Title
Modeling Topic Presence and Covariate Effects in Hierarchical Text Data With Applications to United States Local Health Department Websites

Author
Wang, Jason

Publication Date
2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Modeling Topic Presence and Covariate Effects in Hierarchical Text Data With Applications to United States Local Health Department Websites

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Biostatistics

by

Jason Wang

2021

ABSTRACT OF THE DISSERTATION

Modeling Topic Presence and Covariate Effects in Hierarchical Text Data
With Applications to United States Local Health Department Websites

by

Jason Wang
Doctor of Philosophy in Biostatistics
University of California, Los Angeles, 2021
Professor Robert E. Weiss, Chair

Topic models are probabilistic models used to abstract topical information from collections of text documents. Documents are modeled as probability distributions over latent topics and topics are modeled as probability distributions over words. In a single collection of documents, topics are global, that is, they are shared across the multiple documents in the collection. A nested document collection has documents that are nested inside a higher order structure, for example, stories in a book, articles in a journal, or web pages in a web site. Regular topic models ignore the nesting and treat all documents as distinct. In contrast, a nested document collection, such as web pages nested in web sites, web pages of the same web sites share similarities with each other that they do not share with web pages of other web sites. Regular topic models allow inferences about each web page individually; they are not suited for making inferences about an entire web site.

We propose hierarchical local topic models that place a hierarchical prior on web page topic distributions and explicitly model local topics, or topics that are unique to one web site. The hierarchical prior asserts that web page topic distributions vary around their

web site topic distribution and that web site topic distributions vary around a global topic distribution. Explicitly modeling local topics reduces the number of global topics needed, identifies the local topics and their owning web site, and lets us adjust inferences about how topics are covered.

We propose hierarchical topic presence models that place a sparsity inducing prior on topic distributions; they let us model the presence of topics in web sites, web pages, or both web sites and web pages. Topic presence in a web site can be modeled with logistic regression, as a function of covariates.

We apply hierarchical topic presence models to identify health topics in United States county health department web sites, estimate the percent of web sites that cover particular health topics, and identify demographic predictors of topic presence for human immunodeficiency virus (HIV) and opioid use disorder (OUD) topics.

The dissertation of Jason Wang is approved.

Sae Takada

Marc A. Suchard

Sudipto Banerjee

Robert E. Weiss, Committee Chair

University of California, Los Angeles

2021

To my wife Lianne

for her love and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would first like to express my sincere gratitude to my advisor Dr. Weiss for his advice, patience, and encouragement throughout my academic career. You are always supportive and have given me the freedom to pursue the projects I wanted. Thank you to my committee members Dr. Banerjee, Dr. Suchard, and Dr. Takada for your encouragement and insight that have helped improve my research. Thank you to my friends who have kept me sane in this journey.

Chapter 2 and Chapter 3 of this dissertation have been submitted for publication and are versions of Wang and Weiss (2021b) and Wang and Weiss (2021a), respectively, with both available on `arxiv.com`. Thank you to Dr. Takada for advice and mentoring in the analysis of U.S. county public health department web sites in Chapter 4.

VITA

2010-2013 B.A. (Applied Mathematics and Biochemistry), University of California, Berkeley, Berkeley, California

2014-2016 M.S. (Biostatistics), University of California, Los Angeles (UCLA), Los Angeles, California

2015-2021 Teaching Assistant, UCLA Department of Biostatistics

PUBLICATIONS

Wang, J. & Weiss, R.E. (2021). "Hierarchical Topic Presence Models", *arXiv pre-print*, arXiv:2104.07969

Wang, J. & Weiss, R.E. (2021). "Local and Global Topics in Text Modeling of Web Pages Nested in Web Sites", *arXiv pre-print*, arXiv:2104.01115

CHAPTER 1

Introduction

Digitized text data continues to grow and become more readily available since the advent of the Internet. There are now hundreds of local health department web sites across the United States. Each web site may have hundreds or thousands of web pages and together there are tens of thousands of web pages across all local health department web sites. Local health department web sites provide health information to local communities and cover a range of health topics that reflect the health priorities of a local health department. A public health researcher may be interested in identifying what health topics are covered. However, due to the overwhelming amount of text, it would be too time consuming if not impossible for the researcher to read through all the text to identify those health topics. Thus, there is interest in developing methods to systematically analyze large collections of text. We develop methodology to help the public health researcher survey health departments and understand their priorities through their web sites.

Topic models were first developed as a tool for exploratory analysis of large collections of text documents, such as articles of a journal or blogs of an author. They can reduce a large collection of text with an overwhelmingly large vocabulary to a smaller set of topics. Topic models take a set of documents, identify the set of topics that are covered, and describe how each document covers the set of topics. For each topic, we identify a list of key words that represent the topic and label the topic appropriately given those key words. For example, a topic with key words *food, safety, eat, contamination, check* align with the public health concept of food safety. Then rather than describing documents as a list of words, we can

describe a document with those key words occurring several times as a document that covers food safety. Topic models identify topics and their key words by how often words co-occur within documents. It is likely that *safety* is found in documents where *food* is also found. Methods of finding topics in documents can be applied to web sites and web pages. We define a web site as a collection of web pages, where each web page is a document that contains text.

To apply topic models to text data, we must represent text in a way that we can model them. The bag-of-words representation is quite common. It is a simplifying representation that disregards word order and only retains information about word occurrence and word co-occurrence. Word occurrence tells us what words are in a page and word co-occurrence tells us when different words are in the same page. Each document is reduced to a bag of words. This leaves us with a set of words where duplicate words may exist and order does not matter. For example, the phrase "the words in the bag" can be represented as {*the, words, in, the, bag*} which is the same as {*bag, in, the, the, words*}.

Topic models assert that documents are a mixture of topics and that topics are a mixture of words, where the number of topics is chosen and the number of unique words is determined by the data. Strictly speaking, a topic is characterized by a distribution over words or a list of word probabilities, rather than a list of key words. Similarly, a document is characterized by a distribution over topics or a list of topic probabilities. Document-topic distributions describe how prevalent topics are in documents and topic-word distributions describe how prevalent words are in topics. Suppose we have a collection of 100 documents with a total of 200 unique words across all documents that we model with 20 topics. First, suppose the model identifies a topic with the 5 largest probabilities (0.15, 0.10, 0.10, 0.08, 0.07) for words *food, safety, eat, contamination, check*. The probabilities of these 5 words make up half of all the probability for this topic with the remaining half being shared across 195 words. We could label this topic "food safety" from inspecting its the most probable words. Next, suppose the model finds that the first document has probability 0.60 for the food safety

topic. This means the first documents contains many words from the food safety topic. We could then conclude that the first document covers food safety or that it is dedicated to food safety.

1.1 United States Local Health Department Websites and Motivation for New Models

We collected text data from local health departments across the United States with the goal of identifying health topics and making inferences on how health topics are covered. Each web site contains several web pages. Thus, we have a nested document collection where web pages are documents that are nested within web sites. A local health department web site discusses a set of health topics, where its set of health topics is not necessarily the same as those in another web site.

Regular topics models identifies topics and tell us how web pages cover topics. However, we are interested in how web sites cover topics. Health web sites cover many health topics, and a health web site likely dedicates one or a few pages to a given health topic rather than discuss all health topics across all web pages. A web site covers a topic if at least one web page covers the topic and it does not cover the topic if it instead has many pages with relatively few words from that topic. We define topic coverage at the web site level as whether the web site has a web page dedicated to that topic, which happens if many words on a single page are from that topic. In addition to identifying when web sites cover topics, we are also interested in understanding patterns in how web sites cover topics. We want to know what portion of a web site covers a health topic and what factors are associated with how or whether topics are covered. Each local health department serves a city, county, or multiple counties. Thus, we can collect geographic or demographic variables and study how the variables are associated with web site topic coverage.

1.2 Hierarchical Local Topic Models

Latent Dirichlet allocation (LDA) is the most common form of topic modeling (Blei et al., 2003). It asserts that each word in a document is generated by first drawing a topic at random from the list of topics given the document's document-topic distribution then drawing a word at random from the list of unique words given that topic's topic-word distribution. Regular topic models like LDA do not accommodate the nesting structure of web pages in web sites; they were designed to model a single collection of documents such as a collection of web pages from one web site. Regular topic models can treat all web pages as separate documents or they can treat a web site as a single document by combining text across all web pages of a web site. Treating all web pages as separate documents ignores the fact that web pages of a web site are more similar to one another than they are to web pages of other web sites. Treating a web site as a document greatly reduces the number of observed documents and forces more words to co-occur, leading to topics that may not be interpretable from observing their most probable words. It is more reasonable to treat all web pages as separate documents than to combine web pages of a web site into a single document; however, neither approach directly addresses the structure of web pages nested in web sites.

We propose two extensions to LDA to address the nesting of web pages in web sites. In a single collection of web pages in a web site, there is no structural information that tells us two web pages are more similar to each other than they are to other web pages. In particular, web page topic distributions would all share a global mean topic distribution and are assumed random and vary about that global topic distribution. In a nested collection of web pages in multiple web sites, we expect web pages of the same web site to be more similar to one another than to web pages of other web sites. We propose a hierarchical prior for topic distributions so web page topic distributions vary around a web site topic distribution and web site topic distributions vary around a global topic distribution. The hierarchical prior on topic distributions lets us model similarities in topic distributions between web sites

and between web pages of the same web site.

The second extension comes from our observations of the data. We found that there were often local geographic names or local news items that were specific to each web site and not present in other web sites. Local words form topics that are likely unique to a single web site. Local topics are unique to a single web site, while in contrast global topics, global topics can be shared across multiple web sites. Regular topic models may identify global topics that place high probability on local words; however, they do not explicitly label them as local topics nor do they identify the local topics' owning web sites. We explicitly model local topics to label them as local and to identify the local topics' owning web site. This allows us to model our data with fewer global topics and to adjust our inferences on topic coverage. Hierarchical local topic models incorporate both the hierarchical prior and local topic extensions to LDA.

We specify a single local topic for each web site such that the sum of the global topic probabilities of a web page and its local topic probability is equal to 1. The prevalence of the local topic varies from web site to web site as some health departments mention their county names more often than others. For example, suppose web page 1 of web site A includes local words often while web page 1 of web site B does not have any local words. Specifically, let web page 1 of web site A have probability 0.5 for its local topic and probability 0.25 for a food safety topic and let web page 1 of web site B have probability 0 for its local topic and probability 0.5 for the food safety topic. We consider these two pages similar in web page topic coverage of food safety after removing local words; both pages have an adjusted web page topic coverage of 0.5. If we want to calculate an adjusted web site topic coverage, we can use the largest adjusted web page topic coverage of a web site. Then we can say a web site does not cover a topic if the adjusted web site topic coverage does not meet some threshold value. Explicitly modeling local topics lets us make adjustments to inferences like topic coverage.

1.3 Hierarchical Topic Presence Models

Topic models like LDA and hierarchical local topic models assert that web pages are mixtures over all topics such that the probability of each topic is nonzero. Sparsity inducing priors allow web pages to be a mixture of a subset of the possible topics rather than a mixture of all possible topics. Sparsity inducing priors let us model topic presence, where topic presence is a 1-0 indicator variable that indicates if a topic is present or not. Sparsity inducing priors have been studied and used to model a single collection of documents; however, they have not been specified previously for nested document collections. The nesting of web pages in web sites allows for topic presence modeling at the web site level, web page level, or both.

We propose three hierarchical topic presence models (HTPM) in addition to the null model where topics are always present at the web site and web page level. The first HTPM models topic presence at the web site level while all web pages are mixtures of all the topics present in their respective web sites. The second HTPM models topic presence at the web page level while all web sites are mixtures of all topics. The third HTPM models topic presence at both the web page level and the web site level. Rather than calculate web site topic coverage post-hoc as in LDA or as in hierarchical local topic models, we use web site topic presence to draw inferences about whether a web site covers a particular topic.

Topic presence probability is the probability that a topic is present in a web site or web page. We model topic presence probability in one of two ways. We can let topic presence probabilities be a priori exchangeable, where we estimate the global mean and variance of the probability that a given topic is present across web sites or web pages or we can model topic presence with a logistic regression, where the probability of topic presence is conditional on covariates. In a collection of web pages nested in U.S. local health department web sites, we are interested in modeling web site topic presence as a function of web site covariates, such as geographic or demographic variables. This allows us to draw inferences about how a topic is covered overall and draw inferences about how geographic or demographic variables

are associated with topic coverage.

1.4 Overview of Organization

Chapters 2, 3, and 4 are written as stand alone papers. Each of these three chapters have their own introduction, literature review and background, methodology, results, and discussion sections. Chapter 2 and Chapter 4 each have an appendix section of their own. There are some overlaps between the three chapters. Chapter 2 introduces local topics and discusses local topics in detail. Chapter 3 and Chapter 4 both use local topics; however, they do not discuss local topics in detail. All three chapters use a similar data set of local health department web site data. However, the local health department web site data is recollected for each subsequent chapter. Thus, the data in Chapter 4 is collected most recently. The amount of data used in each subsequent chapter is also larger and more inclusive. Chapter 2 uses a random subset of 20 small web sites, Chapter 3 uses all small web sites, and Chapter 4 uses all county health web sites.

Chapter 2 describes local and global topics and model extensions to explicitly model local topics and a hierarchical prior over topic distributions. Chapter 3 describes modeling topic presence for web pages, web sites, and/or both. Brief overviews of Chapter 2 and 3 are given in Section 1.2 and Section 1.3 respectively. Chapter 4 applies a model from Chapter 3 to study predictors of coverage of HIV and OUD topics in United States county health department web sites. Chapter 4 models all county health department web sites in the United States. It works with the largest data set and provides the most extensive analysis of the three chapters. It details steps in data collection and processing that were not included in Chapter 2 or Chapter 3. Chapter 4 presents hierarchical topic presence models as a web site survey tool for public health researchers. Chapter 5 concludes the dissertation with a discussion.

CHAPTER 2

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

2.1 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 newspapers, blog posts nested within authors, 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 inside the 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 comprise substantial portions of a web site. We do not wish to consider local topics in our desired inferences and so explicitly identifying local topics makes our desired inferences more appropriate. Effectively, local topics are removed from web pages before we make further inferences. 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.

We consider four models indexed by the number of global topics and apply them to web pages as documents. The first model is traditional LDA 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. The hierarchical asymmetric prior lets us model the variability of document-topic distributions between web sites. Additionally, the hierarchical asymmetric prior treats web pages as nested inside web sites. Our third (LT-LDA) and fourth models (HALT-LDA) introduce 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.

Nesting in document collections and local topics have been studied in different data settings (Chang and Blei, 2009; Rosen-Zvi et al., 2004; Yang et al., 2016; Qiang et al., 2017; Chemudugunta et al., 2006; Hua et al., 2020). We discuss the similarities and differences in the context of web pages nested in web sites. Nested document collections can be thought of as a special case of document networks where links are known; web pages of the same web site are linked and web pages of different web sites are not linked (Chang and Blei, 2009; Chen et al., 2015; Guo et al., 2015). Another type of nesting involves secondary documents nested within a primary document (Yang et al., 2016), such as comments nested within a blog post, and we consider this a separate structure. Nested document collections can also be thought of as a collection of different document collections (Hua et al., 2020) where each web site is itself a document collection. Relational topic models model the links between any two web pages and are used to predict links to a newly published web page(Chang and Blei, 2009; Chen et al., 2015; Guo et al., 2015). We do not need to model links between web pages. Some models for nested document collections address nesting by modeling multiple levels of document-topic distributions, but do not explicitly model local topics and their topic-word distributions (Qiang et al., 2017). Under the author model in Rosen-Zvi et al. (2004), local topics are explicitly modeled; however, global topics are not modeled. Under the author-topic model in Rosen-Zvi et al. (2004), global topics are modeled; however, each web page of a given web site shares the same topic distributions and local topics are not modeled. For a single web site or a single document collection, the special words topic model with background distribution (SWB) models a global set of topics, one common web site topic, and a single web page local topic for each web page (Chemudugunta et al., 2006). The common and distinctive topic model (CDTM) extends SWB and removes web page specific local topics to model multiple web sites or multiple document collections (Hua et al., 2020). CDTM models a global set of topics and a separate set of web site local topics for each web

site rather than a single web site local topic for each web site. 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 additionally place 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 infections, 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 in how a web site covers a topic. A health web site contains many web pages that cover different topics, where it dedicates one or a few web pages to a given health topic rather than discusses 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 an extensive local topic coverage should not be allowed to

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 should not average across pages, rather web site coverage should consider the supremum of coverage across pages.

Section 2.2 defines notation and our four models. Section 2.3 discusses computation and inference. Section 2.4 introduces our motivating data set in greater detail and section 2.5 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.

2.2 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 $j = 1, \ldots, M_i$. Words w_{ijh} on a page are indexed by $h = 1, \ldots, N_{ij}$ 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 prior to modeling as in latent Dirichlet allocation. Table 2.1 details notation used in our models.

2.2.1 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 ϕ_k we use the

parameterization

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

where ϕ_k is a V-vector of probabilities $\phi_{k,v}$ such that $\sum_{v=1}^{V} \phi_{kv} = 1$, $0 \leq \phi_{kv} \leq 1$, $c_\beta > 0$ is a scale parameter, and β is a V-vector of parameters β_v such that a priori $E[\phi_k | c_\beta \beta] = \beta$, $\sum_{v=1}^{V} \beta_v = 1$, and $0 \leq \beta_v \leq 1$. Each web page j in web site i has web page-topic distribution denoted by a K vector of probabilities θ_{ij} with a Dirichlet($c_\alpha \alpha$) prior. Topic k has a topic-word multinomial distribution parameterized by a V-vector of probabilities ϕ_k a priori distributed as Dirichlet($c_\beta \beta$). Words have a latent 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 distribution over all K topics, thus, LDA has K global topics and no local topics.

2.2.2 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, θ_{ij}, for page j in web site i is now a $(K + L_i)$-vector of probabilities. The topic-word distribution ψ_{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

Table 2.1: 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
θ_{ij}	Page-topic distribution of web site i web page j
ψ_i	Local topic-word distribution of web site i
ϕ_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 hth word of page j in web site i

for web site i. The LT-LDA model is then

$$\theta_{ij}|c_\alpha\alpha \sim \text{Dirichlet}(c_\alpha\alpha),$$

$$\psi_{il}|c_\gamma\gamma \sim \text{Dirichlet}(c_\gamma\gamma),$$

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

$$z_{ijh}|\theta_{ij} \sim \text{Categorical}(\theta_{ij}),$$

$$w_{ijh}|\Phi_{iz_{ijh}} \sim \text{Categorical}(\Phi_{iz_{ijh}}).$$

The shared prior parameter α 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 θ_{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.

2.2.3 Hierarchical Asymmetric Prior

A symmetric prior Dirichlet($c_\alpha\alpha$) for web page-topic distributions θ_{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 distribution 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 distribution are centered around the web site-topic distribution. We extend the LDA model in section 2.2.1 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 θ_{ij}, such that each web site has a $(K+1)$-vector of parameters α_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 α_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_α and each element of $c_0\alpha_{0,k}$. Combining the hierarchical asymmetric prior with local topics, the HALT-LDA model is

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

2.2.4 Prior Parameter Specification

We place an asymmetric prior on α and a Gamma prior on c_α 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

LDA and LT-LDA, we place priors

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

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

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

We generated 100,000 sets of $c_0\alpha_{0,k}$ for $K = 50$. This generates a largest order statistic for $\alpha_{i,k}$ of 0.09 with a standard deviation of 0.04. At K = 100, the largest order statistic is 0.05 with a standard deviation of 0.02. The largest order statistic from the prior differs from the overall local topic prevalence in our results in section 2.5.2; however, a priori, this result for the highest order statistic was reasonable. Later order statistics were reasonably modeled with Gamma(1,1) We expect each topic to place high probability above 0.02 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, ϕ_k and ψ_i. The priors are fixed such that $c_\beta\beta = c_\gamma\gamma = \{0.05, \ldots, 0.05\}$. Sensitivity analysis in Section 2.7.2 of the web appendix shows that conclusions from HALT-LDA are robust to deviations from our choice of c_β, c_γ, and a_α.

2.3 Computation and Inference for Hierarchical Topic Models

The general goal of inference in hierarchical topic models is to estimate the topic-word distributions, ϕ_k and ψ_i, and web page-topic distributions, θ_{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 ψ_i. We implement HALT-LDA with the data and functions available in the first author's github repository https://github.com/jwanghb/publichealth-web sites and in the supplementary materials.

Let W and Z be ragged arrays of identical structure, with one element w_{ijh} and z_{ijh} for every 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 α be the set of all web site-topic distributions α_i and similarly, let θ, ϕ, and ψ be the sets of all θ_{ij}, ϕ_k, and ψ_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 ϕ_k, ψ_{il}, and θ_{ij} from the posterior. We are left to sample topics z_{ijh} of each word w_{ijh}, scale parameter c_α, and web site-topic distributions α_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_α, web site-topic distribution α_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 α_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 α_i from a Dirichlet$(c_0 \alpha_0 + \sum_{j=1}^{M_i} \lambda_{ij,k})$ (Teh et al., 2006). Parameters c_α and $c_0 \alpha_{0,k}$ are sampled using Metropolis-Hastings.

We estimate conditional means of the multinomial parameters ϕ_k, ψ_i, and θ_{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 of the global topic-word proportions ϕ_k at iteration q is given by the conditional posterior mean

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

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_α 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_α and probability vector α_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_α, α_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 2.7.3 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.

2.4 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, 2012). 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 dataset 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.

