Analysis of Trending Topics and Text-based Channels of Information Delivery in Cybersecurity

TINGMIN WU, Swinburne University of Technology, Australia and CSIRO's Data61, Australia
WANLUN MA and SHENG WEN, Swinburne University of Technology, Australia
XIN XIA, Monash University, Australia
CECILE PARIS and SURYA NEPAL, CSIRO's Data61, Australia
YANG XIANG, Swinburne University of Technology, Australia

Computer users are generally faced with difficulties in making correct security decisions. While an increasingly fewer number of people are trying or willing to take formal security training, online sources including news, security blogs, and websites are continuously making security knowledge more accessible. Analysis of cybersecurity texts from this grey literature can provide insights into the trending topics and identify current security issues as well as how cyber attacks evolve over time. These in turn can support researchers and practitioners in predicting and preparing for these attacks. Comparing different sources may facilitate the learning process for normal users by creating the patterns of the security knowledge gained from different sources. Prior studies neither systematically analysed the wide range of digital sources nor provided any standardisation in analysing the trending topics from recent security texts. Moreover, existing topic modelling methods are not capable of identifying the cybersecurity concepts completely and the generated topics considerably overlap. To address this issue, we propose a semi-automated classification method to generate comprehensive security categories to analyse trending topics. We further compare the identified 16 security categories across different sources based on their popularity and impact. We have revealed several surprising findings as follows: (1) The impact reflected from cybersecurity texts strongly correlates with the monetary loss caused by cybercrimes, (2) security blogs have produced the context of cybersecurity most intensively, and (3) websites deliver security information without caring about timeliness much.

CCS Concepts: • **Security and privacy → Social aspects of security and privacy;**

Additional Key Words and Phrases: Empirical study, trend analysis, cybersecurity topics, news, security blogs

ACM Reference format:
Tingmin Wu, Wanlun Ma, Sheng Wen, Xin Xia, Cecile Paris, Surya Nepal, and Yang Xiang. 2021. Analysis of Trending Topics and Text-based Channels of Information Delivery in Cybersecurity. *ACM Trans. Internet Technol.* 22, 2, Article 52 (October 2021), 27 pages.
https://doi.org/10.1145/3483332

Authors' addresses: T. Wu, Swinburne University of Technology, Australia, CSIRO's Data61, Australia; email: tingminwu.work@gmail.com; W. Ma, S. Wen, and Y. Xiang, Swinburne University of Technology, Australia; emails: wma@swin.edu.au, swen@swin.edu.au, yxiang@swin.edu.au; X. Xia, Monash University, Australia; emails: xin.xia@monash.edu; C. Paris and S. Nepal, CSIRO's Data61, Australia; emails: Cecile.Paris@data61.csiro.au, Surya.Nepal@data61.csiro.au.

1533-5399/2021/10-ART52 $15.00
https://doi.org/10.1145/3483332

1 INTRODUCTION

Humans are playing an indispensable role in cybersecurity and, because of that, are especially targeted in cyber attacks [23, 32]. CybSafe analysis of UK ICO reports that 90% of data breaches were caused by human mistakes in 2019 [19]. Computer users generally have difficulties in making security decisions due to lack of knowledge, cognitive limitation or deviations from rationality [3]. However, they have to deal with sophisticated intrusions when their security software, such as antivirus or firewall, become obsolete [78]. To keep users in the loop is vital as any security measures can leave users more vulnerable when they lose resistance to unknown attacks.

End users are expected to learn more about cyber attacks, security measures and the key techniques that keeps them informed about cyber risks, and they need to take timely actions. Domain knowledge, especially in the cybersecurity field, is not easily turned into cognitive abilities without proper training [6]. Nonetheless, formal security education or training is time-consuming and requires users' undivided attention, and one-size-fits-all trainings hardly keep people engaged as they might have different learning preferences or background knowledge [69]. Such unthoughtful schemes can even cause market losses [59]. Compared to certification programs, online cybersecurity texts give internet users easier access to security knowledge to make correct decisions in the time of cyber incidents.

