there is no separate document for a web site.

Nested document collections may be thought of as a collection of different document collections, where each web site is itself a document collection. Local-global latent Dirichlet allocation (Qiang et al., 2017) models a global topic distribution and a collection level topic distribution but does not model a document level topic distribution. Multiple-corpora latent Dirichlet allocation (MLDA) (Foster et al., 2016) and compound latent Dirichlet allocation (CLDA) (George et al., 2019) model different probabilities of topics by document collection but have the same symmetric prior for each collection's topic distribution. Wang et al. (2009); Shen et al. (2008) model relationships among document collections with word distributions similar but not identical across collections and Wang et al. (2009) allows for collection specific-topics.

For a single document collection, the special words topic model with background distribution (SWB) models a global set of topics, one collection-level topic for common words, and one unique topic for each document (Chemudugunta et al., 2006). The correlated-text-stream model, the global and local topic model, and the common and distinctive topic model, all extend SWB to model multiple document collections (Hong et al., 2011; Liu et al., 2018; Hua et al., 2020). They model, for a nested document collection, a global set of topics and a set of topics for each collection. Hong et al. (2011) and Liu et al. (2018) develop models for tweets, short text documents, and assume that all words in a document come from the same topic. Hua et al. (2020) extends the correlated-text-stream model

such that words in the same document do not all come from the same topic. Other topic models model multiple sets of topic-word distributions (Paul and Girju, 2010; Ge et al., 2015) but do not model a nested document collection.

The common and distinctive topic model (CDTM) (Hua et al., 2020) describes a hierarchical structure of documents nested in collections with symmetric priors and a set of local topics in each collection. However, for our web pages, we are interested in modeling local topics as a nuisance parameter to adjust our inference; thus, we model a single local topic for each web site to simplify our model and avoid searching for an optimal number of local topics. Our models further extend CDTM by placing a more flexible asymmetric or hierarchical asymmetric prior on web page topic distributions.

We show that local topics are not useful for describing words on web pages outside the corresponding local web site. We show this by matching local topics in our HALT-LDA model to global topics in the HA-LDA model and then showing that those matched topics from HA-LDA are not truly global topics but essentially only occur in one web site in the HA-LDA models.

The health department web site data requires additional unique inferences that are not the traditional inferences one would consider when using LDA to analyze a set of reports, newspaper articles, television show transcripts, or books. For the health department web site data we are interested in topic coverage, whether a web site covers a particular topic such as sexually transmitted diseases, emergency preparedness, food safety or heart disease. We are interested in the fraction of web sites that cover a particular topic, and whether a topic is universally covered or not.

Topic coverage has been used to describe the global prevalence of a topic (Song et al., 2009) or the prevalence of a topic in a document (Lu et al., 2011). However, we are interested if a web site covers a topic. A health web site contains many web pages that cover different topics, and typically a health web site dedicates one or a few web pages to a given health topic rather than discussing all health topics across all web pages. Thus, a topic is covered by a web site if a single page covers the topic and we do not consider a topic covered if many pages have relatively few words from that topic. We define topic coverage at the web site level as whether a web site has a page dedicated to that topic, which happens if many or most of the words on a single page are from that topic. Further, local topics may be extensive or may be light on various web pages and extensive local topic coverage should not influence a measure of topic coverage at the page level. Thus using models with explicitly identified local topics, we are able to remove words corresponding to the local topic from a page before

calculating its coverage. An appropriate topic coverage measure at the web page level needs to calculate fraction of coverage of a particular topic ignoring local topics. Web site coverage does not average or sum across pages, rather web site coverage should consider the supremum of coverage across pages.

Section 2 defines notation and our four models. Section 3 discusses computation and inference. Section 4 introduces our motivating data set in greater detail and section 5 lays out our analysis and illustrates the conclusions that are of interest for this data and the conclusions that local models allow for. The paper closes with discussion.

## 2. Topic Models for Nested Web pages

In a collection of web sites, we define a document to be a single web page. Thus, we refer to the document-topic distribution of a web page as the web page-topic distribution. Web sites are indexed by $i = 1, \ldots, M$ and web pages nested within web sites are indexed by $j = 1, \ldots, M_i$, with $M_i$ pages in web site $i$. Words $w_{ijh}$ on a page are indexed by $h = 1, \ldots, N_{ij}$, with $j$th page in web site $i$ having $N_{ij}$ words, and the set of unique words across all web sites and web pages are indexed by $v = 1, \ldots, V$ where $V$ is the number of unique words or the size of the vocabulary. The number of global topics $K$, indexed by $k = 1, \ldots, K$, is assumed fixed and known before modeling as in latent Dirichlet allocation. Table 1 details notation used in our models.

### *2.1. Latent Dirichlet Allocation*

Latent Dirichlet allocation (LDA) asserts that topics are global and their topic-word distributions are drawn from a Dirichlet prior. For Dirichlet distributed parameters $\phi_k$ we use the parameterization

$$\phi_k \sim \text{Dirichlet}(c_\beta \beta),$$

where $\phi_k$ is a $V$-vector of probabilities with $v$th element $\phi_{k,v}$ such that $\sum_{v=1}^{V} \phi_{k,v} = 1$, $0 \leq \phi_{k,v} \leq 1$, $c_\beta > 0$ is a scale parameter, and $\beta$ is a $V$-vector of parameters with $v$th element $\beta_v$ such that $\sum_{v=1}^{V} \beta_v = 1$, $0 \leq \beta_v \leq 1$, and a priori $E[\phi_k | c_\beta \beta] = \beta$. Each web page $j$ in web site $i$ has web page-topic distribution denoted by a $K$ vector of probabilities $\theta_{ij}$ with a Dirichlet($c_\alpha \alpha$) prior. Topic $k$ has a topic-word multinomial distribution parameterized by a $V$-vector of probabilities $\phi_k$ a priori distributed as Dirichlet($c_\beta \beta$). Word $h$ on web page $j$ in web site $i$ has a latent

Table 1: Model notation with definitions.

| Notation | Description |
|---|---|
| $i$ | Web site index, $i = 1, \ldots, M$ |
| $j$ | Web page index, $j = 1, \ldots, M_i$ |
| $h$ | Word index, $h = 1, \ldots, N_{ij}$ |
| $M$ | Number of web sites |
| $M_i$ | Number of pages in web site $i$ |
| $N_{ij}$ | Number of words in page $j$ in web site $i$ |
| $K$ | Number of global topics |
| $L$ | Number of local topics |
| $L_i$ | Number of local topics in web site $i$ |
| $V$ | Number of unique words in the vocabulary |
| $\theta_{ij}$ | Page-topic distribution of web site $i$ web page $j$ |
| $\psi_i$ | Local topic-word distribution of web site $i$ |
| $\phi_k$ | Global topic-word distribution of topic $k$ |
| $w_{ijh}$ | Word $h$ of page $j$ in web site $i$ |
| $z_{ijh}$ | Topic choice of the $h$th word of page $j$ in web site $i$ |

topic $z_{ijh}$. The LDA model is

$$\begin{aligned}
\theta_{ij}|c_\alpha\alpha &\sim \text{Dirichlet}(c_\alpha\alpha),\\
\phi_k|c_\beta\beta &\sim \text{Dirichlet}(c_\beta\beta),\\
z_{ijh}|\theta_{ij} &\sim \text{Categorical}(\theta_{ij}),\\
w_{ijh}|\phi_{z_{ijh}} &\sim \text{Categorical}(\phi_{z_{ijh}}).
\end{aligned}$$

Documents in LDA are characterized by a single distribution over all $K$ topics, thus, LDA has $K$ global topics and no local topics.