2.5 Results

The 10-fold cross validated held-out log likelihoods are plotted against the number of global topics K in Figure 2.1 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

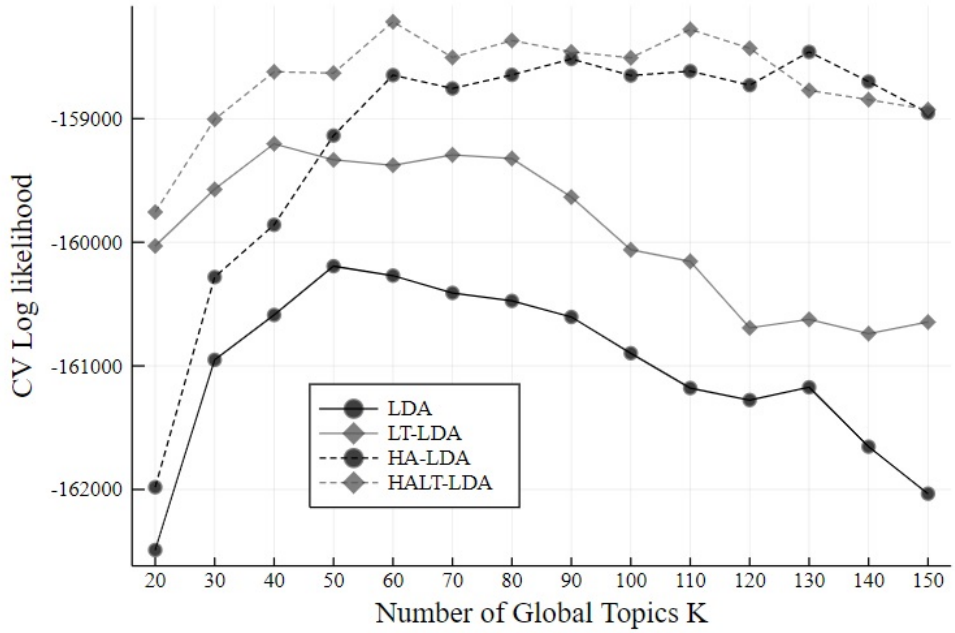

Figure 2.1: Plot of 10-fold cross validated (CV) held-out log likelihood by different number of global topics K.

nested asymmetric prior is a flexible prior that can accommodate 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.

2.5.1 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 $R_{k,v}^{(\text{HA})}$ denote the rank of word v in topic k from HA-LDA and let $R_{i,v}^{(\text{HALT})}$ 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

$$\arg\min_{k} \sum_{v=1}^{V} |R_{i,v}^{(HALT)} - R_{k,v}^{(HA)}|. \tag{2.1}$$

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

$$\arg\min_{k} \sum_{v=1}^{V} (\psi_{i,v}^{(HALT)} - \phi_{k,v}^{(HA)})^2. \tag{2.2}$$

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 2.1 and equation 2.2 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. 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 2.2 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.

2.5.2 Topic Model Output and Applications

Table 2.2 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

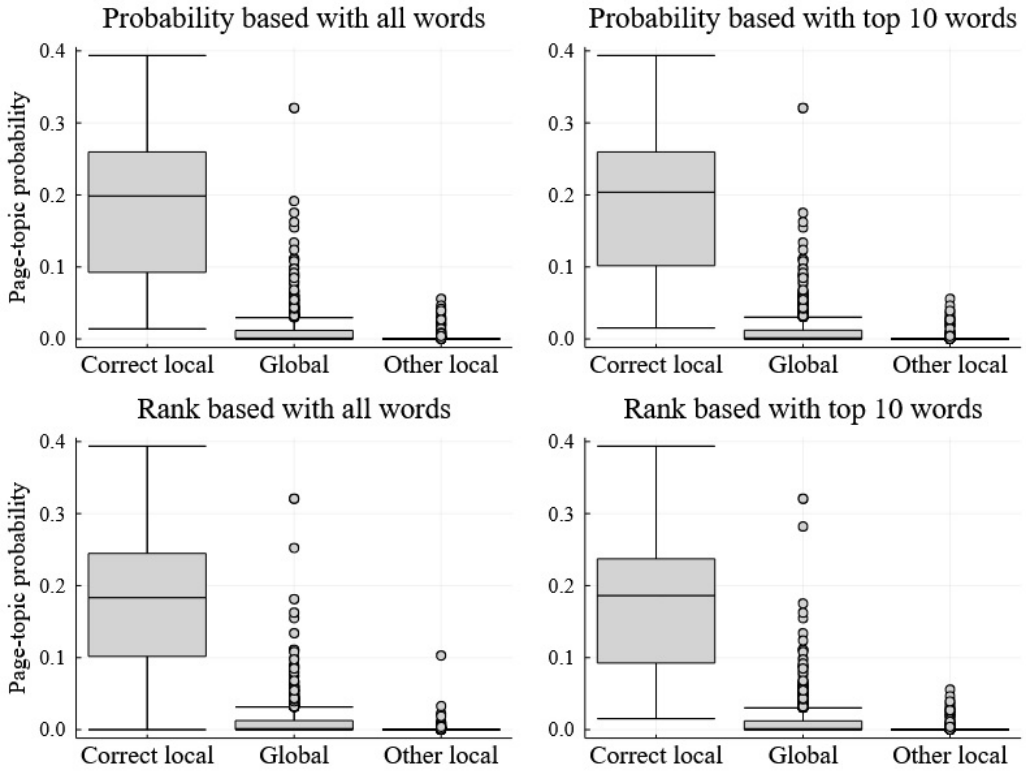

Figure 2.2: Boxplots of the web site average web page-topic distributions $\bar{\theta}_{i\cdot,k} = \frac{1}{M_i}\sum_{j=1}^{M_i}\theta_{ij,k}$ of global topics and matched local topics in HA-LDA. 'Correct local' shows the distribution of $\bar{\theta}_{i\cdot,k}$, where topic k has been matched to 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.

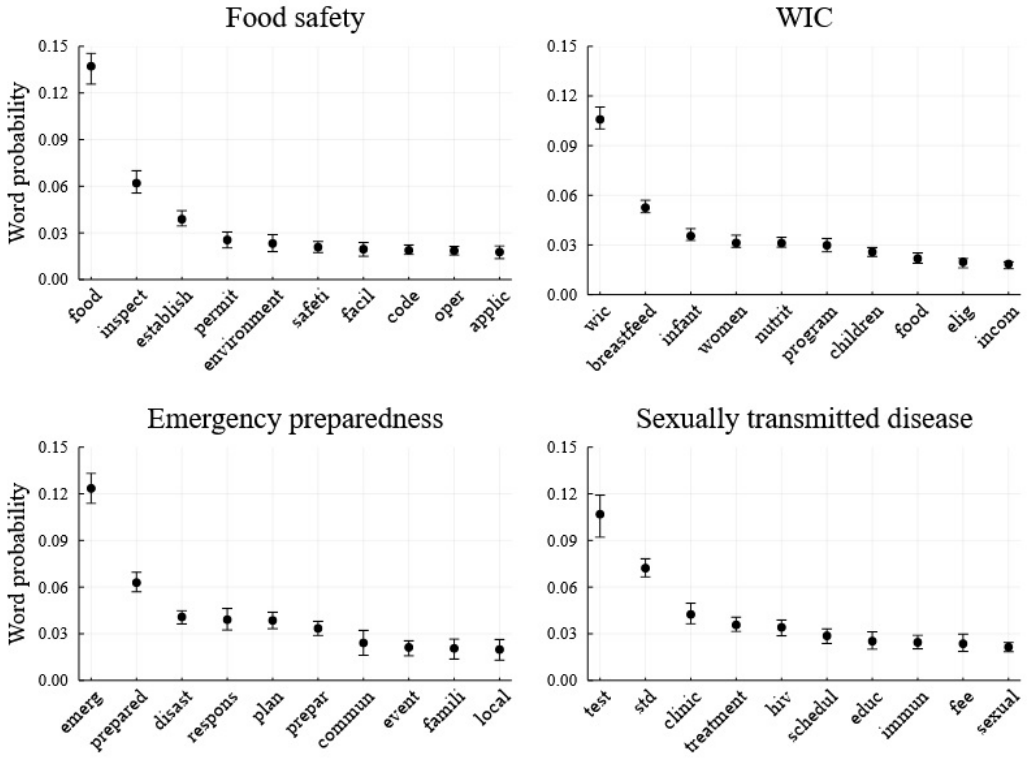

Figure 2.3: Median and 95% intervals of conditional posterior means of word probabilities for the ten most probable words in four health topics.

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 infections. Estimates and 95% intervals of conditional posterior means for word probabilities of these topics' ten most probable words are plotted in Figure 2.3. The word probabilities for the ten most probable words are much larger than the average probability 1/1614.

Table 2.2: 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 2.3.

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*
STI	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*

Table 2.3 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 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 2.2, 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 2.2.

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 site 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

Table 2.3: 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*

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 $\theta_{ij,k}$ across all $j = 1, \ldots, M_i$ pages,

$$\max_j \theta_{ij,k}.$$

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 infections, using estimates from each of

the 1,000 MCMC samples. Plots of ATC are shown in Figure 2.4. 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 infections (STIs), 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 STIs. 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 listing nine clinical services of which one is a screening and tests service. Under the screening and tests service are 5 tests provided of which one is for STIs and one is 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 STIs. 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 STIs. Web site 17 has a page of thirteen frequently asked questions of which one is directly related to STIs. However, testing for STIs is mentioned two additional

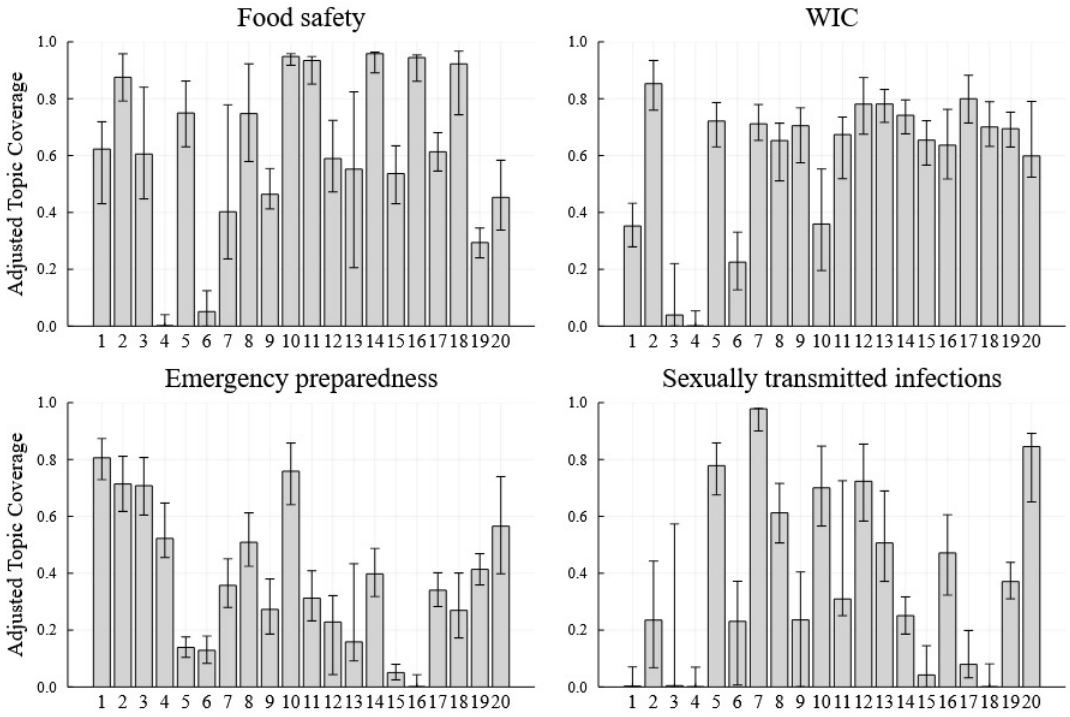

Figure 2.4: Bar plots of adjusted topic coverage for four global topics from Table 2.2. Bar heights are medians and error bars are 95% credible intervals.

times as part of larger answers to questions about services offered. This explains why ATC and the ATC lower bound for STIs 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 having low coverage.

2.6 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 dataset 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 dataset. 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 infections.

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 dataset 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 2.7.1 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 toics.

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 2.3 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.

2.7 Web Appendix

2.7.1 Simulation Study of HALT-LDA and Data Sets with Zero, One, or Two Local Topics

We investigate how the HALT-LDA model with one local topic for each web site models simulated data. We are particularly interested in inferences about local topic probabilities either when local topics are not present or when more than one local topic is present in a web site. All simulated datasets are created with $K = 50$ global topics, $V = 1000$ unique words, $M = 10$ web sites, $M_i = 50$ web pages for all $i = 1, \ldots, 10$, and $N_{ij} = 100$ words for each web page $j = 1, \ldots, 50$. We study 5 scenarios: (1) no local topics, (2) 5 web sites with

one local topic each and no local topics for the remaining web sites, (3) 10 web sites with one local topic each, (4) 5 web sites with one local topic each and 5 with two local topics each, and (5) 10 web sites with two local topics each.

For web sites with one local topic the non-standardized web site average local topic probabilities $\mu_{i,K+1}$ are generated from a Normal$(0.25, 0.05^2)$ truncated at 0 and 1. Web page topic probabilities θ_{ijk} are generated by first sampling an unstandardized local topic probability from Normal$(\mu_{i,K+1}, 0.05^2)$ and unstandardized global topic probabilities from Dirichlet$(\{0.04, \ldots, 0.04\})$ then standardizing such that $\sum_{k=1}^{51} \theta_{ijk} = 1$. We chose an unstandardized mean of 0.25 as 0.2 (0.25/1.25) is a reasonable average local topic probability. For web sites with two local topics, we generate $\mu_{i,K+1}$ and $\mu_{i,K+2}$ from Normal$(0.15, 0.05^2)$ and Normal$(0.1, 0.05^2)$ respectively. Web page topic probabilities are generated similarly with unstandardized local topic probabilities sampled from Normal$(\mu_{i,K+1}, 0.05^2)$ and Normal$(\mu_{i,K+2}, 0.05^2)$. Topic-word probabilities are generated from a Dirichlet$(\{0.01, \ldots, 0.01\})$ to get an average highest order statistic around 0.20 and about 20 words with probability greater than 0.01 in each topic. Topic and word indices are sampled from Multinomial distributions given web page topic probabilities and topic-word probabilities. We generate 100 datasets for each of the 5 variations for a total of 500 simulated datasets.

Web site average local topic probabilities are

$$\frac{1}{50} \sum_{j=1}^{50} \theta_{ij,K+1}$$

for web sites with one local topic or

$$\frac{1}{50} \sum_{j=1}^{50} (\theta_{ij,K+1} + \theta_{ij,K+2})$$

for web sites with two local topics. Estimates of web site average local topic probability are

$$\frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{50}\sum_{j=1}^{50}\bar{\theta}_{ij,K+1}^{(q)}$$

or

$$\frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{50}\sum_{j=1}^{50}(\bar{\theta}_{ij,K+1}^{(q)}+\bar{\theta}_{ij,K+2}^{(q)})$$

where $Q = 500$ MCMC iterations. There are 10 estimates for web site average local topic probability for each of the 500 simulated datasets.

The first row of Figure 2.5 shows histograms of estimated web site average local topic probabilities when local topics are not present. The true web site average local topic probability is 0 and approximately 86% of all estimates are less than 0.005. When local topics are not present, HALT-LDA typically estimates web site average local topic probability near 0. When estimates are greater than 0.005 they typically range between 0.01 and 0.07, much lower than the 0.20 average. Extraneous local topics with probability estimates greater than 0.02 are further investigated. The 3 highest probability words in each of the extraneous local topics are counted in their corresponding web sites and counted in other web sites then averaged among the other web sites.

Define word count ratio as the ratio of corresponding web site count to other web site count. Ratios larger than 1 indicate the extraneous local topic has high probability words found more often in its own web site than in other web sites. All local word count ratios are greater than 1. Thus, when local topics do not exist but are modeled, HALT-LDA identifies a topic as local that places high probability on words found more often in the given web site than in other web sites.

Figure 2.6 plots estimated web site average local topic probability against true web site average local topic probability for scenarios (2) to (5) where local topics exist. The estimated web site average local topic probabilities are close to the true web site average local topic

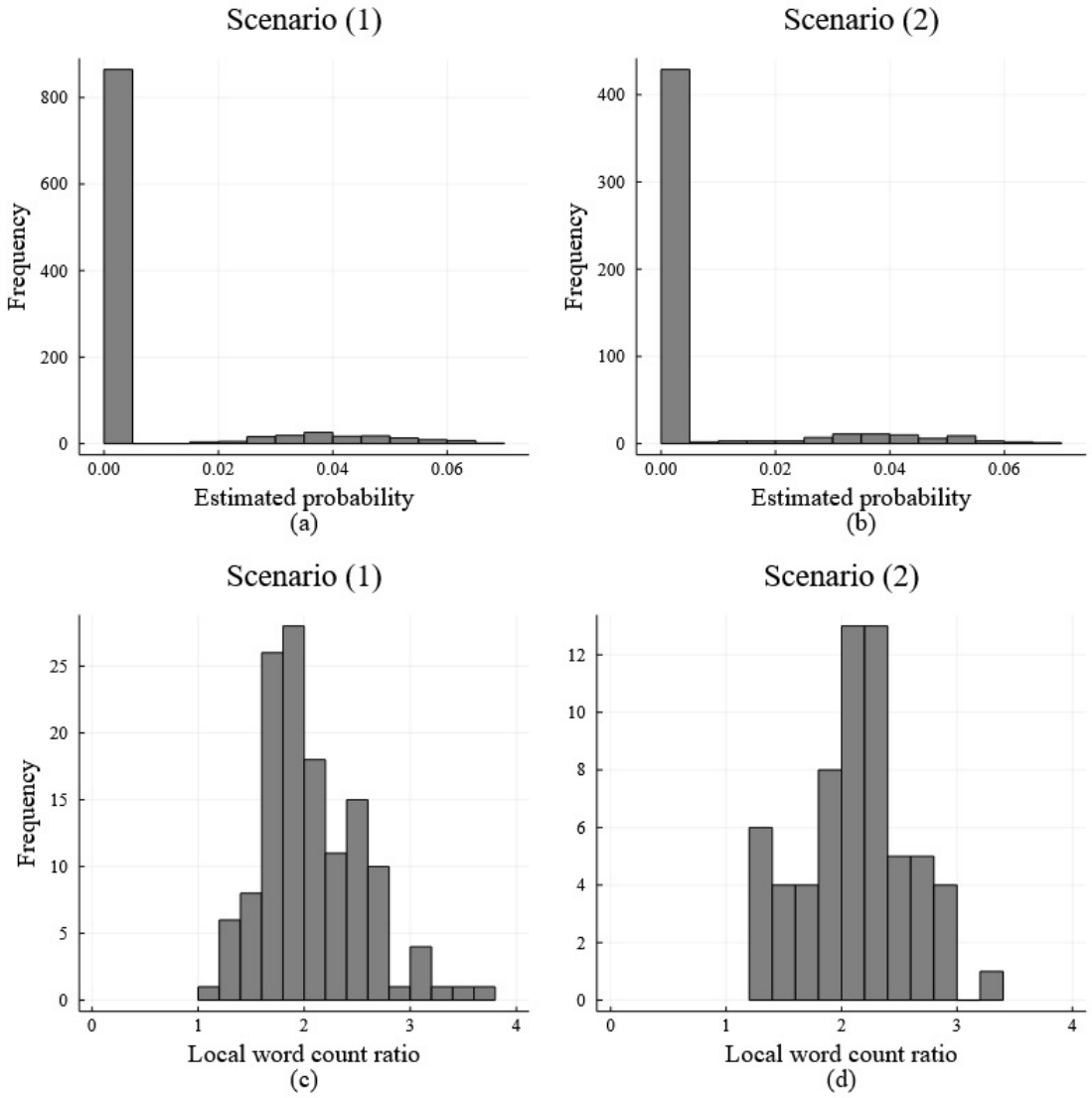

Figure 2.5: (a) Histogram of estimated web site average local topic probabilities for web sites with no local topic in scenario (1), 10×100 estimates are plotted. (b) Histogram of estimated web site average local topic probabilities for web sites with no local topic in scenario (2), 5×100 estimates are plotted. (c) and (d) Histograms of local word count ratios of the highest probability words in extraneous local topics in scenario (1) and (2) respectively.

probability. The bottom two figures indicate that when HALT-LDA models web sites with two local topics, it can merge local topics to a single local topic, when the number of global topics modeled is limited to the true number of global topics. We expect some merged local topics to split into the two local topics with one local topic modeled as a global topic if we were to allow more than 50 global topics.

Among the results of all 400 simulated datasets where local topics do exist, only one web site of the 3500 web sites with local topics shows HALT-LDA incorrectly modeling no local topics when there was indeed a local topic present. The point near (0.25,0) Figure 2.6 (c) is for that local topic. The local topic for that web site was instead identified as a global topic. Inspection of highest probability words of local topics and global topics, as shown in Section 5 of the main text, is recommended to determine that local topics are indeed correctly identified.

