Current State and Modeling of Research Topics in Cybersecurity and Data Science

Tamir Bechor and Bill Jung

Center for Information Systems & Technology, Claremont Graduate University

Abstract

Arguably, the two domains closely related to information technology recently gaining the most attention are 'cybersecurity' and 'data science'. Yet, the intersection of both domains often faces the conundrum of discussions intermingled with ill-understood concepts and terminologies. A topic model is desired to illuminate significant concepts and terminologies, straddling in cybersecurity and data science. Also, the hope exists to knowledge-discover under-researched topics and concepts, yet deserving more attention for the intersection crossing both domains. Motivated by these, then retaining most of the already accepted IMCIC (the International Multi-Conference on Complexity, Informatics, and Cybernetics) 2019 conference paper's content and supplementing it with implicit design activities while conducting the research, this study attempts to take on a challenge to model cybersecurity and data science topics clustered with significant concepts and terminologies, grounded on a text-mining approach based on the recent scholarly articles published between 2012 and 2018. As the means to the end of modeling topic clusters, the research is approached with a text-mining technique, comprised of key-phrases extraction, topic modeling, and visualization. The trained LDA Model in the research analyzed and generated significant terms from the text-corpus from 48 articles and found that six latent topic clusters comprised the key terms. Afterwards, the researchers labeled the six topic clusters for future cybersecurity and data science researchers as follows: Advanced/Unseen Attack Detection, Contextual Cybersecurity, Cybersecurity Applied Domain, Data-Driven Adversary, Power System in Cybersecurity, and Vulnerability Management. The subsequent qualitative evaluation of the articles found the LDA Model supplied the six topic clusters in unveiling latent concepts and terminologies in cybersecurity and data science to enlighten both domains. The main contribution of this research is the identification of key concepts in the topic clusters and text-mining key-phrases from the recent scholarly articles focusing on cybersecurity and data science. By undertaking this research, this study aims to advance the fields of cybersecurity and data science. Besides the main contribution, the additional research contributions are as follows: First, the topic modeling approached using text-mining makes the cybersecurity domain unearth the terminologies that make IST (Information Systems and Technology) researchers investigate further. Secondly, using the result of the study's analysis, IST researchers can decide terms of interest and further investigate the articles that supplied the terms.

Keywords: *Cybersecurity, Data Science, Topic Modeling, Text Mining, Research*

1. Introduction

1.1 General Background:

Like many practical domains, cybersecurity is seeing ever-increasing use of data science, such as machine learning (ML), data mining (DM), and artificial intelligence (AI). As an exemplar, Chen et al. (Chen, Chiang, & Storey, 2012) summarized the

applications, data, analytics, and impacts of "BI&A (Business Intelligence and Application)" in the security and public safety domains.

Both cybersecurity and data science are monumental in terms of significance and popularity, respectively. Armerding (2017) said "over the past decade," cybersecurity has become as important as "military or law enforcement security". Related to such claim, former U.S. President Barack Obama stressed that "cybersecurity is one of the most important challenges we [the U.S.] face as a Nation, and for more than seven years he has acted comprehensively to confront that challenge" (The White House Office of the Press Secretary, 2016). Then, he put effort into action by "directing his Administration to implement a Cybersecurity National Action Plan (CNAP)" (The White House Office of the Press Secretary, 2016). Moreover, the EU (European Union) consider "the [cyber-] security and stability of the net, as well as the integrity of data flows" tremendously significant, as "the digital age" provides enormous benefits in "wealth, knowledge and freedom" (European Union, 2018).

On the other hand, "defined as an interdisciplinary field in which processes and systems are used to extract knowledge or insights from data" (Goulart, 2016), data science is growing huge popularity as firms are recognizing its potential and impacts to their operations (Goulart, 2016). If job demand equates popularity, the popularity of data science can be gauged by the immense demand for data scientists, as Davenport and Patil (2012) calls "Data Scientist: The Sexiest Job of the 21st Century".

However, there are three potential issues to consider. First, the cybersecurity community needs to understand concepts and terminologies of data science applied in the domain. Secondly, both domains would want to avoid the inadvertence of overlooking significant concepts. Lastly, because popular terms tend to attract more attention, both need to circumvent lost opportunities to the less popular constructs worth another looks. Due to these, the community needs to shed light on topic models, projecting significant, related concepts. This will render to the community a summary view of most researched topics or phenomena associated with both domains from recent scholarly literature. It will also become a potential seed to guide information systems and technology (IST) researchers for future research and to ultimately enlighten them to contribute to the existing body of knowledge across both domains.

Actuated by the above, the purpose of this research is to model topics of cybersecurity and data science clustered with significant concepts and terminologies discovered using a text-mining method based on recent scholarly articles published between 2012 and 2018.

Most of this paper's content is based on our IMCIC (the International Multi-Conference on Complexity, Informatics, and Cybernetics) 2019 conference paper and is later supplemented with implicit design activities while conducting the research.

1.2 Theoretical Background: Latent Dirichlet Allocation (LDA) Method

Blei et al. argued the necessity of considering mixture models representing words and documents' exchangeability while extending the de Finetti theorem in that "any

collection of exchangeable random variables has a representation as a mixture distribution - in general an infinite mixture" (Blei, Ng, & Jordan, 2003). Then, they demonstrated to "capture significant intra-document statistical structure via the mixing distribution" (Blei et al., 2003). Furthermore, they argued that, while their paper concentrated on "bag-of-words", the LDA methods were usable to larger bodies of text, such as paragraphs (Blei et al., 2003).

The current paper's research selects LDA as the approach, instead of other methods, for reducing dimensionality of text collections in topic modeling form because of its simplicity and "useful inferential machinery in domains involving multiple levels of structure" (Blei et al., 2003), such as the text-corpus of the current research.

1.3 Topic Modeling

As the dimensionality of applied concepts and terminologies from data science increases and becomes more complex as applied in the cybersecurity, topic modeling produces profound benefits. Blei (2012) reasoned, with more available information, finding and discovering of needed information become harder and also argued for new devices in organizing, searching, and comprehending enormous information volumes. Also, Blei (2012) summarized topic modeling as approaches to organize, understand, search, and summarize a large corpus of digital texts automatically; additionally, this approach can discover the *hidden* themes pervading the collection. In explaining Topic Models Vs. Unstructured Data, Anthes (2010) posited topic models provide potent approaches in exploring and understanding otherwise disorderly information and in discovering latent structures in documents and laying down relations among them.

With the urgency of topic models of cybersecurity and data science and the aforesaid topic modeling benefits, by conducting this scientific research approach using text-mining, we aim to strengthen the aforementioned justifications for research and to contribute to the body of knowledge.

1.4 Research Problems:

This research ultimately aims to address the following primary question:

- In recent scholarly articles on the topic of 'cybersecurity' and 'data science' published between 2012 and 2018, what have been the significant terminologies and other related nomenclature most frequently mentioned around these terminologies?

With the above primary research question raised, the secondary research questions of the current study are as follows:

- How distinguishable are clusters from the topic modeling result? Are they clearly separable, or do they considerably overlap?
- How reliable is the result of document-clustering into the topic models of cybersecurity and data science?

The subsequent organization of this paper is as follows: Section 2 reviews the relevant literature using topic modeling approaches in the cybersecurity and data

science domains. Then, section 3 describes the research methods, followed by section 4 describing the research results. After these results are described, the six topic models resulted from the analysis are evaluated in section 5. Then, section 6 discusses the research implications, and section 7 elucidates the implicit design activities of the research. Finally, the paper ends with a conclusion in section 7.

2. Literature Review

Through the search engine Google Scholar ("Google Scholar," n.d.), we searched relevant literature using the following advanced search terms:

- Entered cybersecurity or cyber security for the field, all of the words
- Entered "topic modeling" for the field, with the exact phrase
- Selected the choice, "anywhere in the article", for the field "where my words occur"
- Entered 2012 and 2018 for the field: "Return articles date between"

After the search hits, we skimmed the title, abstracts, and sections of the articles and selected twelve suitable studies for the literature review.

In chronological order of publication year, Table 1 below lists and summarizes the reviewed articles and breaks down their topic modeling approach, key topics researched, and gap analysis comparing the articles in question to the existing research.

Table 1. Summary of Literature Review and Comparison with the Research in the Current Article

Article	Topic Modeling Approach	Key Topic(s) Researched	Difference Compared to Current Research
Das, Sarkani, and Mazzuchi (2012)	Quantitative security risk assessment model using vulnerability scanners and the impact score and frequency values based on the empirical data derived from NVD	Exploration of a software product evaluation method	Methodological difference in topic modeling
Shuai, Li, Li, Zhang, and Tang (2013)	WL-LDA for better obtainment of results via vector space generation on themes and HT-SVM for better leveraging of the prior knowledge of vulnerability distribution	Automated classification of vulnerability through ML	Key topical difference; Methodological difference in topic modeling – using WL-LDA and HT-SVM to extend LDA
Huang, Kalbarczyk, and Nicol (2014)	LDA to knowledge-discover from big data	Intrusion detection	Key topical difference
Lau, Xia, and Ye (2014)	The CS Gibbs sampling algorithm to apply the probabilistic generative model based on LDA	Cybercriminal networks from online social media	Methodological difference in topic modeling
Aswani, Cronin, Liu, and Zhao (2015)	LDA to cluster topics related with IP address via SSH authentication logs	Classifying SSH logs to identify and differentiate brute-force attackers from normal users	Key topical difference
Samtani, Chinn, and Chen (2015)	LDA as the main method to extract topic clusters and	Hacker assets, such as source code postings,	Key topical difference

	to understand the hacker assets	tutorial postings, and postings with attachments	
Sundarkumar, Ravi, Nwogu, and Govindaraju (2015)	LDA feature selection and DM algorithms	Detection of malware	Key topical difference; Methodological difference in topic modeling – the use of additional DM algorithms
Fang et al. (2016)	LDA to cluster toipcs, DTM to discover trending topics, and ATM to identify the key hackers in each topic cluster	Exploration of key hackers and cyber threats in Chinese hacker communities	Key topical difference; Methodological difference in topic modeling – using DTM and ATM to extend LDA
Lee, Sung, and Kim (2016)	LDA to analyze topic model of information security issues of Korea, the US, and China	Analysis of information security awareness	Key topical difference
Li, Yin, and Chen (2016)	Nonparametric supervised topic model (NSTM)	Identification of high quality carding services in the supply chain of the underground economy and adapting the heterogeneity and precariousness of cybercriminals' customer reviews	Methodological difference in topic modeling
Samtani, Chinn, Larson, and Chen (2016)	LDA to identify topic clusters of hacker code from online hacker forums	Cyberthreat intelligence and malware analysis	Key topical difference
Kolini and Janczewski (2017)	Clustering, topic modeling, and LDA algorithm to find comparison and contrast among the NCSs and latent topics discovered in the NCSs	The 60 national cybersecurity strategies NCSs to compare and contrast among the NCSs and implicit topics found	Key topical difference

Note. The key topical difference in the above table means the main topics of research were different in comparison to the current research.