### *2.2. Local Topics*

Now we introduce $L$ local topics distributed among $M$ web sites, such that each web site $i$ contains $L_i$ local topics and $L = \sum_{i=1}^{M} L_i$. We let $l = 1, \ldots, L_i$ index local topics in web site $i$. The web page-topic distribution, $\theta_{ij}$, for page $j$ in web site $i$ is now a $(K + L_i)$-vector of probabilities. The topic-word distribution $\psi_{il}$ for each local topic is still a $V$ vector of probabilities with a Dirichlet($c_\gamma\gamma$) prior. We define the $(K + L_i) \times V$ array, $\Phi_i = \{\phi_1, \ldots, \phi_K, \psi_{i1}, \ldots, \psi_{iL_i}\}$, as the combined set of global and local topic-word distributions for web site $i$. The LT-LDA model

is then

$$\begin{aligned}
\theta_{ij}|c_\alpha\alpha &\sim \text{Dirichlet}(c_\alpha\alpha),\\
\psi_{il}|c_\gamma\gamma &\sim \text{Dirichlet}(c_\gamma\gamma),\\
\phi_k|c_\beta\beta &\sim \text{Dirichlet}(c_\beta\beta),\\
z_{ijh}|\theta_{ij} &\sim \text{Categorical}(\theta_{ij}),\\
w_{ijh}|\Phi_{iz_{ijh}} &\sim \text{Categorical}(\Phi_{iz_{ijh}}).
\end{aligned}$$

The shared prior parameter $\alpha$ requires that $L_1 = \ldots = L_M$; however, this can be generalized so that each web site $i$ has a separate and appropriate prior for $\theta_{ij}$. In our applications with local topics, we choose $L_i = 1$ for all $i = 1, \ldots, M$ assuming that most web sites have one local topic that places high probability on geographical names and places.

*2.3. Hierarchical Asymmetric Prior*

A symmetric prior Dirichlet($c_\alpha\alpha$) for web page-topic distributions $\theta_{ij}$ is such that $c_\alpha\alpha = d \times \{1, \ldots, 1\}$ for some constant $d$ and describes a prior belief about the sparsity or spread of page-topic distributions. A smaller $d$ describes the prior belief that web pages have high probability for a small number of topics and low probability for the rest, while a larger $d$ describes the prior belief that web pages have more nearly equal probability for all topics. A single asymmetric prior Dirichlet($c_\alpha\alpha$), such that $c_\alpha\alpha = \{d_1, \ldots, d_{K+1}\}$ where not all $d_k$ are equal, accommodates the belief that topics or groups of words with larger $d_k$ will occur more frequently across all pages than topics with smaller $d_k$.

For a nested document collection, we extend the belief that different topics occur more frequently to multiple levels. Thus a given topic will have different probabilities in different web sites, and also, that topic's probability will vary across web pages within a web site. Globally, some topics are more common than others and while we start with a symmetric Dirichlet prior for the unknown global-topic distribution, the global-topic distribution will be asymmetric. Locally, each web site has its own set of common and uncommon topics with the web site-topic distributions centered at the global-topic distribution. Finally each web page within a web site will have their own common and uncommon topics and web page-topic distributions are centered around the web site-topic distribution. We extend the LDA model in section 2.1 by placing a hierarchical asymmetric prior on web page-topic proportions such that web pages nested within web sites share commonalities. We first place a Dirichlet($c_\alpha\alpha_i$) prior on web page-topic distribution $\theta_{ij}$, such that each web site has a $(K+1)$-vector of parameters $\alpha_i$ so that a priori $E[\theta_{ij}|c_\alpha\alpha_i] = \alpha_i$. We next place a Dirichlet($c_0\alpha_0$) prior on

web site-topic distributions $\alpha_i$. The HA-LDA model is

$$\begin{aligned}
\theta_{ij}|c_\alpha\alpha_i &\sim \text{Dirichlet}(c_\alpha\alpha_i),\\
\alpha_i|c_0\alpha_0 &\sim \text{Dirichlet}(c_0\alpha_0),\\
\phi_k|c_\beta\beta &\sim \text{Dirichlet}(c_\beta\beta),\\
z_{ijh}|\theta_{ij} &\sim \text{Categorical}(\theta_{ij}),\\
w_{ijh}|\phi_{z_{ijh}} &\sim \text{Categorical}(\phi_{z_{ijh}}).
\end{aligned}$$

We further place Gamma priors on $c_\alpha$ and each element of $c_0\alpha_{0,k}$. Combining the hierarchical asymmetric prior with local topics, the HALT-LDA model is

$$\begin{aligned}
\theta_{ij}|c_\alpha\alpha_i &\sim \text{Dirichlet}(c_\alpha\alpha_i),\\
\alpha_i|c_0\alpha_0 &\sim \text{Dirichlet}(c_0\alpha_0),\\
\psi_{il}|c_\gamma\gamma &\sim \text{Dirichlet}(c_\gamma\gamma),\\
\phi_k|c_\beta\beta &\sim \text{Dirichlet}(c_\beta\beta),\\
z_{ijh}|\theta_{ij} &\sim \text{Categorical}(\theta_{ij}),\\
w_{ijh}|\Phi_{iz_{ijh}} &\sim \text{Categorical}(\Phi_{iz_{ijh}}).
\end{aligned}$$

### *2.4. Prior Parameter Specification*

We place an asymmetric prior on $\alpha$ and a Gamma prior on $c_\alpha$ in LDA and LT-LDA. Therefore the difference between LDA and LT-LDA is the addition of local topics and the difference between LDA and HA-LDA is the use of a hierarchical asymmetric prior over a single asymmetric prior. We compare these models to study the impact of each extension. We also compare these models to a model with both a hierarchical asymmetric prior and local topics (HALT-LDA). We specify prior parameters to accommodate sparse mixtures of topics in web pages. In LDA and LT-LDA, we place priors

$$\begin{aligned}
c_\alpha &\sim \text{Gamma}(a_\alpha, b_\alpha), \quad a_\alpha = b_\alpha = 1,\\
\alpha &\sim \text{Dirichlet}(\{1/K^*, \ldots, 1/K^*\}),
\end{aligned}$$

where we use the shape-scale parameterization of the Gamma distribution with mean $a_\alpha b_\alpha$ and where $K^* = K$ in LDA and $K^* = K + 1$ in LT-LDA and HALT-LDA. In HA-LDA and HALT-LDA, we treat $c_0\alpha_{0,k}$ as a single parameter and place priors

$$\begin{aligned}
c_\alpha &\sim \text{Gamma}(a_\alpha, b_\alpha), \quad a_\alpha = b_\alpha = 1,\\
c_0\alpha_{0,k} &\sim \text{Gamma}(1, 1).
\end{aligned}$$

The hyperparameters we chose were appropriate for our public health web sites data set and may not be appropriate for all data sets. We expect topic distributions to vary greatly from page to page, even within a web site. This is because though pages of the same web site share a local topic, each page of a web site is likely dedicated to a different health topic. In a collection of local health department web sites, we expect most web sites to cover similar topics. In other words, $\alpha_i$ varies between $i$ but less so than how $\theta_{ij}$ varies between $j$. We simulated Dirichlet draws to check for appropriate choices of hyperparameters. We generated 100,000 sets of $c_0\alpha_0$ for $K = 50$, where elements $c_0\alpha_{0,k}$ are sampled from Gamma(1, 1). This generates an average largest order statistic for $\alpha_{i,k}$ of 0.09 with a standard deviation of 0.04. At $K = 100$, the average largest order statistic is 0.05 with a standard deviation of 0.02. The largest order statistic from the prior is larger than the overall local topic prevalence in our results in section 5.2; however, a priori, this result for the highest order statistic was reasonable.

