Time-based Gap Analysis of Cybersecurity Trends in Academic and Digital Media

MAHDI R. ALAGHEBAND*, Ted Rogers School of Information Technology Management, Ryerson University, Canada
ATEFEH MASHATAN*, Ted Rogers School of Information Technology Management, Ryerson University, Canada
MORTEZA ZIHAYAT*†, Ted Rogers School of Information Technology Management, Ryerson University, Canada

This study analyzes cybersecurity trends and proposes a conceptual framework to identify cybersecurity topics of social interest and emerging topics which need to be addressed by researchers in the field. The insights drawn from this framework allow for a more proactive approach to identifying cybersecurity patterns and emerging threats which will ultimately improve the collective cybersecurity posture of the modern society. To achieve this, cybersecurity-oriented content in both media and academic corpora, disseminated between 2008 and 2018, were morphologically analyzed via text mining. A total of 3,556 academic papers obtained from the top-10 highly reputable cybersecurity academic conferences, and 4,163 news articles collected from the New York Times were processed. The LDA topic modeling followed optimal perplexity and coherence scores resulted in 12 trendy topics. Next, the time-based gap between these trendy topics was analyzed to measure the correlation between media and trendy academic topics. Both convergences and divergences between the two cybersecurity corpora were identified suggesting a strong time-based correlation between these resources. This framework demonstrates the effective use of automated techniques to provide insights about cybersecurity topics of social interest and emerging trends, and informs the direction of future academic research in this field.

CCS Concepts: • **Information systems → Information systems applications**; **Web mining**.

Additional Key Words and Phrases: Cybersecurity Trends, Topic Modeling, Trend Analysis, Academic Context, Digital Media

ACM Reference Format:
Mahdi R. Alagheband, Atefeh Mashatan, and Morteza Zihayat. 2019. Time-based Gap Analysis of Cybersecurity Trends in Academic and Digital Media . *J. ACM* 37, 4, Article 111 (August 2019), 20 pages. https://doi.org/10.1145/1122445.1122456

*All authors contributed equally to this research.
†Corresponding author.

Authors' addresses: Mahdi R. Alagheband, m.alaghband@ryerson.ca, Ted Rogers School of Information Technology Management, Ryerson University, 55 Dundas St. W, Toronto, Canada; Atefeh Mashatan, amashatan@ryerson.ca, Ted Rogers School of Information Technology Management, Ryerson University, 55 Dundas St. W, Toronto, Canada; Morteza Zihayat, mzihayat@ryerson.ca, Ted Rogers School of Information Technology Management, Ryerson University, 55 Dundas St. W, Toronto, Canada.

0004-5411/2019/8-ART111 $15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

The prevalence of the Internet has caused the digitization of society to affect most people's lifestyle over the past three decades. Not only do people use various internet-based gadgets, they adopt the technology-stricken services of banks and governmental organizations. The prevalent usage of the technology results in a wide range of diverse cyber attacks.

Cybersecurity threats and vulnerabilities have seemingly increased in recent years. According to the Identity Theft Resource Center [13], the number of reported data breaches has gone up from 157 in 2005 to 1,091 in 2016. The number of Distributed Denial of Service (DDoS) attacks dramatically increased from 40 Gbps in 2008 to 600 Gbps in 2017 [37]. In North America, the increased activity and ability of malicious actors have caused governments to have a growing consideration for cybersecurity fields. According to the report of Taxpayers for Common Sense, a nonpartisan federal budget watchdog organization, annual unclassified cybersecurity spending by U.S. Federal Agencies increased from approximately 7.5 million dollars in 2007 to approximately 28.3 million dollars in 2016 [24, 40].

Similar regulation was released in Europe: the European Union's (EU) General Data Protection Regulation (GDPR) started regulating data protection and privacy in 2018 [1]. European IT executives conducted a survey of 37 noticeable IT management issues. The survey revealed that cybersecurity ranked fourth in attracting investment and second in technologies that should attract more investment [26]. These regulations represent the fact that understanding the new trends in a volatile context like cybersecurity is of paramount importance.

On the other hand, media and academia each play a pivotal role in different aspects of cybersecurity. Media develops widespread cybersecurity awareness while academic-stricken activities are prerequisites of both technological advancement and investment. Thus, the two disciplines are the main sources of decision making and strategy planning in cybersecurity fields.

News agencies inform societies about cybersecurity threats, vulnerabilities, and preventive solutions. From a business standpoint, the consequences of a data breach are heavily alleviated by the presence of a free press even though the awareness about a given businesses' vulnerabilities could have negative ramifications [26]. It is no surprise that cybersecurity coverage in the media has increased steadily over the past decade. According to the data gathered in this study, the *New York Times*'s (*NYT*) coverage of cybersecurity-related events increased from 81 stories in 2008 to more than 900 in 2019.

From a research point of view, cybersecurity has also become a more popular research topic. The number of papers presented at five prestigious conferences increased from 223 in 2008 to 496 in 2019. While they look to be independent at first, we argue that there are implicit relationships between these resources that, if it can be revealed, present different insights to decision makers.

The importance of cybersecurity in both media and academia motivates us to study how much cybersecurity press coverage corresponds with academic research efforts. Highlighting the trends and the time-gap among them are crucially important as they can perceivably have a big impact on one another. For instance, if the media had not extensively covered the Bitcoin-based Silk Road dark market in 2011, many researchers, particularly cryptographers, would not have been attracted to explore cryptocurrency-driven solutions.

Therefore, in this study, our goal is to understand top cybersecurity trends in academic and media contexts as well as the temporal relationship between the topics that arise in both academic and press discourses. We propose a framework based on well-known machine learning techniques such as topic modeling. The framework collects and processes data from both academic and research resources. Then, it builds a time-based topic model to represent the most noticeable trends in each

year for each resource. Finally, we propose different approaches to analyze the trends and the gaps existing between topics in academia and media. Below, we summarize our contributions:

- We investigate the problem of time-based topic analysis in cybersecurity trends, and propose an end-to-end framework to analyze content and topics.
- We propose two approaches to analyze the gaps and trends between topics presented by two corpora. Specifically, we identify the commonalities and differences of cybersecurity-driven topics between the media and academic fields. The results are informative for researchers, business people, and policymakers in the field of cybersecurity.
- We conduct extensive experiments based on the data collected from a popular digital media source (e.g., *NYT*) and top tier conferences in cybersecurity.

The rest of this paper is organized as follows. Section 2 represents related work. We present our methodology and the conceptual framework in Section 3. Section 4 is dedicated to the conducted experiments which include a topic similarity analysis, gap analysis, and discussion. Finally, the conclusion is presented in Section 5.

2 RELATED WORK

In this section, we review works related to this study. According to the longitudinal nature of our study, the related research includes gap analysis and trend analysis in both academia and media.

2.1 Gap Analysis using Machine Learning

Conducting an analysis of a large collection of documents is challenging. One of the most common techniques for such an analysis is topic modeling. Probabilistic topic modeling algorithms are unsupervised machine learning techniques to automatically analyze a collection of documents. The algorithms find underlying patterns and show the correlation between them. Gap analysis can be tied with topic modeling. Once the topic models are built, by analyzing the similarity and dissimilarity of topics, the gaps can be revealed and explored [5].

In this study, we use one of the most common topic modeling algorithms called *Latent Dirichlet Allocation (LDA)*, introduced by David Blei et al. [11]. LDA is rooted in the hierarchical Bayesian model and assumes that the input documents are generated by a probabilistic process [12, 27]. It decomposes a collection of documents into topics and represents each document with a (weighted or unweighted) subset of topics [10, 15]. LDA begins by randomly designating each word in a document as one out of K topics. Then, it calculates the conditional probabilities for every topic in each document through an iterative process. Beyond heuristic approaches, topic modeling is a process of automatically identifying topics presented in a corpus in order to derive hidden patterns.

The application of topic modeling to perform trend analysis is a well-established approach in text data mining. Advanced and major academic trends have been identified in different fields with the topic modeling technique. For instance, a combination of clustering, bibliometric analysis and text-mining methods are used to identify the academic trends in design research. Academic branches are identified as groups of topics via LDA. *Perplexity* is used to find the optimal number of topics [35]. Lippincott et al. calculated LDA-induced distribution over biomedical research topics [31].