In developing the "quantitative security risk assessment model", while taking a different topic modeling approach than the current study, Das et al. (2012) gathered the empirical data from the National Vulnerability Database (NVD) ("NVD - Home," n.d.) to derive the "impact score, exploitability score, and frequency value", clustered the vulnerabilities via topic modeling, and then labeled the topic groups based on "vulnerability scanner output categories". Then, to achieve the research goal of automated classification of vulnerability through ML, while using NVD vulnerability data, Shuai et al. (2013) introduced "word location information" into the LDA model. However, Shuai et al. (2013) and the current study differ in both key topics and topic modeling approaches. In detecting intrusion, Huang et al. (2014) proposed a "hybrid approach to knowledge discovery from big data" utilizing LDA while they researched a different key topic than the current study. Afterwards, while approached with a different topic modeling method in "mining cybercriminal networks from online social media", Lau et al. (2014) developed a "weakly supervised" method, corroborated by a "probabilistic generative model". Later, Aswani et al. (2015) provided evidence demonstrating the use of LDA in classifying and "topic modeling of SSH [Secure Shell] logs" ("SSH (Secure Shell) Home Page | SSH.COM," n.d.) to detect and distinguish brute-force attacks from legitimate users. Similarly, Samtani et al. (2015) argued that studying "hacker assets" could assist naming cyber-attack tools, obtain implementation, and use knowledge of the assets, and coordinate tools in constructive ways. Compared

to the current study, both Aswani et al. (2015) and Samtani et al. (2015) had differing key topics in their studies.

Then, with both key topical and methodological differences compared to the current study, based on Application Programming Interface (API) call sequences, Sundarkumar et al. (2015) proposed a method employing topic model based on LDA as “feature selection method” to detect malware. Later, taking three different topic modeling approaches, Fang et al. (2016) clustered five topics and justified “Dynamic Topic Model” and “Author Topic Model” broadening LDA to determine each hacker’s topic distribution. Then, with a different key topic than the current study, Lee et al. (2016) approached its research by LDA-based topic modeling using Twitter data and augmented the research by conducting a sentiment analysis. Subsequently, proposing a system in ”identifying high quality carding services in underground economy”, Li et al. (2016) designed a “nonparametric supervised topic model”, a different topic modeling methodology than the current study. Then, Samtani et al. (2016) adapted LDA to discover topic clusters of hacker code from the online forums and to have insight on topic clusters of “attachment postings”. Afterwards, Kolini and Janczewski (2017) surveyed 60 “national cybersecurity strategies”, developed during 2003 and 2016, and used clustering and topic modeling. Both Samtani et al. (2016) and Kolini and Janczewski (2017) had key topical differences compared to the current study.

According to the literature reviewed, the preponderant number of research used LDA as the topic modeling approach, with varying key topic discussions. However, the literature review indicates that there has been no attempted study to data-mine recent scholarly articles with main discussion topics of ‘cybersecurity’ and ‘data science’. Therefore, we believe this is the first attempt through systematic research to elucidate latent themes of cybersecurity and data science from the recent scholarly literature in the form of topic modeling using LDA.

3. Methods

We searched scholarly articles published between 2012 and 2018 with the two main topics: 'cybersecurity' and 'data science'. Initially, few relevant articles were found with the search terms "cybersecurity" and "data science" via widely-known databases, such as ACM Digital Library, Web of Science, and ABI/Inform. Then, we searched Google Scholar (“Google Scholar,” n.d.) using the two terminologies and after finding relevant articles we searched more using a snowball approach. At the end of the snowball search process, we found a total of 50 scholarly articles. However, after validating the relevancy, we excluded two articles, as we mistakenly included one article with its publication year out of range and questioned the fitness of the other article for limited contribution to our topic modeling effort. Therefore, a total of 48 articles became the subsequent text-mining's sources.

The Appendix A lists 48 scholarly articles found through the above search process and then subsequently text-mined; these articles ultimately become the current research's corpus.

To provide the readers the conceptual roadmap of the research, Figure 1 below presents the overall process flow of research methods.

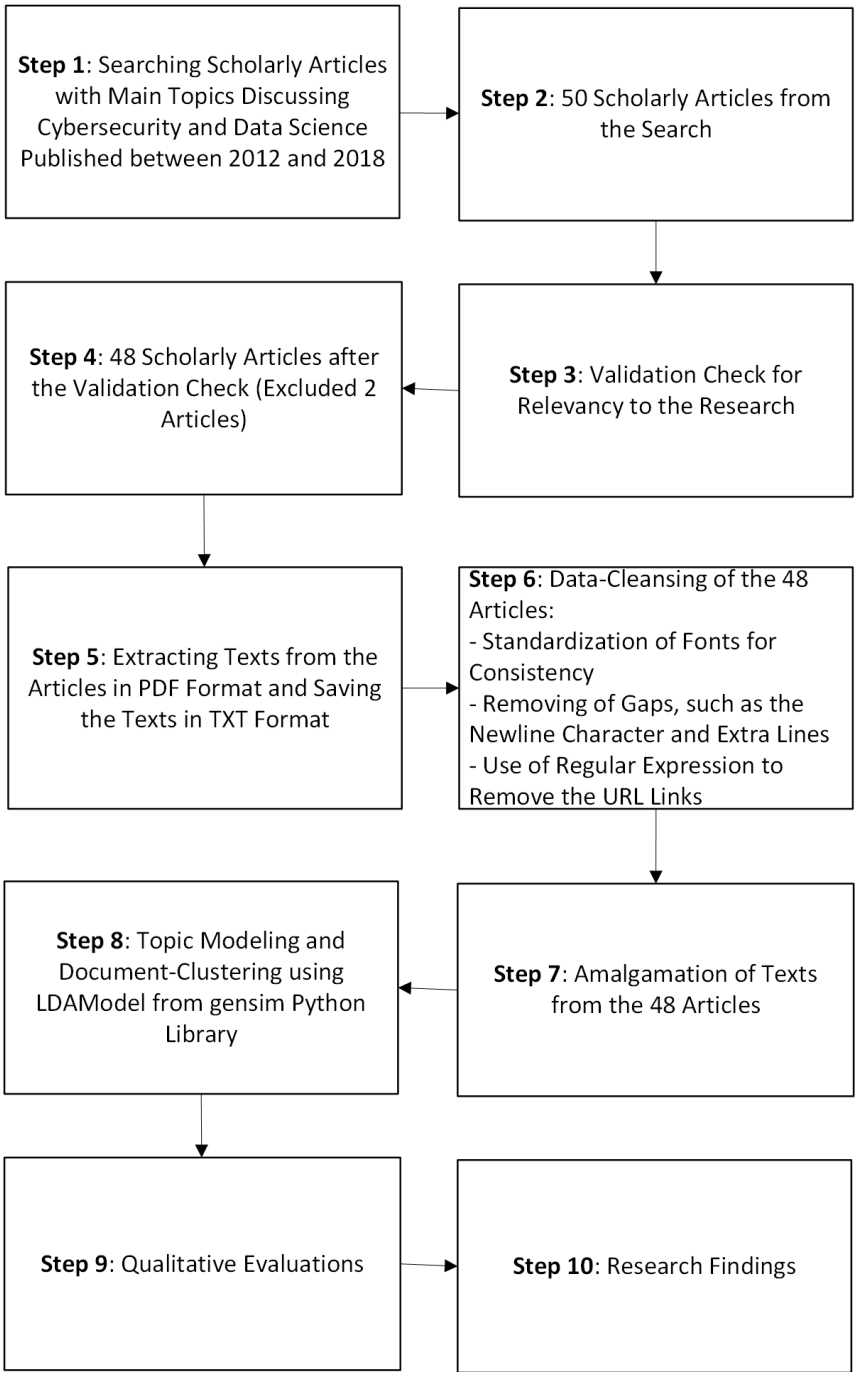


Figure 1. Overall flow of research methods for the current study. The topic model was trained to result most separable topic clusters.

Figure 1: Flowchart of Overall Research Methods

3.1 Text-Mining with Key-phrases Extraction, Topic Modeling, and Document-clustering

After gathering the 48 scholarly articles in PDF format, we used a software converter, PDFMiner (Shinyama, 2014) to extract and convert texts from the PDFs into plain text (.TXT) to process and text-mine the text in Python 2.71 environment. Then, we pre-processed the text data from the articles as described in Figure 1.

We referenced the Python notebook's code example featured at Kaggle (Yasmin, 2017) to run the text-mining with topic modeling, document-clustering, and visualization. We prepared and stored the text data from the plain text (.TXT) files, consisting of the 48 articles' titles, author(s), and main text-corpus, into a DataFrame of pandas Python library (The pandas project, 2017). Then we tokenized the corpus, removed numbers, lemmatized the words in the corpus, computed bigrams and trigrams, removed rare and common tokens, and lastly vectorized the text data. To use and train the LDAModel of Gensim ("gensim," n.d.), the following parameters were set:

- Number of topics: 6
- Chunk size (size of the documents looked at every pass): 10
- Passes: 50 (number of passes through documents)
- Iterations: 400

These parameters were chosen to train Gensim's LDAModel after experimenting with varying numbers of topics, ranging from 3 to 10 topics; 6 seemed to separate the topics well, as a too small number, such as 3, resulted in clusters of too small numbers while a too big number, such as 10, resulted in clusters of too many numbers with the topic clusters overlapping with one another (note: the descriptions of the parameters are from the Python notebook (Yasmin, 2017)).