We identified three sources as the grey literature [65] that users often find for security texts from in their daily life: news, security blogs, and websites. News is published by news agencies as the leading media for general audience. Examples include BBC, USA Today, and so on. Security blogs can be more tailored toward security experts or individuals (including general users) who are interested in cybersecurity. These blogs mainly post security articles consisting of the latest threats, experts' opinions and security solutions for both businesses and individuals to use in practice. Websites include any information provided by authorised organisations (government, research institutes or industries), for the purpose of guiding the readers to behave securely online. This grey literature provide a range of educational materials that can benefit different communities.

The majority of existing analyses have failed to consider all the user-accessible resources of grey literature to provide users with a large selection for informal security learning. This selection could include studies on cyber threats [12, 38] and threat intelligence [1, 76]. Several studies [54, 67, 71] analysed the security knowledge from multiple sources, but the results are outdated now and the data collection was done in a relatively short period (e.g., from 2011 to 2015 in Reference [54]). Additionally, the trendlines of different topics show how they develop and give direction to ongoing studies, but they have barely been analysed before [67, 71]. Some prior research focused on producing security information, but their inferences from information were biased due to lack of timeliness, or were hard to be adopted in the real world [10, 60, 61, 72]. For example, security information sharing informally produces security incident reports, mainly from websites. However, the release requires time to verify whether they meet various standards or not, and might miss the timing of reporting emerging attacks such as zero-day vulnerability [72]. Lack of standardisation also hinders the exchange of security information. Moreover, current security advice is usually too technical to understand or not actionable due to their restrictions (e.g., "never click on links in emails") [60, 61]. Prior works used **Latent Dirichlet Allocation (LDA)** [9] to cluster security questions [13, 54, 83]. However, traditional LDA does not perform well in capturing domain-specific concepts [51]. The topics it generates have low granularity and are hard to distinguish.

To address the issue, we propose a semi-automated classification method to generate broad topics in cybersecurity instead of using LDA-generated topics. We first divide our collected security texts into five datasets according to their sources. For each dataset, we run LDA separately. We find that the generated topics by LDA do not capture the domain-specific concepts and are not

distinct to each other. To derive more meaningful results, we use the term extraction method to generate a set of terms that summarise the categories for each LDA-generated topic. We identify 16 security categories that summarise all those terms. We analyse the popularity and impact of those categories to analyse different cybersecurity trends. More specifically, we compare how the security issues evolve across categories and sources over the last decade. This sheds light into the development of security issues in the past 10 years and reveals how challenges emerge, which in turn can be used in the prediction of unknown threats. The analyses of security issues and differences between the sources also generate patterns in delivering security knowledge to the general public.

Our research focuses on answering three research questions:

RQ1. What are the security issues reported in cybersecurity texts?

We discovered 16 security categories for cybersecurity texts from news, security blogs and websites. They can summarise security issues, including the types of cyber attacks and security techniques. We found that information privacy still remained a dominant topic in the last decade, and this was largely due to criminal offence (including password attack), mobile application attack, and network attack. We also noticed that most articles (83%) discussed multiple security issues (relevant to up to six security categories).

RQ2. How have the security categories varied and evolved over the last decade?

Cybercriminal activity has been the most popular and was discussed in most security articles (65%), followed by the privacy issue, preventive measures (i.e., cybersecurity software, service, and program), with similar popularities at 40%; The increase of the absolute impact of the security categories indicates security incidents evolution in both amount and sophistication over the last decade. Security issues in mobile/application and information privacy gained the largest absolute/relative impact over time. The explosion of ransomware (e.g., WannaCry) brought the absolute impact of *malware/virus* to its peak and exceeded the values of all the other categories. The overall absolute impact of all the security categories strongly correlates with the economic loss caused by cybercrimes. Election security has gained a sudden increase in the absolute impact in 2016, which coincides with the U.S. presidential election campaign.

RQ3. How have the security categories varied and evolved across different sources on cybersecurity over the last decade?