There are several recent studies that used LDA in different contexts. In 2018, Chandelier et al. contrasted content variation between articles in two regional and national newspapers and analyzed their trends in ubiquitous topics [14]. The topic modeling approach for similarity analysis is evaluated in [22]. Further, the authors of [25] used LDA to estimate the time gap across three distinct academic areas including papers, patents, and web news articles on computer science. Having been conceptualized, the results are validated with different evaluation methods. To the

best of our knowledge, the existing literature has not considered the cybersecurity context and the gap analysis between academic and media publications. Moreover, none of these approaches are directly applicable to our context.

2.2 Trend Analysis in Academia

There are a noticeable number of research papers focused on cybersecurity. Li et al. extracted cyberspace research topics using a bibliographic approach on Web of Science (WoS) records and revealed their temporal patterns. The authors quantitatively identified cyberspace trends and landscapes from 1989 to 2016. Interestingly, the keyword *security* has the longest flow, even in comparison with *Internet* and *information* keywords. They showed this with a flow map where the flow of the keyword *security* has a strong evolutionary pattern, being differentiated by keywords such as anonymity and detection since 1989 [30]. However, the analysis was done based on one data source; they have not investigated the time gaps among data sources with respect to topics.

Aside from cybersecurity, there are several recent studies conducted on trend topic structures in other contexts (e.g., healthcare research papers) using machine learning techniques. In Biomedical research, topic modeling can be advantageous as it could result in new treatment hypotheses. In 2018, Coventry and Branley analyzed research trends with regards to cybersecurity in healthcare. They collected a total of 1249 scientific papers from 2014 to 2018, which included the keywords, *cybersecurity* and *healthcare*. They emphasized the threats and consequences currently faced by the healthcare industry and outlined how it can move forward [19]. The research of Coventry and Branley differs from ours, as they focused on healthcare and their methodology is not directly applicable here.

Beykikhoshk et al. proposed temporal modeling and used a hierarchical Dirichlet process-based model to discover time-based topics focused on the autism disorder in medical research literature. They discovered important topics in overlapping time epochs and then constructed a graph to represent the inter-domain similarity [9]. More recently, Beykikhoshk et al. expanded their results on scientific papers about metabolic syndrome. They extended Bayesian mixture models to deal with temporally varying documents [8]. This approach is effective when topics may change over time. Andrei and Arandjelovic identified the research directions in medical research literature. They presented a framework based on Bayesian techniques to extract and monitor changes to the topic structure of a longitudinal document corpus [6].

Lu and Li conducted a comprehensive literature review on Internet-of-Things (IoT) security from 2013 to 2017 to identify current research topics in cybersecurity research as it pertains to IoT devices. They explored major academic databases including IEEE Xplore, Web of Science, ACM, INSPEC, and ScienceDirect. Some emerging technologies were mentioned as research trends associated with IoT security including cloud service, 5G, data privacy, IoT forensics, self-management, and blockchain embedded cybersecurity design [32]. Their approach differs from ours in its focus on current research, its subject matter, methodology, and that no topic modeling method was used.

Topic analysis aids policymakers to obtain cybersecurity trends in academic resources. Topic modeling was used in [29] to analyze National Cybersecurity Strategies. Kolini and Janczewski identify the most frequent topics in the collection of national cybersecurity in different Countries. The strategies of sixty countries from 2003 to 2016 were clustered into hierarchies through the LDA method and labeled by two independent cybersecurity experts. They extrapolated that "defending citizens", "public IT systems", and "critical infrastructure protection" are the most frequent topics in national cybersecurity resources. Also, the authors of [4] introduced an LDA-based search engine to categorize a variety of datasets including police reports, blogs, intelligence reports, security bulletins, and news resources about cybersecurity collected by experts at the University of Leeds.

They focused on the visualization of the structure of the LDA model to support accessibility of information and the end-user [4].

2.3 Trend analysis in media

A wide variety of news is disseminated on the internet on a daily basis. It is important to identify the relationship between topics and track changes. Van Galen and Nicholson applied topic modeling to historical newspaper archives. Since LDA is a relatively fast method and the most common topic modeling algorithm in digital humanities, they applied it to analyze 72 different newspapers, and approximately 15 million articles [41]. Also, the Probabilistic Latent Semantic Analysis (PLSA) method was used to find topic trends in a corpus of Eighteenth-Century American Newspapers [34]. Kawata and Fujiwara applied LDA to extract topics from Nikkei, a Japanese newspaper. They focused on articles written in 2012 and 2013 and visualized the seasonality of topics and the differentiation of the timing of the trending topics [28].

In regards to cybersecurity, since many people are unaware of cybersecurity threats, the role of the media is important for people's awareness even without topic modeling. Farivar analyzed cybersecurity coverage in the media for three years between 2007 and 2010. Through a heuristic approach, he discussed the coverage of illustrious cybersecurity incidents that occurred over that period time [21]. Das et al. delved deeply into the significant types of security and privacy news and their origins which people found salient. They recruited almost 2000 participants to fill out surveys on the selected emerging security and privacy news events. Respondents found that the most noticeable security and privacy news topics were financial data breaches, corporate personal data breaches, high sensitivity systems breaches, and political activist breaches [20].

The automatic processing of news can objectively facilitate the release of trends in mass media. Incidentally, different news sites will have different writing styles and write different types of articles; thus we need some methods (e.g., LDA) to identify and detect news relating to cyber attacks to achieve a comprehensive perspective. For instance, Abdullah et al. classified online news into several types of cyber attack news.

The cyber attack features are one of the important steps in this research as they helped to characterize and distinguish the types of cyber attack news. The variety of attacks based on their type, name, actor, organization affected, platform, and country affected have been categorized. A Conditional Random Field (CRF) classifier, a statistical modeling technique, was trained to build a probabilistic model for data labeling. They found South Korea and Brazil to be the most affected countries and ransomware, malware, and phishing as the most prevalent threat types [2].

Table 1 presents recent related work and the advantages of our solution, which is rooted in our proposed framework for trend analysis. Table 1 compares our approach with other topic modeling research focusing on the cybersecurity trends as well as a few other distinct applications.

3 METHODOLOGY AND CONCEPTUAL FRAMEWORK

In this section, we elaborate on our framework which finds correlations between media and academic corpora. Figure 1 represents our proposed framework for automated topic similarity and gap analysis of cybersecurity contents. The proposed framework consists of four main phases. *i) Data Preparation:* in this phase, we develop different tools to collect data from different resources of academic and media. We apply different data cleaning techniques to remove noisy data and prepare it for trend analysis. *ii) Time-based Topic modeling:* we build topic models for each corpus over time. That is, each corpus will be represented as sets of topics where each set represents the most important topics in a particular year of a corpus. *iii) Trend Analysis:* in this phase, we take advantage of different similarity analysis techniques to investigate the main trends in cybersecurity based on

Table 1. Comparison of recent literature deploying topic modeling

Paper/ Year	Application	Description	Algorithm	Corpus	Gap analysis
Our analysis 2020	**Cybersecurity**	A time-based topic analysis in cybersecurity trends and a proposed end-to-end framework to analyze content and topics.	LDA, a topic-based Gap analysis algorithm	Academia/ Media	√
Abdullah et al. [2] 2018	Cyber security	The classification of online news into several types of cyber attack news and the identification of cyber attack features.	Conditional Random Field (CRF), Latent Semantic Analysis (LSA)	Media	×
Das and Hong [20] 2018	Cyber security	A survey on emergent news events about cybersecurity to understand the typology of news. The broad types of security and privacy news were uncovered.	Manual surveys	Media	×
Kolini [29] 2017	Cybersecurity	An analysis of 60 national cybersecurity strategies during 2003-2016. Consistency and harmonization among them was found.	Hierarchical clustering method-LDA	Academia	×
Li et al. [30] 2017	Cybersecurity	The authors collected academic articles related to cybersecurity from Web of Science and found the most productive years highly burst strength keywords	Theory of technology maturity (no machine learning)	Academia	×
Adams et al. [3] 2018	Cybersecurity	A selection and ranking of attack patterns from Common Attack Pattern Enumeration and Classification (CAPEC) database to decide which type of attack would be successful against systems.	LDA, KL divergence	CAPEC database	×
Bechor and Jung [7] 2018	Cybersecurity	An identification of key concepts from recent scholarly articles focusing on cybersecurity.	LDA	Academia	×
Chandelier et al. [14] 2018	Animal Recolonization	They used STM in a case study of human-wildlife conflicts and emphasized content variation between articles in a regional and a national newspaper. The time trends in topic and gap trends were analyzed.	Structural topic modeling (STM)	Media	√
Chen et al. [17] 2017	Technological patents	This research proposed a topic-based technological forecasting approach to determine the trends of specific topics in massive patent claims.	LDA	Patents	×
Boussalis and Coan [12] 2016	Climate change	A systematic analysis of conservative think tank sceptical discourse in nearly 15 years and clarify the climate specific relation between scientists and politicians	LDA	Public discourses	×

the topic models built in the previous phase. *iv) Gap Analysis:* given the different topic models, we analyze the time gap between the most similar topics in the media and academic corpora.