As the Python notebook demonstrated (Yasmin, 2017), we used pyLDAvis (Mabey, 2015/2018) to visualize the results from the topic modeling. The results from the topic modeling method are discussed in detail in the Results section.

4. Results

4.1 Text-Mining with Key-phrases Extraction, Topic Modeling, and Document-clustering

After the training of Gensim's LDAModel with the aforementioned training parameters, the model analyzed and generated significant terms from the text-corpus. As the model was unsupervised and generated only numerical labels for each topic, we reviewed each of the six topics and provided labels to each as follows:

- Advanced/Unseen Attack Detection
- Contextual Cybersecurity
- Cybersecurity Applied Domain
- Data-Driven Adversary
- Power System in Cybersecurity
- Vulnerability Management

These labels were determined based on the content analysis of the most frequent terms in each cluster. After the following analysis — the sub-topics and the corresponding frequencies resulted from Gensim's LDAModel in the Appendix D and the most salient terms in the pyLDAvis visualization in the Appendix C — and then via internal discussions between the researchers, we finalized naming the labels. After labeling the six topics, we quantified the labels to see which topics were most prevalent and the percent of each topic's tokens. Table 2 below lists the six topics in the percent of tokens' order.

Table 2. Six Topic Clusters and Percent of Tokens

Topic	Percent of Tokens
Advanced/Unseen Attack Detection	22.9%
Contextual Cybersecurity	19.9%
Cybersecurity Applied Domain	18.5%
Data-Driven Adversary	11.7%
Power System in Cybersecurity	7.9%
Vulnerability Management	19%

Note. The topic Advanced/Unseen Attack Detection had the largest size with 22.9% of total tokens, followed by Contextual Cybersecurity (19.9%), Vulnerability Management (19%), and Cybersecurity Applied Domain (18.5%). As the percentages revealed, the proportions of the four aforementioned topics were similar around 20%. The rest of the topics, Data-Driven Adversary (11.7%) and Power System in Cybersecurity (7.9%), combined made about another 20%. Thus, it could be stated that the main topics of the corpus of the 48 scholarly articles with cybersecurity and data science published between 2012 and 2018 in this research were evenly spread and clustered around 5 topics, with Data-Driven Adversary and Power System in Cybersecurity conceptually combined into one topic.

The Appendix B is a pie-chart depicting the above six clusters and their percent of tokens. For further analysis of each of the six topics, we also created a table of the terms and their frequencies in the Appendix D.

Brief analysis of the result of the topic modeling is provided in the alphabetical order of topic names as follows:

Topic 1: Advanced/Unseen Attack Detection: The cluster with 22.9% of total tokens was not obvious to label initially. However, after examining the terms in the topic and also inspecting the abstracts of the articles in the cluster, we determined this cluster's label.

Topic 2: Contextual Cybersecurity: The cluster with 19.9% of total tokens did not initially suggest an obvious label. However, concerning cybersecurity, the terms, such as situational_awareness and contextual_information, seemed to suggest 'context' and 'situation' uniquely applied to the cybersecurity settings.

Topic 3: Cybersecurity Applied Domain: This topic comprised 18.5% of total tokens. After inspecting the top ten terms in the cluster, we concluded the range of terms was a diverse mix of applied domains and labeled it as such.

Topic 4: Data-Driven Adversary: With 11.7% of total tokens, this cluster was dominated by similar 'adversary'-associated terms and shown the cluster was about data-driven adversary in the cybersecurity.

Topic 5: Power System in Cybersecurity: This cluster included 7.9% of total tokens. The most salient term, power_system, with the dominant frequency of .041 within the topic, stood out from the rest of the terms and hinted that the topic was about power system in the cybersecurity context.

Topic 6: Vulnerability Management: The cluster contained 19% of total tokens and was predominated by the vulnerability-related terms. Thus, we decided to label this cluster as Vulnerability Management to remediate cyber threats, such as botnets and malware.

Besides generating the salient terms in each of the six topics as discussed, the LDA topic model also generated the overall, top 30 most salient terms from the entire corpus of the 48 scholarly articles and depicted the six clusters using pyLDAvis (Mabey, 2015/2018) in the Appendix C. These terms help summarize most frequent terms in the text-corpus of the current research.

Also, Gensim's LDAModel document-classified the 48 articles into the six topic clusters. This classification result was used as the data source of the qualitative evaluation in section 5. Table 3 below is the result of the document-clustering into the six topic models.

Table 3. Document-clustering of the 48 articles into the six topic models

Topic	Article
Advanced/Unseen Attack Detection	Abt and Baier (2014) Buczak and Guven (2016) Liu et al. (2015) Meidan et al. (2017) Pajouh, Dastghaibyfard, and Hashemi (2017) Symons and Beaver (2012) Yasakethu and Jiang (2013) Zomlot, Chandran, Caragea, and Ou (2013)
Contextual Cybersecurity	Aleroud and Karabatis (2017) Alguliyev and Imamverdiyev (2014) Czejdo, Iannacone, Bridges, Ferragut, and Goodall (2014) Jones, Bridges, Huffer, and Goodall (2015) Mahmood and Afzal (2013) McKenna, Staheli, Fulcher, and Meyer (2016) Mittal, Das, Mulwad, Joshi, and Finin (2016) Noel, Harley, Tam, Limiero, and Share (2016) Singh and Nene (2013) Vinchurkar and Reshamwala (2012) Zamani and Movahedi (2013) Zuech, Khoshgoftaar, and Wald (2015)
Cybersecurity Applied Domain	Benjamin and Chen (2013) Brundage et al. (2018)

	Chen et al. (2012) Georgescu and Smeureanu (2017) Guarino (2013) He, Tian, Shen, and Li (2015) Joseph, Laskov, Roli, Tygar, and Nelson (2013) Li et al. (2016) Thuraisingham (2015) Thuraisingham et al. (2016)
Data-Driven Adversary	Alsheikh, Lin, Niyato, and Tan (2014) Papernot, Carlini, et al. (2016) Papernot, McDaniel, and Goodfellow (2016) Papernot, McDaniel, Sinha, and Wellman (2016)
Power System in Cybersecurity	Abubakar, Chiroma, Muaz, and Ila (2015) Adhikari, Morris, and Pan (2017) Beaver, Borges-Hink, and Buckner (2013) Borges Hink et al. (2014) Esmalifalak, Nam Tuan Nguyen, Rong Zheng, and Zhu Han (2013)
Vulnerability Management	Beaver, Symons, and Gillen (2013) Camastra, Ciaramella, and Staiano (2013) Carlini, Liu, Kos, Erlingsson, and Song (2018) Fan, Ye, and Chen (2016) Gandotra, Bansal, and Sofat (2014) Hou, Saas, Chen, Ye, and Bourlai (2017) Le, Zincir-Heywood, and Heywood (2016) Mayhew, Atighetchi, Adler, and Greenstadt (2015) Stevanovic and Pedersen (2013)

Note. The 48 articles have been ordered by the topic names in alphabetical order. Within each topic, the articles have been ordered by the authors' names in alphabetical order. The researchers labeled each topic's names after examining the salient terms within each topic.

5. Evaluations of The Results

To evaluate the results, we revisited all 48 articles and validated to see whether the text-mining and the categorization results match with their analysis. We assessed each article's fitness to the topic cluster to which it had been categorized and performed qualitative evaluation. Each sub-section below provides analytic evaluations bifurcated into positive and negative facets of the text-mining results within cybersecurity and data science realms. Table 4 provides the 48 articles' evaluation summary.

5.1 Advanced/Unseen Attack Detection

Evaluated towards Positivity: Pajouh et al. (2017) claimed a classification model's good performance in detecting uncommon and sophisticated attack types, while Zomlot et al. (2013) built prediction models to cope with high false-positive rates from the system sensors in detecting intrusion. Similarly, Buczak and Guven (2016) discussed anomaly detection and unseen attacks in a literature survey, whereas Symons and Beaver (2012) claimed an algorithm's performance improvement in detecting unseen network intrusion. Likewise, Liu et al. (2015) sought to forecast an organization's breach chances contingent upon its network attributes, whilst Abt and Baier (2014) discussed previously unobserved malicious events and activities in using synthetic data.

Evaluated towards Negativity: Meidan et al. (2017) applied ML techniques to identify IoT devices on a network, but the study seemed tangential to the current topic. Likewise, Yasakethu and Jiang (2013) focused the SCADA system protection of the Power System topic.

5.2 Contextual Cybersecurity

Evaluated towards Positivity: Mahmood and Afzal (2013) discussed Big Data analytics built on diverse data types, while Alguliyev and Imamverdiyev (2014) discussed heterogeneous datasets and data correlation used in Big Data applications. Also, Czejdo et al. (2014) integrated external data sources in the cybersecurity data warehouse and explored diverse dataset aspects, when Jones et al. (2015) focused on extracting cybersecurity, contextual concepts. Comparably, Zuech et al. (2015) discussed intrusion detection system (IDS) based on heterogeneous types of big data, whilst Mittal et al. (2016) analyzed real-time tweets to gain threat intelligence in temporal, contextual events. Likewise, McKenna et al. (2016) incorporated contextual data elements in a dashboard development. Then, Noel et al. (2016) mapped vulnerabilities into threats in the post-attack forensics, providing a different context array. Similarly, Aleroud and Karabatis (2017) provided a review with the techniques implementing contextual information for intrusion detection.

Evaluated towards Negativity: Singh and Nene (2013) seemed tangential to the current topic, by centering on IDS, while Vinchurkar and Reshamwala (2012)'s main theme was to review IDS and explain related ML terminologies. Also, Zamani and Movahedi (2013) seemed irrelevant to this cluster because of its theme categorizing ML techniques into either "AI-based" and "computational intelligence-based (CI-based)" methods.

5.3 Cybersecurity Applied Domain

Evaluated towards Positivity: He et al. (2015) covered the security aspects of mobile banking, while Guarino (2013) illustrated cybersecurity and data application in the "digital forensics" domain. Correspondingly, Benjamin and Chen (2013) featured the "hacker communities" domain. Moreover, Brundage et al. (2018) illustrated cybersecurity applied domains, while Georgescu and Smeureanu (2017) used the "black hat hackers community" domain. Furthermore, Li et al. (2016) profiled the "key sellers in the underground economy" domain, whilst Chen et al. (2012) presented a research framework and its applications in various domains.