We expect each topic to place higher probability on a small subset of words but do not expect any words to have high probability across all topics. Therefore, we place a symmetric prior over topic word distributions, $\phi_k$ and $\psi_i$. The choices of 0.01, 0.05, or 0.10 are common for the symmetric prior parameter of word probabilities (Griffiths and Steyvers, 2004; Rosen-Zvi et al., 2004; Grün and Hornik, 2011; Zhou et al., 2012; Guo et al., 2013; Paul and Dredze, 2014). We set priors $c_\beta\beta = c_\gamma\gamma = \{0.05, \ldots, 0.05\}$. We simulated Dirichlet draws to check that 0.05 is reasonable. We generated 100,000 sets of $\phi$ and $\psi$ for $K = 50$, $M = 20$, and $V = 1614$. This generates an average largest order statistic across all $\phi_{k,v}$ and $\psi_{i,v}$ of 0.04 with a standard deviation of 0.01. We expect the 10 most probable words in a topic hold a large portion of the probability relative to the remaining words. The prior we chose reflects that, where, in our generated data, the 10 largest order statistics hold 0.26 total probability, with 0.74 total probability over the remaining 1604 order statistics. Inspecting topics and their most probable words after modeling is recommended regardless of choice. Sensitivity analysis in section A.2 of the web appendix shows that conclusions from HALT-LDA are robust to deviations from our choice of $c_\beta$, $c_\gamma$, and $a_\alpha$.

## 3. Computation and Inference for Hierarchical Topic Models

The general goal of inference in hierarchical topic models is to estimate the topic-word distributions, $\phi_k$ and $\psi_i$, and web page-topic distributions, $\theta_{ij}$. We use Markov chain Monte Carlo (MCMC) to sample from the posterior, where unknown parameters are sequentially sampled conditional on current values of all other unknown parameters. We outline the sampler for the most complex

model, HALT-LDA, where each web site has $L_i = 1$ local topic $\psi_i$. We implement HALT-LDA with the data and functions available in the first author's github repository https://github.com/jwanghb/publichealth-websites and in the supplementary materials.

Let $W$ and $Z$ be ragged arrays of identical structure, with one element $w_{ijh}$ and $z_{ijh}$ for word $h$ in web page $j$ from web site $i$. The $ijh$ element of $W$ corresponding to the $ijh$ word identifies the index from 1 to $V$ of that word, and the corresponding element $Z_{ijh}$ of $Z$ identifies the topic assigned to that word. As $Z$ is latent, it is sampled and will change at every iteration of the MCMC algorithm. Let $\alpha$ be the set of all web site-topic distributions $\alpha_i$ and similarly, let $\theta$, $\phi$, and $\psi$ be the sets of all $\theta_{ij}$, $\phi_k$, and $\psi_i$. Then the joint prior density of all unknown parameters and data is

$$P(W, Z, \phi, \psi, \theta, c_\alpha, \alpha, c_0\alpha_0) = \\ P(W|Z, \phi, \psi)P(Z|\theta)P(\theta|c_\alpha, \alpha)P(\alpha|c_0\alpha_0)P(c_0\alpha_0)P(\phi)P(\psi).$$

Dirichlet-multinomial conjugacy allows us to algebraically integrate out $\phi_k$, $\psi_{il}$, and $\theta_{ij}$ from the posterior. We are left to sample topics $z_{ijh}$ of each word $w_{ijh}$, scale parameter $c_\alpha$, and web site-topic distributions $\alpha_i$ and their prior parameters $c_0\alpha_{0,k}$.

Let $n_{k,v}$, $p_{i,v}$, and $m_{ij,k}$ be counts that are functions of $Z$ and $W$. These counts vary from iteration to iteration as they depend on $Z$. Let $n_{k,v}$ be the total count of word $v$ assigned to topic $k$, let $p_{i,v}$ be the count of word $v$ from the single local topic of web site $i$, and let $m_{ij,k}$ be the count of words from topic $k$ in page $j$ of web site $i$. Let the superscript $^-$ on counts $n^-_{k,w_{ijh}}$, $m^-_{ij,k}$, and $p^-_{i,w_{ijh}}$ indicate that the counts exclude word $w_{ijh}$. Similarly, let $Z^-$ be the set of topic indices $Z$ excluding word $w_{ijh}$. Then the sampling density for $z_{ijh}$ conditioned on scale parameter $c_\alpha$, web site-topic distribution $\alpha_i$, and the remaining topics indices $Z^-$ is

$$P(z_{ijh} = k|Z^-, c_\alpha, \alpha_i, w_{ijh}) \propto \\ (m^-_{ij,k} + c_\alpha\alpha_{i,k}) \times \left(\frac{n^-_{k,w_{ijh}} + \beta_v}{\sum_{v=1}^V n^-_{k,v} + \beta_v}\right)^{1_{k\leq K}} \times \left(\frac{p^-_{i,w_{ijh}} + \gamma_v}{\sum_{v=1}^V p^-_{i,v} + \gamma_v}\right)^{1_{k=K+1}},$$

where $1_{k\leq K}$ is an indicator function that is one if $k$ is a global topic and zero if $k$ is a local topic and $1_{k=K+1} = 1 - 1_{k\leq K}$. To sample web site-topic distribution $\alpha_i$ we use a data augmentation step with auxiliary variables $\lambda_{ij,k}$ with conditional density

$$P(\lambda_{ij,k}|Z, c_\alpha\alpha_{i,k}, \lambda_{-(ij,k)}) = \frac{\Gamma(c_\alpha\alpha_{i,k})}{\Gamma(c_\alpha\alpha_{i,k} + m_{ij,k})}|s(m_{ij,k}, \lambda_{ij,k})|(c_\alpha\alpha_{i,k})^{\lambda_{ij,k}},$$

where $s(\cdot,\cdot)$ is the Stirling number of the first kind. This step allows posterior draws of web site-topic distribution $\alpha_i$ from a Dirichlet($c_0\alpha_0 + \sum_{j=1}^{M_i} \lambda_{ij}$) (Teh et al., 2006). Parameters $c_\alpha$ and $c_0\alpha_{0,k}$ are sampled using Metropolis-Hastings.

We estimate conditional means of the multinomial parameters $\phi_k$, $\psi_i$, and $\theta_{ij}$ for each MCMC sample, as is common in using MCMC sampling in topic models. Let superscript $(q)$ indicate a count, estimate, or sample from iteration $q$ of the MCMC sample. Each iteration $q$ samples a topic index for every word. The conditional estimate for words of the global topic-word proportions $\phi_{k,v}$ at iteration $q$ is given by the conditional posterior mean

$$\bar{\phi}_{k,v}^{(q)} = \frac{c_\beta \beta_v + n_{k,v}^{(q)}}{\sum_{v=1}^{V} c_\beta \beta_v + n_{k,v}^{(q)}}.$$

Similarly, the conditional posterior means for the local topic-word mixture $\psi_{i,v}$ and web page-topic mixtures $\theta_{ij,k}$ at iteration $q$ are

$$\bar{\psi}_{i,v}^{(q)} = \frac{c_\gamma \gamma_v + p_{i,v}^{(q)}}{\sum_{v=1}^{V} c_\gamma \gamma_v + p_{i,v}^{(q)}},$$

$$\bar{\theta}_{ij,k}^{(q)} = \frac{c_\alpha \alpha_{ik} + m_{ij,k}^{(q)}}{\sum_{k=1}^{K+1} c_\alpha \alpha_{ik} + m_{ij,k}^{(q)}}.$$

We perform a 10-fold cross validation to compare fits of LDA, LT-LDA, HA-LDA, and HALT-LDA to the health departments web site data. Each fold splits the data randomly, holding out 20% of the pages from a web site and using the other 80% of pages for MCMC sampling. For each sample $q$ we calculate and save conditional posterior means $\bar{\phi}_{k,v}^{(q)}$ and $\bar{\psi}_{i,v}^{(q)}$ and save the sampled $c_\alpha$ and $\alpha_i^{(q)}$. We save results from 500 MCMC iterations after a burn-in of 1500. We calculate an estimate for scale parameter $c_\alpha$ and probability vector $\alpha_i$ by averaging over the 500 saved samples. We calculate an estimate for topic-word probabilities $\phi_{k,v}$ and $\psi_{i,v}$ by averaging over 500 conditional posterior means. We use the estimates to calculate the held-out log likelihood of held-out pages given $c_\alpha$, $\alpha_i$, $\phi_{k,v}$, and $\psi_{i,v}$. We use the left-to-right particle filtering algorithm for LDA to approximate held-out log likelihoods (Wallach et al., 2009b). Wallach's left-to-right algorithm sequentially samples topic indices and calculates log likelihood components of each word from left to right. The algorithm decomposes the probability of a held-out word to a sum over joint probabilities of a held-out word and topic indices of previous words in the same document. The algorithm has been described by Scott and Baldridge (2013) as a particle-Gibbs method.