Although the four components in Figure 1 are prevalent in topic modeling techniques, our study is the first to combine trend analysis and gap analysis. In the following section, each part of our framework (Figure 1) is explained, and the foundations are mentioned to justify our framework.

3.1 Data Preparation

In the first phase, the most relevant resources of both academia and media have to be selected. The academic dataset is collected from the proceedings of reputable cybersecurity conferences. The conferences are selected from Jianying Zhou's cybersecurity conference ranking list [43], which ranks conferences by the number of participants, the number of accepted papers, the number of submissions, and their H-index. The criteria we used to select the most appropriate proceedings are:

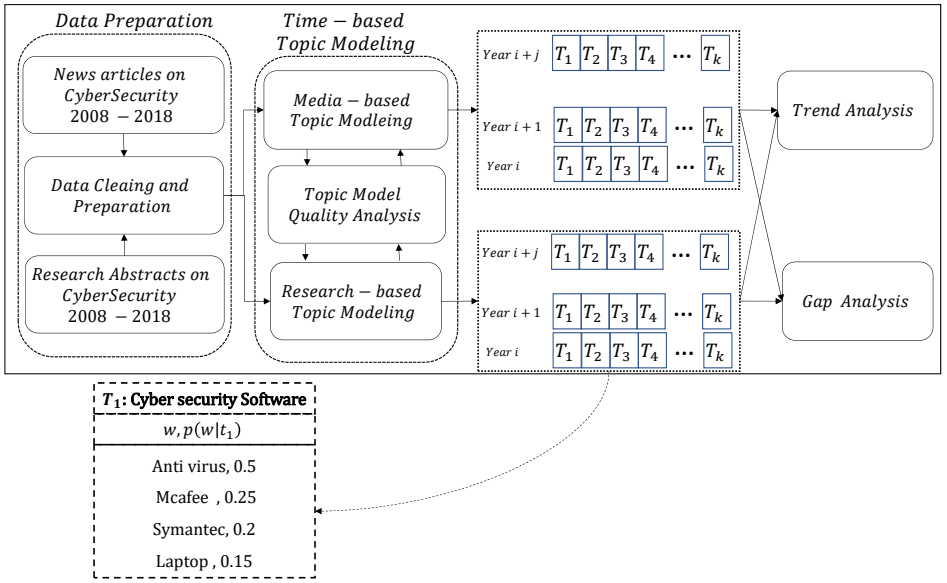


Fig. 1. The proposed framework for trend analysis

- For the sake of comprehensiveness, conferences which cover diverse fields of cybersecurity are chosen; some proceedings with specific sub-fields, e.g., pure cryptography or attack detection, are excluded.
- The data of proceedings have to be available for the past 10 years.

Based on the aforementioned criteria, we selected the following conferences:

- IEEE - Symposium on Security and Privacy (IEEE S&P)
- USENIX Security Symposium (USENIX Security)
- Conference on Computer and Communication Security (ACM CCS)
- Annual Computer Security Applications Conference (ACSAC)
- International Symposium on Research in Attacks, Intrusions, Detection (RAID)
- European Symposium of Research in Computer Security (ESORICS)

We first develop a module to scrape the proceedings of the aforementioned conferences. Then, a web-scrapping crawls into the abstracts of conference papers in the following order:

(1) Apply the DBLP (the computer science bibliography database) for each conference from 2008 to 2018. DBLP provides bibliographic information on selected proceedings including title, authors, and a link to each paper.
(2) We run a script which sequentially gets the papers' title and link from either the publisher or the dedicated website of each conference through DBLP.
(3) The abstract of each paper is extracted through this link.
(4) As for the post-processing step, we double-check the returned information to make sure no paragraphs are missing and all the special characters are represented.

At the same time, adequate news articles on cybersecurity were collected to be used for trend analysis. We use the news aggregator *Lexis Nexis* [42] in order to select related articles since 2008 from the *New York Times*' website. The full-text of articles with at least one of the search tags mentioned in Table 2 are picked out for analysis. The tags are chosen based on the most trending

Table 2. The list of search tags in *New York Times* articles

Information Warfare	Cloud Hacking	Ransomware
Phone Hacking	Data Theft	Spoofing
Computer Network Security	Malicious Software	Internet Privacy
Data Security	Identity Security	Cryptology
Online Security and Privacy	Phishing	Cybercrime
Information Security and Privacy	Information Security Vulnerabilities	

words in cybersecurity. We execute a search query for each calendar year and save the search-results in an HTML file. Each article contains the full-body of the text and a breakdown of the relevance of each search-tag. The following selection criteria are used for the aggregation of most related news articles:

- 80 % relevancy with one tag in Table 2
- 70 % relevancy with the 3 tags in Table 2

These collected articles are parsed using a Python script to produce a CSV from the HTML files. Then, the CSV file is manually checked to ensure that there are no irrelevant phrases or sentences (e.g., the title of books containing cybersecurity-related words, or Twitter and Instagram hashtags).

3.2 Time-based Topic Modeling

The second phase is the time-based topic modeling. As mentioned in Section 2.1, a topic model is a type of statistical model that uncovers the semantic structure of a document set. The specific algorithm applied in this paper is Latent Dirichlet Allocation (LDA). The key assumption of the LDA is that each document within a corpus arises from a distribution of topics, where a topic is defined as a distribution over a fixed number of terms. One of the most important input parameters in topic modeling is the number of topics. To find the right number of topics, we use two well-known measures called *perplexity* and *coherence*. The **perplexity** of a topic model is a measure of how adequately the topics characterize the documents. The **coherence** of a topic model is the degree of similarity exhibited among some words in a topic. The number of topics that represent the model with the highest values for the measures is chosen as the best number of topics. We use the Gensim [39] to run LDA and measure the coherence and perplexity factors for all the topic models built using our framework. In the experiment section, we will present the details of our analysis based on perplexity and coherence values.

Below, we present how we build topic models. Since we are interested in finding topics over time, we apply LDA to articles of each year in each corpus. Once the topic models are built, we need to come up with an approach to represent the topics. Since each topic is represented as a distribution over words, we have a long list of words per topic. We take advantage of Term Frequency-Inverse Document Frequency (TF-IDF) to find the best set of words to represent topics in which the topics over different years can be comparable as well. According to their TF-IDF score, which evaluates how important a word is, the more favorable and noticeable topics are realized in both the academic and media documents. Moreover, we remove the words with low TF-IDF as they are essentially negligible in our similarity analysis and might be misleading.

3.3 Trend Analysis

Trend analysis spots patterns in our collected corpora based on the soft-cosine function which is a way to compute the correlation of two collections of words and then calculates the so-called cosine distance between the two vectors. It represents the number of words that are in common in

both vectors. In this paper, the vectors are topic models in which the first vector is chosen from academia and the second vector represents a topic in media.

Note that there is a subtle difference between ordinary cosine and soft cosine functions. The soft cosine increases the effectiveness of the ordinary cosine function by representing each word in a document as a vector itself. For instance, the ordinary cosine measure would characterize 'Hi poet' and 'Hello Rumi' as having no distance between them, but the soft-cosine measure would return a positive number by adequately representing the fact that 'hi' and 'hello' are synonyms and the closeness between 'Rumi' and 'poet', which are conceptually similar as Rumi was a poet.

Assume that T_1 and T_2 are topic vectors extracted from academic and media corpora, respectively. The soft similarity is computed through the matrix s to indicate the similarity between the two vectors of topics. Matrix s is based on Levenshtein distance [33]; then the soft cosine function is mathematically modeled as follows [38]:

$$Soft - Cosine(T_1, T_2) = \frac{\sum \sum_{i,j=1}^{N} s_{ij} T_{1_i} T_{2_j}}{\sqrt{\sum \sum_{i,j=1}^{N} s_{ij} T_{1_i} T_{1_j}} \cdot \sqrt{\sum \sum_{i,j=1}^{N} s_{ij} T_{2_i} T_{2_j}}}$$

The word-vectors are generated with the fasttext-wiki-news-subwords-300 which is derived from a 2017 dump of Wikipedia articles [23] and included in the Gensim. Thus, our word-vectors and the output of soft-cosine function become available in Gensim, and we can compare the topics generated by our topic models. Then, we begin a trend analysis using the topics generated by our LDA-based topic models and the soft-cosine outputs. We find the most similar topics generated in academia and the most similar topics generated in the *NYT* by using the soft-cosine measure. For the experiments, we present the results and the similarity of topics in detail.