Evaluated towards Negativity: Joseph et al. (2013)'s main theme was about the "adversarial nature" in data, model, and ML, while Thuraisingham et al. (2016) discussed three aspects of the "Science of Cybersecurity" topic. Then, Thuraisingham (2015) concentrated on the issues in big data security and privacy.

5.4 Data-Driven Adversary

Evaluated towards Positivity: Papernot, Carlini, et al. (2016) clearly supported the current cluster, by focusing its discussion on an ML library "in adversarial settings", while Papernot, McDaniel, Sinha, et al. (2016) concentrated on "attacks and defenses" by researching recent findings in "ML security and privacy". Finally, Papernot, McDaniel, and Goodfellow (2016) focused "black-box attacks" leveraging "adversarial examples".

Evaluated towards Negativity: Alsheikh et al. (2014) provided a literature review and emphasized ML approaches in addressing "wireless sensor network" issues more relevant to the Power System in Cybersecurity.

5.5 Power System in Cybersecurity

Evaluated towards Positivity: Adhikari et al. (2017) developed a "cyber-physical test bed" and presented its architecture, while Beaver, Borges-Hink, et al. (2013) evaluated ML approaches to "detect malicious SCADA communications". Also, Borges Hink et al. (2014) discussed the ML approaches for "power system disturbance and cyber-attack discrimination", when Esmalifalak et al. (2013) studied "stealthy false data injection using machine learning in smart grid".

Evaluated towards Negativity: Abubakar et al. (2015) focused on reviewing "current advances" in using cybersecurity "benchmark datasets" related to the cybersecurity domain.

5.6 Vulnerability Management

Evaluated towards Positivity: Stevanovic and Pedersen (2013) discussed the vulnerability issues in detecting "botnet traffic". Correspondingly, Carlini et al. (2018) presented a metric applicable to deep learning models for the vulnerability of the "unintended memorization" and "extraction of secrets", while Mayhew et al. (2015) added values to vulnerability management towards "trustworthiness of documents and actors". Moreover, Camastra et al. (2013) contributed to vulnerability management by studying "the trends of the ML and SC [soft computing] methodologies for ICT [Information and Communication Technology] security".

Evaluated towards Negativity: Beaver, Symons, et al. (2013) seemed limited to "a learning system for discriminating variants of malicious network traffic", while Hou et al. (2017) aimed to detect unknown Android malware. Similarly, Fan et al. (2016) aimed to "detect new malware samples", rather closely linked to the Advanced/Unseen Attack Detection, while Gandotra et al. (2014) mainly argued the families of malware. Finally, Le et al. (2016) used an unsupervised ML technique to data-analyze "unknown traffic to detect botnets".

Overall, the above qualitative review finds a total 69% positivity, meaning the review has agreed 69% that Gensim's LDAModel clustered the articles into the proper topics. The Advanced/Unseen Attack Detection has 75% positivity (six out of eight articles contributing towards positivity), 75% for the Contextual Cybersecurity (nine out of twelve), 70% for the Cybersecurity Applied Domain (seven out of ten), 75% for the Data-Driven Adversary (three out of four), 80% for the Power System in Cybersecurity (four out of five), and 44% for the Vulnerability Management (four out of nine). Thus, compared to the other categories, Gensim's LDAModel seemed inaccurately clustering the articles into the Vulnerability Management, as the model seemed confused with the five studies focused on the Advanced/Unseen Attack Detection. However, with the

overall 69% positivity we find Gensim's LDAModel helps categorize the articles into the six topic clusters of cybersecurity and data science.

Table 4. Evaluation Results of the 48 Articles

Topic	Total Number of Articles	Evaluated towards Positivity	Evaluated towards Negativity	Percent of Positivity
Advanced/Unseen Attack Detection	8	6	2	75%
Contextual Cybersecurity	12	9	3	75%
Cybersecurity Applied Domain	10	7	3	70%
Data-Driven Adversary	4	3	1	75%
Power System in Cybersecurity	5	4	1	80%
Vulnerability Management	9	4	5	44%*
Total	**48**	**33**	**15**	**69%***

Note. Outcomes of qualitative evaluation compared to Gensim's LDAModel's evaluation. "Evaluated towards positivity" means the qualitative review has agreed that Gensim's LDAModel clustered the article into the appropriate topic. In contrast, "evaluated towards negativity" means the qualitative review has *not* agreed that Gensim's LDAModel clustered the article into the appropriate topic. Percent of positivity denotes the number of articles in that topic evaluated towards positivity divided by the total number of articles.

* Rounded up to have no decimal.

6. Discussion

The goal of this research was to model topics of cybersecurity and data science clustered with significant terms and concepts, and the researchers accomplished the goal by the text-mining approach consisted of key-phrases extraction, topic modeling, and visualization.

To answer the primary research question, the researchers searched and collected the 48 scholarly articles published between 2012 and 2018 and then text-mined and analyzed the articles by topic modeling and document-clustering using the LDAModel from the Gensim library ("gensim," n.d.). The findings have been supplied in Table 2 and in the Appendix D. Gensim's LDAModel consequently resulted in the six latent topics, and the appropriate labels for the six clusters were provided, bottomed-up from the sub-topics within each cluster found in Appendix D. Furthermore, the researchers analyzed the topic modeling's result and significant terminologies and provided a qualitative review of the findings.

The result and accompanying analysis of this study also address the two secondary research questions. Regarding the question of the separability, degree of separation, and degree of overlap of clusters from the result, the six clusters in Table 2 were overall well-separated from one another, while, as Appendix D has noted, there are overlaps of some terms appearing in multiple clusters. While Appendix D's notation and Appendix C's visualization help understand this research question, the current research does not provide measurements of the clusters' separations and overlaps. To answer the next question of the reliability of the result, the analysis of the Evaluation reveals that Gensim's LDAModel did not always cluster the source articles into clearly distinguishable topics, particularly for the Vulnerability Management cluster. Some

seem better candidates for labeling with multiple topic clusters. Also, as the review evaluated, some appear misclassified.

One further technical limitation of the research was the topic model document-clustered each article into one topic only. Conversely, the topic model did not support multi-labeling of the articles to the clusters. While multi-labeling may increase the document-clustering's accuracy, it may increase complexity of the labeling and complicate the result's evaluation. Nevertheless, multi-labeling the articles to observe potentially different results is worthwhile.

By providing answers to the aforementioned research questions, this study can now clearly advance the fields of cybersecurity and data science. Regarding this study's contributions, they are twofold. First, the topic modeling approached using text-mining makes the cybersecurity domain unearth the terminologies that make IST researchers investigate further, as Gensim's LDAModel's finding results in the six clusters with the sub-topics of the most frequent terminologies in the selected literature. Thus, the current research's findings become a research seed. Secondly, using the result of the current project's analysis, IST researchers can decide terms of interest and further investigate the articles that supplied the terms. Therefore, the research seed becomes and makes an impact as a guidance for future research direction.

Inspecting each cluster and the sub-topics modeled within the clusters could provide insight worthy of further investigation. For instance, there are six topics, and the ten sub-topics within each topic as shown in Appendix D. Choosing one particular topic, inspecting the sub-topics within the topic, and observing the sub-topics labeled 'Appears under multiple topics' may help the readers link multiple topics and build a model with relations based on the shared sub-topics. Also, we conjecture adding more articles that are not in the 48 articles to the data sources may diversify the concepts discovered and increase opportunities of unearthing concepts deserving more attention, as the current study is limited to the 48 articles.

7. Implicit Mental and Non-Mental Design Activities in the Research

While this study featured the explicit elements that eventually became the sections of the paper -- introduction, literature review, methods, results, evaluations of the results, discussion, and conclusion – however, it also embedded implicit mental and non-mental design activities while conducting the research. These implicit activities are often left out from the final writeup because in general only the explicit elements are required for publishing. Yet, these neglected activities are crucial elements of research and design as they have intimate relationships as shown in Figure 2. Relationship between "Research Design" and "Design Research" (International Institute of Informatics and Cybernetics, IIIC, n.d.).

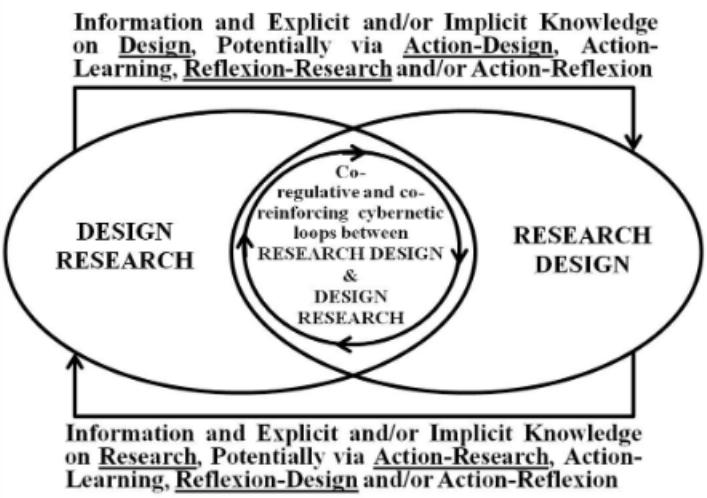


Figure 2: Relationship between "Research Design" and "Design Research" (International Institute of Informatics and Cybernetics, IIIC, n.d.)

In this section, these implicit activities are reflected upon and elaborated as follows.

Table 5. Implicit Mental and Their Corresponding Implicit Non-Mental Activities in the Research

Number	Process Steps in The General Design Cycle	Activity	Implicit Mental Activities	Corresponding Implicit Non-Mental Activities
7.1	Awareness of problem	Finding interesting topics to conduct a research in cybersecurity	What could be good research topics in the cybersecurity domain?	Finding and narrowing down the research topics via a literature survey
7.2	Suggestion	Determining research methods	How to select the research methods?	Machine-learning approach using Natural Language Processing
7.3	Development	Finding databases to search scholarly articles	What are the databases to search the relevant articles?	Consulting a librarian and internal discussions
7.4	Development	Determining the scope and range of the source articles	What is the appropriate time-frame of the recent scholarly articles?	Consulting the scholarly/industry expert
7.5	Development	Performing text-mining	Executing the topic modeling steps (Steps 5 to 8 in Figure 1)	Acquiring the skills required for topic modeling
7.6	Evaluation	Evaluating the results	How can the results from the LDAModel be evaluated?	Revisiting and validating the 48 articles
7.7	Conclusion	Disseminating theories to the body of knowledge	What are the research contributions? -- Reflecting upon the lasting impacts to multiple communities	Discussing a list of contributions internally and making the list explicit in the conclusion

Note. The above activities were sequentially performed while conducting the research.

