# LDA Topic Analysis of a Cybersecurity Textbook

**Travis Scheponik, Alan T. Sherman**
Cyber Defense Lab
University of Maryland, Baltimore County (UMBC)
Baltimore, MD 21250 USA
{tschep1,sherman}@umbc.edu

## Abstract

We perform a Latent Dirichlet Allocation (LDA) topic analysis of Matt Bishop's widely used 1440-page textbook on cybersecurity. To our knowledge, our work is the first time such analysis has been carried out on any single-authored work of this length. This topic analysis might be useful for creating a supplemental interactive guide to the textbook to help readers explore the book more efficiently for particular learning objectives.

**Keywords.** cybersecurity, Latent Dirichlet Allocation (DLA), Matt Bishop, topic analysis.

## 1 Introduction

Learning a new subject can be a daunting task. Traditional textbooks are organized in a fixed fashion that might not match the particular learning needs of the reader. Keyword searches and indices typically are based on string matching without regard to concepts or the reader's needs. We propose that topic modeling of a textbook can produce information that would be useful in creating an interactive guide for the textbook, which could support a reader's learning needs and produce a personalized learning path through the textbook. To this end, as an initial exploratory step, we performed a *Latent Dirichlet Allocation (LDA)* topic analysis of a preprint (Bishop, 2018) of Matt Bishop's widely used textbook on cybersecurity, *Computer Security: Art and Science (CSAS)* (Bishop et al., 2019).

LDA topic analysis is a type of statistical analysis of words in a document that finds latent structure ("latent topics," or for brevity, "topics.") We assume the reader is familiar with this technique, as explained, for example, by Blei, et al. (Blei et al., 2003). One of its applications in machine learning is to find hidden semantic structure.

In a traditional book, the author must commit to a fixed linear order. This order, however, is likely suboptimal for many learners, each of whom may have differing needs and learning objectives. Moreover, there may be relevant hidden structure that is neither in the table of contents nor in the mind of the author. We posit that the latent structure revealed by LDA can be useful in guiding a reader towards their learning objectives.

Our main contribution is the topic modeling of CSAS. To our knowledge, our work is the first to so analyze a large single-authored text. The rest of this paper explains our methods, presents our topic modeling, discusses our results, sketches our vision for an interactive guide, reviews previous work, and presents our conclusions.

## 2 Methods

We describe our source text and methods, including how we preprocessed the text, applied LDA analysis to identify topics, clustered sections by topics, and assigned latent weightings to the original sections by topics.

**The Text.** We analyzed a preprint (Bishop, 2018) of CSAS (Bishop et al., 2019), provided to us by the author in digital form as a series of PDFs with a compressed size of 15.8 MB. This preprint is divided into 29 chapters and 753 sections. We processed each section as an independent document.

**Text Processing.** Referencing the WordNet data source (Miller, 1995), we processed each document in four steps: (1) Parse and tokenize text. (2) Remove short words (less than four characters). (3) Identify English stop words (e.g., "if," "the," "and.") (4) In a second pass, find any words not in WordNet.

**Identifying Topics.** We used the Gensim program (Řehůřek and Sojka, 2010) to identify the most relevant topics based on LDA analysis, setting a limit of 12 distinct topics per document.

**Clustering Sections by Topics.** To explore relationships among sections, we performed a cluster analysis of the original sections by similarity of their latent topics. Specifically, using the Cytoscape tool (Ono), for each topic, for each section we computed an LDA-derived probability that the section deals with the topic.

**Latent Weighting of Sections by Topics.** From our LDA analysis, we tagged each section with a list of topics that match the section, together with the strength of the match. Building on this analysis, and given any set of topics each scored by importance, we can weight each section by the specified topics. Doing so is useful in creating "latent table of contents" for specified learning objectives. As a simple example, we computed such a weighting for each section, by counting the number of topics in the section.

## 3 Results

As summarized in Figures 1–2 and Table 1, our LDA analysis of CSAS produced 887 unique topics from the original 29 chapters and 753 sections. Figure 1 shows the 15 most frequent topics. For example, the five most frequent topics are "system," "security," "user," "access," and "requirement," which match 167, 69, 53, 47, and 43 of the sections, respectively. Most sections matched at most five topics.