3.4 Gap Analysis

Given the similarity values of topics in the previous section, we build a matrix where each cell represents the similarity value of two topics. In our two-dimensional matrix, one dimension represents topics of different years in academia and the other dimension represents the topics of different years in media. Using an extensive correlation analysis, we find the most correlated and similar topics in academia and media. The difference between year of the topic models represents the time gap between two corpus in terms of the particular topic. Algorithm 1 represents the approach to calculate time gaps between two topics.

For each academic topic, we are able to find the most similar topic among all of the media selected topics. Furthermore, we reveal the *time-gap* between the two corpora which represents which one has mentioned the corresponding topic first. In order to only focus on relevant and certain fields of cybersecurity, we set the threshold similarity parameter = 0.3.

4 EXPERIMENTAL RESULTS

In this section, we represent the empirical evaluation of our proposed model.[1] First, the time-gap of the most similar topics between *NYT* articles and academic papers are stated. Second, the volume of coverage related to specific topics is represented.

4.1 Quantitative and Qualitative Analysis of Topics

In this section, we present our analysis on the quality of topic models built for the gap analysis. First, we describe our analysis on finding the best number of topics per year as it is one of the most

[1]Our raw data and experimental results are available in https://www.ryerson.ca/tedrogersschool/cybersecurity-research-lab/research-projects/time-based-gap-analysis-of-cybersecurity-trends-in-academic-and-digital-media/.

Algorithm 1 Gap Analysis

```
Input: Topic models built based on academic and media corpus
Output: Time gaps between topics in academic and media corpus
time-gap ← ∅
for all topic T_{i,j} for a given year i choose a topic j from the academic corpus do
    for all topic T_{m,n} for a given year m choose a topic n from the media corpus do
        Calculate SoftCosine(T_{i,j}, T_{m,n})
    end for
    Find the most similar topic among all T_m topics (minimum similarity of 0.3)
    time-gap[i] ← m − i
end for
Return time-gap
```

important factors impacting the quality of topics. Then, we will discuss the results of a human study to evaluate the quality of the topics from cybersecurity experts' point of view.

Two popular quality measures (i.e., perplexity and coherence) are used to evaluate the quality of topics and choose the right number of topics per year. Figure 2 presents the values of perplexity and coherence measures on both the academic and digital news corpus. Given the different number of topics, we calculate the measures for each year. The higher values of perplexity and coherence imply a higher quality of the topic model. In each chart, the horizontal axe represents the number of topics and the vertical axes represent the values of the measures. We compute the values for the number of topics $\{15, 20, 25, 30, 35, 40, 45, 50\}$. As shown, the values are different each year. In general, the perplexity value decreases when the number of topics increases.

To find the best number of topics per corpus, we find the optimum value by taking both measures into account. For example, for the year 2018 in Figure 2(a), according to perplexity values, *15* is the best number of topics as it represents the highest perplexity. However, the coherence has its lowest value at *15*. According to the figure, the highest value of coherence is achieved with *30* topics. Considering both measures, *30* is selected as the optimum number of topics for 2018 as it represents the highest value for the coherence measure and its perplexity value is slightly lower than its optimum value. Similar analyses are done to choose the number of topics for the other years in both corpora.

Although the latent pace rooted in topic models is worthwhile and useful, the evaluation of assumptions requires human judgements as LDA is an unsupervised learning process. To investigate the quality of the output model, we conduct a human study to measure the semantic meaning of topics and justify our topic modelling approach. We follow a popular methodology proposed by Chang et al. [16] to conduct the human study. Chang et al. present a quantitative method to measure the interpretability of topic models which is prevalent in text mining [16]. They develop the concept of the *word intrusion* to evaluate the latent space in each topic vector (e.g., word set), to find an intruder word. This is a complementary approach to measure semantically coherent topics.

To measure the coherence of topics, 20 cybersecurity experts completed our 40-question survey to evaluate cybersecurity topics and find the intruder words. The survey is designed as follows. We first randomly selected 40 topics. Given a topic vector, the top-4 words representing the topic are chosen. Then, we consider a word with the lowest probability in the topic as the intruder word. In each question of the survey, the five words (top-4 words plus the intruder) are randomly ordered and presented to the expert. The expert should find out which word is the intruder in each topic. Given 20 answers of our experts, we aggregate the results and calculate the precision per topic. On average, our model achieves the performance of *0.73* in terms of the precision.

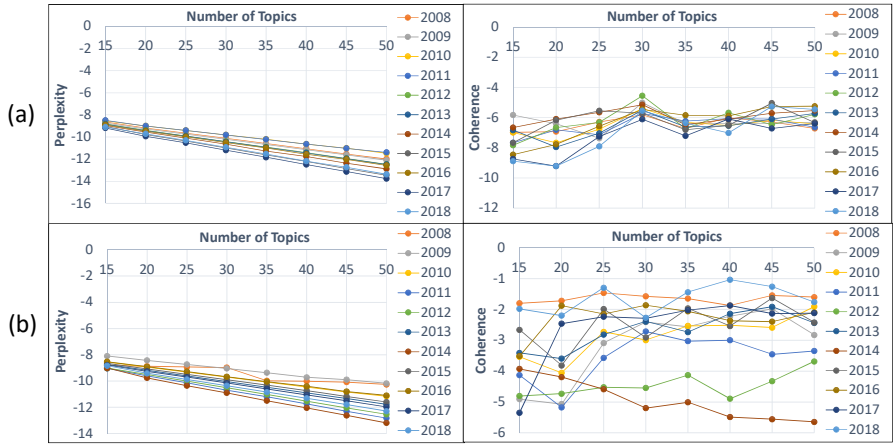


Fig. 2. Topic modeling evaluation to choose the number of topics: (a) Academic Corpus, (b) New York Times Corpus

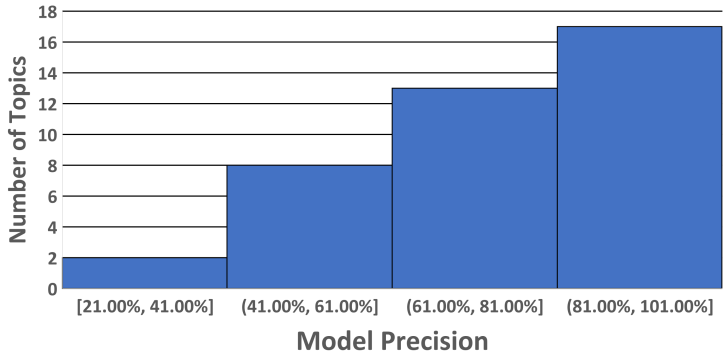


Fig. 3. The histogram chart of model precision of cybersecurity trends

Figure 3 presents the histogram of the survey's results with respect to precision. As shown, the majority of intruder words have been selected by the experts, ensuring that the topic models are interpretable by experts, resulting in the high quality of the topic models. Note that, the topic models are the main component of our gap analysis and therefore these experiments validate the reliability of topic models built for the gap-analysis.

4.2 Time-gap analysis of similar topics

We evaluate the proposed framework using 4,163 news articles and 3,556 academic papers to calculate the time-gap between the two corpora. As shown in Table 3, 'Location', 'Application', 'Privacy', and 'Certificate' topics have the most similarity. The lowest mean similarity belongs to the rows 1 and 10 (social networks and mail messages) with marginal differences.