Almost all the categories are popularly present within the three sources, except the categories *election security* and *false/misleading claim* that are only prevalent in news and webs, respectively. Among the three sources, security blogs have largest popularity and impact over time in the majority of categories. The absolute impact of news and security blogs shows upward trends for all the categories in the 2010s, while *false/misleading claim* has a downtrend in webs. Security issues in mobile/application have been the most influential in news and security blogs during the past 10 years, followed by the privacy issue. Threats in IoT show comparable absolute/relative impact to the privacy issue in news. News and security blogs report security events for the first time at similar speed on most categories. Websites deliver security information without caring much about timeliness, with one third of the articles not specifying the date and the rest having a time lag in posting emerging security issues.

We list our contributions as follows:

- We build a large collection of cybersecurity texts (187,319 articles) from three online sources: news, security blogs, and websites.
- We propose a semi-automated classification with combining the term extraction method and the open card sorting [73] to derive the categories of our texts instead of using LDA-generated topics. We identify 16 security categories to analyse the security issues.

- We conduct an empirical study to analyse the comparison and evolution of the security categories over the last decade as well as across the sources to shed light on the trends of security issues.

The rest of our article is organised as follows. Section 2 reviews the related work. In Section 3, we describe the background of topic modelling and term extraction. Section 4 introduces our research questions and methodology. We present our findings and results in Section 5. Section 6 discusses the implications and threats to validity. We make conclusions and propose the scope of future work in Section 7.

2 RELATED WORK

In this section, we review the existing works on users' selection of security information sources, security perceptions and learning about security.

2.1 Selection of Security Information Sources

There is a proven relationship between security information sources and users' online experience about security and privacy [56, 57]. Users are different in their engagement in security protection scenarios and, thus, have different demands of expertise from the sources [27]. Rader and Wash [54] identified patterns from informal sources of security information that help users seek useful data to behave safely or solve potential risks. Ion et al. [37] compared the security practices from different people. They found that experts mainly suggest regarding security updates and using password managers. In contrast, non-experts mostly suggest clicking only on official websites links and regular password changes. Sauerwein et al. [67] compared public sources of security information. Das et al. [22] studied how different users gain different information from security news. Similarly, Sheshadri et al. [70] analysed privacy news as it impacts users' perception and behaviours. Shillair and Meng [71] compared the impact of different sources in changing users' security behaviours.

Users often seek security information from multiple sources while considering different factors. Nthala and Flechais [50] performed qualitative studies and showed that people applied some measures such as professional level, academic standing and negative experience of the sources. Redmiles et al. [58] found that users measure the trustworthiness of cybersecurity information by the sources for digital security advice and by the content for physical security advice. Nicholson et al. [49] conducted interviews with elderly people about their choices for security-related sources of information and found that they preferred social sources over experts' advice.

In addition, other sources such as app reviews and breach reports also contain some security and privacy issues. Haering et al. [33] introduced an automatic approach to match problem reports in app reviews to bug reports. Xia et al. [81] proposed a method to predict crashing mobile app releases based on mobile app repositories. Similarly, Li et al. [43] applied user review to detect problematic mobile app updates. The study on users' perception toward the Equifax breach found that users tended to undervalue the chance to become victims and procrastinated taking actions even though they recognised the risks [85]. Later, they suggested improving data breach notices in terms of readability, media penetration, format, and risk indication [84, 86]. Murukannaiah et al. [47] proposed a semi-automated approach to identify privacy incidence from different online resources such as new, blogs and social media. Kafali et al. [40] developed a semantic process to measure the gap between security policies and reported beaches and revealed a coverage of 65% between the **U.S. Health Insurance Portability and Accountability Act (HIPAA)** and reported breaches by the **U.S. Department of Health and Human Services (HHS)**. Guo et al. [31] presented a method to leverage crowdsourcing to extract security and privacy requirements from security regulations and breach reports and conducted their evaluation on HIPAA and breach reports by HHS.

Xu et al. [82] conducted a trend analysis of a 12-year breach report dataset. A study examined the trend in information privacy based on research papers [15]. Fagan and Khan [24, 25] studied users' considerations on benefit and risk when they decide whether to follow the security advice or not.