2.7.2 Sensitivity Analysis of HALT-LDA Conclusions to Prior Specifications

We study the effects of changing hyperparameters c_β, c_γ, and a_α on global topic-word distributions $\phi_{k,v}$. The hyperparameters used in the main results are $c_\beta = c_\gamma = 0.05$ and $a_\alpha = 1$. We consider four sensitivity analysis scenarios doubling or halving these parameters. The first sensitivity analysis model (SA1) sets $c_\beta = c_\gamma = 0.025$ and $a_\alpha = 0.5$; the second sensitivity analysis model (SA2) sets $c_\beta = c_\gamma = 0.025$ and $a_\alpha = 2$; the third sensitivity analysis model (SA3) sets $c_\beta = c_\gamma = 0.1$ and $a_\alpha = 0.5$; the fourth sensitivity analysis model (SA4) sets $c_\beta = c_\gamma = 0.1$ and $a_\alpha = 2$. The overall prevalence of local topics in all 5 settings range between 17% and 19% with the main results at 18%, SA1 at 18%, SA2 at 19%, SA3 at 17%, and SA4 at 19%. We match global topics in SA1, SA2, SA3, SA4 to the health topics in Table 2 of the main text with the rank based method using the top 10 words. Table 2.4 shows the 3 highest probability words for the nine health topics and for the matched topics in the four sensitivity analysis models and the prevalence of the topic. Generally, the more prevalent topics are similar across all 5 settings. The breast cancer topic differs the most in

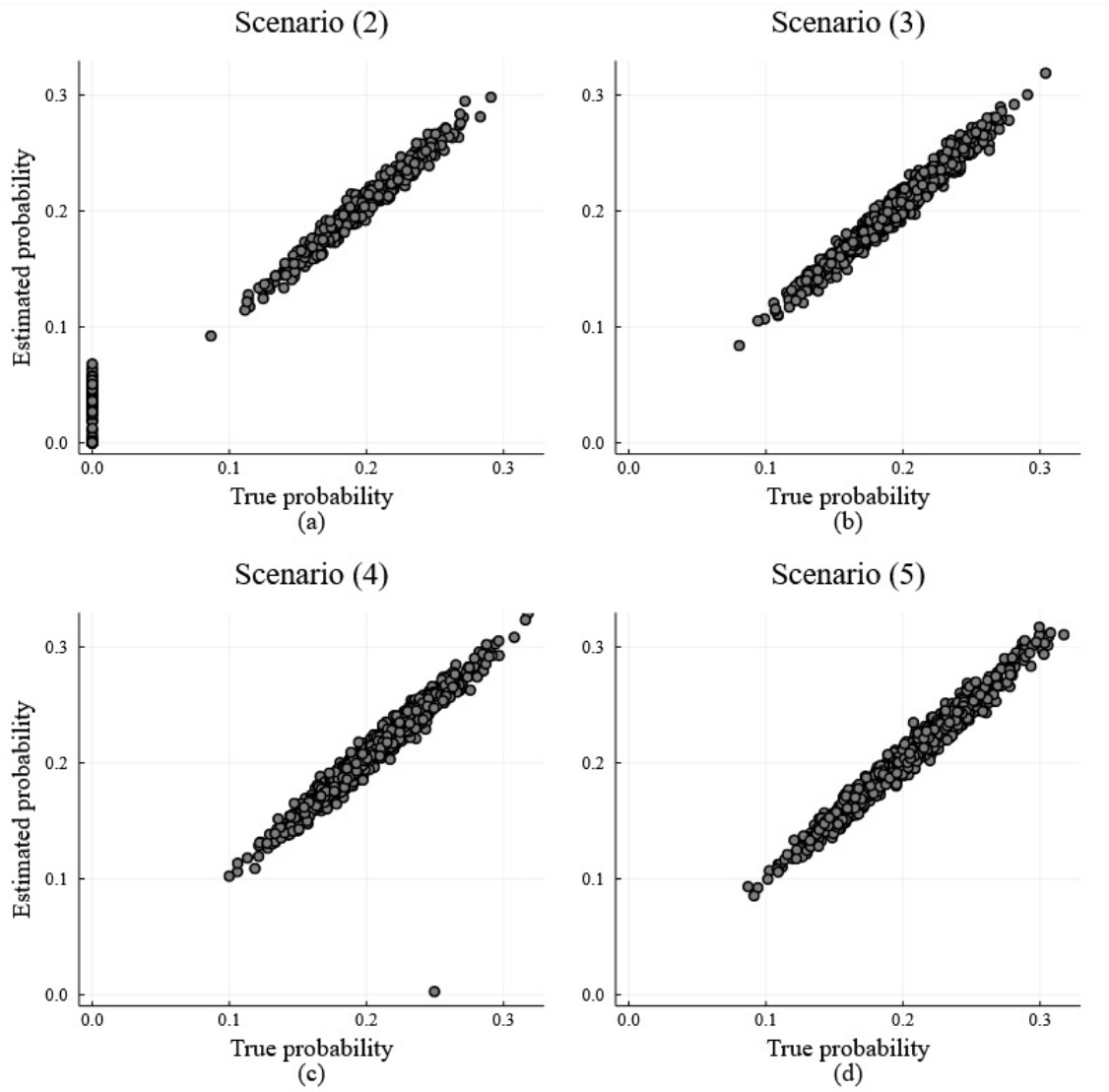

Figure 2.6: Scatterplots of 1000 estimated web site average local topic probabilities vs true web site average local topic probabilities. (a) Scenario (2) where 5 web sites have one local topic each and 5 web sites have no local topics. (b) Scenario (3) where all 10 web sites have one local topic each. (c) Scenario (4) where 5 web sites have two local topics each and 5 web sites have one local topic each. (d) Scenario (5) where all 10 web sites have two local topics each.

the 3 most probable words between the main results and the four sensitivity model results. However, when looking at the 10 most probable words from table 2 in the main text, the breast cancer topic from the main results includes all words *screen, cancer, breast, women* found in the 3 most probable words in SA1, SA2, SA3, and SA4. The topics modeled by HALT-LDA are fairly robust to changes in hyperparameters c_β, c_γ, and a_α.

2.7.3 Left-to-Right Algorithm

We use the left-to-right algorithm (Wallach et al., 2009b) and describe how we adapted it to our HALT-LDA model. To evaluate our models with the left-to-right algorithm we split pages randomly from each web site into 80% for MCMC sampling and the remaining 20% of each web site for evaluation. For each sample q we calculate conditional posterior means $\bar{\phi}_{k,v}^{(q)}$ and $\bar{\psi}_{i,v}^{(q)}$ from the counts in sample q and save the sampled $c_\alpha^{(q)}$ and $\alpha_i^{(q)}$. We calculate an estimate for scale parameter c_α, probability vectors α_i and topic-word probabilities $\phi_{k,v}$ and $\psi_{i,v}$ by averaging over 500 MCMC samples after a burn-in of 1500 samples.

The left-to-right algorithm approximates the probability of a held-out document W_{ij} given topic-word probabilities ϕ and ψ_i and Dirichlet parameters c_α and α_i or $P(W_{ij}|\phi, \psi_i, c_\alpha, \alpha_i)$. We provide pseudocode for a single held-out document but it can be extended to multiple held-out documents by adding outer loops over held-out web pages of each web site. The number of particles R is set to $4 \times 2000/N_{ij}$ as suggested by Wallach et al. (2009b), where N_{ij} is the number of words in web page j in web site i.

1. initialize $ll = 0$
2. for h-th word w_{ijh} in held-out document W_{ij} do
3. initialize $p_{ijh} = 0$
4. for each particle $r = 1, \ldots, R$ do
5. for $h' < h$ do

Table 2.4: Global topics from SA1, SA2, SA3, and SA4 are matched to the health topics shown in Table 2 of the main text using the rank based method with the top 10 words. The 3 highest probability words for the nine health topics in each sensitivity analysis and the main analysis are shown with their respective topic prevalences.

	Main	SA1	SA2	SA3	SA4
Disease prevention	*diseas* (3.3%)	*diseas* (3.2%)	*diseas* (3.3%)	*diseas* (3.5%)	*prevent* (3.7%)
	prevent	*caus*	*peopl*	*peopl*	*caus*
	risk	*peopl*	*caus*	*caus*	*risk*
Food safety	*food* (2.9%)	*food* (1.8%)	*food* (1.5%)	*food* (1.9%)	*food* (1.7%)
	inspect	*establish*	*inspect*	*establish*	*establish*
	establish	*permit*	*establish*	*inspect*	*permit*
WIC	*wic* (2.7%)	*wic* (2.9%)	*wic* (2.4%)	*wic* (2.7%)	*wic* (2.6%)
	breastfeed	*breastfeed*	*breastfeed*	*breastfeed*	*breastfeed*
	infant	*infant*	*infant*	*infant*	*infant*
Vaccinations	*immun* (2.0%)	*vaccin* (2.3%)	*vaccin* (2.4%)	*immun* (2.3%)	*vaccin* (2.2%)
	vaccin	*immun*	*immun*	*vaccin*	*immun*
	adult	*adult*	*adult*	*adult*	*adult*
Breast cancer	*test* (1.9%)	*screen* (0.9%)	*cancer* (1.0%)	*screen* (1.2%)	*cancer* (1.1%)
	women	*cancer*	*screen*	*women*	*women*
	clinic	*women*	*breast*	*cancer*	*screen*
Emergency preparedness	*emerg* (1.8%)	*emerg* (1.8%)	*emerg* (2.1%)	*emerg* (1.9%)	*emerg* (2.2%)
	prepar	*prepar*	*prepar*	*prepar*	*prepar*
	disast	*disast*	*disast*	*disast*	*disast*
Hospital care	*care* (1.7%)	*care* (1.8%)	*care* (1.8%)	*care* (1.7%)	*care* (2.0%)
	patient	*inform*	*inform*	*priva*	*provid*
	provid	*priva*	*priva*	*medic*	*patient*
Sexually transmitted disease	*test* (1.5%)	*test* (1.5%)	*test* (0.7%)	*test* (0.8%)	*test* (1.3%)
	std	*std*	*std*	*std*	*std*
	clinic	*hiv*	*hiv*	*hiv*	*hiv*
Family program	*child* (1.4%)	*child* (1.7%)	*child* (1.4%)	*child* (1.5%)	*child* (1.3%)
	children	*program*	*famili*	*famili*	*famili*
	famili	*famili*	*children*	*program*	*parent*

6. sample $z_{ijh'}^{(r)}$ from multinomial on 1:K+1 where

$$P(z_{ijh'}^{(r)} = k | w_{ijh'}, \phi, \psi_i, c_\alpha, \alpha, \{z_{h''}^{(r)}\}_{h'' \neq h', h'' < h}) \propto (m_{ij,k}^{\prime(r)-} + \alpha_i) \phi_{k,w_{ijh}}^{1_{k \leq K}} \psi_{i,w_{ijh}}^{1_{k=K+1}},$$

for $k = 1, \ldots, K+1$ where $m_{ij,k}^{\prime(r)-}$ is the count of words from topic k for all $h'' < h$ where $h'' \neq h'$.

7. end for

8. $p_{ijh} = p_{ijh} + \sum_{k=1}^{K+1} P(w_{ijh}, z_{ijh}^{(r)} = k | \phi, \psi_i, c_\alpha, \alpha)$

9. sample $z_{ijh}^{(r)}$ from multinomial$_{1:K+1}$ where

$$P(z_{ijh}^{(r)} = k | w_{ijh}, \phi, \psi_i, c_\alpha, \alpha, \{z'_h \text{ where } h' < h\}) \propto (m_{ij,k}^{(r)-} + \alpha_i) \phi_{k,w_{ijh}}^{1_{k \leq K}} \psi_{i,w_{ijh}}^{1_{k=K+1}},$$

for $k = 1, \ldots, K+1$ where $m_{ij,k}^{(r)-}$ is the count of words from topic k for all $h' < h$.

10. end for

11. $p_{ijh} = p_{ijh}/R$

12. $ll = ll + \log(p_{ijh})$

13. end for

14. $\log(P(W_{ij} | \phi, \psi_i, c_\alpha, \alpha_i)) \approx ll$

CHAPTER 3

Hierarchical Topic Presence Models

3.1 Introduction

Probabilistic topic models have been used to abstract topical information from collections of text documents by modeling documents as a mixture of K latent topics where each topic is itself a mixture of the V unique words in a vocabulary. A topic is characterized by a vector of word probabilities and a 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. Topic models such as latent Dirichlet allocation (LDA) are constructed under a Dirichlet-multinomial framework, where words in a document follow a multinomial distribution with a Dirichlet prior (Blei et al., 2003; Chang and Blei, 2009). More recently, Zhou et al. (2012) introduced the Poisson factor analysis (PFA) framework which models word counts with a Poisson likelihood. Zhou and Carin (2015) demonstrate computational advantages of PFA models over LDA models. We discuss and propose novel topic models in the Poisson factor analysis framework.

Our work is motivated by a text data set of web pages nested within local health department web sites in the United States. We treat web pages as separate documents nested in web sites. We are interested in identifying health topics covered by health department web sites, how frequently topics are covered and which topics are or are not covered in individual web sites. As is usual in PFA and LDA models, we label topics by inspecting the most frequent words in the topic.

Some development of models for nested or clustered documents has occurred with LDA.

Some models address nesting by modeling multiple levels of document-topic distributions (Qiang et al., 2017). Some models explicitly model topics that are unique to documents in a given cluster (Hua et al., 2020; Wang and Weiss, 2021b). Wang and Weiss (2021b) further proposed hierarchical priors on document-topic distributions to accommodate the belief that which topics are more common vary from web site to web site. In contrast, there has been little development of PFA for nested documents.

Different public health department web sites will likely contain different subsets of topics. Some health topics will be present in most web sites while other health topics may be rarer or may be of specific interest depending on demographic or geographic characteristics of the local health department. Thus, we propose modeling topic presence conditional on covariates.

Sparsity inducing priors model documents as a mixture of a subset of the possible topics and can be implemented by introducing unknown topic presence binary indicators for whether a given topic contributes words to a particular document. Topic presence has previously been introduced in non-nested document collections (Williamson et al., 2010; Zhou et al., 2012; Archambeau et al., 2015; Gan et al., 2015). Zhou et al. (2012) proposed the sparse Gamma-Gamma Poisson factor analysis, also known as the negative binomial focused topic model (NB-FTM) (Zhou and Carin, 2015) in the PFA framework. The NB-FTM needs to be adapted for nested documents. Further, prior researchers have not modeled topic presence as functions of document covariates.

Nesting of web pages in web sites allows for topic presence modeling at the web site level, the web page level, or both. We propose three topic presence models (TPM). The first TPM has topic presence at the web site level while all web pages are mixtures of the topics present in their respective web sites. The second TPM is for web pages only while web sites are mixtures of all topics. The third model has TPMs for both web site and web page. Web site (web page) topic presence is a vector of unknown binary variables that identifies the subset of topics in a web site (web page) – topics must be present at the web site level to be present

in a web page nested in the web site.

Previous topic presence models have modeled topic presence as a priori independent where the unknown probabilities of topic presence have fully known priors. We extend this to allow topic presence probabilities to be a priori exchangeable, where we estimate the global mean and variance of the probability of a given topic's presence across web sites (or web pages). We extend this model and consider a logistic regression model for topic presence where probability of topic presence is modeled conditional on covariates.

When we model multiple web sites, pages of a web site are likely to include common local words or phrases such as names and locations that are not commonly found in the web pages of other web sites. These local words form a *local topic* that is unlikely to be found on other web sites. Local topics are unique to a web site while global topics can be present in multiple web sites. Local topics have been introduced to LDA models (Hua et al., 2020; Wang and Weiss, 2021b) and Wang and Weiss (2021b) showed that including local topics reduces the number of global topics needed without sacrificing performance.

In topic models, typically the number of topics K is a parameter to be specified. The number of topics K can be modeled however Zhou and Carin (2015) suggest that sufficiently large K provides a good approximation to models with K unknown. We take K as a parameter that we tune.

We derive a Gibbs sampler for inference after suitable data augmentation. We need two families of auxiliary random variables distributed as the Chinese restaurant table (CRT) distribution to sample topic parameters from conditional posterior Gamma distributions (Teh et al., 2006; Zhou et al., 2012). We introduce families of Pólya-Gamma distributed (Polson et al., 2013) auxiliary random variables to allow us to sample our logistic regression coefficients in the topic presence models as Gibbs steps.

We consider several hierarchical PFA models, with and without local topics. We also consider six topic presence models: at the web site level we consider using covariates to predict presence, an exchangeable prior, and topics always present. At the web page level

we consider exchangeable and topics always present. We compare models with perplexity, a measure of predictive fit (Wallach et al., 2009b; Zhou et al., 2012) that we extend to our hierarchical settings. We provide a quick automated approach using the most probable words of a topic to check if our models correctly capture patterns in web site topic presence.

The next section 3.2 presents the LDA and PFA models in our context then section 3.3 extends PFA to be hierarchical PFA with local topics and hierarchical topic presence models. Section 3.4 presents our analysis of local health department web sites and the paper closes with discussion.

3.2 Poisson Factor Analysis

We first present notation for the Poisson factor analysis (PFA) model in the context of our hierarchical data set and extend PFA to include local topics. Let $i = 1, \ldots, M$ index web sites and let $j = 1, \ldots, N_i$ index web pages nested in web sites with N_i web pages in web site i.

For PFA, we treat individual web pages as separate documents. Let $k = 1, \ldots, K$ index global topics where K is set in advance. The vocabulary or set of unique words in a document collection is known and has length V and we let $v = 1, \ldots, V$ index words in the vocabulary. Poisson factor analysis (PFA) models word counts with a Poisson likelihood. Let z_{ijkv} be the latent count of word v from topic k in web page j of web site i. Let ϕ_{kv} be the probability of word v in topic k and let θ_{ijk} be the weight of topic k in web page j of web site i such that θ_{ijk} is the expected count of words from topic k in web page j of web site i. Then $\phi_{kv}\theta_{ijk}$ is the expected count of word v from topic v in web page j of web site i, and PFA models latent counts $z_{ijkv}|\phi_{kv}, \theta_{ijk} \sim \text{Poisson}(\phi_{kv}\theta_{ijk})$.

We model one local topic for each web site. Let global topic word probability vectors be $\phi_k = (\phi_{k1}, \ldots, \phi_{kV})'$ and let the local topic word probability vector be $\psi_i = (\psi_{i1}, \ldots, \psi_{iV})'$ for web sites $i = 1, \ldots, M$, such that $\psi_{iv} > 0$ and $\sum_{v=1}^{V} \psi_{iv} = 1$. Only web pages of web

site i can have non-zero topic weight for local topic ψ_i. Define $\Phi_i = (\phi_1, \ldots, \phi_K, \psi_i)'$ to be the $(K+1) \times V$ matrix of word probability vectors for all global topics plus web site i's one local topic word probability vector. Then Φ_{ikv} is the probability of word v in topic k in web site i, where $k = 1, \ldots, K+1$ where $k \leq K$ indexes the K global topics while $k = K+1$ is the local topic for web site i. Extend the definitions of θ_{ijk} and z_{ijkv} to have k run from 1 to $K+1$. The PFA local topic model (PFA-LT) models z_{ijkv} as

$$z_{ijkv} | \Phi_{ikv}, \theta_{ijk} \sim \text{Poisson}(\Phi_{ikv} \theta_{ijk}). \tag{3.1}$$

From now on, for models with local topics, k runs from 1 to $K+1$ while for models without local topics, k runs from 1 to K.

3.3 Poisson Factor Analysis with Local Topics and Hierarchical Topic Presence

Topic presence is a web site or web page binary variable that indicates whether a topic is present or not in the web page or web site. We can model web site topic presence, web page topic presence, both, or neither. Let $b_{ik} = 1$ indicate that topic k is present in web site i and let $c_{ijk} = 1$ indicate that topic k is present in web page j of web site i. When b_{ik} and c_{ijk} are both included in our model, topic k is present in web page j of web site i only if both $b_{ik} = c_{ijk} = 1$. The number of words in web page j of web site i is $z_{ij\bullet\bullet} = \sum_{kv} z_{ijkv}$. We model topic weights $\theta_{ijk} \geq 0$ conditional on global topic weight parameters r_k, web site topic presence b_{ik} and web page topic presence c_{ijk} such that

$$\theta_{ijk} | r_k, b_{ik}, c_{ijk} \sim \text{Gamma}(r_k b_{ik} c_{ijk} z_{ij\bullet\bullet}, 1). \tag{3.2}$$

Thus, $\theta_{ijk} = 0$ with probability 1 if $b_{ik} = 0$ or $c_{ijk} = 0$. The gamma density in (3.2) has mean equal to the variance as for smaller θ_{ijk}, we want smaller variance and for larger θ_{ijk}

we want larger variation. The scale parameter in (3.2) is 1 as there is an arbitrary scaling involved which is unnecessary for modeling the counts.

The number of words $z_{ij\bullet}$ is a scaling factor to increase or decrease θ_{ijk} as for a given r_k, web pages with more words will have larger θ_{ijk} compared to web pages with fewer words. Omitting $z_{ij\bullet}$ in (3.2) would require a factor indexed by ij to model the web page word count $z_{ij\bullet}$. As size $z_{ij\bullet}$ is at best an ancillary statistic, we condition on $z_{ij\bullet}$ in (3.2). Conditional on the total $z_{ij\bullet}$ of a set of KV independent Poisson random variables (PRVs), the set of PRVs are distributed as multinomial. However, KV is very large, the probabilities are small and the Poisson approximation to the multinomial distribution will be quite accurate. In modeling counts z_{ijkv} as Poisson in (3.1), we do not directly condition on $z_{ij\bullet}$ but only indirectly in (3.2), so the Poisson approximation should be quite acceptable.

We place a gamma hyperprior on the r_k for $k = 1, \ldots, K+1$,

$$\begin{aligned} r_k | r_0 &\sim \text{Gamma}(r_0, 1), \\ r_0 &\sim \text{Gamma}(d_{r0}, e_{r0}), \end{aligned}$$

where r_0 is a prior mean global topic weight with fixed prior hyperparameters d_{r0} and e_{r0}. We place a Dirichlet prior on word probability vectors ϕ_k and ψ_i such that

$$\begin{aligned} \phi_k &\sim \text{Dirichlet}(\alpha_\phi 1_V), \\ \psi_i &\sim \text{Dirichlet}(\alpha_\psi 1_V), \end{aligned}$$

where α_ϕ and α_ψ are fixed hyperparameters and 1_V is a ones vector of length V and the Dirichlet$(c1_V)$.

3.3.1 Models for Topic Presence Probabilities

Web site topic presence b_{ik} is given a Bernoulli(π_{ik}) prior, where π_{ik} is the probability of topic k being present in web site i. We consider three prior specifications for π_{ik} and b_{ik}: topics are always (A) present; an exchangeable (E) prior across topics on the probability that a topic is present, and a structured (S) prior on π_{ik} where we use covariates and logistic regression to model topic presence.