We provide a brief summary of the algorithm applied to HALT-LDA in section A.3 of the web appendix. Held-out log likelihoods are averaged over the cross-validation sets and used to identify a reasonable choice for the number of global topics $K$ and to compare between the LDA, LT-LDA, HA-LDA, and HALT-LDA. We analyze a final HALT-LDA model with 1,000 samples after a burn-in of 1,500 samples.

## 4. Health Department Web Site Data

The National Association of County and City Health Officials maintains a directory of local health departments (LHD) in the United States that includes a URL for each department web site (National Association of County and City Health Officials, 2018). We scrape each web site for its textual content using Python and Scrapy (van Rossum, 1995; ScrapingHub, 2018). All web sites were scraped during November of 2019. We remove text items that occur on nearly every page, such as titles or navigation menus. Pages with fewer than 10 words are removed. Common English stop words, such as 'the', 'and', 'them', and non-alphabet characters are removed, and words are *stemmed*, e.g. 'coughing' and 'coughs' are reduced to 'cough'. Uncommon words, which we define as words occurring in fewer than 10 pages across all web sites, are removed. Due to computation time of MCMC sampling, a subset of 20 web sites with fewer than 100 pages each were randomly selected to use in our analyses. The data set analyzed had 124,491 total words with $V = 1614$ unique words across 923 pages. At $K = 60$ it takes approximately 65 minutes to run 1000 total iterations with HALT-LDA with an Intel Core i7-6700 processor.

## 5. Results

The 10-fold cross validated held-out log likelihoods are plotted against the number of global topics $K$ in Figure 1 for the four models: LDA, LT-LDA, HA-LDA, and HALT-LDA. For every fixed number of global topics $K$, our extensions LT-LDA, HA-LDA and HALT-LDA outperform LDA. At smaller $K$, because they also include 20 local topics, HALT-LDA and LT-LDA allow more total topics compared to HA-LDA and LDA. Thus, we expect and see that models with local topics perform better at a smaller number of global topics $K$. The consistent improvement in log likelihood from LDA to LT-LDA indicates that local topics exist and that web pages in a web site do indeed share a local topic. However, the improvement from HA-LDA to HALT-LDA decreases as $K$ increases. This is because the nested asymmetric prior is a flexible prior that can accommodate

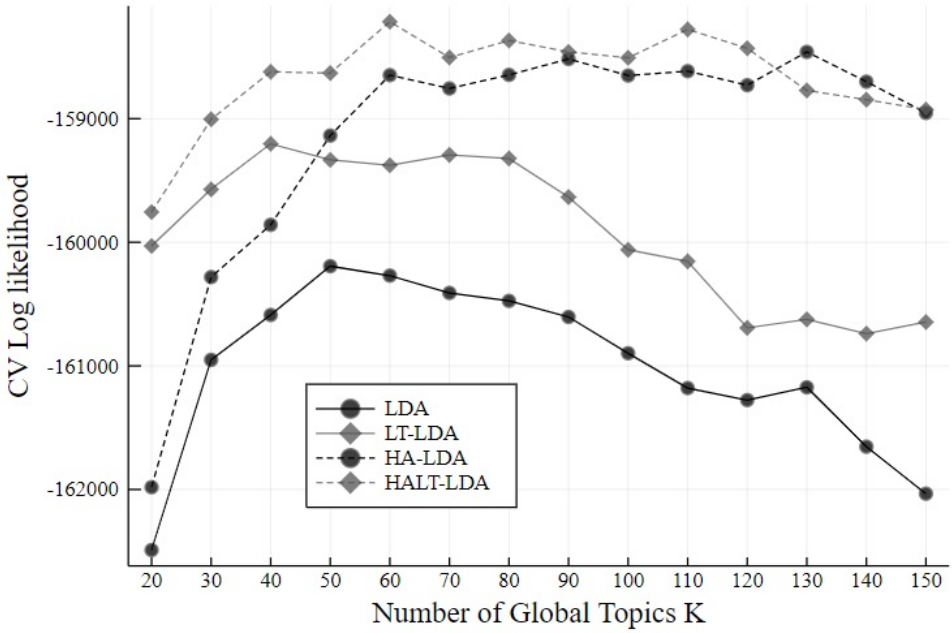


Figure 1: Plot of 10-fold cross validated (CV) held-out log likelihood by different number of global topics $K$.

local topics though it does not formally identify specific topics as local. It allows pages of a web site to share commonalities, such as high probability in its local topic and low probability in local topics of other web sites. The HALT-LDA cross-validated log likelihoods peak slightly higher and at smaller $K$, while HA-LDA peaks at larger $K$. Both these models support a larger number of topics than their counterparts without a hierarchical asymmetric prior. The results suggests that LT-LDA, HA-LDA, and HALT-LDA model web pages nested in web site better than LDA, and local topics allow us to specify a smaller number of global topics with similar or better performance. In later inference for the public health departments, we are not interested in the local topics except to remove words corresponding to local topic from pages before further calculations. Therefore, it is much more useful to use the LT models which automatically identify local topics to more easily make inferences only about global topics.

### *5.1. Matching and Comparing Local Topics*

We match local topics in HALT-LDA with $K = 60$ to global topics in HA-LDA with $K = 90$ to illustrate the existence of local topics and their high prevalence within a single web site relative to their prevalence in other web sites. We

choose $K = 60$ for HALT-LDA where log likelihood peaks and choose $K = 90$ where HA-LDA performs nearly at its peak at $K = 130$ but is closer to HALT-LDA in total number of topics. We compare two methods for matching topics; a rank based method and a probability based method. The rank based method finds topics in HA-LDA that have similar sets of word ranks as a local topic in HALT-LDA while the probability based method finds topics in HA-LDA that have similar word probabilities as a local topic in HALT-LDA. Let $R_{k,v}^{(\mathrm{HA})}$ denote the rank of word $v$ in topic $k$ from HA-LDA and let $R_{i,v}^{(\mathrm{HALT})}$ denote the rank of word $v$ in local topic $i$ from HALT-LDA. For the rank based method, the matched topic index in HA-LDA for local topic $i$ is

$$\arg\min_{k} \sum_{v=1}^{V} |R_{i,v}^{(HALT)} - R_{k,v}^{(HA)}|. \tag{1}$$

Define $\psi_{i,v}^{(HALT)}$ as the local topic-word probability for web site $i$ and word $v$ in HALT-LDA and define $\phi_{k,v}$ as the topic-word probability for topic $k$ and word $v$ in HA-LDA. By the probability based method, the matched topic index in HA-LDA for local topic $i$ is

$$\arg\min_{k} \sum_{v=1}^{V} (\psi_{i,v}^{(HALT)} - \phi_{k,v}^{(HA)})^2. \tag{2}$$

Topics generally place higher probability on a small subset of words while placing small probability on the majority of words. We may want to consider only the most probable subset of words in our calculations in equation 1 and equation 2 if we define topics by their most probable words. Thus, we consider limiting the summations to the subset of most common words. Define $T_i^{(10)}$ as the indices of the top 10 words from local topic $i$ in HALT-LDA. Then the calculations for rank based and probability based matching are respectively

$$\arg\min_{k} \sum_{v \in T_i^{(10)}} |R_{i,v}^{(HALT)} - R_{k,v}^{(HA)}|,$$

$$\arg\min_{k} \sum_{v \in T_i^{(10)}} (\psi_{i,v}^{(HALT)} - \phi_{k,v}^{(HA)})^2.$$

We estimate topic-word probabilities by averaging across 1,000 conditional posterior means and match using those estimates. The ranks are computed using these posterior mean probabilities. For each web site $i$, we matched one topic

in HA-LDA to local topic $i$ in HALT-LDA. Thus, there are 20 matched local topics in HA-LDA, one for each web site. For a given web site, we refer to the matched local topic that belongs to the web site as the *correct local* topic and the remaining 19 matched local topics as *other local* topics.