2.2 Security Learning

Security education or training has attracted a large amount of research. This includes the studies on phishing attack prevention [34, 53, 79], browser warnings [4, 75] and password protection [26, 29, 68]. Safa and Solms [66] shed light on how security knowledge can help reduce the risks of cyber incidents. Some users suffered, because they overestimated their knowledge of security [18]. Abu et al. [2] developed mental models for users to help them protect their privacy, e.g., with E2E encryption. Stevens et al. [74] did threat modelling in different enterprise scenarios and showed its efficacy in security defence. Chen et al. [14] designed a desktop game to teach a series of security practices that users can apply in the real world. Wu et al. [80] studied users' understanding of security texts and built a corpus to explain security terms. Golla et al. [30] studied users' understanding of security warnings and designed password-reuse notifications based on their perceptions.

While most of the existing works focused on modelling users' security behaviours or developing the tools for security education, there is no recent study analysing the trends of security topics on a large-scale easily accessible online texts. Compared to formal security training, the grey literature about cybersecurity (e.g., newspapers, personal stories, online forums, professional guidelines) are more approachable and diverse to seek help and read regularly. The resources deliver important information somehow enable users to make good security decisions. We need an empirical study to exhaust the cybersecurity texts and understand what security issues they report and how they evolve over time as well as difference between the sources. Such a study can help improve informal security learning for end users and forecast innovative cyber attacks.

3 BACKGROUND

3.1 Topic Modelling

Topic modelling is a machine learning technique to discover the topics for a collection of documents based on text-mining [8]. LDA [9] is one of the most common algorithms for topic modelling and has been used in previous studies to identify topics in different areas [45, 54]. LDA is a probabilistic model that for each document, gives a set of topic probabilities. Each topic is a set of words with different weights [9]. The model considers word occurrences and co-occurrences within a document as well as across different documents in the whole corpus.

3.2 Term Extraction

Term extraction is an important subtask of information retrieval in various linguistic areas [52]. The purpose of term extraction is to locate the terms that contain informational content from a set of documents. TermSuite [17] is an open-source toolbox that can identify (multi-word) term variants based on syntactic and morphological patterns. The termhood is measured by the relative frequency in a domain-specific corpus as well as a general corpus. The candidate terms are selected with measure values higher than 2, a threshold recommended in Reference [17].

4 STUDY SETUP

4.1 Research Questions

RQ1. What are the security issues reported in security texts?

Security texts deliver news and articles about cybersecurity for a range of technology enthusiasts and general users. They explain the attack techniques and distribute security tips, guidelines and advice for both businesses and home computer users. Categorising the security texts can help identify the security issues. The analysis of the issues sheds light on the challenges faced by researchers and practitioners to advance the development of threat intelligence to protect the security and privacy of online users. In addition, the security issues identification creates the knowledge patterns for users when they seek security advice online.

RQ2. How have the security categories varied and evolved over the last decade?

It is critical to update the topic analysis with the most recent posts. Although there are similar works that have studied security topics, their results are not useful anymore, since they have been outdated by at least five years. The worldwide financial loss caused by cybercriminals are predicted to be $6 billion per year in 2021, increasing from $3 billion in 2015 [28]. Intrusions become more sophisticated and hackers employ more advanced techniques. In the analysis of the latest trends and drawing a big picture for the security issues, we are the first to identify the security categories systematically beyond using LDA. By doing research on the differences and similarities between LDA-generated topics and our defined security categories, we provide more distinct security topics with a more comprehensive analysis.

RQ3. How have the security categories varied and evolved across different sources on cybersecurity over the last decade?

Different sources can deliver security information in different ways. News articles are generally published by authorised newspapers and report the latest security events. Security blogs also report the latest news on cybersecurity but might give more insights into the key techniques used from research or technical papers. Websites mainly come from organisations such as universities and banks. They commonly focus on providing informal security information such as security advice for educational purposes.

Understanding different topics from distinct sources can help us cater to the needs of users with different backgrounds. Users might also be concerned about different attacks or data breaches to various degrees. Different sources have different preferences over featured articles and techniques. Compared to RQ2, RQ3 mainly aims to analyse how the topics evolve across different sources. This analysis can help in informing users where to acquire sufficient security knowledge from and in detecting the emerging trends over platforms.

4.2 Research Methodology

To answer the three research questions, we collected real-world media texts and conducted comparative analyses that could informally provide end users with security knowledge. We focused on the topic trends across different security categories as well as different sources to provide insights into how security issues evolve.