The negative *time-gap* in Table 3 means that the corresponding topic words appeared earlier in the academic communities. Except for the last set {Mail, Message, Messaging} in the 10th row, academia have almost pioneered all trends. This minor exception happened as {Mail, Message, Messaging} topic words are general-purpose and meaningful even outside of the cybersecurity

Table 3. Time-gap between most similar topics

	Topic Words	Number of Topics	Mean Similarity	Mean time-gap (year)
1	facebook, twitter, google, youtube, snapchat, social	35	0.371	-1.25
2	android, apple mobile, phone	46	0.412	-1.95
3	internet, web, webpage, site, online, packet, traffic, network	116	0.389	-0.440
4	application, app, apps	48	0.420	-.02
5	browser, certificate, site	35	0.413	+.02
6	private, privacy	42	0.406	-1.48
7	location	13	0.4701	-1.65
8	malware, worm, ransomware, virus	62	0.382	-0.462
9	cloud, data, storage	71	0.386	-1.74
10	mail, message, messaging	18	0.369	+0.83

context. Although {Android, Apple, Mobile, Phone} topics have a remarkable mean similarity, they have been noticed in the academic corpus almost 2 years earlier. Furthermore, the topic words of rows 4 and 5 not only have a high and almost same mean similarity, but also they are published at nearly the same time.

4.3 Volume of Coverage

The Volume of Coverage (VoC) criterion assists us in perceiving the discrepancies of keywords in academic and media contexts. The distribution of topic models is summarized in Table 4. There are distinct analytical differences among various cybersecurity aspects. First, it indicates that the topics {Application, App, Apps} relating to application vulnerabilities are mentioned in cybersecurity conferences more than twice as often as media.

Furthermore, there is a striking difference between academic papers (2.56%) and media articles (21.1%) regarding {Facebook, Twitter, Google, Youtube, Snapchat, Social} topics. This indicates that topics relating to popular social networks are more highly represented in the media. Further, the VoC of {Internet, Web, Webpage, Site, Online, Packet, Traffic, Network} shows that topics relating to the Internet have the highest amount of coverage and roughly the same VoC in media and academia. Conversely, {Kernel, Rootkit, Rootkits, Root} keywords are not appealing in both academia and media, and they have the least coverage among all topics.

The tenth row of Table 4 indicates that topics relating to digital certificates, {Browser, Certificate, Site}, receive significant and approximately equal coverage from both academics and the media. Also, on average, {Privacy} is the most noticeable cybersecurity aspect for researchers and journalists. {Location}, {Cloud}, and the software-threatening tools {Malware, Worm, Ransomware, Virus} have attracted a large amount of interest from academia more than media. On the flip side, {Mail, Message, Messaging} and {Bitcoin, Blockchain, Cryptocurrency} are notably covered in the press.

Moreover, the variation of each topic word for 10 years since 2008 is drawn in Figures 2, 3 and 4. Figure 4(a) represents a stark contrast between the two corpora. Although ‘privacy’ has been on average of utmost importance for both researchers and journalists, academic support has noticeably been with less fluctuation. It is clear that there has been a slight increase in academic privacy-driven subjects. Although ‘privacy’ has faced steep fluctuations in the press, the total number of ‘privacy’

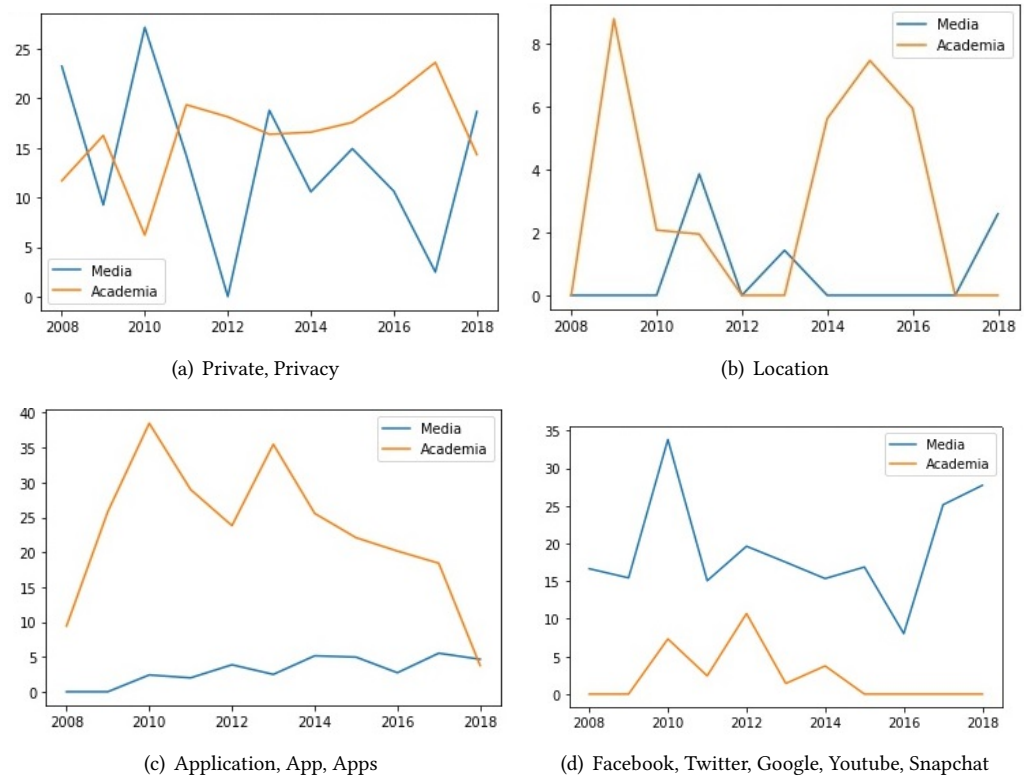
Table 4. The comparison of topic words' average Volume of Coverage (VoC) in academic and media corpora from 2008 to 2018

	Representative	Topic Words	Average VoC in Academia	Average VoC in Media	#Figure
1	Privacy	private, privacy	18.04%	15.00%	Figure 4(a)
2	Location	location	3.18%	0.79%	Figure 4(b)
3	Application	application, app, apps	25.20%	11.5%	Figure 4(c)
4	Social Network	facebook, twitter, google, youtube, snapchat, social	2.56%	21.1%	Figure 4(d)
5	Rootkit	kernel, rootkit, rootkits, root	0.56%	0%	Figure 5(a)
6	Malware	malware, worm, ransomware, virus	11.44%	4.51%	Figure 5(b)
7	Cloud	cloud	5.02%	0.478%	Figure 5(c)
8	Blockchain	bitcoin, blockchain, cryptocurrency	1.70%	5.48%	Figure 5(d)
9	Internet	internet, web, webpage, site, online, packet, traffic, network	32.59%	32.39%	Figure 6(a)
10	Certificate	browser, certificate, site	8.03%	7.11%	Figure 6(b)
11	Android	android, apple mobile, phone	9	11.45	Figure 6(c)
12	Mail	mail, message, messaging	2.89%	11.5%	Figure 6(d)

mentions has become approximately equal on both sides in 2018. When it comes to 'Location', although both academia and media have faced a sharp fall and rise, both curves followed different patterns in Figure 4(b). The highest and the second-largest percentages of mentioning happened in academic papers while the *NYT* cited almost none in that period of time (2008-2010 and 2014-2017).

Note that 'Location' is not as common as the most other topic words in both data sets. Figure 4(c) shows that making mention of {Application, App, Appa} in academia vastly outnumbered media for all ten years. However, recently, 'application' has had the same rate of mentioning since the sudden reduction in academia. Further, popular internet-based applications are summarized in Figure 4(d). In media, despite irregular fluctuations, the topic words have steadily risen from 17% to 27%. The lowest rate in academia and the highest rate in media have been the same, almost 10%, in 2012 and 2016, respectively.