Topics can be always (A) present at the web site level such that $\pi_{ik} \equiv 1$ and therefore $b_{ik} \equiv 1$ for all web sites i and topics k. In the exchangeable (E) prior, all websites have probability $\pi_{ik} \equiv \pi_k$ with $\pi_k | d_\pi, e_\pi \sim \text{Beta}(d_\pi, e_\pi)$ prior on π_k and the π_k's are exchangeable. We parameterize the Beta prior parameters $d_\pi \equiv d_\pi(\mu_\pi, \sigma_\pi^2)$ and $e_\pi \equiv e_\pi(\mu_\pi, \sigma_\pi^2)$ in terms of the mean $\mu_\pi = d_\pi/(d_\pi + e_\pi)$ and variance $\sigma_\pi^2 = \mu_\pi(1-\mu_\pi)/(d_\pi + e_\pi + 1)$ of Beta(d_π, e_π) and place beta priors on the new parameters

$$\mu_\pi \sim \text{Beta}(d_{\mu\pi}, e_{\mu\pi}), \tag{3.3}$$

$$\sigma_\pi^2 \sim \text{Beta}(d_{\sigma\pi}, e_{\sigma\pi}), \tag{3.4}$$

where $d_{\mu\pi}$, $e_{\mu\pi}$, $d_{\sigma\pi}$, and $e_{\sigma\pi}$ are fixed hyperparameters.

The structured (S) prior models π_{ik} as functions of Q web site covariates $X_i = (X_{i1}, \ldots, X_{iQ})'$ for web site i including the intercept and let $\beta_k = (\beta_{k1}, \ldots, \beta_{kQ})'$ be the Q-vector of regression coefficients for topic k. The structured prior sets $\pi_{ik} = g(X_i'\beta_k) = \exp(X_i'\beta_k)/(1 + \exp(X_i'\beta_k))$ where $g(a) = \exp(a)/(1+\exp(a))$ is the inverse logit link. We place a Normal(β_0, Σ) prior on β_k, where β_0 is a prior mean Q-vector and $\Sigma_{Q\times Q}$ is the prior covariance matrix. We place a Normal($\mu_0, \sigma_0^2 I_Q$) prior on β_0, where I_Q is the Q-dimension identity matrix, σ_0 is a scalar, and $\mu_0 = (\mu_{01}, \ldots, \mu_{0Q})'$ is a mean vector of length Q. We let Σ be a diagonal covariance matrix with diagonal elements $\sigma_1^2, \ldots, \sigma_Q^2$ indexed by $q = 1, \ldots, Q$ and place a Gamma(d_σ, e_σ) prior on $1/\sigma_q^2$. Hyperparameters μ_0, σ_0^2, d_σ, and e_σ are fixed.

We can similarly apply the same (A), (E), and (S) prior specifications at the web page

level. Web page topic presence c_{ijk} is given a Bernoulli(η_{ijk}) prior, where η_{ijk} is the probability of topic k being present in web page j of web site i. Topic indicator could be always present at the web page level such that $\eta_{ijk} \equiv 1$ and $c_{ijk} \equiv 1$ for all web pages j, web sites i, and topics k. The exchangeable prior sets $\eta_{ijk} = \eta_k$ for all web pages j and web sites i and prior $\eta_k|\mu_\eta, \sigma_\eta^2 \sim \text{Beta}(d_\eta, e_\eta)$ and as at the web site level, we reparameterize in terms of the mean $\mu_\eta = d_\eta/(d_\eta + e_\eta)$ and variance $\sigma_\eta^2 = \mu_\eta(1-\mu_\eta)/(d_\eta + e_\eta + 1)$ and set priors $\mu_\eta \equiv \mu_\eta(\mu_\eta, \sigma_\eta^2) \sim \text{Beta}(d_{\mu\eta}, e_{\mu\eta})$ and $\sigma_\eta^2 \equiv \sigma_\eta^2(\mu_\eta, \sigma_\eta^2) \sim \text{Beta}(d_{\sigma\eta}, e_{\sigma\eta})$ and $d_{\mu\eta}$, $e_{\mu\eta}$, $d_{\sigma\eta}$, and $e_{\sigma\eta}$ are fixed hyperparameters. If we are interested in web page covariate effects, we can place a structured prior on web page topic presence. However, web page covariates are likely to be less available than web site covariates, or web page covariates may be the same as web site covariates. The health departments web site data only has web site covariates. Thus we place structured priors at the web site level only and do not consider structured priors for web page topic presence further.

We thus consider six combinations of web site and web page topic presence models denoted by a two letter sequence: AA, EA, SA, AE, EE, SE, with first letter denoting the web site topic presence model, A, E, or S and the second letter denoting the web page topic presence model, A or E. In our model naming, we add local topics to these models and indicate the addition with the addition -LT.

3.3.2 Gibbs Sampling

We describe a Gibbs sampling procedure for the most complicated SE-PFA-LT model. Let a dot '$\bullet$' in subscripts indicate a sum across an index, for example $z_{ij\bullet\bullet}$ is the count of words in web page j of web site i. Let $h = 1, \ldots, z_{ij\bullet\bullet}$ index individual words in web page j of web site i and let $w_{ijh} \in \{1, \ldots, V\}$ and $t_{ijh} \in \{1, \ldots, K+1\}$ be the known word index and latent topic index of the hth word in web page j of web site i. Let $\zeta_{ijkv} = \Phi_{ikv}\theta_{ijk}/(\sum_{k'=1}^{K+1} \Phi_{ik'v}\theta_{ijk'})$ be the probability of topic k in web page j of web site i given word v such that $\sum_{k=1}^{K+1} \zeta_{ijkv} = 1$. Rather than conditionally sample latent counts z_{ijkv}, we sample topic index t_{ijh} for word

w_{ijh} conditional on topic weights θ_{ijk} and topic word probabilities $\Phi_{kw_{ijh}}$

$$t_{ijh}|\Phi_{kw_{ijh}}, w_{ijh}, \sim \text{Multinomial}(\{\zeta_{ij1w_{ijh}}, \ldots, \zeta_{ij(K+1)w_{ijh}}\})$$

for all words in all web pages. Latent counts z_{ijkv} at each iteration of the Gibbs sampler are deterministic functions of the t_{ijh} and w_{ijh}.

Given the z_{ijkv} and other parameters, global topic probability vector ϕ_k, local topic probability vector ψ_i and topic weight θ_{ijk} are conditionally independent and sampling is straightforward due to conjugacy with conditional densities

$$\begin{aligned}
\phi_k|\{z_{\cdot\cdot kv}\}_v &\sim \text{Dirichlet}(\{\alpha_\phi + z_{\cdot\cdot k1}, \ldots, \alpha_\phi + z_{\cdot\cdot kV}\}), \\
\psi_i|\{z_{i\cdot(K+1)v}\}_v &\sim \text{Dirichlet}(\{\alpha_\psi + z_{i\cdot(K+1)1}, \ldots, \alpha_\psi + z_{i\cdot(K+1)V}\}), \\
\theta_{ijk}|r_k, b_{ik} = c_{ijk} = 1, z_{ijk\cdot} &\sim \text{Gamma}(r_k z_{ij\cdot\cdot} + z_{ijk\cdot}, 0.5),
\end{aligned}$$

and $\theta_{ijk} = 0$ if $b_{ik} = 0$ or $c_{ijk} = 0$.

Conjugacy gives a convenient conditional density for the prior web page topic presence probability η_k

$$\eta_k|c_{\cdot\cdot k}, d_\eta, e_\eta \sim \text{Beta}(d_\eta + c_{\cdot\cdot k}, e_\eta + N_\cdot - c_{\cdot\cdot k}).$$

Parameters d_η and e_η are functions of mean μ_η and variance σ_η^2 and we use two Metropolis-Hastings (Hastings, 1970) steps to sample μ_η and σ_η^2. To sample web site topic presence b_{ik} and web page topic presence c_{ijk}, marginalize over θ_{ijk} conditional on $z_{ijk\cdot} = 0$ otherwise if $z_{ijk\cdot} > 0$ then $b_{ik} = c_{ijk} = 1$. When $z_{ijk\cdot} = 0$ sample

$$\begin{aligned}
b_{ik}|\pi_{ik}, r_k, \{c_{ijk}\}_j &\sim \text{Bernoulli}\left(\frac{\pi_{ik}\prod_{j=1}^{N_i}(1-0.5)^{c_{ijk}r_k z_{ij\cdot\cdot}}}{1-\pi_{ik}+\pi_{ik}\prod_{j=1}^{N_i}(1-0.5)^{c_{ijk}r_k z_{ij\cdot\cdot}}}\right), \\
c_{ijk}|\eta_k, b_{ik}, r_k &\sim \text{Bernoulli}\left(\frac{\eta_k(1-0.5)^{b_{ik}r_k z_{ij\cdot\cdot}}}{1-\eta_k+\eta_k(1-0.5)^{b_{ik}r_k z_{ij\cdot\cdot}}}\right).
\end{aligned}$$

Sampling for r_k, r_0 proceeds by introducing two families of non-negative integer-valued auxiliary random variables $\{l_{ijk}\}$ and $\{\ell_k\}$ that are conditionally distributed as the Chinese restaurant table (CRT) distribution. These auxiliary variables ensure conjugacy for sampling r_k and r_0. The CRT has two parameters, z, a non-negative integer, and real valued $r > 0$. Then if $l|z, r \sim \text{CRT}(z, r)$, l has probability mass function

$$P(l = \lambda|z, r) = \frac{\Gamma(r)}{\Gamma(r + z)}|s(z, \lambda)|r^{\lambda},$$

where $s(\bullet, \bullet)$ denotes Stirling numbers of the first kind. Then l can be sampled as a sum of independent Bernoulli random variables, $l = \sum_{m=1}^{z} y_m$, where

$$y_m \sim \text{Bernoulli}\left(\frac{r}{m - 1 + r}\right).$$

Define auxiliary variables $l_{ijk}|z_{ijk\bullet}, r_k, z_{ij\bullet\bullet} \sim \text{CRT}(z_{ijk\bullet}, r_k z_{ij\bullet\bullet})$ and $\ell_k \sim \text{CRT}(\sum_{i=1}^{M} \sum_{j=1}^{N_i} l_{ijk}, r_0)$ distribution. Then conditionally sample r_k and r_0 as

$$r_k|\sum_{i=1}^{M}\sum_{j=1}^{N_i} l_{ijk}, r_0 \sim \text{Gamma}\left(r_0 + \sum_{i=1}^{M}\sum_{j=1}^{N_i} l_{ijk}, \frac{1}{1/e_r - \sum_{i=1}^{M}\sum_{j=1}^{N_i} b_{ik}c_{ijk}z_{ij\bullet\bullet}\ln(1 - 0.5)}\right),$$

$$r_0|\sum_{k=1}^{K+1} \ell_k \sim \text{Gamma}\left(d_{r0} + \sum_{k=1}^{K+1} \ell_k, \frac{1}{1/e_{r0} - \sum_{k=1}^{K+1}\ln(1 - u_k)}\right),$$

$$\text{where } u_k = \frac{-\sum_{i=1}^{M}\sum_{j=1}^{N_i} b_{ik}c_{ijk}z_{ij\bullet\bullet}\ln(1 - 0.5)}{1/e_r - \sum_{i=1}^{M}\sum_{j=1}^{N_i} b_{ik}c_{ijk}z_{ij\bullet\bullet}\ln(1 - 0.5)}.$$

Sampling for β_k conditions on auxiliary Pólya-Gamma (PG) random variables $\{\omega_{ik}\}$. This augmentation step ensures conjugacy for sampling β_k. Let $\omega \sim PG(b, c)$, then we can express ω as an infinite sum of independent Gamma$(b, 1)$ variables g_m, such that

$$\omega \overset{D}{=} \frac{1}{2\pi^2}\sum_{m=1}^{\infty} \frac{g_m}{(m - 1/2)^2 + c^2/(4\pi^2)}.$$

We approximate samples from the Pólya-Gamma distribution as a truncated sum of Gamma variables. Zhao et al. (2017) uses a Pólya-Gamma augmentation step with a truncation level of 20 to sample coefficients in modeling word presence in topics. We find that this truncation level also works well for our topic presence models with structured priors. We introduce auxiliary variable $\omega_{ik}|X_i'\beta_k \sim \text{PG}(1, X_i'\beta_k)$ and conditionally sample β_k by

$$\begin{aligned}
\beta_k|\{\omega_{ik}\}_i, \beta_0, \Sigma &\sim \text{Normal}(\mu_k^*, \Sigma_k^*), \\
\text{where } \Sigma_k^* &= (X'\text{diag}(\{\omega_{ik}\}_i)X + \Sigma^{-1})^{-1}, \\
\mu_k^* &= \Sigma_k^*(X'\kappa_k^* + \Sigma^{-1}\beta_0), \\
\kappa_k^* &= \{b_{1k} - 0.5, \ldots, b_{Mk} - 0.5\}.
\end{aligned}$$

Prior mean coefficient vector β_0 has a conditional Normal posterior distribution and prior precision σ_q^2 has a conditional Gamma posterior distribution

$$\begin{aligned}
\beta_{0q}|\{\beta_{kq}\}_k, &\sim \text{Normal}\left(\frac{(1/\sigma_0)\mu_{0q} + (1/\sigma_q)\sum_{k=1}^K \beta_{kq}}{1/\sigma_0 + K/\sigma_q}, (1/\sigma_0 + K/\sigma_q)^{-2}\right), \\
\sigma_q^2|\{\beta_{kp}\}_k, \mu_{0q} &\sim \text{Gamma}\left(d_\sigma + K/2, (1/e_\sigma + \sum_{k=1}^K (\mu_{0q} - \beta_{kq})^2/2)^{-1}\right).
\end{aligned}$$

3.3.3 Model Evaluation

We randomly select 80% of words in each web page to be our training set and hold out the remaining 20% to evaluate our models. We keep 1000 samples after a burn in of 10,000 samples to calculate perplexity. Let superscript $s = 1, \ldots, S$ index Gibbs samples from the posterior and let $y_{i'j'v}$ be the count of held-out words v in web page j' in web site i'. We define perplexity, the exponentiated log predictive probability, as

$$\text{Perplexity} = \exp\left(-\frac{1}{y_{\cdots}}\sum_{i'=1}^M \sum_{j'=1}^{N_{i'}} y_{i'j'v} \log f_{i'j'v}\right),$$

where

$$f_{i'j'v} = \frac{\sum_{s=1}^{S}\sum_{k=1}^{K}\phi_{kv}^{(s)}\theta_{i'j'k}^{(s)} + \psi_{i'}^{(s)}\theta_{i'j'(K+1)}^{(s)}}{\sum_{s=1}^{S}\sum_{k=1}^{K}\sum_{v=1}^{V}\phi_{kv}^{(s)}\theta_{i'j'k}^{(s)} + \psi_{i'}^{(s)}\theta_{i'j'(K+1)}^{(s)}}$$

is the predicted probability of word v of web page j in web site i. We repeat this random partitioning, MCMC sampling, and perplexity calculation for 5 cross validation sets and average over the 5 perplexity values for a given model.

3.4 Analyzing Web Content of Local Health Department Web Sites

We analyze text data from local health department (LHD) web sites in the United States listed on the National Association of City and County Health Officials directory (National Association of County and City Health Officials, 2012). Only web sites whose web address contain the text string 'health', 'hd', or 'ph' were included. We restrict our analysis to small web sites defined as having at most 100 web pages where each web page has from 50 to at most 1000 words. We do not scrape web pages that are files such as .doc or .pdf files, which are often forms to be filled out. We are more interested in what is intended for people to read while browsing the web. There are 108 LHD web sites that meet this criteria. We scraped websites for textual content using Python and Scrapy in April 2020. We remove text items that occur on nearly every page of a web site, such as titles or navigation menus. 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 defined as words occurring in fewer than 20 web pages, are removed. The dataset analyzed has 1,061,926 total words with $V = 3{,}544$ unique words across 5,863 web pages.

We include a web site region covariate that indicates whether a LHD is from a state in the Northeast, South, Midwest, or West. There are fewer than 10 LHD in either Northeast (8)

and West (5) regions, and therefore we combined them into a new Northeast/West region. There are $Q = 3$ web site level covariates. There are 70 web sites from LHD in Midwest states, 25 web sites from LHDs in Southern states, and 13 web sites from LHDs in either Western or Northeastern states. We set $X_i = \{1, 0, 0\}$ to indicate that web site i is from the Midwest region. Similarly, $X_i = \{0, 1, 0\}$ and $X_i = \{0, 0, 1\}$ indicates web site i is from the South or Northeast/West region. Coefficients β_{k1}, β_{k2}, and β_{k3} correspond to intercepts for the Midwest, South, and West regions respectively. Given this specification for covariates, we are interested in the differences between regions or $\beta_{k1} - \beta_{k2}$, $\beta_{k1} - \beta_{k3}$, and $\beta_{k2} - \beta_{k3}$ for global topics k.

3.4.1 Prior Specifications

We model web pages nested in local health department web sites with 5 topic presence models, EA-PFA-LT, AE-PFA-LT, EE-PFA-LT, SA-PFA-LT, and SE-PFA-LT and compare it to a reference AA-PFA-LT model where topics are always present at both web page and web site levels. We use the same hyperparameters in all models. We choose prior for shape parameter r_0 such that $d_{r0} = 0.01$ and $e_{r0} = 1/.01$. We choose priors for topic word probability vectors ϕ_k and ψ_k such that $\alpha_\phi = 0.05$ and $\alpha_\phi = 0.05$ to encourage topics to place small probability on most words and large probability on a few words. We set coefficient hyperparameters $\mu_0 = (0, 0, 0)'$, $\sigma_0 = 0.5$, $d_\sigma = 1$, and $e_\sigma = 1$ in centering the prior at the prior belief that there are no region effects and picking a prior variance that supports that a typical global topic is neither present in nearly all web sites nor unique to one web site but rather somewhere in between. This is reflected in our specifications for the exchangeable prior on web site topic presence in EA-TPM-LT. We specify a prior Beta($d_{\mu\pi} = 10$, $e_{\mu\pi} = 10$) prior on the prior mean of global web site topic probability π_k and we specify a prior Beta($d_{\sigma\pi} = 1$, $e_{\sigma\pi} = 5$) prior on the prior variance of global web site topic probability π_k. We set hyperparameters for page topic presence probability η_k such that $d_{\mu\eta} = 1$, $e_{\mu\eta} = K - 1$, $d_{\sigma\eta} = 1$, $e_{\sigma\eta} = K - 1$ in our analysis as we expect most web

pages to have one or a few topics present.

3.4.2 Model Comparisons

We compare our models at $K = 25, 50, 100, 200, 300, 400, 500, 600$. We further compare SA-PFA-LT and SE-PFA-LT with their no local topic counterparts SA-PFA and SE-PFA. Figure 3.1 plots the average held-out perplexity at different number of global topics K for all six models with local topics. All models perform similarly with AE-PFA-LT performing slightly worse overall. Perplexity of all models continue to improve at $K = 600$ however the difference between perplexity at $K = 500$ and $K = 600$ is less than 2. Further increasing K increases computation time and may only improve the fit slightly. We model our full data with $K = 500$ global topics in our analysis. Figure 3.2 compares perplexity between SA-PFA-LT and SE-PFA-LT against their no local topic counterparts, SA-PFA and SE-PFA. Models without local topics require more global topics to perform as well as models with local topics. The four models begin to perform similarly at $K = 400$, where all models begin to show little perplexity improvement for each 100 increase in K. We model our data with local topics in the analysis as we do not want to model covariate effects of local topics.

3.4.3 Analysis of Regional Effects

We consider the regional effects modeled with SA-PFA-LT as we are mainly interested in web site topic presence and do not want to model covariate effects of local topics. Table 3.1 shows the 5 most probable words in 10 local topics. Nearly all local topics include geographical names among the 5 most probable words. Other high probability words in local topics are those that occur in news bulletins or other text that appears in multiple web pages of a web site. We choose a subset of global topics from the $K = 500$ global topics to review. The topics in the subset must meet three criteria, significance, frequency, and being a health topic. First, we are interested in whether topic presence differs between regions or whether

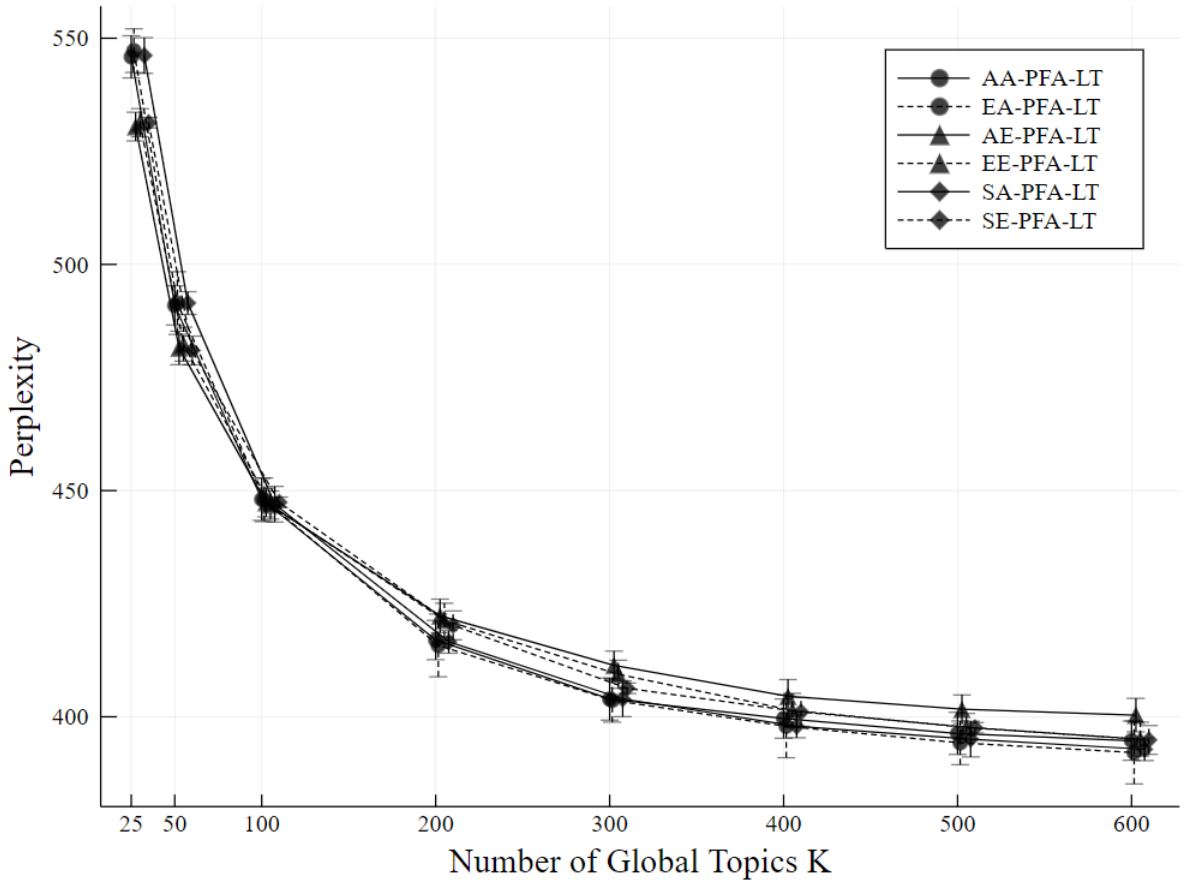

Figure 3.1: Perplexity by number of global topics K comparison between models (lower is better).