4.2.1 Data Collection. We collected our cybersecurity texts from three sources of grey literature: news, security blogs, and websites. We mainly focused on the easily accessible online articles that computer users read to gain security knowledge. We only collected articles from the year 2000 to the date of paper writing.

We developed a crawler in Python by leveraging Beautiful Soup library [63]. We only extracted text contents (including titles) for topic analysis. We stripped out images, videos, and meaningless contents (e.g., navigation menu and contact information). The publication dates of the articles were extracted from the search results or taken out of the text contents for trend analysis.

We selected the sources based on their popularity, impact and relevance. As summarised in Table 1, we applied the criteria to include and exclude the security texts. In the following, we

Table 1. Inclusion/Exclusion Criteria for the Security Texts

Inclusion criteria	Exclusion criteria
• Published after the year 2000. • Written in English. • Full text available. • Search results of the specific keywords (news and webs).	• Written in a language other than English. • No full text available. • Content not relevant to cybersecurity. • Content is technical to general users (e.g., articles that contains code only).

explain how we selected the articles and search results from each source. Details of the sources can be found in our supplementary material.[1]

News. We selected the newspapers published in English with top circulation (>100,000) worldwide. We included all the 16 news sources used in a similar study [54]. In addition, we added three more news sources that have become more prevalent recently, e.g., Herald Sun (circulation: 303,140 in 2018), and Tech News World (Reader purchase >$100 billion per year).

To identify the contents on cybersecurity, we applied 27 keywords as filters to search for relevant articles only. We included the 25 terms used in Reference [54], and added two new keywords ("cybersecurity" and "cyber attack") in the set. According to Google Trends data, people have searched the keyword "cybersecurity" seven times more frequently in the past few years. We manually went through the found articles though most of them were applicable.

During collection, four news sources were removed, because they restricted reading the articles and required subscription or purchasing membership plans, e.g., The Globe and Mail. We combined the search results of all the keywords and removed the duplicates. Altogether, 68,066 articles were collected from 15 newspapers.

Security blogs. We then collected texts from the blogs on cybersecurity that provided the latest security news or articles for computer users with various levels of tech knowledge. The blogs can feature threat intelligence to educate their audience in taking protective measures against cyber attacks. We started the collection of security blogs from the sources used in Reference [44]. We also applied the list and extracted security articles in our previous study [80]. We further extended the security blogs by checking recommendations on reputable websites. We selected the source if it had been recommended three times or have 1 million followers, e.g., "The Hacker News" has more than 2 million followers on Facebook. We also included a few security blogs hosted by governments such as AUSCERT.[2] Forty-one of 42 blogs remained after removing the ones with non-text posts, such as commands, attached files, and images. We also manually verified the contents to confirm their relevance to cybersecurity. In total, we collected 109,587 articles from the blogs.

Webs. We extended the domain of web pages used in the existing studies to cover all the applicable ones. A similar study [54] collected the web pages with which organisations delivered information or instructions on cybersecurity to their employees to help them be aware of risks and behave safely online. We classified the organisations that provide this information into three types: governmental (federal/state government agencies), industrial (telecommunications companies, social network companies and banks), and academic (universities and research agencies). In addition to the web pages used in Reference [54], we collected more pages from the top-ranked organisations in Australia and divided them into the above-mentioned classes.

We applied the 45 keywords used for web page search in a study [54], combined the search results and removed the duplicates. We also removed the ones that were empty or not

[1] https://github.com/ktd4869/Security_trend_analysis/blob/main/TOIT_supplementary_material.pdf.
[2] https://www.auscert.org.au/resources/blogs-publications/.

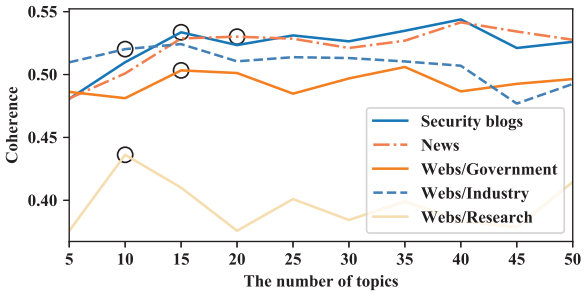

Fig. 1. The coherence of the models generated with the different number of topics (5, 10, . . . , 50). The optimal number of topics is highlighted with a circle marker.