Web site averages, $\bar{\theta}_{i\cdot,k} = \frac{1}{M_i}\sum_{j=1}^{M_i}\theta_{ij,k}$, of web page-topic distributions are calculated by averaging estimates across pages of a web site. Thus in HA-LDA there are 20 averages that correspond to *correct local* topics, 380 averages that correspond to *other local* topics, and 1400 averages that correspond to the remaining global topics. Figure 2 plots boxplots of web site average probabilities for *correct local* topics, *other local* topics, and global topics plotted in between as a reference. The first row shows the probability based methods and the second row shows the rank based methods. The first column are methods using all words and the second column using top 10 words. There is extreme localization of local topics in HA-LDA regardless of topic matching method. *Correct local* topics typically have high web site average probabilities, global topics have lower averages, and *other local* topics have the lowest averages, with most nearly 0.

### *5.2. Topic Model Output and Applications*

Table 2 lists the ten most probable words for the most prevalent global topic and for another 9 health topics from among the top 20 highest probability topics in HALT-LDA for $K = 60$. We label each topic after inspecting its most probable words. The prevalence column shows the average probability of a topic across all web pages and web sites. The most prevalent (5.4%) topic has top words *inform, provid, contact, please, requir, call, need, must, click, may* that generally describe getting information and contacting the public health department. The cumulative prevalence of all 60 global topics is 82%, with 18% in local topics. Thus, the local topic in each web site generally accounts for a large proportion of text. Four health topics we use in our later analysis are food safety, Special Supplemental Nutrition Program for Women, Infants, and Children (WIC), emergency preparedness, and sexually transmitted disease. Estimates and 95% intervals of conditional posterior means for word probabilities of these topics' ten most probable words are plotted in Figure 3. The word probabilities for the ten most probable words are much larger than the average probability 1/1614.

Table 3 lists the five most probable words for each of the $M = 20$ local topics. Most local topics contain a geographical name or word among its top five words. The local topic in web site 7 has top words related to food sanitation inspection because web site 7 contains 14 pages dedicated to reports for monthly inspections and another 16 pages related to food protection and food sanitation

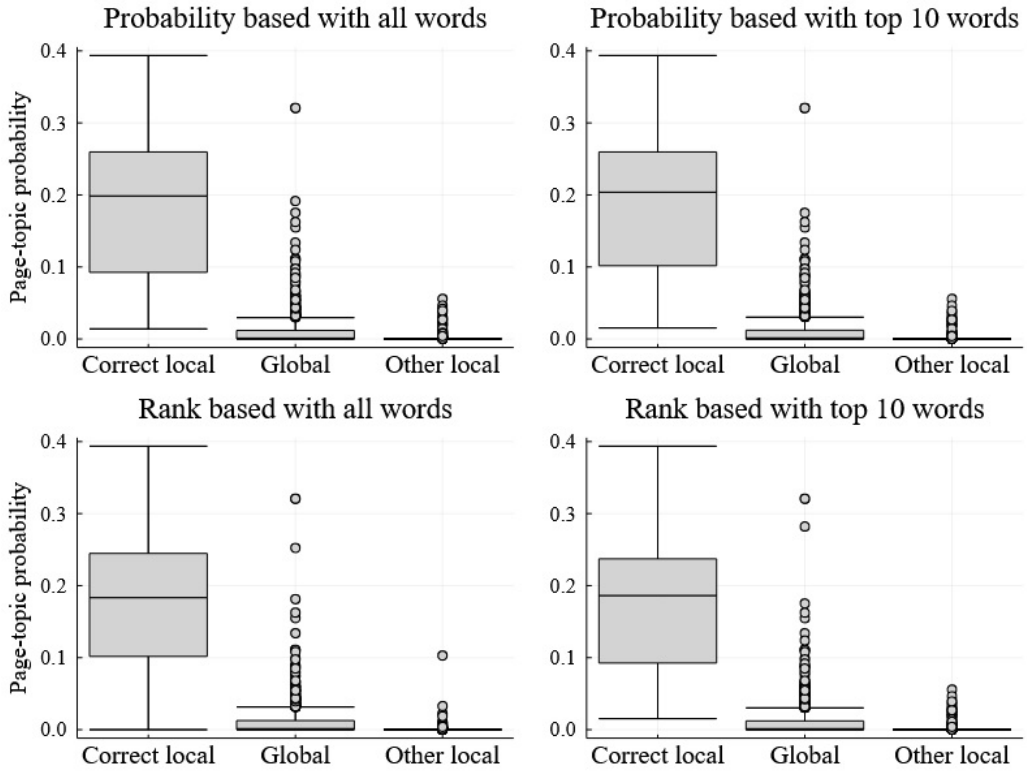


Figure 2: Boxplots of the web site average web page-topic distributions $\bar{\theta}_{i\cdot,k} = \frac{1}{M_i}\sum_{j=1}^{M_i}\theta_{ij,k}$ of global topics and matched local topics in HA-LDA. 'Correct local' shows the distribution of $\bar{\theta}_{i\cdot,k}$, where topic $k$ has been matched to web site $i$'s local topic in HALT-LDA. 'Other local' shows the distribution of $\bar{\theta}_{i\cdot,k}$, where topic $k$ is a local topic but not the matched local topic. Global shows the distribution of $\bar{\theta}_{i\cdot,k}$ for the remaining topics $k$.

Table 2: The ten highest probability words for the most common topic (General) and nine health topics from HALT-LDA for $K = 60$. Topic labels in the first column are manually labeled and the prevalence is the average probability across all web pages and web sites. Means and 95% credible intervals for the probabilities of the words for the 4 health topics in boldface are plotted in Figure 3.

| Label | Prevalence | Top 10 words |
|---|---|---|
| General | 5.4% | *inform, provid, contact, pleas, requir, call, need, must, click, may* |
| Disease prevention | 3.3% | *diseas, prevent, risk, caus, use, includ, year, effect, peopl, also* |
| **Food safety** | 2.9% | *food, inspect, establish, permit, environment, safeti, facil, code, oper, applic* |
| **WIC** | 2.7% | *wic, breastfeed, infant, women, nutrit, program, children, food, elig, incom* |
| Vaccinations | 2.0% | *immun, vaccin, adult, children, child, schedul, flu, appoint, clinic, diseas* |
| Breast cancer | 1.9% | *test, women, clinic, screen, famili, pregnanc, plan, breast, cancer, exam* |
| **Emergency preparedness** | 1.8% | *emerg, prepared, disast, respons, plan, prepar, commun, event, famili, local* |
| Hospital Care | 1.7% | *care, patient, provid, medic, nurs physician, treatment, visit, hospit, includ* |
| **Sexually transmitted disease** | 1.5% | *test, std, clinic, treatment, hiv, schedul, educ, immun, fee, sexual* |
| Family Program | 1.4% | *child, children, famili, parent, program visit, home, babi, help, hand* |