Figure 2 shows clusters of sections, sorted by cluster size. Each cluster is laid out around a topic (target) to which the sections (sources) map. For each topic and section, LDA computes a probability measure of how strongly the section maps to the topic.

Table 1 gives an example of a latent table of contents, based on the simplistic criteria of maximizing the number of topics. Specifically, the right column of Table 1 lists in order the five sections that match the most of the 887 topics. By comparison, the left column lists the first five sections of CSAS.

## 4 Discussion

We discuss a variety of issues from our study, including possible applications, computer work, alternate approaches, and limitations of our study.

**Applications.** The hidden semantic structure uncovered by our LDA analysis promises to be a useful foundation on which to build tools to help guide learners through CSAS, in ways that are sensitive to their needs and objectives. Jill Watson (Goel and Polepeddi) describes an intelligent tutor with some attractive properties that might be incorporated into an interactive table of contents built from our LDA analysis. In Section 5, we provide initial thoughts on how an interactive guide could direct a learner toward a specified learning objective.

Table 1: An example of a latent table of contents for CSAS. The left column lists the first five sections of CSAS. The right column lists the five sections that match the most topics. The integers in parentheses give the number of matching topics.

| **Original Contents** | **Weighted Latent Contents** |
|---|---|
| 1.1 The Basic Components (4) | 1.7.2 People Problems (8) |
| 1.1.1 Confidentiality (3) | 5.4.2 McLean's System Z (8) |
| 1.1.2 Integrity and Mechanism (2) | 4.3 The Role of Trust (8) |
| 1.1.3 Availability of Security (3) | 5.2.4.1 The get-read Rule (7) |
| 1.2 Threats (5) | 5.2.1 Informal Description (5) |

**Computer Work.** The most time-consuming aspect of our experimental work was dealing with a myriad of low-level issues in processing the raw source files. We used the Python pdftotext package to process the PDF files. However, we needed to write a regular expression to detect the string, "Y/z ∈", which caused the parser to fail. When processing complex PDFs, to reduce programming effort, it is helpful to use well tested software packages.

**Alternative Approaches.** We chose to use LDA rather than *Latent Semantic Indexing (LSI)* (Papadimitriou et al., 1998) or lda2vec (Moody, 2016). LSI is poorly suited because we analyzed CSAS at the section versus paragraph level, and many sections deal with more than one topic. The lda2vec approach is poorly suited because we analyzed only one textbook.

**Limitations.** A possible limitation of our approach is that we do not consider the author's original intent. Also, a few topics (e.g., "system") appear with