Table 2. Articles Statistics per Dataset after Sanitisation

Dataset	#total articles collected	#articles after santisation	#words per article	
			Mean	SD
News	68,066	51,685	822	873
Security blogs	109,587	108,354	906	2,248
Webs/Governmental	14,047	9,618	655	1,533
Webs/Industrial	25,917	16,810	716	1,475
Webs/Academic	1,430	852	813	1,514

security-related. The final collected dataset contained 41,394 articles from 41 webs (17 governmental, 15 industrial, and 9 academic).

4.2.2 Generation of Topics. To identify what is being discussed in cybersecurity texts, we applied a topic modelling algorithm to extract/generate topics from our collected articles. We used LDA to extract topics from the security texts. We ran LDA on each dataset separately, since LDA is proven to be biased with large datasets [35]. We implemented the algorithm in Python by using its "gensim" library [62]. We identified the optimal number of topics based on topic coherence [64]. To identify the optimal number of the topics generated by LDA per dataset, we swept the 5 to 50 interval with steps of 5 and generated separate models accordingly. We used topic coherence to measure the performance of the models [64]. Figure 1 depicts the results of the models. The optimal one is marketed with a circle in each case. Our rule was to pick the model with the highest coherence. However, if the coherence did not increase much (difference $\leq$0.01) after the first peak, then we kept the first highest value. For instance, the coherence for security blogs only rose by 0.01 from 15 to 40 topics, so we took 15 as the optimal topic number. If two models had similar coherence (e.g., news models with 15 and 20 topics), then we manually compared the generated topics and selected more distinct one (e.g., news model with 20 topics). We manually read the generated topics and the articles per topics, and removed two topics in news model, since both the topics and the articles are irrelevant to cybersecurity.

Articles sanitisation. Based on the generated topics, we removed the articles that were not related to security. As each generated topic was presented as a list of words, we inferred the conceptually specific topics by reading and understanding the combinations. We selected the topics whose all words were irrelevant to cybersecurity. We then manually read most of the articles (>70%) from each of those topics and removed the ones whose contents were irrelevant. The statistics about the sanitised dataset is demonstrated in Table 2.

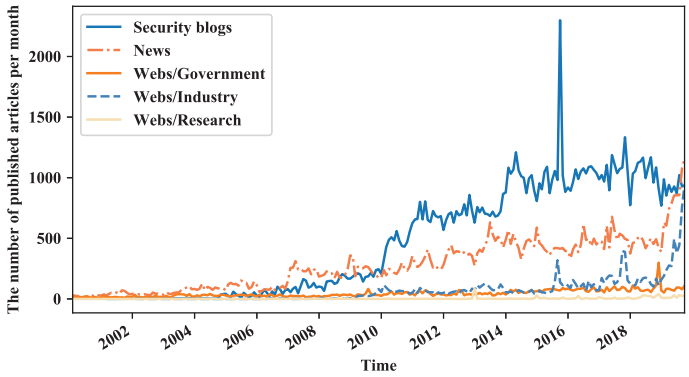

Fig. 2. The number of articles published from 2000 to now.

Figure 2 shows the number of published articles per month in the five datasets in the last 20 years. Most of the cybersecurity texts came to surface after 2010, with a regular posting afterwards. We see that security blogs account for the majority of the grey literature on security. With around half of the security blogs in post numbers, news volume shows a dramatic increase in the last year. Websites form a relative small portion, with a gentle growth over the time. In 2016 and 2018, two peaks can be spotted in the industrial websites curve. It has also jumped since last year. We observed that the number of published articles per month in the 2000s is far smaller than the number in the later 10 years. Therefore, we mainly focused on the analysis of trends in the 2010s in our study.