As stated in Figure 5(a), there is a significant difference in the percentage of *'kernel', 'rootkit', 'rootkits', and 'root'*. They seem to be especially technical key words because they were never mentioned in our selected media corpus. In Figure 5(b), media and academia started from the common point in 2008 and finished with almost the same results, but their numbers followed different patterns. Generally, academic cybersecurity resources mentioned more malware variety. Figure 5(c) summarized 'cloud' as a topic trend. Excluding the first two years, the cloud approach (e.g., cloud computing, cloud networking, etc.) has been the more dominant approach in academia, while the average of media is a mere .0478%. Figure 5(d) shows that overall, there has been a



(a) Private, Privacy

(b) Location

(c) Application, App, Apps

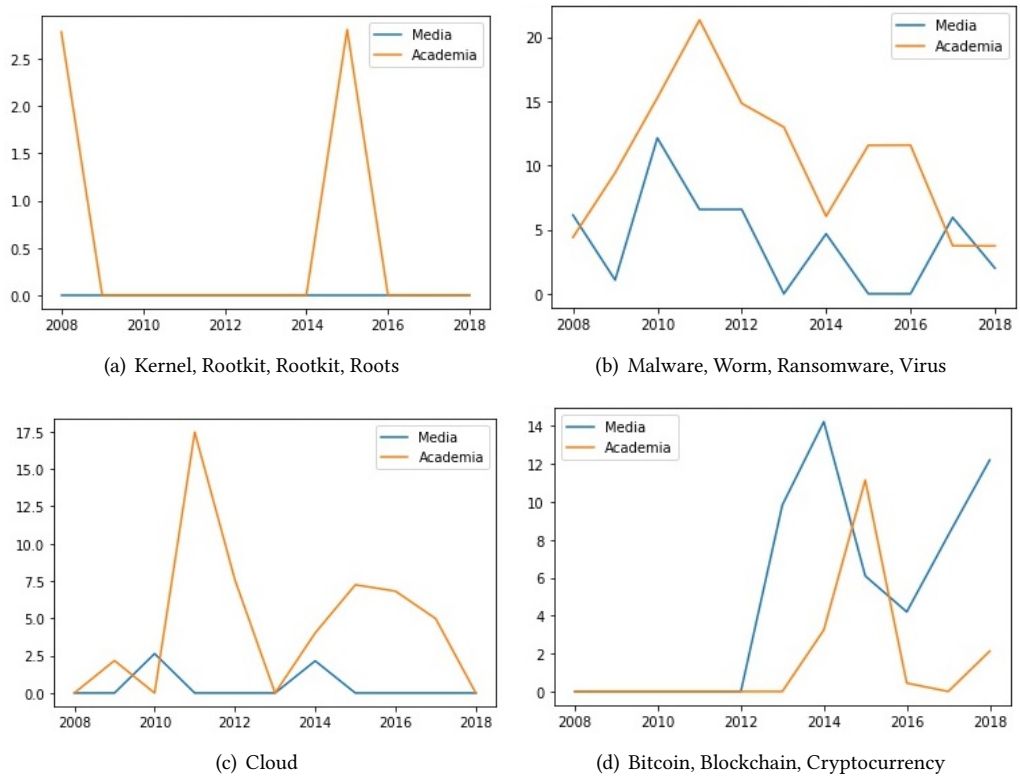
(d) Facebook, Twitter, Google, Youtube, Snapchat

Fig. 4. The percentage of documents characterized by different topic words (Part1) Vertical axis: the volume of coverage, Horizontal axis: time

strong upward trend in the number of {Bitcoin, Blockchain, Cryptocurrency}-related topics in both corpora. Although media has more mentions since 2016, we will hear more about these topics before long in academic and media contexts.

Although the mean VoC of {internet, web, webpage, site, online, packet, traffic, network} are similar in the two curves of Figure 6(a), media had strikingly more coverage from 2010 to 2017 than academic resources. Also, media coverage has witnessed two sharp annual increases from 1% to 34% in 2009 and from 26% to 49% in 2011. Regarding certificate, browser, and site, after continuous and fast fluctuation from 2008 to 2017, seen in Figure 6(b), the percentages of academic and media documents have recently plateaued. The highest percentage is 22% in 2013 for media and 13% in 2009 for scientific resources. Therefore, researchers noticeably paid more attention and earlier than journalists.

According to Figure 6(c), except for the spike (32%) in 2011, both media and academia have witnessed almost the same behaviour with regards to 'Android', 'Apple', 'Mobile', and 'Phone' topics. Figure 6(d) has two different parts with distinct patterns in two periods of time. Between 2008 and 2013, the average of media has been quadruple that of academic papers. They have a focal point in 2014 and again follow distinctively different patterns between 2013 and 2018. All things considered, {Mail, Message, Messaging} were mentioned more in cybersecurity-related articles of media.



(a) Kernel, Rootkit, Rootkit, Roots

(b) Malware, Worm, Ransomware, Virus

(c) Cloud

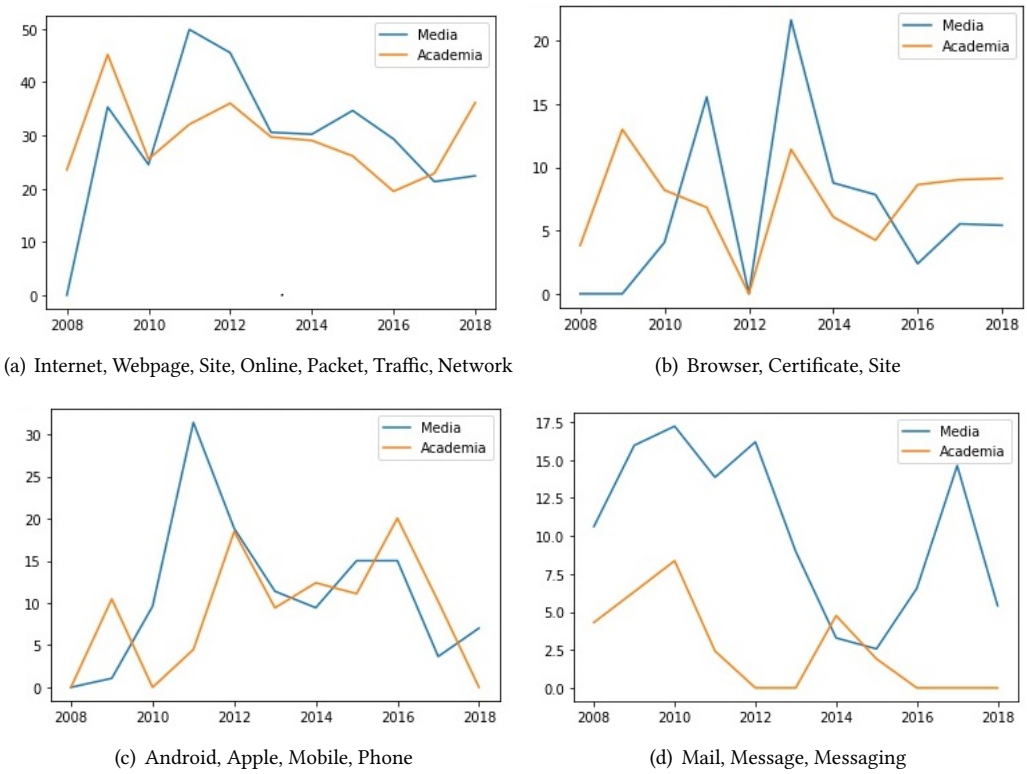
(d) Bitcoin, Blockchain, Cryptocurrency

Fig. 5. The percentage of documents characterized by different topic words (Part2) Vertical axis: the volume of coverage, Horizontal axis: time

4.4 Discussion

In order to better make sense of cybersecurity issues, we discuss more key findings and implications. In Figure 7, the scatter diagram depicts the trend distribution between academic papers and media articles. In fact, it is the combination of two charts. The right side topics are mostly mentioned in media, while the left side topics are more remarkable in academia. The vertical axis is the average of the two VoC mentioned in Table 4. Almost 67% of topics have been mentioned more in academic corpus with widely varied coverage, ranging from 0.3% (Rootkit) to 32% (Internet).

Finally, Figure 8 represents the topic correlation as a heat map. It shows the fluctuation in interest on cybersecurity trends. The lighter spots symbolize the more correlated cybersecurity topics and the time-gap between media and academia. The three vertical light lines in 2011, 2013, and 2016 represent the highest correlations of cybersecurity discussions in the *NYT* in the last decade. Moreover, the bright points emphasize considerable correlations between corresponding years for a few topics. Comparing Figure 8 with the time-variant graphs in Figures 4, 5, and 6 shows that the time-gap between trendy topics are explicitly highlighted. For instance, the location of (*NYT*, Academic) = (2011, 2009-2010) point, which is a strong bright spot in Figure 8, characterizes almost a 1.5 year time-gap between the corpora. Interestingly, the topics which peaked in 2009-2010 in academic contexts, including *Mail*, *Android*, *Certificate*, *Internet*, *Privacy*, and *Location*, experienced a second peak in the media corpus in 2011, after almost 1.5 years. Conversely, the high mentioning



(a) Internet, Webpage, Site, Online, Packet, Traffic, Network

(b) Browser, Certificate, Site

(c) Android, Apple, Mobile, Phone

(d) Mail, Message, Messaging

Fig. 6. The percentage of documents characterized by different topic words (Part3) Vertical axis: the volume of coverage, Horizontal axis: time

of *Privacy* and *Social Network* in media in 2010, resulted in an increase in academia in 2011-2012 (see the (*NYT*, Academic) = (2010, 2011-2012) point).