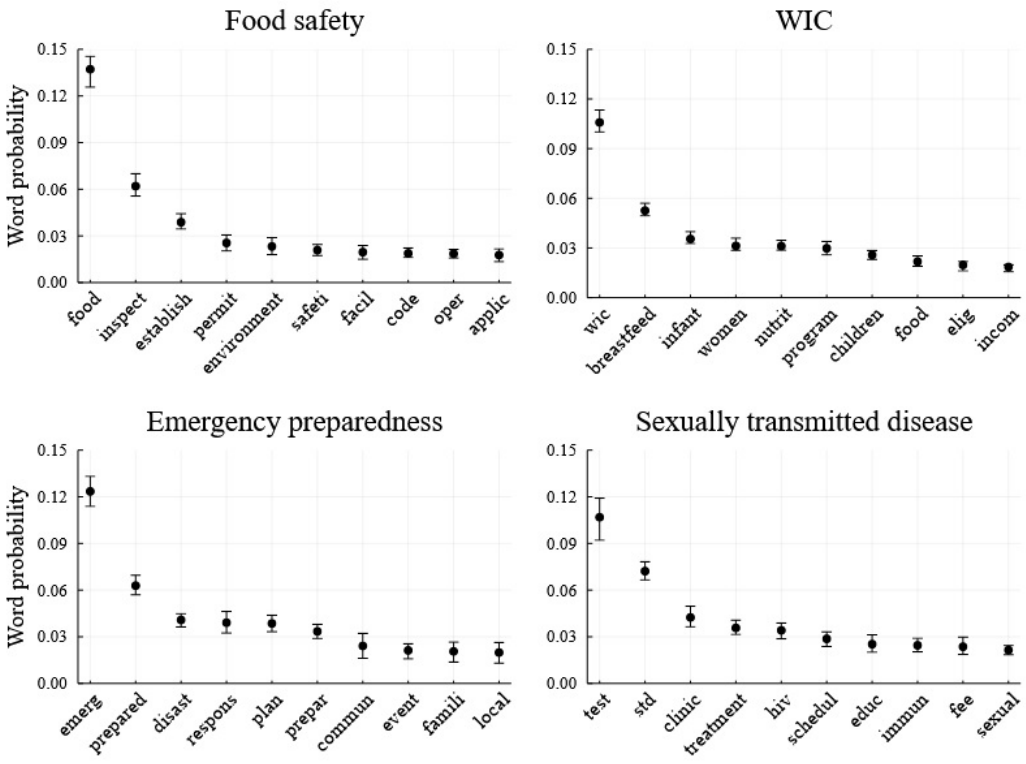


Figure 3: Median and 95% intervals of conditional posterior means of word probabilities for the ten most probable words in four health topics.

Table 3: Top five highest probability words in local topics from HALT-LDA for $K = 60$. Most local topics include a geographical name or word among the top five words.

| | Location | State | Top 5 Words (local topic) |
|---|---|---|---|
| 1 | Elkhorn Logan Valley | Nebraska | *month, nation, awar, elvphd, day* |
| 2 | Sandusky County | Ohio | *sanduski, ohio, fremont, street, read* |
| 3 | Ford County | Illinois | *ford, program, illinoi, bird, press* |
| 4 | Loup Basin | Nebraska | *loupbasin, loup, basin, nebraska, program* |
| 5 | Wayne County | Missouri | *center, wayn, creat, homestead, back* |
| 6 | Greene County | Iowa | *green, medic, center, care, therapi* |
| 7 | Bell County | Texas | *report, inspect, food, retail, octob* |
| 8 | Moniteau County | Missouri | *moniteau, missouri, center, requir, map* |
| 9 | Williams County | Ohio | *phasellu, sed, dolor, fusc, odio* |
| 10 | Harrison and Clarksburg | West Virginia | *alert, harrison, clarksburg, subscrib, archiv* |
| 11 | Oldham County | Kentucky | *oldham, kentucki, click, local, resourc* |
| 12 | Boyle County | Kentucky | *boyl, bag, item, bed, home* |
| 13 | Dallas County | Missouri | *buffalo, routin, dalla, food, inspect* |
| 14 | Shelby County | Tennessee | *sschd, ohio, shelbycountyhealthdeptorg, email, shelbi* |
| 15 | Taney County | Missouri | *averag, normal, assur, commun, exposur* |
| 16 | Monroe County | Missouri | *monro, phone, email, map, fax* |
| 17 | Three Rivers District | Kentucky | *river, three, district, kentucki, local* |
| 18 | Central District | Nebraska | *central, district, permit, resourc, island* |
| 19 | Levy County | Florida | *florida, updat, weekli, month, april* |
| 20 | Ozark County | Missouri | *ozark, contact, info, home, box* |

out of a total of 86 pages. The local topic in web site 13 has top words related to food sanitation inspection because 11 of its 30 pages mention food inspections. In Table 2, food safety is a global topic that shares similar words. We further investigate the food safety topic later in our analysis. Web site 9 is the only web site with several pages containing placeholder text, i.e. lorem ipsum or nonsensical Latin, which account for the top words in its local topic. Web site 15 has two large pages each with about 3000 words describing job openings which account for the top words in its local topic. Other than the local topic in web site 7 and 13, no other local topic is similar to the global topics in Table 2.

Web sites 7, 9, and 15 have global topics that appear to be local topics for these web sites. The global topic with top words *taney, report, commun, anim, outreach* may be a second local topic for web site 15 as it is related to a common news block in several web pages. Similarly the global topic with top words

*william, ohio, dept, divis, inform* and the global topic with top words *nbsp, bell, district, texa, director* may be second local topics for web sites 9 and 7. These three global topics were less prevalent within the respective web sites than the local topics discovered by the model. Additionally, we found two other global topics with top words *green, center, medic, foundat, jefferson* and *shall, section, ordin, dalla, person* that may be second local topics for web sites 6 and 13. The global topic with top words *green, center, medic, foundat, jefferson* has nearly the prevalence within web site 6 as the local topic of web site 6. The global topic with top words *shall, section, ordin, dalla, person* is more prevalent in web site 13 than the local topic of web site 13. However, the identified local topic with top words *buffalo, routin, dalla, food, inspect* has more local words specific to web site 13 than the global topic. Our model either identifies the most prevalent local topic or the local topic with more local words.

We model public health web sites using topic models to understand how local health departments cover health topics online. In a web site, multiple health topics may be covered and it is more reasonable to dedicate a single or handful of web pages to a given health topic rather than have every web page discuss all health topics. Rather than comparing web site average probabilities of a given topic, we compare topic coverage. Informally, topic coverage measures whether a web site has at least one dedicated page on a given topic. Formally, we define coverage of topic $k$ in web site $i$ as the largest web page-topic probability $\theta_{ijk}$ across all $j = 1, \ldots, M_i$ pages,

$$\max_j \theta_{ijk}.$$

We use topic coverage to help identify common health topics that may be missing in a web site.

We found that pages in web sites repeat common text, such as geographic names and words, events and news, or contact information. These words have high probability in local topics and local topics account for the largest proportion of web page-topic probability across all web sites. Additionally, the probability of local topics vary between web sites. Thus, we adjust for local topic content on web pages when comparing coverage of (global) health topics. For example, a web page with 20% probability for its local topic and a 40% probability for the heart disease topic and a web page with 40% probability for its local topic and 30% probability for the heart disease topic should both be viewed as pages 50% dedicated to the heart disease topic. The adjusted topic coverage (ATC) for topic

$k$ in web site $i$ is therefore

$$\text{ATC}_{ik} = \max_j \frac{\theta_{ij,k}}{1 - \theta_{ij,K+1}}.$$

We calculate the adjusted topic coverage for four common health topics, food safety, WIC, emergency preparedness and sexually transmitted disease, using estimates from each of the 1,000 MCMC samples. Plots of ATC are shown in Figure 4. We use ATC to identify common health topics that may be missing from individual health web sites and in particular investigate web sites where the lower bound of ATC is below 0.05.

Web sites 4 and 6 have ATC lower bounds below 0.05 for food safety and none of their web pages cover food safety. We noted that web sites 7 and 13 have a local topic that shares some high probability words with the food safety topic. However, the ATC for food safety for both web sites are still moderate, between 0.23 and 0.78 in web site 7 and between 0.20 and 0.82. For WIC, web site 4 has the lowest ATC and none of its web pages cover WIC. Web site 3 has ATC lower bound below 0.05 for WIC. The web site mentions WIC in two pages; however, they are not pages dedicated to WIC. One page has 16 frequently asked questions with one related to WIC and another page is an overview of the health department and mentions WIC among other programs and services. Web site 16 has the lowest ATC for emergency preparedness and, upon inspection, none of its 23 web pages covered emergency preparedness. Web site 15 contains a resource page with multiple sections with one section directing the reader to emergency preparedness web sites outside of web site 15.