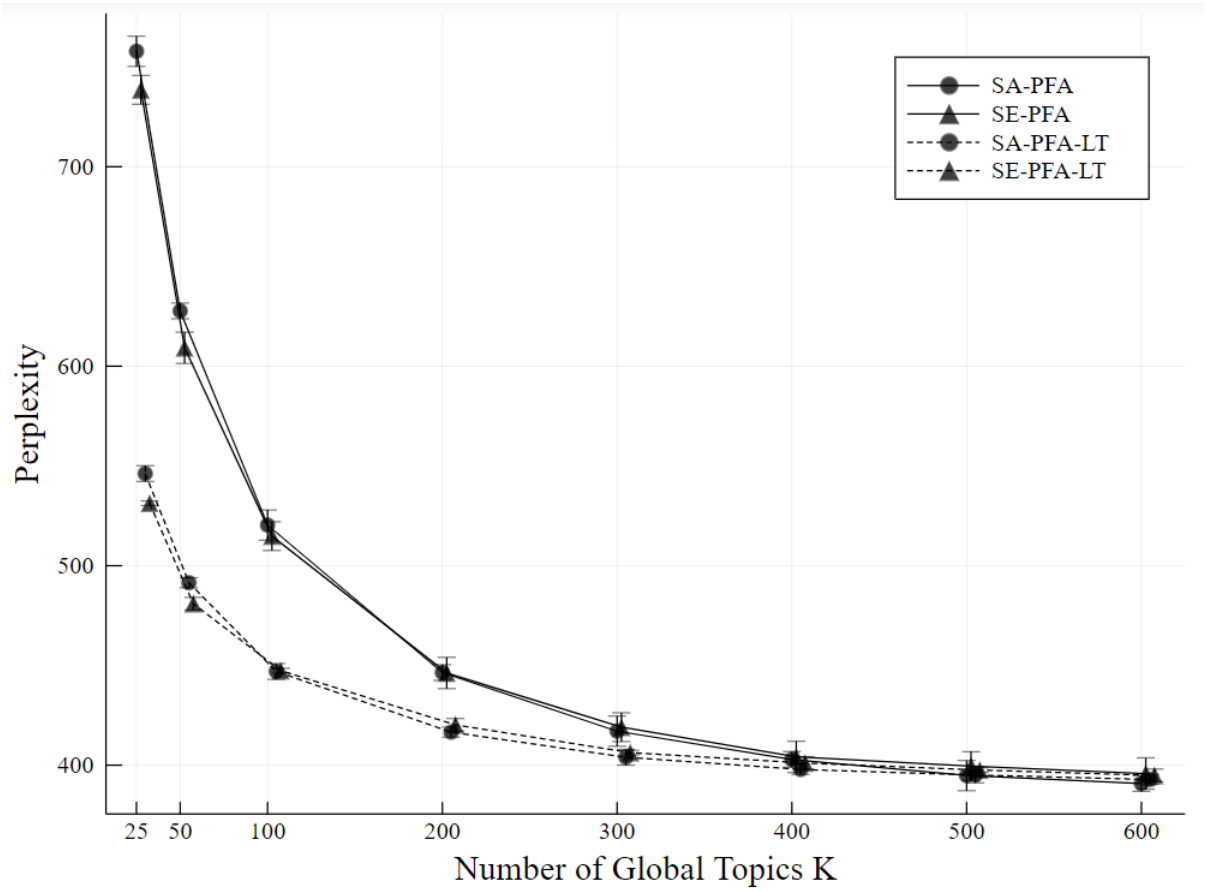

Figure 3.2: Perplexity comparison between topic presence models with a structured prior on web site topic presence.

Table 3.1: Five most probable words of 10 local topics from SA-PFA-LT with $K = 500$ global topics. Most local topics include a geographical name or word among its top five words. Multi-county*: Logan, Morgan, Phillips, Sedgwick, Washington, and Yuma counties.

County	State	Region	Top 5 Words
Taylor	Florida	South	*florida, taylor, program, environment, link*
Wakulla	Florida	South	*wakulla, water, florida, control, mosquito*
Effingham	Illinois	Midwest	*effingham, illinoi, test, idph, new*
Livingston	Illinois	Midwest	*livingston, news, covid, current, comment*
Vermilion	Illinois	Midwest	*vermilion, illinoi, cdc, resourc, click*
Shannon	Missouri	Midwest	*inspect, shannon, food, center, emin*
Hocking	Ohio	Midwest	*hock, ohio, program, map, safeti*
Noble	Ohio	Midwest	*nobl, ohio, provid, resourc, respons*
La Paz	Arizona	West	*paz, vaccin, arizona, comment, dose*
Multi-county*	Colorado	West	*colorado, nchd, northeast, nchdorg, morgan*

$\beta_{kq} - \beta_{kq'}$ is significantly positive or negative for global topics k and separate regions q and q'. Differences are significant when the 95% sampling interval is all positive or all negative. Second, The topic must be present in at least 20 web sites and present in at most 88 of $M = 108$ web site. Third, the topic must be health related. There are 101 topics that meet the significance criteria, 75 topics that further meet the frequency criteria, and 45 topics that meet all three criteria. We select 5 topics to review. We label them in Table 3.2 and show their 10 most probable words and the posterior mean (95% posterior interval) of their total web site presence $b_{\cdot k}$.

We carefully label each topic to avoid confusion when two topics are similar. Similar or related topics may share common most probable words. There were no topics that shared a similar set of most probable words with the tickborne diseases topic or the foodborn illness topic. There were two topics that are related to the CDC guidance topic; a general CDC topic with most probable words *prevent, diseas, control, center, cdc, protect, reduc, accord, main, measur* and a CDC web links topic with most probable words *http, wwwcdcgov, index-html, pdf, link, htm, indexhtm, indexphp, ncov, imag*. The WIC nutrition and breastfeeding topics are similar in that both are related to childcare. However, the WIC nutrition topic

Table 3.2: The 10 most probable words of 5 global health topics from SA-PFA-LT with $K = 500$ global topics. Abbreviations used are Center for Disease and Control Prevention (CDC) and Special Supplemental Nutrition Program for Women, Infants, and Children (WIC). The third column lists posterior mean and 95% posterior interval of the percentage of web sites with the corresponding topic present.

Topic	Top 10 Words	Total Presence
Tickborne diseases	*tick, diseas, lyme, bite, remov, deer, tickborn, skin, transmit, attach*	47(42,54)
Foodborn illness	*ill, foodborn, noroviru, outbreak, vomit, guidelin, suspect, diarrhea, contamin, clean*	50(42,58)
CDC guidance	*cdc, guidanc, recommend, updat, healthcar, guidelin, faq, advisori, disinfect, worker*	73(69,77)
WIC nutrition	*wic, infant, nutrit, women, breastfeed, children, elig, food, pregnant, incom*	85(83,86)
Breastfeeding	*breastfeed, mother, breast, support, peer, babi, counselor, milk, pump, mom*	73(69,79)

is specifically about the WIC nutrition program while the breastfeeding topic is specifcally about breastfeeding. There is a third related mother/pregnant women topic with most probable words *women, pregnanc, pregnant, infant, prenat, birth, matern, babi, mother, outcom*. The pregnancy topic does not have most probable words for nutrition or breastfeeding.

Table 3.3 summarizes the covariate effects in these 5 topics. The estimates are averaged over MCMC samples and intervals are 95% MCMC intervals. The tickborne diseases topic has most probable words *tick, diseas, lyme, bite, remov, deer, tickborn, skin, transmit, attach* and is present in LHD web sites in the Midwest and West/Northeast more often than they are in LHD web sites in the South. The foodborn illness topic has most probable words *ill, foodborn, noroviru, outbreak, vomit, guidelin, suspect, diarrhea, contamin, clean* and is present in LHD web sites in the Midwest and West/Northeast more often than they are in LHD web sites in the South. The CDC guidance topic has most probable words *cdc, guidanc, recommend, updat, healthcar, guidelin, faq, advisori, disinfect, worker* and is present in LHD web sites in the West/Northeast more often than they are in LHD web sites in the South. However, the difference is borderline significant with a 95% interval of (-3.44,-0.05) compar-

Table 3.3: Summary of regional differences $\hat{\beta}_{kq} - \hat{\beta}_{kq'}$ on the logit scale. Estimates are averages over 1,000 MCMC samples saving every 10th sample and after a burn-in of 25,000 samples. An * indicates covariate effect is significant (one sided) at significance level 0.025. The MW-S column indicates the regional difference between Midwest and South. The MW-W/NE column indicates the regional difference between Midwest and West/Northeast. The S-W/NE column indicates the regional difference between South and West/Northeast.

Topic	MW-S	MW-W/NE	S-W/NE
Tickborne diseases	2.29(0.93,3.83)*	-0.30(-1.57,0.91)	-2.59(-4.38,-0.72)*
Foodborn illness	2.17(0.74,3.86)*	0.01(-1.48,1.31)	-2.16(-4.30,-0.38)*
CDC guidance	0.84(-0.17,1.85)	-0.79(-2.48,0.60)	-1.63(-3.44,-0.05)*
WIC nutrition	0.47(-0.66,1.48)	2.18(0.94,3.47)*	1.71(0.30,3.23)*
Breastfeeding	0.52(-0.68,1.65)	1.43(0.15,2.68)*	0.91(-0.53,2.56)

ing South to West/Northeast. The Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) nutrition topic has most probable words *wic, infant, nutrit, women, breastfeed, children, elig, food, pregnant, incom* and is present in LHD web sites in the Midwest and South more often than they are in LHD web sites in the West/Northeast. The breastfeeding topic has most probable words *breastfeed, mother, support, breast, babi, peer, counselor, milk, wic, pump* and is present in LHD web sites in the Midwest more often than they are in LHD web sites in the West/Northeast.

SA-PFA-LT models how covariates are associated with web site topic presence. We want to check if our model correctly captures these web site topic presence patterns. However, doing so manually by reading through all web pages and web sites is time consuming, thus, we describe a quick automated approach to checking using the V^* most probable words in a topic. For topic k we check the portion of web sites in a region with at least one web page with all $V^{(*)}$ most probable word. For example, for $V^* = 2$, the tickborne diseases topic is present in web site i if at least one page in web site i contains both words *tick* and *lyme*. Many topics can be described by a few most probable words. Thus, we let V^* be the number of words with probability greater than 0.1 in a given topic. We confirm that the 1 or 2 most probable words among the 5 health topics we further analyze are not identical to that of any

Table 3.4: Counts and portions of web sites with at least one web page containing the V^* probable words in a topic, where V^* is the number of words in a topic with probability greater than 0.1.

Topic	Midwest (N = 70)	South (N = 25)	West/Northeast (N = 13)	V^*
Tickborne diseases	35(50.0%)	11(44.0%)	7(53.8%)	1
Foodborn illness	42(60.0%)	10(40.0%)	9(69.2%)	2
CDC guidance	65(92.9%)	21(84.0%)	13(100.0%)	1
WIC nutrition	62(88.6%)	19(76.0%)	6(46.2%)	1
Breastfeeding	48(68.6%)	12(48.0%)	6(46.2%)	2

other topic. Table 3.4 shows the counts and percentages of web sites containing at least one page with V^* most probable words in each region. Our model indicates that the tickborne diseases topic is more prevalent in the Midwest and West/Northeast than in the South. We see the same pattern in Table 3.4, where 50% (35/70) of web sites in the Midwest and 53.8% (7/13) of web sites in the West/Northeast have at least one page with the word *tick* while 44.0% (11/25) of web sites in the South have at least one page with the word *tick*. Similarly, for the other four topics, our logistic model results reflect the quick automated check results.

3.4.4 Analysis of Tickborne Diseases Topic

The regression results from SA-PFA-LT indicate that the tickborne disease topic is more prevalent in the Midwest and West/Northeast than in the South. This is supported by our quick automated check and further supported in a 2018 CDC report of vectorborne diseases (Rosenberg et al., 2018). The report showed that from 2004-2016 the states with the top quintile of reported cases of tickborne disease are from the Midwest and Northeast. We further look into the model results for the tickborne disease topic and identify web sites that are missing the topic. More formally, we search for web sites i where 97.5% or more of MCMC samples of $b_{ik}^{(s)} = 0$. This approach finds 24 web sites missing the tickborne disease topic; 9 from the Midwest, 12 from the South, and 3 from the West/Northeast.

We further check individual web sites from the 3 West/Northeast web sites. These three web sites belong to the La Paz County Health Department in Arizona, the Cambridge Public Health Department in Massachusetts, and the Weber-Morgan Health Department in Utah. In the web pages we collected for these three web sites, we found no web pages with the word *tick*. The La Paz County Health Department and Weber-Morgan Health Department are from the West region where tickborne disease is not as prevalent as in the Northeast. Upon closer inspection, we found no current online web pages from La Paz County Health Department's web site related to tickborne diseases. However, we did find a web page related to mosquitos and the Zika Virus. We found two PDF links on Weber-Morgan Health Department's web site with the word *tick*. One is a pet disaster kit checklist, and the other is a large list of reportable diseases in Utah. These pages were not collected as they are PDF files. There was no dedicated informational page on tickborne diseases on Cambridge Public Health Department's web site; however, we found one news article about inviting residents to participate in a tick monitoring project. At the time of web scraping, this web page was not available to scrape. We were not able to find an archive of the news article around the date of scraping in April 2020. Given the data we collected and modeled, SA-PFA-LT correctly identified these web sites as not having the tickborne disease topic present.

3.5 Discussion

We proposed novel topic presence models with local topics to model topic presence at two different levels in a nested document collection and apply our work to a collection of web pages nested in small web sites from local health departments in the U.S. We discussed three priors that can be placed on topic presence probabilities at web sites or web pages and showed that all topic presence models perform similarly. Thus, there is no sacrifice in fit when topic presence modeling is desired.

Our AE-PFA model is similar to the sparse Gamma-Gamma PFA. However, our AE

model uses an exchangeable prior on web page topic presence probabilities rather than an independent prior where the η_ks are known a priori. Also, we include a scaling factor of $z_{ij\bullet}$ in equation (3) while other models under the PFA construction do not; including the scaling factor adjusts for different word counts in different documents.

SA-PFA-LT and SE-PFA-LT model web site topic presence probabilities conditional on web site covariates. We modeled the full data set with SA-PFA-LT to make inference on health topics and inference on regional effects on web site topic presence. Among 500 possible topics we found many health topics where there were significant regional effects and further reviewed 5 health topics. We found that it is important to carefully label topics as some topics are related. After checking for related topics and distinguishing between them in labeling, we made inferences on which regions were more likely than others to have one of the health topics present. We went further and checked several web sites that were missing the tickborne disease topic. Our model correctly identified three web sites in the West/Northeast that were missing the topic. Although, one of the three web sites did have a web page related to tick monitoring news, it was not available at the time of web scraping. Our analysis is limited to what is available online at the time. The limitation is highlighted when making inference on topic presence in a specific web site, while making inference on regional patterns allows us to leverage data from multiple web sites.

CHAPTER 4

Predictors of Health Topic Coverage in U.S. County Health Department Web Sites

4.1 Introduction

Developing health priorities for local health departments is complex, involving developing a mission, vision, and values in collaboration with various stakeholders (National Association of County and City Health Officials, 2021). Local health department web sites provide health information to local communities and cover a range of health topics that reflect the health priorities of a local health department. We are interested in understanding whether sociodemographic variables correlated with the prevalence of health conditions are also correlated with how topics are covered in county health department web sites. In particular, we are interested in identifying human immunodeficiency virus (HIV) related topics and opioid use disorder (OUD) related topics and we are interested in identifying sociodemographic predictors of their coverage in U.S. county health web sites.

We investigate HIV and OUD topics as HIV and OUD are some of the leading causes of deaths in the United States. An estimated 1.2 million people in the U.S. are living with HIV where there were about about 36,400 (13.3 per 100,000) estimated new HIV infections in 2018. Though the rate of new infections have remained stable compared to 2014, HIV remains one of the leading causes of deaths among 15-64 year olds in the United States (Centers for Disease Control and Prevention, 2021). Opioid overdose deaths in the U.S. have increased from 2014 to 2019 (National Institute on Drug Abuse, 2021). Drug overdose is

the leading cause of injury-related death in the United States with over 70% of those deaths involving an opioid (National Institute on Drug Abuse, 2021).

New HIV rates are higher among African Americans than any other race/ethnicity (Centers for Disease Control and Prevention, 2020a). Thus, we anticipate that web sites would cover HIV topics more often in counties that have a higher proportion of African American/black people.

OUD is more common in populations under 65 years of age (Hedegaard et al., 2018). It is also more common among white populations; though, in recent years black populations in large metro areas have had increasing numbers of opioid overdose death rates (Lippold et al., 2019). Thus, we might expect web sites to cover OUD topics more often in counties that have a higher proportion of people under 65 years of age and counties that have a higher proportion of white people.

While some differences in topic coverage is expected, HIV and OUD remain important health issues across the United States. Topics related to either should be covered across all county health web sites regardless of county demographic.

No previous publications have studied demographic predictors of health topic coverage in county health web sites. We develop methods to survey U.S. county health web sites and identify predictors of health topic coverage.

4.2 Methods

We identified 196 county health departments with their own dedicated web site through the National Association of City and County Health Officials' (NACCHO) local health department directory (National Association of County and City Health Officials, 2012). We developed software using Python and Scrapy to collect text data from health department web sites by crawling all web pages of a web site and scraping text from web pages (van Rossum, 1995; ScrapingHub, 2018). Specifically, we scraped text that was tagged as a para-

graph text in HyperText Markup Language (HTML). We identified forms, e.g. .doc or .pdf files, and calendar and event web pages to avoid scraping. We acquired county population data from the 2019 American Community Survey (ACS) 5-year results (U.S. Census Bureau, 2020). We summed count variables over counties for multi-county health departments. We collected total population, total number of families, count of non-Hispanic black population, count of Hispanic population of any race, population 25 years and older, count of 25 years or older population with a high school education, count of families below federal poverty line, and count of population 65 years and older.

We developed software to further process scraped web pages and words in web pages. We processed words by lower-casing and stemming all words, e.g. *diseased* and *diseases* are reduced to *disease*. We removed duplicate web pages and Spanish translations of web pages. We removed paragraph text occurring in over 25% of all web pages in a web site. We removed frequent words occurring in over 50% of web pages and further removed non-health related words occurring in over 10% of web pages. We removed rare words occurring in fewer than 50 web pages or fewer than 10 web sites. We removed state names, calendar days and months, and words of one or two characters in length. We removed words that were not in NLTK's English dictionary (Bird et al., 2009) or the Medical Subject Headings (MeSH) thesaurus (National Library of Medicine, 2021). After processing web pages and words, we removed web pages with fewer than 5 words and web sites with fewer than 10 web pages.

We transformed the ACS variables to log of total population (log population), percent of population that is non-Hispanic black (% black), percent of population that is Hispanic of any race (% Hispanic), percent of 25 years and older population that have a high school education or higher (% HS grad), percent of families below the federal poverty line (% poverty), and percent of the population 65 years and older (% over 65). We standardize predictors by subtracting the mean across all departments and dividing by the standard deviation. We chose total population as we expect more populated counties to cover more health topics on their web sites. We chose variables of racial composition, educational at-

tainment, poverty status, and age as they are common predictors of health outcomes. HIV disproportionately affects African American/black people and OUD disproportionately affects younger and white people. People living in poverty have limited access to healthcare and cannot afford some basic necessities that may lead to circumstances that increase risk for health conditions such as HIV (Centers for Disease Control and Prevention, 2016). Counties with worse economic prospects are also more likely to have higher rates of opioid prescriptions, opioid-related hospitalizations, and drug overdose deaths (Ghertner and Groves, 2018). There is less research focused on educational attainment and its relationship and impact on HIV and OUD, however, we are interested in studying education attainment's association with web site coverage of HIV and OUD topics.

We used the bag-of-words (BoW) representation of words to model text data. The BoW representation is a simplifying representation that disregards word order and only retains information about word co-occurrence. For example, the phrase "the words in the bag" can be represented as *the, words, in, the, bag* which is the same as *bag, in, the, the, words*. We just know that these words occur together. A topic model asserts that documents are mixtures of latent topics, represented by probability distributions over topics. Topics are probability distributions on words. Topic models are used to abstract word probabilities of latent topics. Higher probability words better describe a topic compared to lower probability words and higher probability words are often words that regularly appear on web pages. We can then label a topic by inspecting its highest probability words. For example, a topic with the five most probable words *opioid, overdos, naloxon, drug, addict* can be labeled an "opioid use disorder" topic. We identified HIV and OUD related topics by checking for whether *hiv* or *opioid* was among the ten most probable words in each topic. We labeled all topics appropriately according to their 10 most probable words.