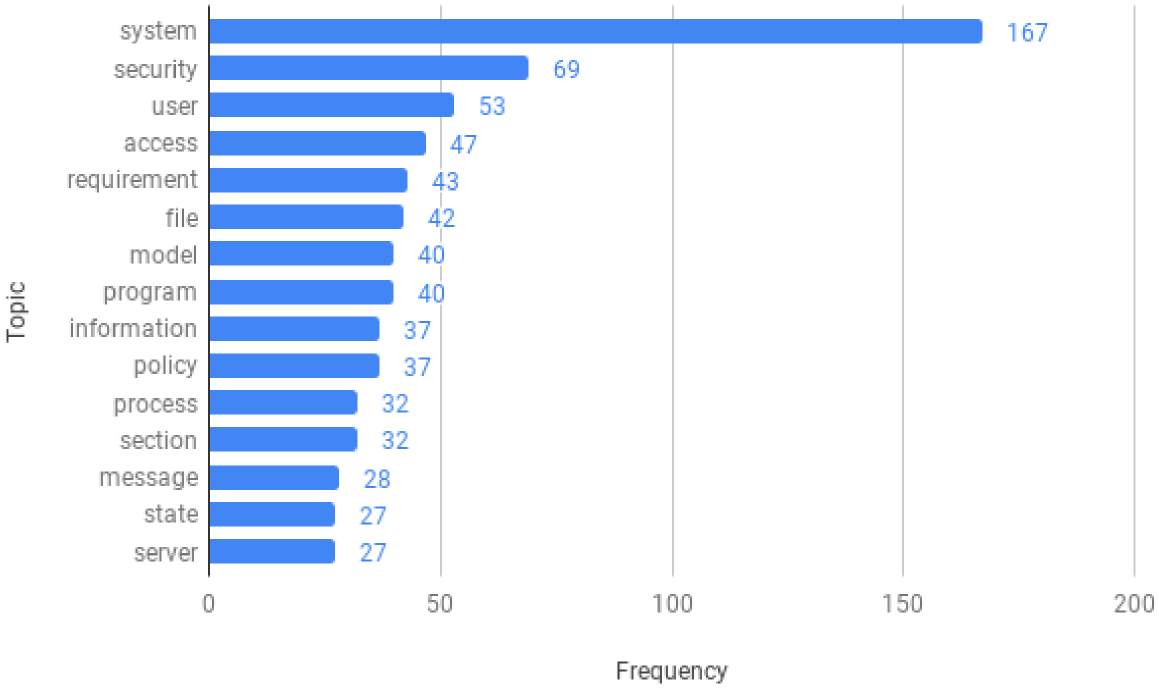

Figure 1: The 15 most frequent latent topics found by LDA in CSAS

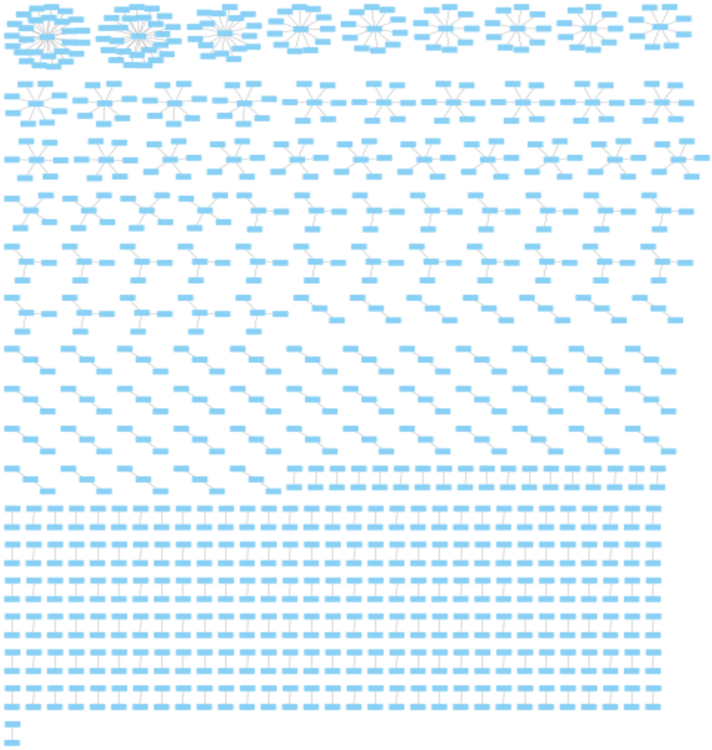

Figure 2: Clusters of original sections by latent topics found in CSAS. This figure shows a sequence of subgraphs sorted by size. At the center of each subgraph is a topic. Connected to this center are the sections that are related to the topic, as measured by a probability computed by LDA.

very high frequency, which might cause overfitting of the model.

## 5 Building an Interactive Guide

We briefly outline our preliminary ideas of how to build an interactive tool that, building on our LDA analysis, guides a reader toward their learning objectives. The tool accepts an arbitrary query and returns a list of sections that address the query. Such a tool might also be useful to authors to gain new insights about the structure of the book they are writing or revising.

The tool might work as follows. It depends on the LDA analysis, which includes a set of sections tagged by topics and their associated probabilities, and processed by a Bayesian classifier. Upon receiving a query from the user, the tool uses the classifier to determine the relevant topics and their weights in the query. These topics are then used to produce a list of sections, selected by the relevant topics and their weights.

## 6 Previous Work

Alghamdi and Alfalqi (Alghamdi and Alfalqi, 2015) survey several text-mining and topic-labeling algorithms, including LDA.