For sexually transmitted diseases (STDs), web sites 1, 3, 4, 15, and 18 have ATC less than 0.05. Web sites 3, 4, and 18 did not have web pages covering STDs. Web site 1 did contain a health information web page with fourteen different drop down menus, each for a different topic. Among the fourteen was an "STD & HPV Resource List" menu. Web site 15 has a web page with nine sections for different clinical services of which one is a screening and tests service. There are five tests under the screen and tests services, where one is for STD testing and the other for HIV/AIDS screening. Web sites 6, 9, and 17 additionally have ATC lower bounds below 0.05. Web site 6 has a page that lists eighteen services that their women's health clinic offers of which one is testing for STDs. Web site 9 has a page that gives an overview of their reproductive health and wellness clinic and lists services offered. One of the services is testing and treating STDs. Web site 17 has a page of thirteen frequently asked questions of which one is directly related to STDs. However, testing for STDs is mentioned two additional times as part of larger answers to questions about services offered. This

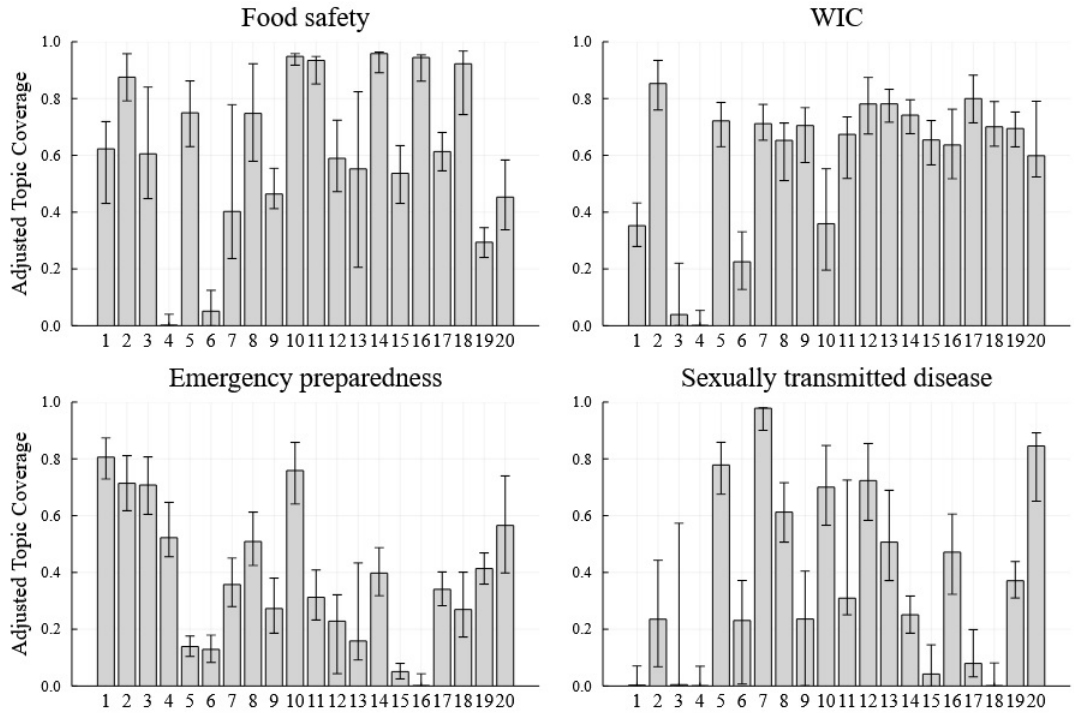


Figure 4: Bar plots of adjusted topic coverage for four global topics from Table 2. Bar heights are medians and error bars are 95% credible intervals.

explains why ATC and the ATC lower bound for STDs in web site 17 is the highest of these eight web sites.

All web sites with ATC lower bound less than 0.05 did not cover the corresponding topic, only linked to an outside resource, or contained a larger page that briefly mentions the topic. ATC looks at a web page's probability of a given topic relative to the cumulative probability of all global topics. Under this metric, a web site with a web page covering several global topics may be considered to have low coverage.

## 6. Discussion

We introduced and defined local topics as topics that are unique to one web site or group of web pages. Local topics may be common in a nested document collection and we show that in our data set nearly all local topics included geographical names among their most probable words. We conclude that local topics exist and have high topic probabilities in our data set. We proposed two

extensions HA and LT as well as their combination to accommodate the locality and inference in models with nested documents and local topics.

Adding either or both extensions improves cross-validated log likelihood compared to LDA, and HA-LDA performs better than LT-LDA for larger numbers $K$ of global topics. Combining both extensions, HALT-LDA has a higher peak log likelihood than HA-LDA. However, the peaks are similar between the two and we do not conclude that one outperforms the other in log likelihood. Instead, these two models perform similarly and are both better than LDA or LT-LDA. A more notable difference is that HALT-LDA performs well at a smaller number of global topics $K$. As computation time is largely dependent on the number of topics each word may be drawn from, it is advantageous to use HALT-LDA because it uses smaller $K$ to reach similar performance as HA-LDA.

The key benefit of explicitly modeling local topics is that inference and interpretation are much easier. The model directly identifies local topics and we can infer what proportion of a web page is composed of its local topic. This proportion varies across web sites and web pages. Thus, when comparing coverage of global topics across web sites we should adjust for the probability of local topics. We compared adjusted topic coverage (ATC) of common health topics across web sites and identified web sites that did not cover food safety, WIC, emergency preparedness, and sexually transmitted disease.

Our goal in modeling nested documents is to study global topics and make comparisons about their distributions within groups of documents. Models should accommodate strong localizations of topics and the addition of local topics and a hierarchical asymmetric prior are useful. However, it may be difficult to determine a priori the number of local topics to introduce. We assumed a single local topic for each web site, which is reasonable for a set of web sites each dedicated to public health in a specific location. However, we noted that 5 web sites in our data set appear to have two local topics. We study 5 scenarios in which simulated web sites have none, one, or two local topics in the section A.1 of the web appendix. When local topics are modeled when they do not exist the probability of that local topic is typically small and further, HALT-LDA identifies a local topic that gives high probability to words that occur more often in the local topic's corresponding web site and do not occur as often in the other web sites. When two local topics exist, HALT-LDA almost always merges the two topics into a single local topic. However, this is when the number of global topics $K$ in HALT-LDA matches the number of global topics used to generated the data. When a larger $K$ is set we expect the merged local topic to split as shown in our analysis of 20 web sites with $K = 60$ global topics.

The intervals of conditional posterior means for the highest probability words in topics essentially check for label switching. Word probabilities for the same word in different common global topics were distinct; if switching were occurring, the 95% intervals for the word would overlap in the two topics. Thus, the 95% intervals of the conditional posterior means would be large. The word probabilities shown in Figure 3 did not fluctuate much which would suggest there was no label switching. For example, if Food safety and WIC had label-switched, then the 95% intervals for "food" would extend from 0.03 to 0.12 in both topics and similarly "wic" would extend from less than 0.01 to 0.10 in both topics.

## Supplemental Materials

*Web Appendix*

Web appendix file that includes our simulation study, sensitivity analysis, and brief overview of the left-to-right algorithm applied to HALT-LDA.

*Code for HALT-LDA*

Julia code used to implement HALT-LDA is given in supplementary materials. The file name is TopicModelHealthWebsites.jl. The file and an example using the 20 web site data (dataanddict.jld) can also be found on the first author's github repository https://github.com/jwanghb/publichealth-websites.

*Data of the 20 Web Sites*

The 20 web site data reported in this paper. The file name is dataanddict.jld.

## Declaration of Interest

Declarations of interest: None.