For each activity above, the implicit mental activities and their corresponding implicit non-mental counterparts are discussed in the frame of Figure 3. 'The General Design Cycle' (Hevner & Chatterjee, 2010) below.

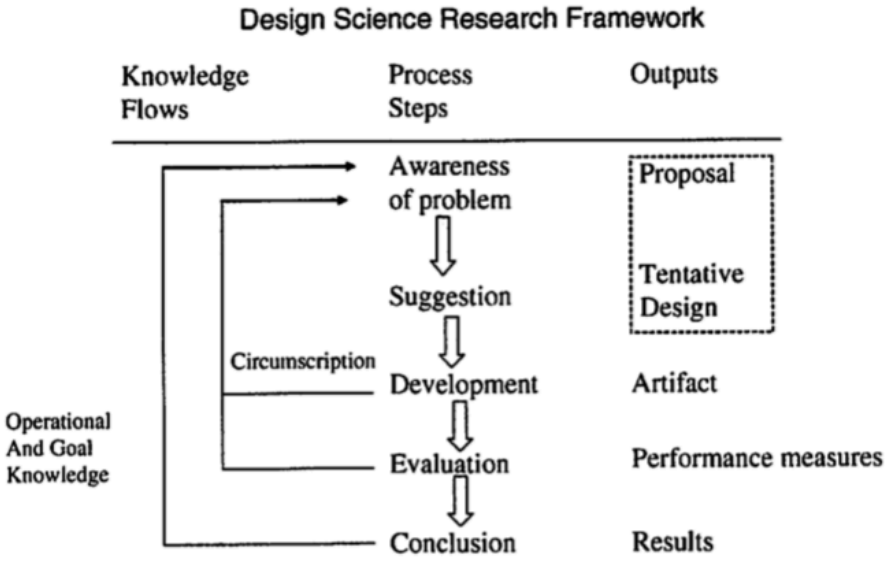


Figure 3: The General Design Cycle
(Hevner & Chatterjee, 2010)

7.1. Finding Interesting Topics to Conduct a Research in Cybersecurity

What could be good research topics in the cybersecurity domain?: As the cybersecurity domain is fast-moving, dynamically changing, the authors thought that the more recent data would be the better source. Also, potential research contribution as the crucial ingredient of good research topics was discussed between the authors. To satisfy that end, the authors agreed that the research needed to be beneficial for both the research and practice, as follows:

- The research: What topics have gained the most attention from the scholarly community?
- The practice: Can these recent scholarly topics suggest the next breakthrough in cybersecurity?

Finding and narrowing down the research topics via a literature survey: The conversation initiated when the co-author – the Ph.D. student – was looking for potential research topics in the cybersecurity domain. The lead author – the advisor of the student -- suggested conducting a systematic literature review on recent scholarly articles in the cybersecurity domain.

7.2. Determining Research Methods

How to select the research methods?: To find good research topics, a widely accepted method, such as the conventional, systematic literature search (Vom Brocke et al., 2009), could be adapted. However, this approach relies on too much manual

human analysis and may prone to errors from such a labor-intensive nature of work. The authors discussed and agreed that an automated way using data science can benefit the study and scale well for a large number of articles.

Machine-learning approach using Natural Language Processing: To automate the process, the authors particularly investigated topic modeling and LDA approach. For the LDA topic modeling, the authors experimented with the Gensim library by importing and running code in the Jupyter Notebook (Project Jupyter, n.d.).

7.3. Finding Databases to Search Scholarly Articles

What are the databases to search the relevant articles?: As texts are mined ultimately from the source articles, the authors discussed to use reputable databases to search scholarly articles.

Consulting a librarian and internal discussions: Subsequently, the authors consulted an experienced librarian, and then she suggested the databases, such as ACM Digital Library, Web of Science, and ABI/Inform. Also, they had internal discussions and evaluated the use of Google Scholar as well.

7.4. Determining the Scope and Range of the Source Articles

What is the appropriate time-frame of the recent scholarly articles?: In the research, the authors selected the scholarly articles published between 2012 and 2018. They discussed and agreed that cybersecurity is a recent phenomenon, thus recent articles could provide more relevant topics. Also, instead of widening the time period, they settled with a smaller window of seven years between 2012 and 2018.

Consulting the scholarly/industry expert: To find such appropriate time-frame, the authors relied on the experience of the lead author who was an industry expert.

7.5. Performing Text-Mining

Executing the topic modeling steps (Steps 5 to 8 in Figure 1): Steps 5 to 8 required skills in coding Python and using the Gensim library.

Acquiring the skills required for topic modeling: The co-author was experienced in Python programming. He acquired the skills to use the Gensim and experimented the topic modeling library before running the text-mining.

7.6. Evaluating the Results

How can the results from the LDAModel be evaluated?: The obtained results from the LDAModel were machine-generated. Ensuing manual analyses were needed to validate the results.

Revisiting and validating the 48 articles: Each of the 48 articles was read and evaluated towards positivity and negativity. Then, the percent of positivity was calculated.

7.7. Disseminating Theories to the Body of Knowledge

What are the research contributions? -- Reflecting upon the lasting impacts to multiple communities: Theories to the research community as well as contributions to the closely-related communities were illuminated.

Discussing a list of contributions internally and making the list explicit in the conclusion: After narrowing down the impacted communities, the authors brainstormed and named the contributions to the research, business, and technology communities.

8. Conclusion

The main contribution of this research project is the identification of key concepts in the topic clusters and text-mining key-phrases from the recent scholarly articles focusing on cybersecurity and data science. The approach is unique because of the application of probabilistic topic modeling (e.g. LDA Model) of most frequent terms from the articles. Also, the identification of the key concepts empowers IST researchers to further survey the areas unearthed.

Regarding contributions to the broader audience, the research contributes to multiple communities:

- Research: Towards achieving the goal of building a theory in the cybersecurity domain, the research has supplied a classification model in theory building, and this becomes a precursor to building a model with defining relationships in theory building process .
- Business: The research presents the logical, scientific topic model, and the outcomes. Professionals can apply these findings to understand the most frequent terms from the research and correlate with the counterparts in the real-world to discover deeper insight.
- Technology: The research has provided a topic modeling approach using text-mining and analytics using a well-received Python library specializing topic modeling ("gensim," n.d.). This method benefits the technology sector by illustrating a sounding approach to discover relevant, frequent terms in two related disciplines.

In this research, we used the popular LDA model (Blei et al., 2003) to perform the topic modeling. We encourage fellow IST researchers to adapt other models to perform topic modeling to see whether outcomes would be different. Also, both cybersecurity and data science are wide-ranging disciplines with numerous sub-topics within each discipline. Relating sub-topics from each of these disciplines makes studies more challenging. Perhaps focusing one topic cluster from the current research, such as Vulnerability Management, would provide IST researchers opportunities to conduct more focused research. This research has a couple implications for future research. First,

the most frequent terms show future researchers the key-phrases in each cluster and enable them to deep-dive into more focused research arenas. Secondly, the documents clustered into the six clusters can guide fellow researchers to conduct focused literature reviews in their pursuing topics and become the seed for their future research.

9. Appendixes

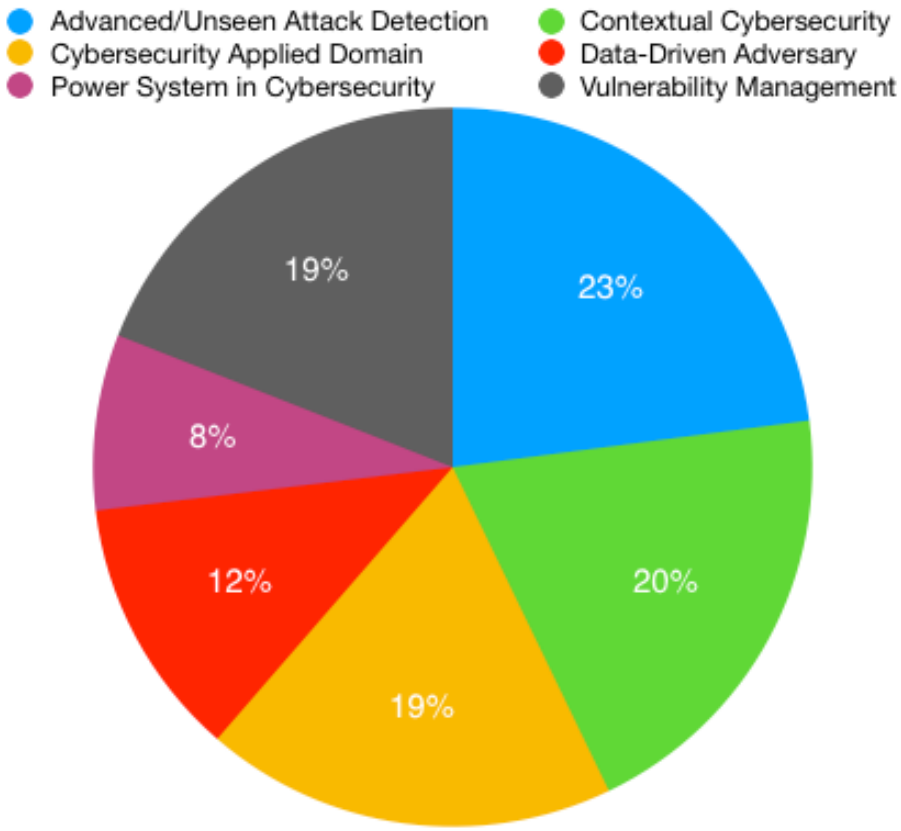
Appendix A. The 48 Scholarly Articles from the Search and Subsequently Text-Mined for Topic Modeling