A web site covers a topic if that topic is present in the web site. Website topic presence is a variable that indicates whether a topic is present in a web site. If a web site does not have a particular topic present then words in all web pages of the web site do not

come from that topic. We modeled our data using a hierarchical topic presence model (HTPM). An HTPM models the presence of topics in web sites. The model uses a logistic regression to model the presence of each latent topic in each web site as a function of county demographic predictors. We placed a prior distribution on regression coefficients such that the regression coefficient for a particular predictor for any topic is normally distributed with an unknown global mean and standard deviation. Thus, we can make inferences on overall effects of predictors on web site topic coverage as well as inference for individual topics. We investigated regression results for HIV and OUD related topics. The HTPM is a Bayesian model and we implemented the model using Markov chain Monte Carlo (MCMC). We report posterior medians of odds ratios (OR) and 95% posterior intervals (PI). We report one-sided p-values (p) as the posterior probability that the coefficient is positive or negative, which ever is smaller.

A key tuning parameter of HTPM is choosing the number of topics. This choice determines how many topics or different word distributions the model will identify. We chose the number of topics by (1) checking and comparing predictive performance of the model at $10, 25, 50, 100, 200, 300$, and 400 topics and (2) examining the 10 most probable words in health topics for each model, as most topics can be described by their few most probable words. We used 80% of words in each web page to model the data and held out 20% of words to calculate predictive performance as the average log predictive probability of held-out words across all web pages. We determined reasonable values for the number of topics by identifying where predictive performance improvements from increasing the number of topics begin to slow. We further inspected the most probable words of health topics in the model we identified as having relatively good predictive performance. The full model with complete prior specification and computational details is given in the web appendix.

4.3 Results

The flow chart in Figure 4.1 summarizes the steps of processing web pages and text to create the dataset used in our analysis. There are 3604 unique words and 2,871,940 total words across 23,570 web pages nested in 193 web sites after cleaning and processing the text data. Table 4.1 summarizes word and web page counts, the unstandardized county demographic predictors, and untransformed county population.

Table 4.1: Summary of county demographic variables, web page counts, and word counts.

	Single county N = 171		Multi-county N = 22		All N = 193	
Variable	Mean	SD	Mean	SD	Mean	SD
Population(1000s)	239.3	873.8	151.1	241.7	229.2	826.6
% Black	8.6	12.2	5.6	11.1	8.3	12.1
% Hispanic	8.6	11.6	6.1	5.6	8.3	11.1
% HS grad	35.1	7.7	32.9	5.2	34.8	75.0
% Poverty	10.7	4.6	9.0	3.4	10.5	4.5
% Seniors	19.0	5.4	20.2	4.3	19.1	5.3
Number of web pages	122.2	153.8	121.9	109.1	122.1	149.4
Words per web page	124.1	133.2	104.4	123.2	121.8	132.2

A.Figure 1 in the web appendix shows that predictive performance improvements from increasing the number of topicsare minimal after 200 topics while presence of unhelpful topics increases in models with over 300 topics. We continued our analysis using HTPM with 300 topics. We visually inspected the 10 most probable words of topics and found numerous health topics with coherent sets of 10 most probable words. A full list of health topics and their 5 most probable words is given in the web appendix.

We found five topics where *hiv* was one of the 10 most probable words and two topics with *opioid* as one of the 10 most probable words. Table 4.2 shows the 10 most probable words in all seven topics and the posterior median percent (95% interval) of web sites that cover the topic. The infectious disease surveillance topic is covered by 78% (75%,82%) of

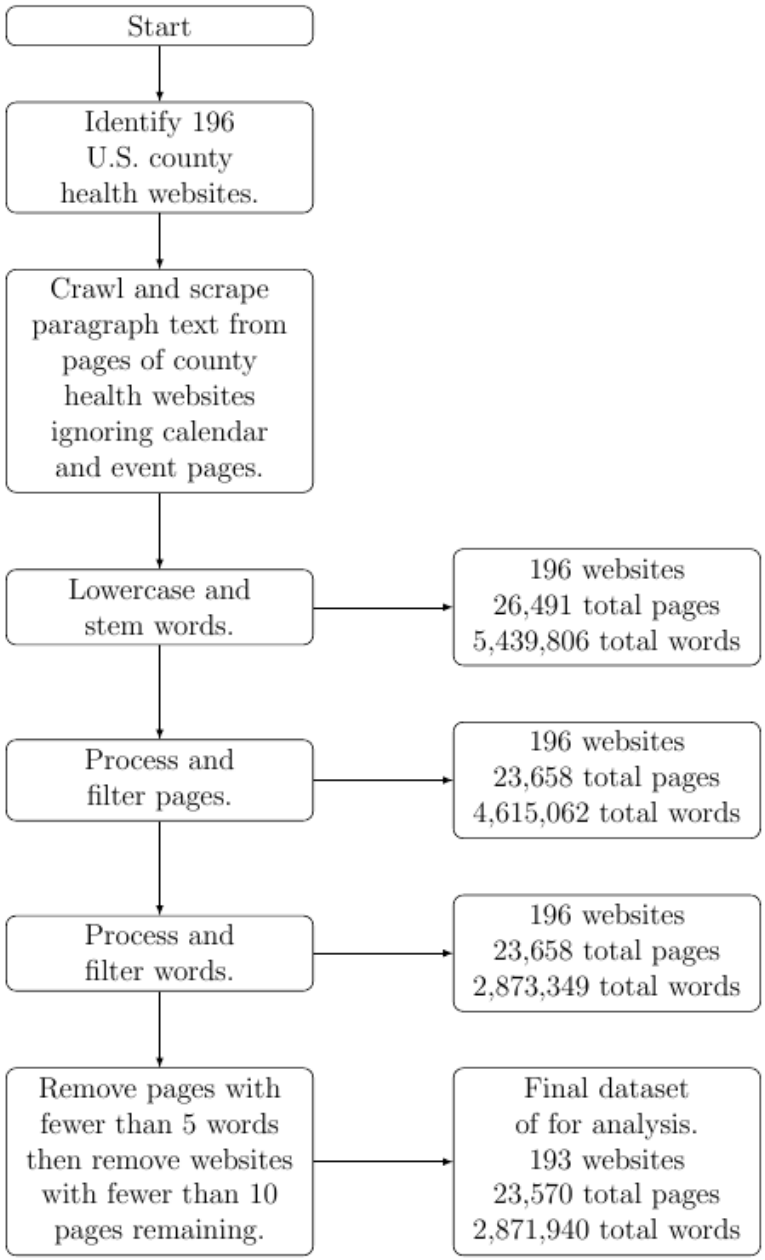

Figure 4.1: Flowchart of text data collection and processing to construct the dataset for analysis.

web sites, including most probable words *tuberculosi* and *hepat*, stemmed from "tuberculosis" and "hepatitis". Unlike the infectious disease surveillance topic, the HIV statistics disease surveillance topic is more specific to HIV as HIV is the only disease among the 10 most probable words. The HIV statistics surveillance topic is covered by 40% (35%,48%) of web sites. The general sexually transmitted infections (STI) topic is covered by 87% (82%,92%) of web sites and it is not specific to HIV. It includes another STIs, *syphili*, among its most probable words and it does not include *aid* among its most probable words. The HIV care and prevention topic is covered by 74% (69%,78%) of web sites and it is specific to HIV with its two most probable words being *hiv* and *aid*. This topic also has the most words specific to HIV with *hiv, aid, prep,* and *prophylaxi*. Pre-exposure prophylaxis (PrEP) is a preventative measure against HIV (Centers for Disease Control and Prevention, 2020b) and *prep* and *prophylaxi* are the processed form of *PrEP* and *prophylaxis* respectively. Neither *prep* nor *prophylaxi* are found in the 10 most probable words of any other topic. The STI screening and treatment services topic is covered by 97% (94%,98%) of web sites, more than any other HIV related topic. The OUD topic has *opioid* as its most probable word and *naloxon* as its third most probable word, while no other topic has *naloxon* in its 10 most probable words. Naloxone is stemmed to *naloxon* and is a treatment for opioid overdose (National Institute on Drug Abuse, 2020). The OUD topic is covered by 38% (33%,44%) of web sites. The substance use disorder topic is not specific to OUD with *opioid* being the 10th most probable word. Other drugs *alcohol* and *marijuana* have a higher word probability than *opioid* in the topic. The substance use disorder topic is covered by 70% (66%,75%) of web sites.

Table 4.3 and Table 4.4 report the posterior median odds ratio estimates and 95% intervals for the demographic predictors for the seven HIV or OUD topics. Log population is significant and positively associated with the coverage of all but the STI screening and treatment services topic, which is already covered in nearly all web sites. Overall, counties with higher log population and percent black cover more topics and counties with higher

Table 4.2: The 10 most probable words in order of probability for HIV related topics and OUD related topics and with median percent coverage (95% PI) across all web sites.

Topic	Top 10 Words	Percent Coverage
Infectious Disease Surveillance	*diseas, epidemiolog, infecti,tuberculosi, surveil, hiv, aid, prevent, communic, hepat*	78(75,82)
Disease Surveillance (HIV statistics)	*diseas, surveil, hiv, monthli, aid, prevent, archiv, statist, morbid, infecti*	40(35,48)
Sexually Transmitted Infections (general)	*hiv, sexual, test, transmit, treatment, sex, partner, syphili, infect, condom*	87(82,92)
HIV Care and Prevention	*hiv, aid, test, prep, prevent, treatment, prophylaxi, care, drug, statu*	74(69,78)
STI Screening and Treatment Services	*clinic, immun, famili, diseas, test, hiv, sexual, care, nurs, educ*	97(94,98)
Opioid Use Disorder	*opioid, overdos, naloxon, drug, addict, pain, prescript, treatment, prescrib, mat*	38(33,44)
Substance Abuse	*drug, abus, substanc, alcohol, prevent,prescript, marijuana, medic, treatment, opioid*	70(66,75)

percent Hispanic, percent HS grad, percent poverty, and percent over 65 cover fewer topics.

Percent black is positively associated with the coverage of two HIV topics, infectious disease surveillance and HIV care and prevention. A one standard deviation increase in percent black is associated with a 50% ($p = 0.001$) increase in odds of covering infectious disease surveillance and a 54% ($p = 0.002$) increase in odds of covering HIV care and prevention in county health department web sites. Percent over 65 is negatively associated with the coverage of the OUD topic and a one standard deviation increase in percent over 65 is associated with a 28% ($p = 0.012$) decrease in odds of a web site covering OUD. Percent HS grad is negatively associated with the coverage of the substance use disorder topic, where a one standard deviation increase in percent HS grad is associated with a 34% ($p = 0.004$) decrease in odds of a web site covering substance use disorder. No other demographic predictor is significantly associated with the coverage of HIV and OUD related topics topics.

There are no significant demographic differences in the coverage of the HIV statistics disease surveillance, general sexually transmitted infections (STI), and STI screening and treatment services topics. Both the general sexually transmitted infections topic and the STI screening and treatment services topic are relatively common and covered by most web sites. The HIV statistics disease surveillance topic is relatively uncommon and covered by fewer web sites than the other HIV related topics.

4.4 Discussion

We used a hierarchical topic presence model to discover latent topics and inspected the most probable words of each topic to label topics and identify topics related to HIV and OUD. We identified five topics related to HIV and two topics related to OUD. By inspection of the other most probable words, we found the HIV care and prevention topic and the OUD topic to be most relevant to HIV and OUD respectively.

We expected log population overall to be positively associated with most topics. The

Table 4.3: Estimates and 95% intervals of odds ratio for predictors log population, % black, and % Hispanic for the seven topics related to HIV or OUD. Predictors were standardized to mean zero and standard deviation one before modeling

Topic	Log population	% Black	% Hispanic
Infectious Disease Surveillance	1.51* (1.20,1.99)	1.50* (1.08,2.22)	1.18 (0.85,1.7)
Disease Surveillance (HIV statistics)	1.46* (1.16,1.9)	1.11 (0.81,1.45)	0.99 (0.73,1.37)
Sexually Transmitted Infections (general)	1.43* (1.07,1.88)	1.40 (0.99,2.14)	1.08 (0.78,1.67)
HIV Care and Prevention	1.42* (1.11,1.91)	1.54* (1.09,2.30)	1.12 (0.82,1.59)
STI Screening and Treatment Services	1.34 (0.99,1.74)	1.15 (0.73,1.77)	0.99 (0.67,1.51)
Opioid Use Disorder	1.32* (1.03,1.70)	1.01 (0.76,1.36)	0.76 (0.56,1.02)
Substance Abuse	1.34* (1.03,1.74)	0.97 (0.72,1.32)	0.96 (0.71,1.35)
Overall (across all topics)	1.35* (1.29, 1.42)	1.07* (1.02, 1.12)	0.91* (0.87, 0.95)

Table 4.4: Estimates and 95% intervals of odds ratio for predictors % HS grad, % poverty, and % over 65 for the seven topics related to HIV or OUD. Predictors were standardized to mean zero and standard deviation one before modeling

Topic	% HS grad	% Poverty	% Over 65
Infectious Disease Surveillance	1.00 (0.72,1.37)	1.10 (0.83,1.49)	1.06 (0.8,1.49)
Disease Surveillance (HIV statistics)	0.78 (0.56,1.10)	0.88 (0.66,1.15)	0.85 (0.63,1.15)
Sexually Transmitted Infections (general)	0.82 (0.58,1.18)	1.18 (0.89,1.63)	1.08 (0.77,1.55)
HIV Care and Prevention	0.84 (0.61,1.14)	1.06 (0.81,1.42)	1.03 (0.77,1.40)
STI Screening and Treatment Services	0.83 (0.57,1.20)	1.00 (0.71,1.37)	0.95 (0.69,1.33)
Opioid Use Disorder	0.82 (0.62,1.13)	0.97 (0.74,1.24)	0.72* (0.53,0.94)
Substance Abuse	0.66* (0.47,0.90)	0.94 (0.73,1.24)	0.81 (0.62,1.08)
Overall (across all topics)	0.82* (0.78, 0.86)	0.95* (0.9, 0.99)	0.92* (0.89, 0.96)

overall effect and the individual effects for each topic confirmed this. We found percent black was a significant predictor positively associated with coverage of HIV care and prevention. African American/black populations are disproportionately affected by HIV and have the highest rate of new infections in 2018 with 45.4 per 100,000 people followed by Hispanic/Latinos with 22.4 (Centers for Disease Control and Prevention, 2020a). Thus, it is reasonable that counties with higher percent black populations cover HIV related topics on their web sites more often than counties with smaller percentages. Hispanic populations also have a high rate of new HIV infections; however the rate is not nearly as high as black populations and percent Hispanic was not a significant predictor for HIV topics. There is high overall coverage of STI screening and treatment services and general sexually transmitted infections topics and no significant associations between coverage and race, education, poverty, or age. Most county health web sites presumably cover STI testing and treatment generally and coverage is not associated with demographic predictors outside log population.

The infectious disease surveillance topic has other diseases among its most probable words. The word *tuberculosi* has the highest probability among the three diseases *tuberculosi, hiv/aid,* and *hep*. Like HIV, Tuberculosis (TB) also disproportionately affects African American/black populations (Deutsch-Feldman et al., 2021). The difference in how infectious diseases affect black populations is reflected in county health web site coverage of infectious disease surveillance. Interestingly, HIV statistics disease surveillance shares most probable words with both the infectious disease surveillance and HIV/AIDS care and prevention topics; however coverage of the topic is not significantly associated with percent black. The topic is covered by fewer web sites overall and its coverage is significantly associated with log population. Most web sites covering HIV statistics disease surveillance belong to the largest counties.

We found that percent over 65 was a significant predictor and negatively associated with the coverage of OUD. Opioid overdose deaths largely occur in populations under 65 years of age (Hedegaard et al., 2018). The difference in opioid overdose death rates between age

groups is reflected in county health web site coverage of OUD. Percent over 65 was not a significant predictor for the coverage of substance use disorder. However, percent HS grad is significant and positively associated with the coverage of substance use disorder. Substance use among 26 year and older populations is higher among those with at least a high school degree. In 2019, an estimated 16.3% of people without a high school degree used illicit drugs while between 17 and 21 percent of people with a high school degree or higher used illicit drugs (Substance Abuse and Mental Health Services Administration, 2020).

4.4.1 Public Health Implications

We described novel methodology for collecting text data from web sites, filtering text content to analyze health content, and modeling web site text data with web site predictors to abstract health topics and identify predictors of topic coverage. Without physically reading through all web pages across all web sites, we discovered topics related to HIV/AIDS and OUD and differentiated between them by inspecting the ten most probable words of each topic. We used the regression results to identify predictors of HIV and OUD topic coverage in county health web sites. We confirmed a lower coverage of HIV care and prevention in counties with a lower percent black population. We confirmed an overall low coverage of the OUD topic and a lower coverage in counties with a lower percent over 65 population.

We found that the overall coverage of 38% (33%,44%) for the OUD topic to be relatively low compared to other health topics. It may be worth covering an OUD topic across all county health departments regardless of a counties age demographic as opioid overdose deaths in the U.S. have rapidly increased from 2014 to 2019 (National Institute on Drug Abuse, 2021). Despite HIV incidence remaining stable in 2018 compared to 2014 (Centers for Disease Control and Prevention, 2020a), HIV remains a major public health issue in the United States. HIV care and prevention should be covered by health departments regardless of a county's racial makeup.

4.4.2 Limitations

There are limitations regarding the collection and analysis of text data on county health web sites. Many local health departments in the Northeast U.S. are at the city level and many local health departments did not provide county health web site but instead provided a county government web site. We did not collect such web sites. Thus, our collection of county health web sites is not entirely representative of the U.S. The data that we collected in March 2021 represents a snapshot of the health information web sites provide at that specific time. Thus, our analysis is limited to data from a specific date. We analyze HIV and OUD related topics to counteract this limitation. HIV and OUD have been important health issues in the U.S. for many years. Thus, we do not expect that the coverage of these topics have changed dramatically over the recent months as compared to for example COVID-19 topics.

4.4.3 Other Applications of Data and Methodology

We used the data to analyze HIV and OUD related topics and how they are covered in county health web sites. There were many more health topics, with about a quarter of the 300 topics being health related. The data and analysis can be used to analyze other health topics; however, some additional relevant predictors may be desirable. For example analyzing topics related to maternal health or childcare may require the addition of percent of families with children, percent of population under a certain age, or median age as predictors. Our study methodology can also be adapted to analyze different text data sets. The methodology abstracts topics from a nested collection of text with aims to understand how web sites cover different topics. Both a change in topic to analyze or a change in data set would likely require respecification of model tuning parameters, such as the number of topics and prior parameters.

4.5 Web Appendix

4.5.1 Hierarchical Topic Presence Model

Topic models are used to abstract topic information from a collection of text documents. Text documents are reduced to a bag-of-words such that word co-occurrence is retained but word ordering is lost. Topic models then assert that documents are mixtures of topics and that topics are mixtures of words. We describe the hierarchical topic presence model used in the main text to abstract and make inferences about human immunodeficiency virus (HIV) and opioid use disorder (OUD) topics.

We describe the HTPM in context of modeling text in web pages nested in web sites. Our dataset is comprised of M web sites indexed by $i = 1, \ldots, M$, each with N_i web pages indexed by $j = 1, \ldots, N_i$, with vocabulary or set of unique words of size V indexed by $v = 1, \ldots, V$. The tuning parameter K is the number of global topics indexed by $k = 1, \ldots, K$. The K global topics may be shared across multiple web sites. Each web site $i = 1, \ldots, M$ has a single local topic, a topic unique that web site. Thus, there are K global topics and M local topics. For each web site i, we let $k = 1, \ldots, K$ index global topics $1, \ldots, K$ and $k = K + 1$ index the local topic of the web site. Local topics are not useful in abstracting health topics shared across web sites. Local topics place high probability on words that occur frequently in a web site and infrequently in other web sites. Local topics often place high probability on geographic names and locations or common words from local news. We are not interested in analyzing such topics, thus, we model them as nuisance parameters, and this reduces the number of global topics needed. In the main text, we refer to global topics as topics and do not reference local topics.

The HTPM we describe here is the same as the SE-PFA-LT model detailed in Wang and Weiss (2021a) though with different data set, predictors, and prior specifications for coefficients. The HTPM models word counts z_{ijkv}, which are latent counts of words v from topic k in web page j in web site i. Let topic weight θ_{ijk} be the expected total count of

Table 4.5: Model notation with definitions.

Notation	Description
i	Website index, $i = 1, \ldots, M$ number of web sites
j	Page index, $j = 1, \ldots, N_i$ number of web pages
k	Global topic index, $k = 1, \ldots, K$ number of global topics
v	Vocabulary word index, $v = 1, \ldots, V$ number of unique words
θ_{ijk}	Expected count of words from topic k in web page j in web site i
ψ_{iv}	Probability of word v in local topic in web site i
ϕ_{kv}	Probability of word v in global topic k
z_{ijkv}	Count of word v from topic k in web page j in web site i

words from topic k in web page j in web site i. A topic is a probability distribution over words. The unknown probability vector for global topic k is $\phi_k = (\phi_{k1}, \ldots, \phi_{kV})'$ and the probability vector corresponding to the local topic in web site i is $\psi_i = (\psi_{i1}, \ldots, \psi_{iV})'$. Define $\Phi_i = (\phi_1, \ldots, \phi_K, \psi_i)'$ to be the $(K+1) \times V$ matrix of word probability vectors for all global topics and web site i's local topic. Then Φ_{ikv} is the probability of word v in topic k in web site i, where $k = 1, \ldots, K+1$. Table 4.5 summarizes our model notation.