According to the last quarter of all figures in both corpora in Section 4, we will hear increasingly more about privacy issues before long. Unlike the other selected topics, privacy is merely a notion, it can be accompanied by different applications or frameworks. Therefore, it is highly predictable that the privacy-preserving solutions will receive striking media and academic attention (e.g., privacy preserving data mining, application, authentication protocol and cloud frameworks).

Apart from privacy, the changes in media trends during the last decade are appealing. The social network applications (e.g., *Facebook*, *Twitter*) were the most mentioned topics in media from 2008 to 2010. Between 2016 and 2018, *Bitcoin*, *Cryptocurrency*, and *Blockchain* have experienced the most amplification. Our trends analysis is consistent with the top trends of the Gartner report about emerging technologies [36], meaning the cryptocurrency-oriented topics will receive special consideration in the press. Also, other applications of blockchain technology, beyond cryptocurrency, will be increasingly mentioned.

Furthermore, academic and media contexts implicitly affect each other. On the one hand, the press can stimulate academic researchers to delve into some fields. For instance, both academia and media have not been interested in Bitcoin until 2012. However, the Bitcoin white paper and first-ever block of Bitcoin were released in 2008 and 2009 respectively. Our results represent that researchers' consideration coincided with Silk Road, which was disseminated by media. In 2012,

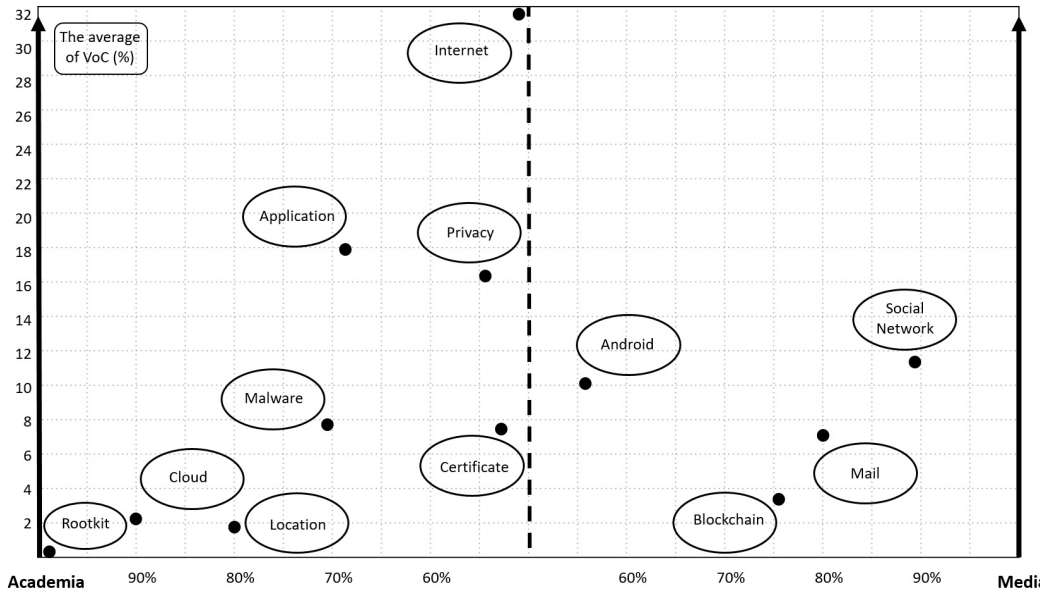


Fig. 7. The scatter diagram of trends distributed between the academic and media resources (Each key word is a representative of a set of topics specified in Table 4.)

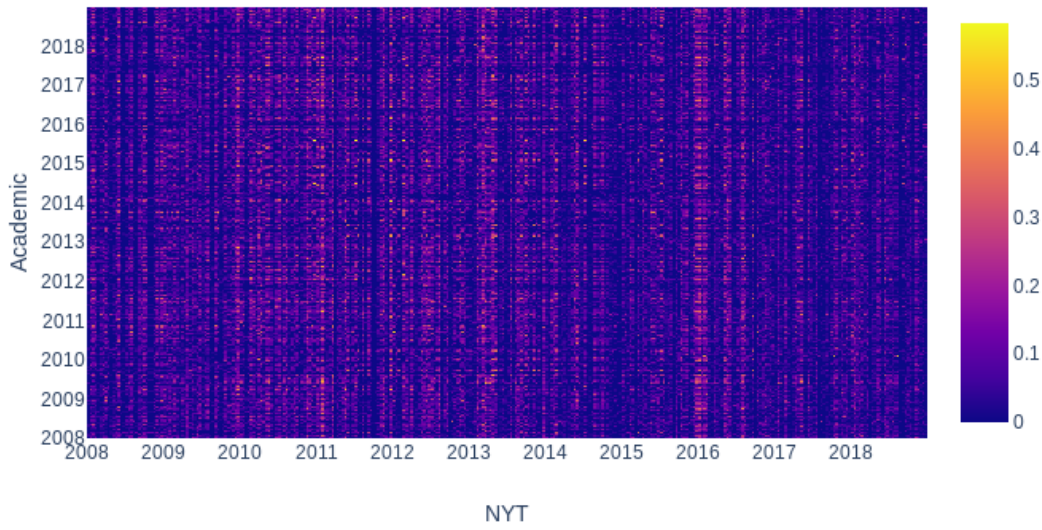


Fig. 8. The heat-map graph for the relationship between popularity of media and academia corpora

media began striking attention on Bitcoin likely because the bitcoin-driven Silk Road dark market released in late 2011 [18]. Silk Road was the first sophisticated dark market to anonymously trade illegal things using Bitcoin. Then, academia perceived the importance of the technologies behind Bitcoin and increasingly advanced them from 2013 on-wards. On the other hand, excluding the *Internet* topic as it is a general topic, *Rootkit* and *Certificate* are the most prevalent in academic

papers between 2008 and 2010. This widespread support of researchers caused the increasing coverage of media in the four years following 2010. The browsers' security and digital certificate have been noticeably considered from 2010 to 2014 in the media corpus. Thus, the recently prevalent using of the *https* protocol, which is based on a certificate, is rooted in mutual collaboration of journalists and researchers.

5 CONCLUSION AND FUTURE WORK

In this paper, we proposed a time-based trend analysis framework to investigate the relationship between two prominent text resources, academia and media, in the context of cybersecurity trends. The proposed framework builds topic models over the last 10 years of cybersecurity-oriented top conferences as well as the news articles published in the *New York Times*. We highlighted the top trends and identified the time-gap between most trending topics in cybersecurity both in academia and media. We also investigated the volume of coverage in both academic and media corpora. Finding such gaps can be used by decision makers and businesses to proactively deal with the new cybersecurity trends. Moreover, it helps journalists and researchers to have a more focused study on cybersecurity topics that matter.

There are some approaches for future studies. Increasing levels of cybercrime is strong evidence of the importance of cybersecurity-driven content. First, the gap analysis of cybersecurity trends between newspaper corpora and cybersecurity whistle-blowers (e.g., Edward Snowden, WikiLeaks) can unexpectedly reveal some noteworthy points. Moreover, the correlation between cybersecurity-based standards and patents with either academic papers or the whistle-blowers' reports could yield interesting insights. Such insights can be further studied for causality analysis. Third, our proposed framework can be applied for other applications or even some specific and targeted sections of cybersecurity.

6 ACKNOWLEDGEMENT

This work was supported in part by Natural Sciences and Engineering Research Council of Canada (NSERC) under Discovery Grant (RGPIN-2019-06150, PI: Atefeh Mashatan, RGPIN-2018-05041, PI: Morteza Zihayat). The authors would like to acknowledge the efforts of Reaz Huq and members of the Ryerson University Cybersecurity Research Lab for their support during this study. The authors would like to thank the Associate Editors as well as the anonymous reviewers for their valuable feedback and insightful suggestions which helped the authors to greatly improve the paper.

REFERENCES

[1] European Parliament : Council of the European Union. 2016. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).* Publications Office of the European Union. Retrieved 2018 from https://publications.europa.eu/en/publication-detail/-/publication/3e485e15-11bd-11e6-ba9a-01aa75ed71a1/language-en

[2] Mohamad Syahir Abdullah, Anazida Zainal, Mohd Aizaini Maarof, and Mohamad Nizam Kassim. 2018. Cyber-Attack Features for Detecting Cyber Threat Incidents from Online News. In *Cyber Resilience Conference (CRC).* IEEE, Xplore, Malaysia, 1–4.