Using texts from Wikipedia, Hoffman et al. (Hoffman et al., 2010), Haruechaiyasak and Damrongrat (Lancichinetti et al., 2014), and Lancichinetti et al. (Haruechaiyasak and Damrongrat, 2008) optimize topic modeling algorithms for machine performance.

Much of the previous work on topic modeling studies document repositories of multi-authored texts on diverse topics from Twitter (Twitter) and Wikipedia (Wikipedia contributors). For example, Tong and Zhang (Tong and Zhang, 2016) analyze data sets from Twitter and Wikipedia to improve text mining and labeling of related topics. Tong also analyzes a single-authored series of tweets by CEO Tim Cook (account, 2020) on business. By contrast, we focus on creating new learning paths through a single-authored scholarly work.

As studied by Kling and Star (Kling and Star, 1998), a *learning path* is a sequence of activities that guide a learner to gain a specified learning goal Kling and Star (Kling and Star, 1998). Along this path, to assess the learner's understanding, it is helpful to ask the learner questions (see Haddi et al. (Haddi et al., 2008)). Online platforms—such as Lynda (lynda.com), Udemy (udemy.com), and Udacity (udacity.com)—offer learning paths for a variety of topics.

## 7 Conclusion

Using LDA methods that could be applied to any textbook, we performed a topic analysis of CSAS. Next, we plan to build on this initial step to create and assess an interactive tool that can create a personalized learning path to guide any reader through the textbook.

## Acknowledgments

We thank Matt Bishop for providing the source files for a preprint edition of CSAS. Sherman and Schep-onik were supported in part by the National Science Foundation under SFS grants DGE-1241576, 1753681, and 1819521, and by the U.S. Department of Defense under CySP grants H98230-17-1-0387, H98230-18-1-0321, and H98230-19-1-0308.

## References

Tim CookVerified account. 2020. Tim cook (@tim$_{c}ook$).

Rubayyi Alghamdi and Khalid Alfalqi. 2015. A survey of topic modeling in text mining. *International Journal of Advanced Computer Science and Applications*, 6.

Matt Bishop. 2018. Computer Security: Art and Science. Preprint in digital form.

Matt Bishop, Elisabeth Sullivan, and Michelle Ruppel. 2019. *Computer Security: Art and Science*. Addison-Wesley.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Ashok K Goel and Lalith Polepeddi. Jill watson: A virtual teaching assistant for online education.

Adil Haddi, Mohammed Ramdani, and M. Bellafkih. 2008. The suitable learning path for a learner. Information Processing and Management of Uncertainty in Knowledge Based Systems,.

Choochart Haruechaiyasak and Chaianun Damrongrat. 2008. Article recommendation based on a topic model for wikipedia selection for schools. pages 339–342.

Matthew Hoffman, Francis R. Bach, and David M. Blei. 2010. Online learning for latent dirichlet allocation. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 856–864. Curran Associates, Inc.

Rob Kling and Susan Leigh Star. 1998. Human centered systems in the perspective of organizational and social informatics. *ACM SIGCAS Computers and Society*, 28(1):22–29.

Andrea Lancichinetti, M. Irmak Sirer, Jane X. Wang, Daniel Acuna, Konrad Körding, and Luís A. Nunes Amaral. 2014. A high-reproducibility and high-accuracy method for automated topic classification.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Christopher E. Moody. 2016. Mixing dirichlet topic models and word embeddings to make lda2vec. *CoRR*, abs/1605.02019.

Keiichiro Ono. Cytoscape. [Online; accessed 25-May-2020].

Christos H. Papadimitriou, Hisao Tamaki, Prabhakar Raghavan, and Santosh Vempala. 1998. Latent semantic indexing: A probabilistic analysis. In *Proc. of Symposium on Principles of Database Systems (PODS)*, pages 159–168. ACM Press. Early work to probabilistically analyse the mechanism behind LSA.

Zhou Tong and Haiyi Zhang. 2016. A text mining research based on LDA topic modelling. volume 6, pages 201–210.

Twitter. Twitter. [Online; accessed 30-May-2020].

Wikipedia contributors. Wikipedia. [Online; accessed 30-May-2020].

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. pages 45–50.

*Submitted to EMNLP 2020 (June 3, 2020).*