The HTPM models word count z_{ijkv} as

$$z_{ijkv} | \Phi_{ikv}, \theta_{ijk} \sim \text{Poisson}(\Phi_{ikv}\theta_{ijk}).$$

Define b_{ik} as web site topic presence of topic k in web site i. Topic k is present in web site i when unknown parameter $b_{ik} = 1$ and not present in web site i when $b_{ik} = 0$. Similarly, web page topic presence $c_{ijk} = 1$ if topic k in web page j of web site i is present. A topic is only truly present in a web page when both $b_{ik} = 1$ and $c_{ijk} = 1$. Let r_k be a global topic weight parameter for topic k. Let the subscript $\cdot$ indicate a sum across an index, such that $z_{ij\cdot\cdot}$ is the count of words in web page j of web site i. The HTPM models topic weight θ_{ijk} as

$$\theta_{ijk} | r_k, b_{ik}, c_{ijk} \sim \text{Gamma}(z_{ij\cdot\cdot} r_k b_{ik} c_{ijk}, 1)$$

where a variable distributed Gamma(a, b) has mean ab. Topic weight θ_{ijk} is non-zero only when both $b_{ik} = 1$ and $c_{ijk} = 1$.

The HTPM models web site topic presence b_{ik} with a logistic regression. Let X_i be web site i's Q-vector of standardized predictors and let β_k be the corresponding Q-vector of predictor effects for topic k. Predictors are standardized by subtracting the mean across all departments and dividing by the standard deviation. Then the HTPM models b_{ik} as Bernoulli,

$$b_{ik}|\beta_k \sim \text{Bernoulli}\left(\frac{\exp(X_i'\beta_k)}{1+\exp(X_i'\beta_k)}\right).$$

The model sets hierarchical normal priors on the β_k's with unknown grand mean β_0 and diagonal covariance matrix $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_Q^2)$

$$\beta_k|\beta_0, \Sigma \sim \text{Normal}(\beta_0, \Sigma),$$

with hyperpriors for β_0 and σ_q^{-2}

$$\beta_0 \sim \text{Normal}(\mu_0, \sigma_0^2 I_Q),$$
$$\sigma_q^{-2} \sim \text{Gamma}(d_\sigma, e_\sigma),$$

where $\mu_0, \sigma_0, d_\sigma,$ and e_σ are known hyperparameters and I_Q is the $Q \times Q$ identity matrix. We used $Q = 7$ predictors including the intercept in the main analysis.

The HTPM models web page topic presence c_{ijk} with an exchangeable prior. Let η_k be the prior probability that topic k is present in any web page j in any web site i. Then the HTPM models c_{ijk} as Bernoulli,

$$c_{ijk}|\eta_k \sim \text{Bernoulli}(\eta_k),$$

with prior on η_k

$$\eta_k \sim \text{Beta}(d_\eta, e_\eta).$$

We parameterize $d_\eta \equiv d_\eta(\mu_\eta, \sigma_\eta^2)$ and $e_\eta \equiv e_\eta(\mu_\eta, \sigma_\eta^2)$ in terms of mean $\mu_\eta = d_\pi/(d_\pi + e_\pi)$ and variance $\sigma_\eta^2 = \mu_\eta(1-\mu_\eta)/(d_\pi + e_\pi + 1)$ of the $\text{Beta}(d_\eta, e_\eta)$ and place Beta priors on mean and variance

$$\mu_\eta \sim \text{Beta}(d_{\mu\eta}, e_{\mu\eta}),$$
$$\sigma_\eta^2 \sim \text{Beta}(d_{\sigma\eta}, e_{\sigma\eta}),$$

where $d_{\mu\eta}$, $e_{\mu\eta}$, $d_{\sigma\eta}$, and $e_{\sigma\eta}$ are known hyperparameters.

The HTPM places hierarchical Gamma priors on global topic weight parameter r_k for $k = 1, \ldots, K+1$,

$$r_k | r_0 \sim \text{Gamma}(r_0, 1) \text{ for } k = 1, \ldots, K+1,$$
$$r_0 \sim \text{Gamma}(d_{r0}, e_{r0}),$$

where d_{r0} and e_{r0} are known hyperparamters and sets Dirichlet priors on topics' word probability vectors,

$$\phi_k \sim \text{Dirichlet}(\alpha_\phi 1_V),$$
$$\psi_i \sim \text{Dirichlet}(\alpha_\psi 1_V),$$

where α_ϕ and α_ψ are fixed hyperparameters. We refer to this model as HTPM in the the main text, while Wang and Weiss (2021a) uses the term HTPM to describe this model and

variations of this model. We follow the Markov chain Monte Carlo sampling procedure outlined in Wang and Weiss (2021a) to draw posterior samples for inference and implement the procedure with Julia (Bezanson et al., 2017). We drew a total of 60,000 samples, dropped the first 50,000 samples as burnin, and kept the last 10,000 with a thinning of 10 for a total of 1,000 saved samples.

4.5.1.1 Prior Specification

We set prior means and variances for the overall predictor effects β_0 to be $\mu_0 = (0, 0, 0, 0, 0, 0, 0)$ and $\sigma_0^2 = 1$. We set the Gamma parameters $d_\sigma = 16$ and $e_\sigma = 1$ for $q = 1$ corresponding to the intercept term and we set Gamma parameters $d_\sigma = 2$ and $e_\sigma = 1/8$ for $q > 1$ for the standardized population and demographic predictors. We choose this specification because a one unit increase in any predictor is a one standard deviation change therefore we expect coefficients a priori less than 1 and we want to place more uncertainty in the intercepts.

A priori we expect that web pages will have few topics so we set Beta parameters to be $d_{\mu\eta} = 1$, $e_{\mu\eta} = K - 1$, $d_{\sigma\eta} = 1$, and $e_{\sigma\eta} = K - 1$. We set a vague Gamma prior with parameters $d_{r0} = 0.01$ and $e_{r0} = 100$ for r_0. We set $\alpha_\phi = \alpha_\psi = 0.05$ to encourage topics to place small probability on most words and larger probability on a few words.

4.5.2 Tuning the Number of Topics

We compare the HTPM at $K = 10, 25, 50, 100, 200, 300$ and 400. We randomly select 80% of words in each web page to be our training set and save the remaining 20% as the test or held-out set to evaluate our models. We use the 1000 saved samples to calculate the log predictive probability of the held out words. Let superscript $s = 1, \ldots, S$ index samples from the posterior and let $y_{i'j'\cdot v}$ be the count of held-out word v in web page j' in web site

i'. We define perplexity, the exponentiated log predictive probability, as

$$\text{Perplexity} = \exp(-\frac{1}{y_{\cdots}} \sum_{i'=1}^{M'} \sum_{j'=1}^{N'_{i'}} \sum_{v=1}^{V} y_{i'j'v} \log f_{i'j'v})$$

where

$$f_{i'j'v} = \frac{\sum_{s=1}^{S} \sum_{k=1}^{K} \phi_{kv}^{(s)} \theta_{i'j'k}^{(s)} + \psi_{i'}^{(s)} \theta_{i'j'(K+1)}^{(s)}}{\sum_{s=1}^{S} \sum_{k=1}^{K} \sum_{v=1}^{V} \phi_{kv}^{(s)} \theta_{i'j'k}^{(s)} + \psi_{i'}^{(s)} \theta_{i'j'(K+1)}^{(s)}}$$

is the predicted probability of word v of web page j in web site i averaged across S samples and K topics. A.Figure 4.2 plots perplexity as a function of the number of global topics K.

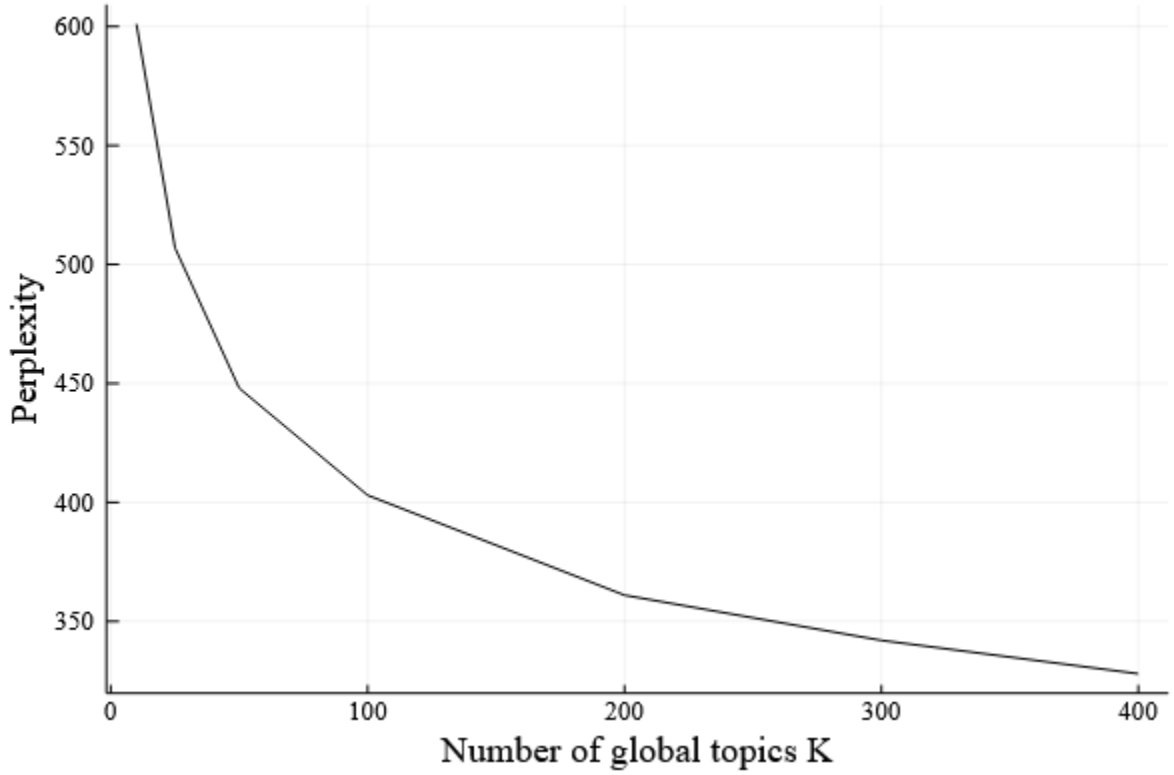

Figure 4.2: Perplexity (lower is better) plotted as a function of number of global topics K.

4.5.3 Comparing Websites With and Without Pages Containing Most Probable Words in a Topic

We compare web sites with and without web pages containing the most probable words in a topic to validate our regression results in the main analysis. When a predictor is positively associated with the presence of a topic we would expect that predictor to be larger in web sites with a web page containing the most probable words of that topic. We plot a boxplot of the distribution of standardized predictors for web sites that have at least one web page with both the two most probable words of a topic (*web sites with top 2*) and for web sites that do not (*web sites without top 2*). We repeat this for each of the four topics, infectious disease surveillance, HIV care and prevention, opioid use disorder, and substance use disorder, where we found a significant association between topic coverage and at least one predictor other than log population. For the four topics, A.Figure 4.3 shows boxplots of standardized predictors for *web sites with top 2* and *web sites without top 2* for all 6 predictors. When a predictor is positively associated with the coverage of a particular topic we expect that predictor to be generally larger in *web sites with top 2* than in *web sites without top 2* and vice versa for predictors that are significant and negatively associated with topic coverage.

Percent black was positively associated with coverage of the infectious disease surveillance topic and the HIV care and prevention topic. A.Figure 4.3 shows that percent black was generally larger in *web sites with top 2* than in *web sites without top 2* for both topics. Percent seniors was negatively associated with coverage of the opioid abuse and percent HS grad was negatively associated with coverage of the general drug abuse topic. Percent seniors was smaller in *web sites with top 2* than in *web sites without top 2* for the opioid abuse topic and percent HS grad was lower in *web sites with top 2* than in *web sites without top 2* for the general drug abuse topic. Log population is positively associated for all four topics and log population is generally higher in *web sites with top 2* than in *web sites without top 2*.

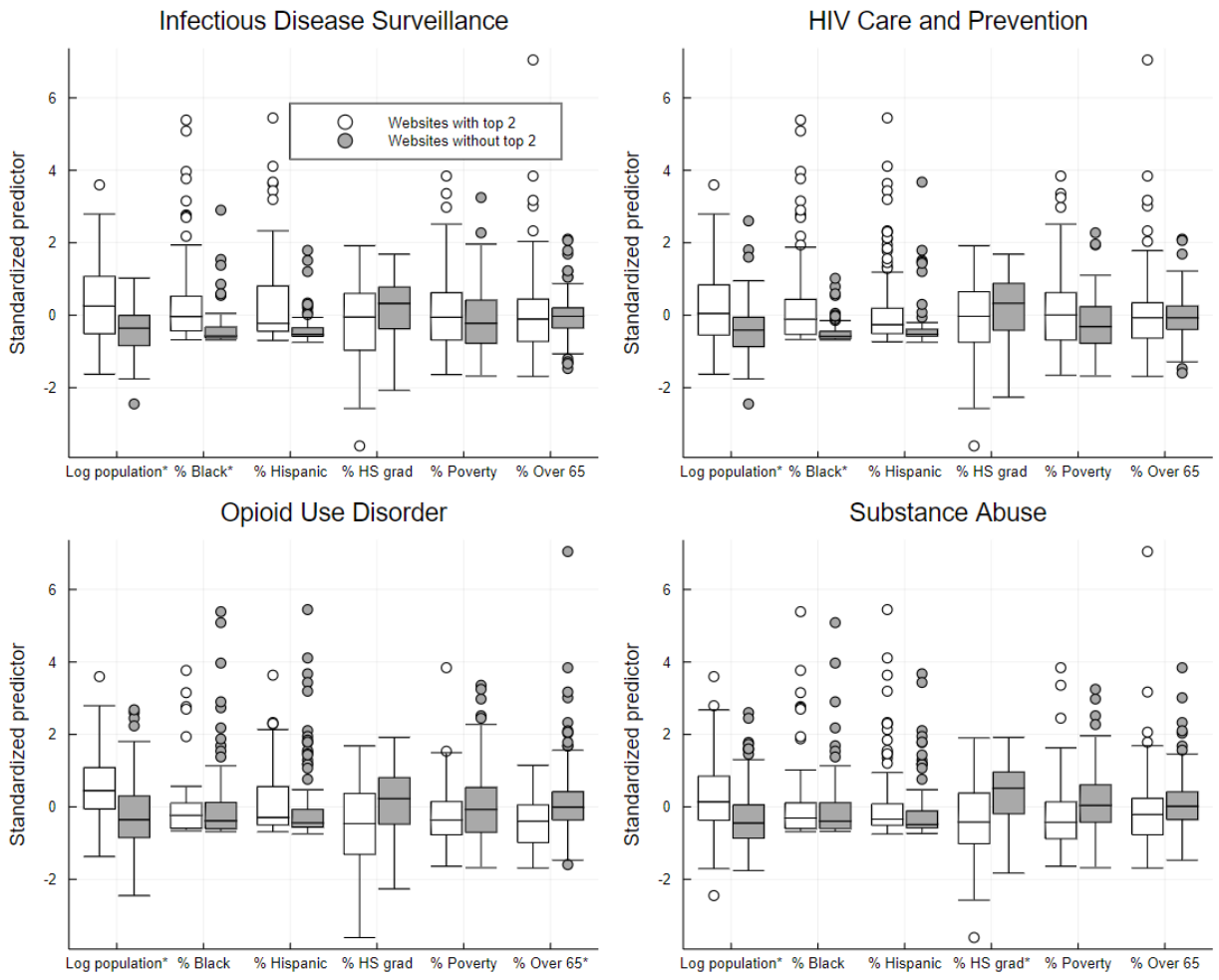

Figure 4.3: Boxplot of standardized predictors for the four topics where we found significant association between topic coverage and at least one demographic predictor. For each topic and predictor we plot boxplots for web sites that have at least one web page with both the two most probable words of a topic (*web sites with top 2*, clear boxplot) and for web sites that do not (*web sites without top 2*, grey boxplot). An * next to the predictor name indicates that it is a significant predictor of topic coverage.

4.5.4 Other Health Topics

The main text discusses 7 topics that are related to either HIV or to OUD. We identified another 84 health topics from among the 300 global topics allowed by HTPM. We show the 5 most probable words and the proportion (95% posterior interval) of web sites with the topic present for the 84 health topics in A.Table 4.6. A.Table 4.7 shows median odds ratio and 95% posterior intervals corresponding to the 6 predictors for the 84 health topics.

Topics with a similar set of 5 most probable words were distinguished by inspecting the next 5 most probable words. Beach advisory topics 1 and 4 were similar, flu shot topics 27 and 67 were similar, animal control topics 40 and 48 were similar, and pregnancy topics 71 and 78 were similar.

Topic 1 has 10 most probable words *beach, advisori, sampl, monitor, bacteria, qualiti, enterococci, healthi, indic, enter* and topic 4 has 10 most probable words *advisori, beach, bacteria, island, test, level, lift, recommend, fish, simon.* St. Simons Island is located in Glynn county, Georgia and its beach advisories were recently lifted in March 2021. These two beach advisory topics split into two separate beach advisory topics at $K = 300$ global topics. Topic 27 has 10 most probable words *influenza, pandem, flu, virus, season, avian, viru, antivir, surveil, respiratori* and topic 67 has 10 most probable words *flu, vaccin, shot, season, influenza, protect, older, high, cdc, recommend.* Topic 67 is related to flu shot recommendations for certain age groups while topic 23 is a more general flu shot topic. Topic 40 has 10 most probable words *anim, pet, dog, rabi, cat, vaccin, wild, alert, domest, owner* and topic 48 has 10 most probable words *anim, rabi, wild, pet, raccoon, bite, vaccin, dog, cat, test.* Topic 40 has more words related to domestic pet with words *pet, domest,* and *owner* while topic 48 only has *pet* and also includes a wild animal, *raccoon*, among its 10 most probable words. Topic 71 has 10 most probable words *birth, pregnanc, famili, women, contracept, method, sexual, condom, counsel, reproduct* and topic 78 has 10 most probable words *babi, pregnanc, pregnant, women, birth, prenat, infant, mother, care, famili.* Topic

71 is more related to family planning and preventing unwanted pregnancy while topic 78 is more related to prenatal care.

Topic 6 combines two separate topics of Halloween and Infectious Disease. This is likely because both topics may have included *treat* among its most probable words. Treat has a different meaning in both topics. However, our model does not incorporate word meaning or context, thus, cannot distinguish between the two usages of *treat*.

Table 4.6: The 5 most probable words in order of probability for health topics that were not discussed in the main text. The right column is median percent coverage (95% PI) across all web sites. Topics are ordered from lowest percent coverage to highest. COVID is an abbreviation for the 2019 novel coronavirus disease, SIDS is an acronym for sudden infant death syndrome, and WIC refers to the Special Supplemental Nutrition Program for Women, Infants, and Children. * indicates topic required further inspecting next 5 most probable words to distinguish from another topic that shared a similar set of 5 most probable words.