	Authors	Title	Year
1	Abt and Baier (2014)	A Plea for Utilizing Synthetic Data When Performing Machine Learning-Based Cyber-Security Experiments	2014
2	Abubakar et al. (2015)	A Review of the Advances In Cyber Security Benchmark Datasets for Evaluating Data-Driven-Based Intrusion Detection Systems	2015
3	Adhikari et al. (2017)	WAMS Cyber-Physical Test Bed for Power System, Cybersecurity Study, and Data Mining	2017
4	Aleroud and Karabatis (2017)	Contextual Information Fusion for Intrusion Detection: A Survey and Taxonomy	2017
5	Alguliyev and Imamverdiyev (2014)	Big Data: Big Promises for Information Security	2014
6	Alsheikh et al. (2014)	Machine Learning in Wireless Sensor Networks: Algorithms, Strategies, and Applications	2014
7	Beaver, Borges-Hink, et al. (2013)	An Evaluation of Machine Learning Methods to Detect Malicious SCADA Communications	2013
8	Beaver, Symons, et al. (2013)	A Learning System for Discriminating Variants of Malicious Network Traffic	2013
9	Benjamin and Chen (2013)	Machine Learning for Attack Vector Identification in Malicious Source Code	2013
10	Brundage et al. (2018)	The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation	2018
11	Buczak and Guven (2016)	A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection	2016
12	Camastra et al. (2013)	Machine Learning and Soft Computing for ICT Security: An Overview of Current Trends	2013
13	Carlini et al. (2018)	The Secret Sharer: Measuring Unintended Neural Network Memorization and Extracting Secrets	2018
14	Chen et al. (2012)	Business Intelligence and Analytics: From Big Data to Big Impact	2012
15	Czejdo et al. (2014)	Integration of External Data Sources with Cyber Security Data Warehouse	2014
16	Esmalifalak et al. (2013)	Detecting Stealthy False Data Injection Using Machine Learning in Smart Grid	2014
17	Fan et al. (2016)	Malicious Sequential Pattern Mining for Automatic Malware Detection	2016
18	Gandotra et al. (2014)	Malware Analysis and Classification: A Survey	2014
19	Georgescu and Smeureanu (2017)	Using Ontologies in Cybersecurity Field	2017
20	Guarino (2013)	Digital Forensics as a Big Data Challenge	2013
21	He et al. (2015)	Understanding Mobile Banking Applications' Security Risks through Blog Mining and the Workflow Technology	2015
22	Borges Hink et al. (2014)	Machine Learning for Power System Disturbance and Cyber-Attack Discrimination	2014
23	Hou et al. (2017)	Deep Neural Networks for Automatic Android Malware Detection	2017
24	Jones et al. (2015)	Towards a Relation Extraction Framework for Cyber-Security Concepts	2015
25	Joseph et al. (2013)	Machine Learning Methods for Computer Security	2013
26	Le et al. (2016)	Data Analytics on Network Traffic Flows for Botnet Behaviour Detection	2016

27	Li et al. (2016)	Identifying High Quality Carding Services in Underground Economy Using Nonparametric Supervised Topic Model	2016
28	Liu et al. (2015)	Cloudy with a Chance of Breach: Forecasting Cyber Security Incidents	2015
29	Mahmood and Afzal (2013)	Security Analytics: Big Data Analytics for Cybersecurity: A Review of Trends, Techniques and Tools	2013
30	Mayhew et al. (2015)	Use of Machine Learning in Big Data Analytics for Insider Threat Detection	2015
31	McKenna et al. (2016)	Bubblenet: A Cyber Security Dashboard for Visualizing Patterns	2016
32	Meidan et al. (2017)	Profiliot: A Machine Learning Approach for IoT Device Identification Based on Network Traffic Analysis	2017
33	Mittal et al. (2016)	Cybertwitter: Using Twitter to Generate Alerts for Cybersecurity Threats and Vulnerabilities	2016
34	Noel et al. (2016)	Cygraph: Graph-Based Analytics and Visualization for Cybersecurity	2016
35	Pajouh et al. (2017)	Two-Tier Network Anomaly Detection Model: A Machine Learning Approach	2017
36	Papernot, Carlini, et al. (2016)	Cleverhans V2. 0.0: An Adversarial Machine Learning Library	2016
37	Papernot, McDaniel, and Goodfellow (2016)	Transferability in Machine Learning: From Phenomena to Black-Box Attacks Using Adversarial Samples	2016
38	Papernot, McDaniel, Sinha, et al. (2016)	Sok: Towards the Science of Security and Privacy in Machine Learning	2016
39	Singh and Nene (2013)	A Survey on Machine Learning Techniques for Intrusion Detection Systems	2013
40	Stevanovic and Pedersen (2013)	Machine Learning for Identifying Botnet Network Traffic	2013
41	Symons and Beaver (2012)	Nonparametric Semi-Supervised Learning for Network Intrusion Detection: Combining Performance Improvements with Realistic In-Situ Training	2012
42	Thuraisingham et al. (2016)	A Data Driven Approach for the Science Of Cyber Security: Challenges and Directions	2016
43	Thuraisingham (2015)	Big Data Security and Privacy	2015
44	Vinchurkar and Reshamwala (2012)	A Review of Intrusion Detection System Using Neural Network and Machine Learning	2012
45	Yasakethu and Jiang (2013)	Intrusion Detection via Machine Learning for SCADA System Protection	2013
46	Zamani and Movahedi (2013)	Machine Learning Techniques for Intrusion Detection	2013
47	Zomlot et al. (2013)	Aiding Intrusion Analysis Using Machine Learning	2013
48	Zuech et al. (2015)	Intrusion Detection and Big Heterogeneous Data: A Survey	2015

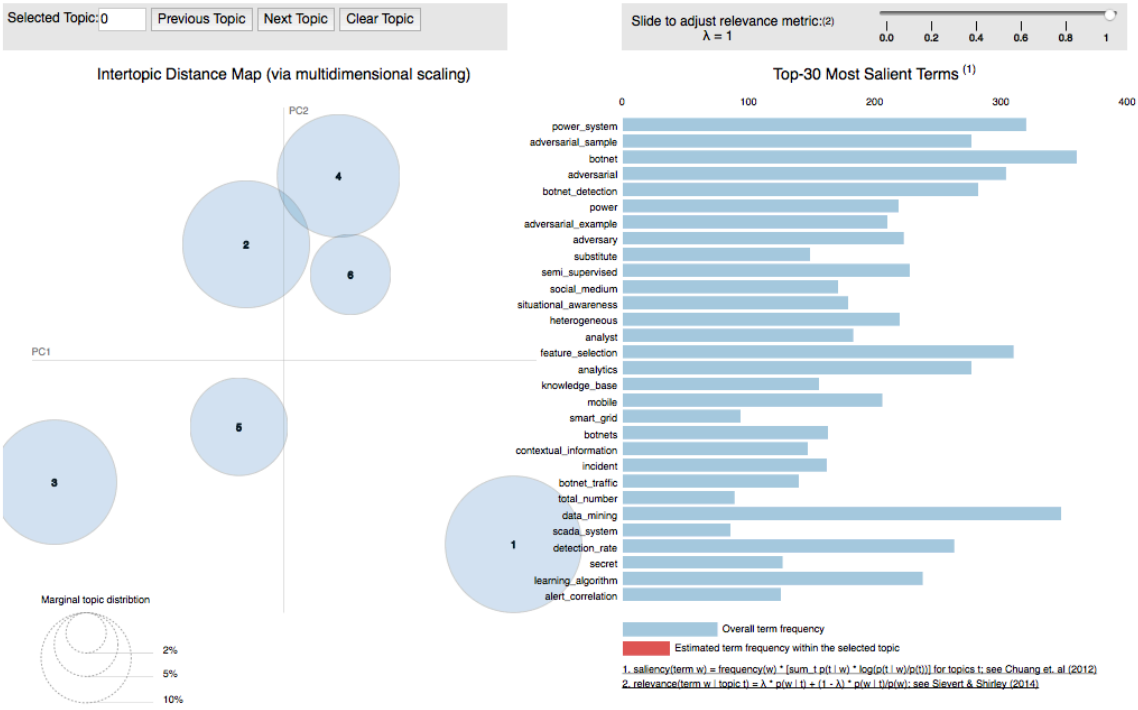
Note. The articles are ordered by author name in ascending order.

Appendix B. 6 Topic Clusters and Percent of Tokens



Note. The topic cluster "Advanced/Unseen Attack Detection" had the largest proportion among the topic clusters with 22.9% of total tokens, followed by Contextual Cybersecurity (19.9%), Vulnerability Management (19%), and Cybersecurity Applied Domain (18.5%). The remaining topics, Data-Driven Adversary (11.7%) and Power System in Cybersecurity (7.9%), combined made approximately another 20%. If the two topics, Data-Driven Adversary and Power System in Cybersecurity, were to merge into one topic, the main topics of the corpus of the 48 scholarly articles with cybersecurity and data science published between 2012 and 2018 in this research would have formed around 5 topics.

Appendix C. Result of pyLDAvis Visualization with the top 30 most salient terms



Note. The main 6 topics are clustered and visualized on the left, while the top 30 most salient terms from the entire corpus of the scholarly articles are depicted on the right.