[3] Stephen Adams, Bryan Carter, Cody Fleming, and Peter A Beling. 2018. Selecting system specific cybersecurity attack patterns using topic modeling. In *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE).* IEEE, 490–497.

[4] Noura Al Moubayed, David Wall, and A Stephen McGough. 2017. Identifying Changes in the Cybersecurity Threat Landscape using the LDA-Web Topic Modelling Data Search Engine. In *International Conference on Human Aspects of Information Security, Privacy, and Trust.* Springer, Vancouver, Canada, 287–295.

[5] Rubayyi Alghamdi and Khalid Alfalqi. 2015. A survey of topic modeling in text mining. *International Journal of Advanced Computer Science and Applications (IJACSA)* 6, 1 (2015), 147–153.

[6] Victor Andrei and Ognjen Arandjelović. 2016. Identification of promising research directions using machine learning aided medical literature analysis. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2471–2474.

[7] Tamir Bechor and Bill Jung. 2019. Current State and Modeling of Research Topics in Cybersecurity and Data Science. *Systemics, Cybernetics and Informatics* 17 (2019), 129–156.

[8] Adham Beykikhoshk, Ognjen Arandjelović, Dinh Phung, and Svetha Venkatesh. 2018. Discovering topic structures of a temporally evolving document corpus. *Knowledge and Information Systems* 55, 3 (2018), 599–632.

[9] Adham Beykikhoshk, Ognjen Arandjelović, Svetha Venkatesh, and Dinh Phung. 2015. Hierarchical Dirichlet process for tracking complex topical structure evolution and its application to autism research literature. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 550–562.

[10] David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*. ACM, New York, NY, USA, 113–120.

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

[12] Constantine Boussalis and Travis G Coan. 2016. Text-mining the signals of climate change doubt. *Global Environmental Change* 36 (2016), 89–100.

[13] ID Theft Center. 2016. *Identity Theft Resource Center Breach Statistics 2005-2016.* Technical Report. Identity Theft Resource Center,. Retrieved 2019 from https://www.idtheftcenter.org/images/breach/Overview2005to2016Finalv2.pdf

[14] Marie Chandelier, Agnès Steuckardt, Raphael Mathevet, Sascha Diwersy, and Olivier Gimenez. 2018. Content analysis of newspaper coverage of wolf recolonization in France using structural topic modeling. *Biological conservation* 220 (2018), 254–261.

[15] Allison June-Barlow Chaney and David M Blei. 2012. Visualizing topic models. In *Sixth international AAAI conference on weblogs and social media*. The AAAI Press, Dublin, Ireland, 481–485.

[16] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*. 288–296.

[17] Hongshu Chen, Guangquan Zhang, Donghua Zhu, and Jie Lu. 2017. Topic-based technological forecasting based on patent data: A case study of Australian patents from 2000 to 2014. *Technological Forecasting and Social Change* 119 (2017), 39–52.

[18] Nicolas Christin. 2013. Traveling the Silk Road: A measurement analysis of a large anonymous online marketplace. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, New York, NY, USA, 213–224.

[19] Lynne Coventry and Dawn Branley. 2018. Cybersecurity in healthcare: A narrative review of trends, threats and ways forward. *Maturitas* 113 (2018), 48–52.

[20] Sauvik Das, Joanne Lo, Laura Dabbish, and Jason I Hong. 2018. Breaking! A Typology of Security and Privacy News and How It's Shared. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1.

[21] Cyrus Farivar. 2009. A brief examination of media coverage of cyberattacks (2007-Present). *The virtual battlefield: Perspectives on cyber warfare* 3, 182-188 (2009), 182–188.

[22] Richard Fothergill, Paul Cook, and Timothy Baldwin. 2016. Evaluating a Topic Modelling Approach to Measuring Corpus Similarity. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (23-28). European Language Resources Association (ELRA), Paris, France, 23–28.

[23] GitHub. 2017. *Data repository for pretrained NLP models and NLP corpora.* Github. Retrieved 2017 from https://github.com/RaRe-Technologies/gensim-data/releases/tag/patent-2017

[24] Public Safety Canada government. 2018. *National Cybersecurity Strategy, Canada's Vision for Security and Prosperity in the Digital Age, 2018.* The Canada government. Retrieved 2018 from https://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/ntnl-cbr-scrt-strtg/ntnl-cbr-scrt-strtg-en.pdf

[25] Do-Heon Jeong and Min Song. 2014. Time gap analysis by the topic model-based temporal technique. *Journal of informetrics* 8, 3 (2014), 776–790.

[26] Leon Kappelman, Vess Johnson, Russell Torres, Chris Maurer, and Ephraim McLean. 2019. A study of information systems issues, practices, and leadership in Europe. *European Journal of Information Systems* 28, 1 (2019), 26–42.

[27] Spyridon Kavvadias, George Drosatos, and Eleni Kaldoudi. 2019. An Online Service for Topics and Trends Analysis in Medical Literature. In *World Congress on Medical Physics and Biomedical Engineering 2018*. Springer, Singapore, 481–485.

[28] Shinya Kawata and Yoshi Fujiwara. 2016. Constructing of network from topics and their temporal change in the Nikkei newspaper articles. *Evolutionary and Institutional Economics Review* 13, 2 (2016), 423–436.

[29] Farzan Kolini and Lech J Janczewski. 2017. Clustering and Topic Modelling: A New Approach for Analysis of National Cyber security Strategies.. In *PACIS*. Association for Information Systems AISeL, Malaysia, 126.

[30] Zili Li, Li Zeng, and Zhigang Luo. 2017. Identifying the Landscape and Trends in Cyberspace Research: From 1989 to 2016. In *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*. IEEE, IEEE Xplore, Orlando, FL, 22–29.

[31] Tom Lippincott, Diarmuid Ó. Séaghdha, Lin Sun, and Anna Korhonen. 2010. Exploring Variations Across Biomedical Subdomains. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 689–697. http://dl.acm.org/citation.cfm?id=1873781.1873859

[32] Yang Lu and Li Da Xu. 2018. Internet of Things (IoT) cybersecurity research: a review of current research topics. *IEEE Internet of Things Journal* 6, 2 (2018), 2103–2115.

[33] Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM computing surveys (CSUR)* 33, 1 (2001), 31–88.

[34] David J Newman and Sharon Block. 2006. Probabilistic topic decomposition of an eighteenth-century American newspaper. *Journal of the American Society for Information Science and Technology* 57, 6 (2006), 753–767.

[35] Binling Nie and Shouqian Sun. 2017. Using text mining techniques to identify research trends: A case study of design research. *Applied Sciences* 7, 4 (2017), 401.

[36] Kasey Panetta. 2017. *Top trends in the gartner hype cycle for emerging technologies, 2017.* Technical Report. Gartner. Retrieved 2019 from https://www.gartner.com/smarterwithgartner/top-trends-in-the-gartner-hype-cycle-for-emerging-technologies-2017/

[37] Paul Bowen C.F. Chui Philippe Alcoy, Steinthor Bjarnason. 2016. *Insight to the Global Threat Landscape, NETSCOUT Arbor's 13th Annual Worldwide Infrastructure Security Report.* Technical Report. NETSCOUT Arbor company. Retrieved 2019 from https://pages.arbornetworks.com/rs/082-KNA-087/images/13th_Worldwide_Infrastructure_Security_Report.pdf

[38] Grigori Sidorov, Alexander Gelbukh, Helena Gómez-Adorno, and David Pinto. 2014. Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas* 18, 3 (2014), 491–504.

[39] Bhargav Srinivasa-Desikan. 2018. *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras.* Packt Publishing Ltd, Birmingham, UK.

[40] Adam Stone. 2017. *How much does federal government spend on cybersecurity?* Fifth Domain. Retrieved 2019 from https://www.fifthdomain.com/civilian/2017/09/01/how-much-does-federal-government-spend-on-cybersecurity/

[41] Quintus Van Galen and Bob Nicholson. 2018. In Search of America: Topic modelling nineteenth-century newspaper archives. *Digital Journalism* 6, 9 (2018), 1165–1185.

[42] David A Weaver and Bruce Bimber. 2008. Finding news stories: a comparison of searches using LexisNexis and Google News. *Journalism & Mass Communication Quarterly* 85, 3 (2008), 515–530.

[43] Jianying Zhou. 2018. *Computer Security Conference Ranking and Statistic.* TEXAS A and M University. Retrieved 2018 from http://faculty.cs.tamu.edu/guofei/sec_conf_stat.htm