	Topic	Top 10 Words	Percent Coverage
1	Beach Advisory (bacteria)*	*beach, advisori, sampl, monitor, bacteria*	34(29,40)
2	Antibiotics	*antibiot, pharmaci, resist, infect, antimicrobi*	45(38,54)
3	Zika Virus	*zika, viru, travel, pregnant, mosquito*	45(39,52)
4	Beach Advisory (St. Simons)*	*advisori, beach, bacteria, island, test*	46(41,52)
5	Beach Samples	*beach, sampl, qualiti, swim, standard*	48(41,56)
6	Halloween + Infectious Disease	*ebola, halloween, west, decor, treat*	49(36,66)
7	Meningitis	*mening, prevent, vaccin, bacteri, infect*	50(43,59)
8	MMR Vaccine	*measl, mump, rash, vaccin, diseas*	51(44,60)
9	Immunizations	*pertussi, hpv, vaccin, cough, whoop*	53(47,62)

Continued on next web page

Table 4.6 – *Continued from previous web page*

	Topic	Top 10 Words	Percent Coverage
10	Physical Therapy	*therapi, physic, occup, therapist, skill*	54(49,61)
11	Infectious Diarrhea	*noroviru, diarrhea, ill, contamin, vomit*	58(52,65)
12	COVID Guidance	*guidanc, covid, social, gather, distanc*	60(54,65)
13	Bed Bugs	*bed, bug, pest, rodent, infest*	60(53,67)
14	Head Lice	*head, lice, scabi, treatment, hair*	60(50,71)
15	Food/Product Recall	*recal, product, food, alert, due*	61(55,69)
16	Mental Behaviour	*mental, behavior, treatment, substanc, youth*	62(58,67)
17	Bioterrorism	*bioterror, anthrax, smallpox, agent, biolog*	63(55,73)
18	Radon	*radon, level, test, lung, kit*	64(56,72)
19	Reopening (COVID)	*reopen, phase, capac, outdoor, guidanc*	65(58,74)
20	Tickborne Disease	*tick, diseas, bite, lyme, fever*	66(59,75)
21	COVID Cases	*case, covid, total, confirm, announc*	66(63,69)
22	Cases (outbreak)	*case, outbreak, cdc, identifi, confirm*	67(62,72)
23	Flu	*influenza, pandem, flu, virus, season*	67(59,77)
24	Diabetes	*diabet, type, prevent, prediabet, risk*	68(62,74)
25	Sudden Infant Death Syndrome	*sleep, safe, babi, infant, sid*	69(60,79)
26	COVID (general)	*covid, test, isol, confirm, quarantin*	69(63,76)
27	Flu Shot (general)*	*flu, ill, shot, season, complic*	70(65,76)
28	Hepatitis	*hepat, infect, liver, viru, men*	70(64,77)
29	General Medical Conditions	*medic, disord, condit, genet, special*	71(62,80)
30	Construction Site	*paint, level, dust, exposur, soil*	71(63,79)

Continued on next web page

Table 4.6 – *Continued from previous web page*

	Topic	Top 10 Words	Percent Coverage
	Exposure		
31	Ecigarettes	*ecigarett, product, vape, youth, tobacco*	72(64,80)
32	Mental Health/ Suicide	*mental, support, suicid, crisi, behavior*	72(65,78)
33	Masks (COVID)	*covid, mask, distanc, spread, social*	73(69,77)
34	Coping/ Mental Health	*talk, stress, feel, problem, cope*	74(67,80)
35	COVID Phase	*covid, phase, worker, expand, distribut*	74(69,79)
36	Face Cover	*face, cover, mask, wear, cloth*	74(67,82)
37	Quarantine	*quarantin, isol, test, trace, monitor*	75(67,82)
38	Severe Pain	*pain, sever, occur, common, headach*	75(70,81)
39	Virus Spread	*viru, infect, spread, ill, diseas*	76(72,80)
40	Animal Control (pets)*	*anim, pet, dog, rabi, cat*	76(70,83)
41	Blood Pressure	*blood, pressur, heart, high, cholesterol*	78(69,87)
42	Disease Spread (touch)	*hand, avoid, sick, touch, cough*	78(74,82)
43	Overweight	*physic, obes, healthi, weight, childhood*	78(70,87)
44	Chemical Contamination	*chemic, contamin, harm, lake, drink*	79(72,85)
45	Mosquitoborne Disease	*mosquito, bite, viru, repel, nile*	79(74,85)
46	Poison	*poison, prevent, childhood, children, blood*	79(73,86)

Continued on next web page

Table 4.6 – *Continued from previous web page*

	Topic	Top 10 Words	Percent Coverage
47	Foodborne Illness	*ill, foodborn, food, waterborn, investig*	79(72,87)
48	Animal Control (general)*	*anim, rabi, wild, pet, raccoon*	80(74,85)
49	Hand Washing	*hand, wash, clean, soap, sanit*	80(74,85)
50	Vaccine Dose Priority	*vaccin, dose, given, older, recommend*	80(76,85)
51	Passenger Safety	*seat, car, safeti, child, passeng*	80(73,87)
52	Mosquito Breeding	*mosquito, contain, tire, stand, cover*	80(74,87)
53	Breast Cancer	*cancer, breast, cervic, women, screen*	80(77,84)
54	Germs	*bacteria, germ, contamin, spread, kill*	82(76,87)
55	Heart Disease	*heart, diseas, american, men, stroke*	82(75,89)
56	Cough/Fever	*cough, fever, breath, nose, ill*	83(79,88)
57	Dental Care	*dental, oral, sealant, teeth, children*	83(77,91)
58	Chronic Conditions	*diseas, condit, chronic, sever, heart*	83(77,89)
59	West Nile	*west, nile, viru, mosquito, bird*	84(78,89)
60	Second Vaccine Dose	*vaccin, dose, second, administ, care*	85(80,90)
61	COVID Test	*covid, test, guidanc, case, vaccin*	86(83,88)
62	Smoking	*smoke, smokefre, secondhand, act, prohibit*	86(73,97)
63	Rabies	*rabi, anim, bat, bite, rabid*	87(80,94)
64	Breastfeeding	*breastfeed, support, mother, babi, infant*	87(83,91)
65	Carbon Monoxide	*carbon, monoxid, burn, window, fire*	88(79,95)
66	Side Effects	*effect, side, reaction, protect, safe*	88(80,93)
67	Flu Shot (recommendation)*	*flu, vaccin, shot, season, influenza*	88(84,91)

Continued on next web page

Table 4.6 – *Continued from previous web page*

	Topic	Top 10 Words	Percent Coverage
68	Emergency Kit	*prepar, kit, emerg, suppli, weather*	89(83,93)
69	Septic Tank	*septic, tank, permit, instal, soil*	90(85,93)
70	Infectious Disease Spread	*infect, diseas, spread, risk, becom*	90(87,93)
71	Pregnancy (contraceptives)*	*birth, pregnanc, famili, women, contracept*	91(87,95)
72	Blood Test	*blood, test, screen, pressur, lab*	92(88,95)
73	Vaccination (General)	*vaccin, immun, diseas, tetanu, hepat*	92(88,95)
74	Vaccine Clinic	*vaccin, clinic, immun, consent, older*	92(89,95)
75	Primary Care	*care, primari, child, adult, famili*	93(90,96)
76	Tobacco	*tobacco, smoke, quit, cessat, smoker*	93(89,96)
77	Medical Care	*medic, care, patient, physician, treatment*	94(91,96)
78	Pregnancy (prenatal care)*	*babi, pregnanc, pregnant, women, birth*	94(92,96)
79	Physical Excercise	*physic, particip, exercis, life, session*	95(90,98)
80	WIC Eligibility	*nutrit, food, incom, particip, benefit*	95(93,97)
81	Tuberculosis	*tuberculosi, test, treatment, skin, infect*	95(91,98)
82	WIC (general)	*women, children, infant, healthi, pregnant*	96(94,98)
83	Emergency Preparedness (disaster)	*emerg, disast, prepared, prepar, natur*	97(96,99)
84	Emergency Preparedness(response)	*emerg, prepared, medic, respond, coordin*	99(97,100)

Table 4.7: Median odds ratio for each standardized regression coefficient for health topics not discussed in the main text. * indicates one-sided significance at level 0.025. The topics are ordered the same as in A.Table 4.6.

	Topic	Log population	% Black	% Hispanic	% HS grad	% Poverty	% Over 65
	Overall	1.35*	1.07*	0.91*	0.82*	0.95*	0.92*
1	Beach Advisory (bacteria)	1.44*	0.95	1.07	0.77	0.76*	1.04
2	Antibiotics	1.39*	1.08	0.93	0.76	0.86	0.88
3	Zika Virus	1.48*	1.30	0.88	0.76	0.99	0.85
4	Beach Advisory (St. Simons)	1.49*	0.96	1.05	0.71*	0.93	1.19
5	Beach Samples	1.45*	1.02	0.88	0.79	0.85	1.08
6	Halloween Disease Spread	1.30*	1.04	0.95	0.75	0.94	0.89
7	Meningitis	1.33*	1.07	0.72*	0.88	1.01	0.75
8	MMR Vaccine	1.46*	1.19	0.92	0.78	1.03	0.92
9	Immunizations	1.44*	1.10	0.98	0.73	0.99	1.01
10	Physical Therapy	1.09	1.10	0.90	0.83	0.88	0.84
11	Infectious Diarrhea	1.41*	1.08	0.65*	0.70*	0.93	0.86
12	COVID Guidance	1.44*	1.12	0.90	0.67*	0.90	0.94
13	Bed Bugs	1.24	1.04	0.76	1.08	0.91	0.69*
14	Head Lice	1.40*	0.89	0.76	0.93	0.96	0.74
15	Food/Product Recall	1.28	0.88	0.78	0.97	0.94	0.82
16	Mental Behaviour	1.34*	1.06	0.82	0.71*	0.95	0.77
17	Bioterrorism	1.36*	0.97	0.98	0.91	0.89	0.99
18	Radon	1.47*	0.91	0.71*	0.92	0.87	0.99
19	Reopening (COVID)	1.38*	0.85	0.82	0.69*	0.83	1.04
20	Tickborne Disease	1.34*	0.95	0.70*	1.02	0.97	0.91

Continued on next web page

Table 4.7 – *Continued from previous web page*

Topic	Log population	% Black	% Hispanic	% HS grad	% Poverty	% Over 65
21 COVID Cases	1.34*	1.09	0.85	0.71*	0.95	0.90
22 Cases (outbreak)	1.38*	1.20	0.76	0.73*	0.81	0.96
23 Flu (general)	1.44*	0.99	0.91	0.96	0.91	0.97
24 Diabetes	1.37*	1.13	0.94	0.85	1.13	0.86
25 Sudden Infant Death Syndrome	1.39*	0.93	0.82	0.99	0.98	0.95
26 COVID (general)	1.30*	0.89	0.77	0.78	0.97	0.85
27 Flu Shot (general)	1.52*	1.27	0.95	0.79	0.90	0.98
28 Hepatitis	1.41*	1.21	0.88	0.76	0.96	0.86
29 General Medical Conditions	1.30*	1.08	0.79	1.09	1.05	0.85
30 Construction Site Exposure	1.39*	0.90	0.74	0.90	0.97	0.83
31 Ecigarettes	1.41*	1.09	1.01	0.75	1.01	0.99
32 Mental Health/Suicide	1.28	1.01	0.79	0.65*	0.88	0.74*
33 Masks (COVID)	1.36*	0.97	0.89	0.66*	0.83	0.87
34 Coping/Mental Health	1.25	0.96	0.74	0.83	1.01	0.94
35 COVID Phase	1.32*	1.01	0.76	0.63*	0.87	1.03
36 Face Cover	1.39*	1.10	0.83	0.81	0.95	0.99
37 Quarantine	1.34*	1.13	1.04	0.69*	0.97	0.78
38 Severe Pain	1.38*	1.06	0.79	0.82	0.87	0.72*
39 Virus Spread	1.36*	1.12	1.01	0.88	0.98	1.08
40 Animal Control (pets)	1.38*	1.16	0.94	0.89	0.98	1.09
41 Blood Pressure	1.36*	1.19	0.88	0.89	1.06	1.12
42 Disease Spread (touch)	1.48*	1.11	0.89	0.74	0.94	1.05

Continued on next web page

Table 4.7 – *Continued from previous web page*

Topic	Log population	% Black	% Hispanic	% HS grad	% Poverty	% Over 65
43 Overweight	1.31	1.10	1.14	0.83	1.05	1.04
44 Chemical Contamination	1.36*	1.25	0.78	0.79	1.02	0.97
45 Mosquitoborne Disease	1.44*	1.09	0.99	0.90	0.98	0.86
46 Poison	1.24	1.10	0.97	0.96	1.18	0.89
47 Foodborne Illness	1.30*	1.04	0.86	0.83	1.01	0.86
48 Animal Control (general)	1.49*	1.37	0.92	0.81	0.95	0.97
49 Hand Washing	1.35*	1.07	0.73*	0.75	0.85	0.95
50 Vaccine Dose Priority	1.29	0.85	0.93	0.87	0.90	1.03
51 Passenger Safety	1.26	1.06	0.86	0.87	1.07	0.95
52 Mosquito Breeding	1.42*	1.27	1.05	0.92	0.93	0.87
53 Breast Cancer	1.34*	1.24	0.94	0.80	1.16	0.97
54 Germs	1.39*	1.21	0.77	0.89	0.91	0.95
55 Heart Disease	1.39*	1.22	0.92	0.96	1.11	1.03
56 Cough/Fever	1.31*	1.13	1.05	0.80	0.93	0.98
57 Dental Care	1.28*	1.17	0.99	0.83	1.06	1.01
58 Chronic Conditions	1.42*	1.17	0.92	0.77	0.96	0.85
59 West Nile	1.35*	1.08	0.98	0.88	0.94	0.69*
60 Second Vaccine Dose	1.25	0.84	0.97	1.01	0.93	1.04
61 COVID Test	1.47*	1.04	0.81	0.70*	0.86	0.88
62 Smoking	1.29	1.09	0.84	0.86	0.95	0.80
63 Rabies	1.39*	1.30	0.96	0.88	1.01	1.00
64 Breastfeeding	1.29	0.97	0.87	0.85	1.11	1.15
65 Carbon Monoxide	1.37*	1.09	0.88	0.84	0.90	0.90

Continued on next web page

Table 4.7 – *Continued from previous web page*

Topic	Log population	% Black	% Hispanic	% HS grad	% Poverty	% Over 65
66 Side Effects	1.24	1.10	0.84	0.84	1.03	0.96
67 Flu Shot (recommendation)	1.35*	0.90	0.93	0.80	0.88	0.97
68 Emergency Kit	1.37*	0.97	0.88	0.76	0.89	0.93
69 Septic Tank	1.23	1.17	0.73	1.01	0.92	0.99
70 Infectious Disease Spread	1.32	1.21	1.02	0.98	1.10	1.08
71 Pregnancy (contraceptives)	1.41*	1.29	1.03	0.83	1.06	1.01
72 Blood Test	1.23	1.07	0.72	1.00	1.05	0.74
73 Vaccination (General)	1.28	1.11	0.91	0.93	1.01	1.08
74 Vaccine Clinic	1.36*	1.03	0.81	0.72	0.87	0.89
75 Primary Care	1.27	0.98	1.04	0.93	0.98	1.04
76 Tobacco	1.32*	1.15	0.91	0.88	1.06	1.02
77 Medical Care	1.38*	1.10	1.03	0.88	1.01	1.05
78 Pregnancy (prenatal care)	1.26	1.06	0.89	0.98	1.06	1.01
79 Physical Excercise	1.29	1.06	0.85	0.81	0.94	0.92
80 WIC Eligibility	1.32	1.21	0.95	0.90	1.05	1.02
81 Tuberculosis	1.29	1.18	0.97	0.88	0.98	0.92
82 WIC (general)	1.36*	1.16	0.97	0.82	1.04	0.96
83 Emergency Preparedness (disaster)	1.39*	1.15	0.94	0.80	0.97	0.91
84 Emergency Preparedness (response)	1.35	1.10	0.91	0.82	0.96	0.92

CHAPTER 5

Discussion

This dissertation introduced methodology for the public health researcher to survey United States local health departments through investigation of departments' web sites. Defining all possible health topics and their key words is difficult and time consuming for a researcher to do themself, by hand. It is possible the researcher may not identify all health topics or may not identify an appropriate set of key words for every health topic. Then having the researcher read all web pages to identify which health topics are covered by each web page or web site is even more time consuming with many opportunities for mistakes in identifying topic coverage. We demonstrated our methodology with data from U.S. county health department web sites and automatically identified health topics, topic coverage, and predictors of topic coverage in local health department web sites, without knowing a priori what health topics existed in the data and without having to have a human read all web pages.

Our work presented an important public health application; however, our methods can be applied to any nested document collection. For example, a social media researcher wanting to study social media posts from a group of users but is uncertain what topics are discussed may find our methods useful. Users can post multiple times and each user has a different set of characteristics. Thus, posts are nested within users and we have user level predictors. Posts are analogous to web pages and users are analogous to web sites. We could then draw inferences about the associations between user characteristics and user topic coverage. If we instead have post level predictors, we can model post topic coverage as a function of post

level predictors. We can apply the same models to other nested document collections such as articles nested in journals, articles nested in newspapers, blogs nested in authors, or posts nested in forums.

Bibliography

Archambeau, C., Lakshminarayanan, B., and Bouchard, G. (2015). Latent IBP compound Dirichlet allocation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):321–333.

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98.

Bird, S., Loper, E., and Ewan, K. (2009). Natural language processing with python. *O'Reilly Media Inc.*

Blei, D. M. and Lafferty, J. D. (2005). Correlated topic models. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, NIPS, pages 147–154, Cambridge, MA, USA. MIT Press.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Boyd-Graber, J. L., Hu, Y., and Mimno, D. M. (2017). Applications of topic models. *Foundations and Trends in Information Retrieval*, 11(2-3):143–296.

Centers for Disease Control and Prevention (2016). Today's HIV/AIDS epidemic. *CDC Fact Sheet.*

Centers for Disease Control and Prevention (2020a). Estimated HIV incidence and prevalence in the United States 2014-2018. *HIV Surveillance Supplemental Report*, 25.

Centers for Disease Control and Prevention (2020b). PrEP (pre-exposure prophylaxis). https://www.cdc.gov/hiv/basics/prep.html.

Centers for Disease Control and Prevention (2021). 20 leading causes of

death and injury. *Web-based Injury Statistics Query and Reporting System*. https://www.cdc.gov/injury/wisqars/LeadingCauses.html.

Chang, J. and Blei, D. (2009). Relational topic models for document networks. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 81–88, Clearwater Beach, Florida USA.

Chemudugunta, C., Smyth, P., and Steyvers, M. (2006). Modeling general and specific aspects of documents with a probabilistic topic model. NIPS 2006, pages 241–248, Cambridge, MA, USA. MIT Press.

Chen, N., Zhu, J., Xia, F., and Zhang, B. (2015). Discriminative relational topic models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(5):973–986.

Deutsch-Feldman, M., Pratt, R. H., Price, S. F., Tsang, C. A., and Self, J. L. (2021). Tuberculosis - United States, 2020. *Morbity Mortality Weekly Report*, 70:409–414.

Gan, Z., Chen, C., Henao, R., Carlson, D., and Carin, L. (2015). Scalable deep Poisson factor analysis for topic modeling. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1823–1832, Lille, France.

Ghertner, R. and Groves, L. (2018). The opioid crisis and economic opportunity: Geographic and economic trends. *ASPE Research Brief*.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235.

Grün, B. and Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, 40(13):1–30.

Guo, W., Wu, S., Wang, L., and Tan, T. (2015). Social-relational topic model for social networks. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 1731–1734, New York, NY, USA. Association for Computing Machinery.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Hedegaard, H., Minino, A. M., and Warner, M. W. (2018). Drug overdose deaths in the United States, 1999-2017. *NCHS Data Brief No. 329*.

Hua, T., Lu, C.-T., Choo, J., and Reddy, C. K. (2020). Probabilistic topic modeling for comparative analysis of document collections. *ACM Transactions on Knowledge Discovery from Data*, 14(2).

Lippold, K. M., Jones, C. M., Olsen, E. O., and Giroir, B. P. (2019). Racial/ethnic and age group differences in opioid and synthetic opioid–involved overdose deaths among adults aged $\geq$ 18 years in metropolitan areas — United States, 2015–2017. *Morbity Mortality Weekly Report*.

Liu, Y., Niculescu-Mizil, A., and Gryc, W. (2009). Topic-link LDA: Joint models of topic and author community. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML 2009, pages 665–672, New York, NY, USA. ACM.

Lu, Y., Mei, Q., and Zhai, C. (2011). Investigating task performance of probabilistic topic models: An empirical study of PLSA and LDA. *Information Retrieval*, 14(2):178–203.

National Association of County and City Health Officials (2012). Directory of local health departments. https://www.naccho.org/membership/lhd-directory.

National Association of County and City Health Officials (2021). Strategic planning guide.

National Institute on Drug Abuse (2020). Opioid overdose reversal with naloxone (Narcan, Evzio). https://www.drugabuse.gov/drug-topics/opioids/opioid-overdose-reversal-naloxone-narcan-evzio.

National Institute on Drug Abuse (2021). Overdose death rates. *Trends and Statistics*. https://www.drugabuse.gov/drug-topics/trends-statistics/overdose-death-rates.

National Library of Medicine (2021). Medical subject headings. https://www.nlm.nih.gov/mesh/meshhome.html.

Paul, M. and Dredze, M. (2014). Discovering health topics in social media using topic models. *PloS One*, 9:e103408.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Pólya–gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349.

Qiang, S., Wang, Y., and Jin, Y. (2017). A local-global LDA model for discovering geographical topics from social media. In *APWeb and WAIM Joint Conference on Web and Big Data*, pages 27–40, Cham. Springer International Publishing.

Roberts, M. E., Stewart, B. M., Tingley, D., and 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*, pages 1–20.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI 2004, pages 487–494, Arlington, Virginia, USA. AUAI Press.

Rosenberg, R., Lindsey, N., Fischer, M., Gregory, C., Hinckley, A., Mead, P., Paz-Bailey, G., Waterman, S., Drexler, N., Kersh, G., Hooks, H., Partridge, S., Visser, S., Beard, C., and

Petersen, L. (2018). Vital signs : Trends in reported vectorborne disease cases — United States and territories, 2004–2016. *Morbidity and Mortality Weekly Report*, 67(17):496–501.

Scott, J. and Baldridge, J. (2013). A recursive estimate for the predictive likelihood in a topic model. In Carvalho, C. M. and Ravikumar, P., editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 527–535, Scottsdale, Arizona, USA. PMLR.

ScrapingHub (2018). Scrapy 1.8 documentation. https://scrapy.org/.

Song, Y., Pan, S., Liu, S., Zhou, M. X., and 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*, CIKM'09, pages 1757–1760, New York, NY, USA. Association for Computing Machinery.

Substance Abuse and Mental Health Services Administration (2020). 2019 national survey on drug use and health detailed tables. https://www.samhsa.gov/data/report/2019-nsduh-detailed-tables.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

U.S. Census Bureau (2020). 2019 American community survey 5-year public use data. https://www.census.gov/data/developers/data-sets/acs-5year.2019.html.

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

Wallach, H. M., Mimno, D., and McCallum, A. (2009a). Rethinking LDA: Why priors matter. In *Proceedings of the 22Nd International Conference on Neural Information Processing Systems*, NIPS 2009, pages 1973–1981, USA. Curran Associates Inc.

Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. (2009b). Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML 2009, pages 1105–1112, New York, NY, USA. ACM.

Wang, J. and Weiss, R. E. (2021a). Hierarchical topic presence models. *arXiv preprint at arXiv:2104.07969*.

Wang, J. and Weiss, R. E. (2021b). Local and global topics in text modeling of web pages nested in web sites. *arXiv preprint at arXiv:2104.01115*.

Williamson, S., Wang, C., Heller, K. A., and Blei, D. M. (2010). The IBP compound Dirichlet process and its application to focused topic modeling. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML 2010, pages 1151–1158, Madison, WI, USA.

Yang, Y., Wang, F., Jiang, F., Jin, S., and Xu, J. (2016). A topic model for hierarchical documents. In *2016 IEEE First International Conference on Data Science in Cyberspace (DSC)*, pages 118–126.

Zhao, H., Du, L., and Buntine, W. (2017). A word embeddings informed focused topic model. In *Proceedings of the Ninth Asian Conference on Machine Learning*, volume 77 of *Proceedings of Machine Learning Research*, pages 423–438.

Zhou, M. and Carin, L. (2015). Negative binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):307–320.

Zhou, M., Hannah, L., Dunson, D., and Carin, L. (2012). Beta-negative binomial process and Poisson factor analysis. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 1462–1471, La Palma, Canary Islands.