Appendix D. Topics Modeled with 10 Most Frequent Terms within 6 Topics

Topic: **Advanced/Unseen Attack Detection** (22.9%; 23% in the pie-chart)	Summary: This topic cluster reveals latent terms related attack types that are not seen before, such as semi_supervised (as attacks are unseen before, there is need for review of manual human analysis), false_alarm (the researchers conjecture that there will be much false alarms associated with these types of attacks), synthetic data (data is not of natural origin), unknown attack, and time_series (because attacks are unseen before, collecting time series-based data will be fundamental in detecting these types of attack). Therefore, the researchers label this topic cluster as Advanced/Unseen Attack Detection.	
Sub-Topics	**Frequency**	**Notes**
semi_supervised	0.010	
data_set	0.009	Appears under multiple topics.
malicious_activity	0.008	
incident	0.007	
false_alarm	0.007	
detection_rate	0.006	
synthetic_data	0.006	Appears under multiple topics.
unknown_attack	0.006	
training_testing	0.005	
time_series	0.005	
Topic: **Contextual Cybersecurity** (19.9%; 20% in the pie-chart)	Summary: The terms closely associated with contextual data analysis, such as heterogeneous, situational_awareness, knowledge_base, contextual_information, correlation, and alert_correlation appear under this topic cluster. Therefore, the researchers label this topic cluster as Contextual Cybersecurity.	
Sub-Topics	**Frequency**	**Notes**
feature_selection	0.012	

heterogeneous	0.010	
situational_awareness	0.009	
analyst	0.009	
data_mining	0.008	Appears under multiple topics.
knowledge_base	0.008	
contextual_information	0.007	
correlation	0.007	
data_set	0.006	
alert_correlation	0.006	Appears under multiple topics.
Topic: **Cybersecurity Applied Domain** (18.5%; 19% in the pie-chart)	Summary: This topic cluster is named as Cybersecurity Applied Domain because the terms, such as mobile, social_medium, computer_security, and banking, are prevalent.	
Sub-Topics	**Frequency**	**Notes**
analytics	0.011	
mobile	0.010	
social_medium	0.009	
computer_security	0.006	
data_driven	0.006	
hacker	0.006	
text	0.006	
banking	0.006	
social_network	0.006	
hacker_community	0.006	
Topic: **Data-Driven Adversary** (11.7%; 12% in the pie-chart)	Summary: Except the data science-related terms, the terms related to adversary prevail in this topic cluster. Thus, the researchers label this topic cluster as Data-Driven Adversary.	
Sub-Topics	**Frequency**	**Notes**
adversarial_sample	0.024	
adversarial	0.023	
adversarial_example	0.017	
adversary	0.015	
substitute	0.013	
learning_algorithm	0.010	
oracle	0.008	
substitute_model	0.008	
model_trained	0.008	
logistic_regression	0.007	
Topic: **Power System in Cybersecurity** (7.9%; 8% in the pie-chart)	Summary: This topic cluster is dominated by industrial terms related to national infrastructure for utility. Therefore, the researchers label it as Power System in Cybersecurity.	
Sub-Topics	**Frequency**	**Notes**
power_system	0.041	
power	0.024	
smart_grid	0.012	
total_number	0.011	
scada_system	0.011	
command	0.009	
data_mining	0.009	Appears under multiple topics.
injection	0.009	
measurement	0.007	
cyber_crime	0.007	
Topic: **Vulnerability Management** (19%; 19% in the pie-chart)	Summary: This topic cluster is predominated by the terms associated vulnerabilities or threats, such as botnet, malware, and detection. Thus, the researchers label this topic cluster as Vulnerability Management.	
Sub-Topics	**Frequency**	**Notes**
botnet	0.019	
botnet_detection	0.015	
botnets	0.008	

botnet_traffic	0.007	
naive_bayes	0.007	
secret	0.007	
detection_rate	0.006	Appears under multiple topics.
evasion	0.006	
numeric	0.006	
malware_detection	0.006	

Note. The six topics are listed in alphabetical order. The summaries of each topic cluster are provided and also the terms appearing in multiple topics, such as data_mining, data_set and detection_rate, are noted in the table. The column "Topic" means each of the six topic clusters originally resulted in numeric value from Gensim's LDAModel and subsequently labeled by the researchers; "Summary" means a summary of each topic cluster denoting what each one represents, approached using a bottom-up analysis of the constituent sub-topics; "Sub-topics" mean ten most frequent terms within each topic cluster discovered by Gensim's LDAModel; "Frequency" means a percent of the sub-topic in the distinct terms of the entire text-corpus; and "Notes" mean the sub-topic appears in multiple topics.

Acknowledgements

We thank Dr. Nagib Callaos Conference Chair of IMCIC 2019 for inviting us to publish this paper. Also, we thank Dr. Terry Ryan at Claremont Graduate University and Dr. Conrad Shayo at California State University San Bernardino for peer-reviewing our IMCIC 2019 conference paper and specially Dr. Lorne Olfman, Director of Center for Information Systems & Technology, Claremont Graduate University, for his support for this research and recommending the peer-reviewers.

References

Abt, S., & Baier, H. (2014). A Plea for Utilising Synthetic Data when Performing Machine Learning Based Cyber-Security Experiments (pp. 37–45). ACM Press. https://doi.org/10.1145/2666652.2666663

Abubakar, A. I., Chiroma, H., Muaz, S. A., & Ila, L. B. (2015). A Review of the Advances in Cyber Security Benchmark Datasets for Evaluating Data-Driven Based Intrusion Detection Systems. *Procedia Computer Science*, *62*, 221–227. https://doi.org/10.1016/j.procs.2015.08.443

Adhikari, U., Morris, T., & Pan, S. (2017). WAMS Cyber-Physical Test Bed for Power System, Cybersecurity Study, and Data Mining. *IEEE Transactions on Smart Grid*, *8*(6), 2744–2753. https://doi.org/10.1109/TSG.2016.2537210

Aleroud, A., & Karabatis, G. (2017). Contextual information fusion for intrusion detection: a survey and taxonomy. *Knowledge and Information Systems*, *52*(3), 563–619. https://doi.org/10.1007/s10115-017-1027-3

Alguliyev, R., & Imamverdiyev, Y. (2014). Big data: big promises for information security. In *Application of Information and Communication Technologies (AICT), 2014 IEEE 8th International Conference on* (pp. 1–4). IEEE.

Alsheikh, M. A., Lin, S., Niyato, D., & Tan, H.-P. (2014). Machine learning in wireless sensor networks: Algorithms, strategies, and applications. *IEEE Communications Surveys & Tutorials*, *16*(4), 1996–2018.

Anthes, G. (2010). Topic models vs. unstructured data. *Communications of the ACM*, *53*(12), 16–18.

Armerding, T. (2017, January 31). Obama's cybersecurity legacy: Good intentions, good efforts, limited results. Retrieved August 22, 2018, from https://www.csoonline.com/article/3162844/security/obamas-cybersecurity-legacy-good-intentions-good-efforts-limited-results.html

Aswani, K., Cronin, A., Liu, X., & Zhao, H. (2015). Topic modeling of SSH logs using latent dirichlet allocation for the application in cyber security. In *Systems and Information Engineering Design Symposium (SIEDS), 2015* (pp. 75–79). IEEE.

Beaver, J. M., Borges-Hink, R. C., & Buckner, M. A. (2013). An Evaluation of Machine Learning Methods to Detect Malicious SCADA Communications (pp. 54–59). IEEE. https://doi.org/10.1109/ICMLA.2013.105

Beaver, J. M., Symons, C. T., & Gillen, R. E. (2013). A learning system for discriminating variants of malicious network traffic (p. 1). ACM Press. https://doi.org/10.1145/2459976.2460003

Benjamin, V. A., & Chen, H. (2013). Machine learning for attack vector identification in malicious source code. In *Intelligence and Security Informatics (ISI), 2013 IEEE International Conference on* (pp. 21–23). IEEE.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

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

Borges Hink, R. C., Beaver, J. M., Buckner, M. A., Morris, T., Adhikari, U., & Pan, S. (2014). Machine learning for power system disturbance and cyber-attack discrimination (pp. 1–8). IEEE. https://doi.org/10.1109/ISRCS.2014.6900095

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., … Filar, B. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. *ArXiv Preprint ArXiv:1802.07228*.

Buczak, A. L., & Guven, E. (2016). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Communications Surveys & Tutorials*, *18*(2), 1153–1176. https://doi.org/10.1109/COMST.2015.2494502

Camastra, F., Ciaramella, A., & Staiano, A. (2013). Machine learning and soft computing for ICT security: an overview of current trends. *Journal of Ambient Intelligence and Humanized Computing*, *4*(2), 235–247. https://doi.org/10.1007/s12652-011-0073-z

Carlini, N., Liu, C., Kos, J., Erlingsson, Ú., & Song, D. (2018). The Secret Sharer: Measuring Unintended Neural Network Memorization & Extracting Secrets. *ArXiv Preprint ArXiv:1802.08232*.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: from big data to big impact. *MIS Quarterly*, 1165–1188.

Czejdo, B. D., Iannacone, M. D., Bridges, R. A., Ferragut, E. M., & Goodall, J. R. (2014). Integration of external data sources with cyber security data warehouse (pp. 49–52). ACM Press. https://doi.org/10.1145/2602087.2602098

Das, R., Sarkani, S., & Mazzuchi, T. A. (2012). Fast Abstract: Software Selection Based on Quantitative Security Risk Assessment. In *High-Assurance Systems Engineering (HASE), 2012 IEEE 14th International Symposium on* (pp. 171–172). IEEE.

Davenport, T. H., & Patil, D. J. (2012, October 1). Data Scientist: The Sexiest Job of the 21st Century. Retrieved August 22, 2018, from https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Esmalifalak, M., Nam Tuan Nguyen, Rong Zheng, & Zhu Han. (2013). Detecting stealthy false data injection using machine learning in smart grid (pp. 808–813). IEEE. https://doi.org/10.1109/GLOCOM.2013.6831172

European Union. (2018, August 22). Cyber-Security - EU Global Strategy - European Commission. Retrieved August 22, 2018, from /globalstrategy/en/cyber-security

Fan, Y., Ye, Y., & Chen, L. (2016). Malicious sequential pattern mining for automatic malware detection. *Expert Systems with Applications*, *52*, 16–25. https://doi.org/10.1016/j.eswa.2016.01.002

Fang, Z., Zhao, X., Wei, Q., Chen, G., Zhang, Y., Xing, C., … Chen, H. (2016). Exploring key hackers and cybersecurity threats in Chinese hacker communities. In *Intelligence and Security Informatics (ISI), 2016 IEEE Conference on* (pp. 13–18). IEEE.

Gandotra, E., Bansal, D., & Sofat, S. (2014). Malware Analysis and Classification: A Survey. *Journal of Information Security*, *05*(02), 56–64. https://doi.org/10.4236/jis.2014.52006

gensim: topic modelling for humans. (n.d.). Retrieved March 3, 2018, from https://radimrehurek.com/gensim/tut2.html

Georgescu, T. M., & Smeureanu, I. (2017). Using Ontologies in Cybersecurity Field. *Informatica Economica*, *21*(3/2017), 5–15. https://doi.org/10.12948/issn14531305/21.3.2017.01

Google Scholar. (n.d.). Retrieved September 24, 2017, from https://scholar.google.com/

Goulart, J. (2016, January 20). Data Scientist: The Number One Job In America. Retrieved August 22, 2018, from https://blog.edx.org/the-importance-of-data-science-in-the-21st-century/?track=blog

Guarino, A. (2013). Digital forensics as a big data challenge. In *ISSE 2013 securing electronic business processes* (pp. 197–203). Springer.