## Funding Sources

Weiss was partially supported by the Center for HIV Identification, Prevention, and Treatment (CHIPTS) NIH grant P30MH058107. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

## References


Blei, D.M., Lafferty, J.D., 2005. Correlated topic models, in: Proceedings of the 18th International Conference on Neural Information Processing Systems, MIT Press, Cambridge, MA, USA. pp. 147–154. doi:10.5555/2976248.2976267.

Blei, D.M., Lafferty, J.D., 2006. Dynamic topic models, in: Proceedings of the 23rd International Conference on Machine Learning, p. 113–120. doi:10.1145/1143844.1143859.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. Journal of Machine Learning Research 3, 993–1022. doi:10.1162/jmlr.2003.3.4-5.993.

Boyd-Graber, J.L., Hu, Y., Mimno, D.M., 2017. Applications of topic models. Foundations and Trends in Information Retrieval 11, 143–296. doi:10.1561/1500000030.

Chang, J., Blei, D., 2009. Relational topic models for document networks, in: Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, PMLR, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA. pp. 81–88.

Chemudugunta, C., Smyth, P., Steyvers, M., 2006. Modeling general and specific aspects of documents with a probabilistic topic model, MIT Press, Cambridge, MA, USA. p. 241–248. doi:10.7551/mitpress/7503.003.0035.

Chen, N., Zhu, J., Xia, F., Zhang, B., 2013. Generalized relational topic models with data augmentation, pp. 1273–1279.

Chen, N., Zhu, J., Xia, F., Zhang, B., 2015. Discriminative relational topic models. IEEE Transactions on Pattern Analysis and Machine Intelligence 37, 973–986. doi:10.1109/TPAMI.2014.2361129.

Foster, A., Li, H., Maierhofer, G., Shearer, M., 2016. An extension of standard latent Dirichlet allocation to multiple corpora. SIAM Undergraduate Research Online 9. doi:10.1137/15S014599.

Ge, T., Pei, W., Ji, H., Li, S., Chang, B., Sui, Z., 2015. Bring you to the past: Automatic generation of topically relevant event chronicles, in: ACL (1), The Association for Computer Linguistics. pp. 575–585.

George, C.P., Xia, W., Michailidis, G., 2019. Analyses of multi-collection corpora via compound topic modeling. Machine Learning, Optimization, and Data Science 11943.

Griffiths, T.L., Steyvers, M., 2004. Finding scientific topics. Proceedings of the National Academy of Sciences 101, 5228–5235. doi:10.1073/pnas.0307752101.

Grün, B., Hornik, K., 2011. topicmodels: An R package for fitting topic models. Journal of Statistical Software 40, 1–30. doi:10.18637/jss.v040.i13.

Guo, W., Li, H., Ji, H., Diab, M., 2013. Linking tweets to news: A framework to enrich short text data in social media, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Sofia, Bulgaria. pp. 239–249. URL: https://aclanthology.org/P13-1024.

Guo, W., Wu, S., Wang, L., Tan, T., 2015. Social-relational topic model for social networks, in: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, Association for Computing Machinery, New York, NY, USA. p. 1731–1734. doi:10.1145/2806416.2806611.

Hong, L., Dom, B., Gurumurthy, S., Tsioutsiouliklis, K., 2011. A time-dependent topic model for multiple text streams, pp. 832–840. doi:10.1145/2020408.2020551.

Hua, T., Lu, C.T., Choo, J., Reddy, C.K., 2020. Probabilistic topic modeling for comparative analysis of document collections. ACM Transactions on Knowledge Discovery from Data 14. doi:10.1145/3369873.

Liu, H., Ge, Y., Zheng, Q., Lin, R., Li, H., 2018. Detecting global and local topics via mining twitter data. Neurocomputing 273, 120–132. doi:10.1016/j.neucom.2017.07.056.

Liu, Y., Niculescu-Mizil, A., Gryc, W., 2009. Topic-link LDA: Joint models of topic and author community, in: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, New York, NY, USA. pp. 665–672. doi:10.1145/1553374.1553460.

Lu, Y., Mei, Q., Zhai, C., 2011. Investigating task performance of probabilistic topic models: An empirical study of PLSA and LDA. Information Retrieval 14, 178–203. doi:10.1007/s10791-010-9141-9.

National Association of County and City Health Officials, 2018. Directory of local health departments. Https://www.naccho.org/membership/lhd-directory.

Paul, M., Dredze, M., 2014. Discovering health topics in social media using topic models. PloS One 9, e103408. doi:10.1371/journal.pone.0103408.

Paul, M.J., Girju, R., 2010. A two-dimensional topic-aspect model for discovering multi-faceted topics, in: Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010.

Qiang, S., Wang, Y., Jin, Y., 2017. A local-global LDA model for discovering geographical topics from social media, in: APWeb/WAIM. doi:10.1007/978-3-319-63579-8_3.

Roberts, M.E., Stewart, B.M., Tingley, D., Airoldi, E.M., 2013. The structural topic model and applied social science, in: Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation, pp. 1–20.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P., 2004. The author-topic model for authors and documents, in: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, AUAI Press, Arlington, Virginia, USA. p. 487–494. doi:10.5555/1036843.1036902.

van Rossum, G., 1995. Python tutorial. Technical Report CS-R9526. Centrum voor Wiskunde en Informatica (CWI). Amsterdam.

Scott, J., Baldridge, J., 2013. A recursive estimate for the predictive likelihood in a topic model, in: Carvalho, C.M., Ravikumar, P. (Eds.), Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, PMLR, Scottsdale, Arizona, USA. pp. 527–535.

ScrapingHub, 2018. Scrapy 1.8 documentation. Https://scrapy.org/.

Shen, Z.Y., Sun, J., Shen, Y.D., 2008. Collective latent Dirichlet allocation, in: 2008 Eighth IEEE International Conference on Data Mining, IEEE. pp. 1019–1024.

Song, Y., Pan, S., Liu, S., Zhou, M.X., Qian, W., 2009. Topic and keyword re-ranking for LDA-based topic modeling, in: Proceedings of the 18th ACM Conference on Information and Knowledge Management, Association for Computing Machinery, New York, NY, USA. p. 1757–1760. doi:10.1145/1645953.1646223.

Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M., 2006. Hierarchical Dirichlet processes. Journal of the American Statistical Association 101, 1566–1581. doi:10.1198/016214506000000302.

Terragni, S., Fersini, E., Messina, E., 2020. Constrained relational topic models. Information Sciences 512, 581–594. URL: https://www.sciencedirect.com/science/article/pii/S0020025519308850, doi:https://doi.org/10.1016/j.ins.2019.09.039.

Wallach, H.M., Mimno, D., McCallum, A., 2009a. Rethinking LDA: Why priors matter, in: Proceedings of the 22nd International Conference on Neural Information Processing Systems, Curran Associates Inc., USA. pp. 1973–1981. doi:10.5555/2984093.2984314.

Wallach, H.M., Murray, I., Salakhutdinov, R., Mimno, D., 2009b. Evaluation methods for topic

models, in: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, New York, NY, USA. pp. 1105–1112. doi:10.1145/1553374.1553515.

Wang, C., Blei, D., Heckerman, D., 2008. Continuous time dynamic topic models, in: Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence, p. 579–586.

Wang, C., Thiesson, B., Meek, C., Blei, D., 2009. Markov topic models, in: Artificial intelligence and statistics, PMLR. pp. 583–590.

Yang, Y., Wang, F., Jiang, F., Jin, S., Xu, J., 2016. A topic model for hierarchical documents, in: 2016 IEEE First International Conference on Data Science in Cyberspace (DSC), pp. 118–126. doi:10.1109/DSC.2016.97.

Zhang, J., Song, Y., Zhang, C., Liu, S., 2010. Evolutionary hierarchical Dirichlet processes for multiple correlated time-varying corpora, in: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1079–1088.

Zhou, M., Hannah, L., Dunson, D., Carin, L., 2012. Beta-negative binomial process and Poisson factor analysis, in: Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, PMLR. pp. 1462–1471.