He, W., Tian, X., Shen, J., & Li, Y. (2015). Understanding Mobile Banking Applications' Security risks through Blog Mining and the Workflow Technology.

Hevner, A., & Chatterjee, S. (2010). *Design Research in Information Systems: Theory and Practice* (2010 edition). New York ; London: Springer.

Hou, S., Saas, A., Chen, L., Ye, Y., & Bourlai, T. (2017). Deep Neural Networks for Automatic Android Malware Detection (pp. 803–810). ACM Press. https://doi.org/10.1145/3110025.3116211

Huang, J., Kalbarczyk, Z., & Nicol, D. M. (2014). Knowledge discovery from big data for intrusion detection using LDA. In *Big data (BigData Congress), 2014 IEEE international congress on* (pp. 760–761). IEEE.

International Institute of Informatics and Cybernetics, IIIC. (n.d.). Journal of Systemics, Cybernetics and Informatics. Retrieved February 27, 2019, from http://www.iiisci.org/journal/sci/IssueCFP.asp

Jones, C. L., Bridges, R. A., Huffer, K. M. T., & Goodall, J. R. (2015). Towards a Relation Extraction Framework for Cyber-Security Concepts (pp. 1–4). ACM Press. https://doi.org/10.1145/2746266.2746277

Joseph, A. D., Laskov, P., Roli, F., Tygar, J. D., & Nelson, B. (2013). *Machine Learning Methods for Computer Security (Dagstuhl Perspectives Workshop 12371)* (p.). Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik GmbH, Wadern/Saarbruecken, Germany. https://doi.org/10.4230/dagman.3.1.1

Kolini, F., & Janczewski, L. (2017). Clustering and Topic Modelling: A New Approach for Analysis of National Cyber security Strategies. *PACIS 2017 Proceedings, Malaysia*.

Lau, R. Y., Xia, Y., & Ye, Y. (2014). A probabilistic generative model for mining cybercriminal networks from online social media. *IEEE Computational Intelligence Magazine*, *9*(1), 31–43.

Le, D. C., Zincir-Heywood, A. N., & Heywood, M. I. (2016). Data analytics on network traffic flows for botnet behaviour detection. In *Computational Intelligence (SSCI), 2016 IEEE Symposium Series on* (pp. 1–7). IEEE.

Lee, T.-H., Sung, W.-K., & Kim, H.-W. (2016). A Text Mining Approach to the Analysis of Information Security Awareness: Korea, United States, and China. In *PACIS* (p. 69).
Li, W., Yin, J., & Chen, H. (2016). Identifying high quality carding services in underground economy using nonparametric supervised topic model.
Liu, Y., Sarabi, A., Zhang, J., Naghizadeh, P., Karir, M., Bailey, M., & Liu, M. (2015). Cloudy with a Chance of Breach: Forecasting Cyber Security Incidents. In *USENIX Security Symposium* (pp. 1009–1024).
Mabey, B. (2018). *pyLDAvis: Python library for interactive topic model visualization. Port of the R LDAvis package*. Jupyter Notebook. Retrieved from https://github.com/bmabey/pyLDAvis (Original work published 2015)
Mahmood, T., & Afzal, U. (2013). Security analytics: Big data analytics for cybersecurity: A review of trends, techniques and tools. In *Information assurance (ncia), 2013 2nd national conference on* (pp. 129–134). IEEE.
Mayhew, M., Atighetchi, M., Adler, A., & Greenstadt, R. (2015). Use of machine learning in big data analytics for insider threat detection. In *Military Communications Conference, MILCOM 2015-2015 IEEE* (pp. 915–922). IEEE.
McKenna, S., Staheli, D., Fulcher, C., & Meyer, M. (2016). BubbleNet: A Cyber Security Dashboard for Visualizing Patterns. *Computer Graphics Forum*, *35*(3), 281–290. https://doi.org/10.1111/cgf.12904
Meidan, Y., Bohadana, M., Shabtai, A., Guarnizo, J. D., Ochoa, M., Tippenhauer, N. O., & Elovici, Y. (2017). ProfilIoT: a machine learning approach for IoT device identification based on network traffic analysis (pp. 506–509). ACM Press. https://doi.org/10.1145/3019612.3019878
Mittal, S., Das, P. K., Mulwad, V., Joshi, A., & Finin, T. (2016). CyberTwitter: Using Twitter to generate alerts for cybersecurity threats and vulnerabilities. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 860–867). https://doi.org/10.1109/ASONAM.2016.7752338
Noel, S., Harley, E., Tam, K. H., Limiero, M., & Share, M. (2016). CyGraph: Graph-Based Analytics and Visualization for Cybersecurity. In *Handbook of Statistics* (Vol. 35, pp. 117–167). Elsevier. https://doi.org/10.1016/bs.host.2016.07.001
NVD - Home. (n.d.). Retrieved August 4, 2018, from https://nvd.nist.gov/
Pajouh, H. H., Dastghaibyfard, G., & Hashemi, S. (2017). Two-tier network anomaly detection model: a machine learning approach. *Journal of Intelligent Information Systems*, *48*(1), 61–74. https://doi.org/10.1007/s10844-015-0388-x
Papernot, N., Carlini, N., Goodfellow, I., Feinman, R., Faghri, F., Matyasko, A., … others. (2016). cleverhans v2. 0.0: an adversarial machine learning library. *ArXiv Preprint ArXiv:1610.00768*.
Papernot, N., McDaniel, P., & Goodfellow, I. (2016). Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *ArXiv Preprint ArXiv:1605.07277*, 13.
Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. (2016). Towards the science of security and privacy in machine learning. *ArXiv Preprint ArXiv:1611.03814*.
Project Jupyter. (n.d.). Project Jupyter. Retrieved February 28, 2019, from https://www.jupyter.org
Samtani, S., Chinn, K., Larson, C., & Chen, H. (2016). AZSecure Hacker Assets Portal: Cyber threat intelligence and malware analysis. In *Intelligence and Security Informatics (ISI), 2016 IEEE Conference on* (pp. 19–24). IEEE.
Samtani, S., Chinn, R., & Chen, H. (2015). Exploring hacker assets in underground forums. In *Intelligence and Security Informatics (ISI), 2015 IEEE International Conference on* (pp. 31–36). IEEE.
Shinyama, Y. (2014). PDFMiner. Retrieved April 8, 2018, from https://euske.github.io/pdfminer/index.html
Shuai, B., Li, H., Li, M., Zhang, Q., & Tang, C. (2013). Automatic classification for vulnerability based on machine learning. In *Information and Automation (ICIA), 2013 IEEE International Conference on* (pp. 312–318). IEEE.
Singh, J., & Nene, M. J. (2013). A Survey on Machine Learning Techniques for Intrusion Detection Systems, *2*(11), 7.
SSH (Secure Shell) Home Page | SSH.COM. (n.d.). Retrieved August 4, 2018, from https://www.ssh.com/ssh/
Stevanovic, M., & Pedersen, J. M. (2013). Machine learning for identifying botnet network traffic. *Networking and Security Section, Department of Electronic Systems, Aalborg University, Tech. Rep.*
Sundarkumar, G. G., Ravi, V., Nwogu, I., & Govindaraju, V. (2015). Malware detection via API calls, topic models and machine learning. In *Automation Science and Engineering (CASE), 2015 IEEE International Conference on* (pp. 1212–1217). IEEE.
Symons, C. T., & Beaver, J. M. (2012). Nonparametric semi-supervised learning for network intrusion detection: combining performance improvements with realistic in-situ training. In *Proceedings of the 5th ACM workshop on Security and artificial intelligence* (pp. 49–58). ACM.
The pandas project. (2017). Python Data Analysis Library — pandas: Python Data Analysis Library. Retrieved April 8, 2018, from https://pandas.pydata.org/
The White House Office of the Press Secretary. (2016, February 9). FACT SHEET: Cybersecurity National Action Plan. Retrieved August 22, 2018, from https://obamawhitehouse.archives.gov/the-press-office/2016/02/09/fact-sheet-cybersecurity-national-action-plan
Thuraisingham, B. (2015). Big Data Security and Privacy (pp. 279–280). ACM Press. https://doi.org/10.1145/2699026.2699136

Thuraisingham, B., Kantarcioglu, M., Hamlen, K., Khan, L., Finin, T., Joshi, A., … Bertino, E. (2016). A data driven approach for the science of cyber security: Challenges and directions. In *Information Reuse and Integration (IRI), 2016 IEEE 17th International Conference on* (pp. 1–10). IEEE.

Vinchurkar, D. P., & Reshamwala, A. (2012). *A Review of Intrusion Detection System Using Neural Network and Machine Learning*. IJESIT.

Vom Brocke, J., Simons, A., Niehaves, B., Riemer, K., Plattfaut, R., Cleven, A., & others. (2009). Reconstructing the giant: on the importance of rigour in documenting the literature search process. In *Ecis* (Vol. 9, pp. 2206–2217).

Yasakethu, S., & Jiang, J. (2013). Intrusion detection via machine learning for SCADA system protection. In *Proceedings of the 1st International Symposium for ICS & SCADA Cyber Security Research* (pp. 101–5).

Yasmin. (2017, September 14). LDA and T-SNE Interactive Visualization. Retrieved April 3, 2018, from https://www.kaggle.com/ykhorramz/lda-and-t-sne-interactive-visualization

Zamani, M., & Movahedi, M. (2013). Machine learning techniques for intrusion detection. *ArXiv Preprint ArXiv:1312.2177*.

Zomlot, L., Chandran, S., Caragea, D., & Ou, X. (2013). Aiding intrusion analysis using machine learning. In *Machine Learning and Applications (ICMLA), 2013 12th International Conference on* (Vol. 2, pp. 40–47). IEEE.

Zuech, R., Khoshgoftaar, T. M., & Wald, R. (2015). Intrusion detection and Big Heterogeneous Data: a Survey. *Journal of Big Data*, *2*(1). https://doi.org/10.1186/s40537-015-0013-4