4.2.3 Security Categories Identification. We carefully examined each topic generated by LDA and reviewed the texts, but found the topics were not still satisfactory, because they could not cover the dataset completely and had overlapped excessively. To solve this issue and provide more in-depth insights, we further identified the security categories with term extraction and did manual category identification by card sorting instead of using topics generated by LDA [73]. We extracted terms from the articles of each topic separately by TermSuite [17]. We only kept the candidate terms with measure values higher than 2, a threshold recommended in Reference [17]. As a result, we have a collection of terms to replace and represent each topic generated by LDA, with a reasonable number of terms per topic ($Mean$: 46, SD : 23).

Category identification. We identified the security categories of our texts based on the lexical semantics of the generated terms. After removing the duplicates, we applied open card sorting [73] to categorise the 810 terms across all the topics. We randomly selected 100 terms and classified them into different categories, and then applied those to the rest and kept identifying new categories. In total, 16 categories were identified whose details are given in Table 3. We borrowed the abbreviation styling method in Reference [54] for the topics.

We further applied the 16 categories to all the terms, where each term assigned to a maximum of three categories. Three researchers from our faculty who had expertise in cybersecurity performed a manual classification. Each term was classified following the rule of majority voting [48]. We used Cohen's Kappa [16] to measure the agreement between each pair of labellers. The resultant values (all >0.93) indicate strong agreements between the labellers. An expert review was conducted to ensure the validity of the classification. We recruited two experts in a governmental research lab who had at least three years of experience in the cybersecurity field for this purpose. For each expert, we generated a 200-term sample (25%) to review, while our researchers were sitting next

Table 3. Sixteen Manually Classified Security Categories

Category	Definition	Example terms
CycmnAc cybercriminal activity	The malicious activity where the hacker group leverages computer techniques for illegal purposes.	malicious action, hacker, law enforcement action
CysePrg cybersecurity program	The cybersecurity venue or event hosted by an authorised organisation, e.g., awareness training, foundations learning, risk assessment.	cybersecurity conference, CISO Forum, consumer education
ElecSe election security	The protection of elections and voting infrastructure from cyberattack, e.g., tampering with or infiltration of voting machines and equipment, election office networks and practices, and voter registration databases.	voting security, election system, electronic voting machine
FMClm false /misleading claims	The deceptive advertising claimed by business online, illegal claims about product quality, condition, or price	deceptive claim/advertising, online complaint assistant
IdtFncFrd identity theft /financial fraud	Criminals gain unauthorised access to steal credentials to cause unintended charges.	data breach, financial crimes, credit card fraud
InfPry information privacy	Actions that harm or protect users' privacy preferences and personally identifiable information.	customer privacy/data, GDPR, privacy protection
IoTThr IoT threat	Security threats in IoT devices, software and network connected to the internet.	firmware, mobile device, industrial control systems
MalVr malware/virus	Malicious software developed to harm computers or networks.	spyware, adware, worm, trojan
MbAppSe mobile/application security	Security solutions or attacks at the software level, e.g., android apps.	fake android app, mobile security, mobile-threat report
NatSe national security	The security and defence of a nation-state, e.g., its citizens, economy, and institutions, which is regarded as a duty of government.	cyberespionage, national cybersecurity, transnational crime
NetAtk network attack	Malicious attempts to gain the unauthorised privilege of network or cause service disruption.	DDoS attack, zombie bot, remote code execution
PwdEnc password/encryption	Password/data protection and encryption.	MFA, RSA encryption
SeSwServ security software /services	Software or services designed to help users against attacks, e.g., antivirus products, educational services	antivirus, MalwareBytes, SIEM, security company,malware filter
SeUdVnb security update /vulnerability	Security weakness exploited by hackers to perform malicious activity. Security update fixes the system or application bugs.	flaw, patch, security bulletin, Microsoft Exploitability Index
SpmPh spam/phishing	Scammers spread unsolicited messages online or in social media with malicious links to steal sensitive information or infect computers.	scam, identity parameter, spam, junk/phishing email
WbAtk web-based attack	Malicious action on web browsers, extensions and content management, e.g., leveraging third-party plugins to perform code injection.	SQL injection, web extension, drive-by download

to the expert to respond to any questions based on the think-aloud protocol [42]. After our explanations, there was only one error correction, that merely added one term to one more category.

We built a corpus for each category as a set of terms. Each term was added into the corpus of its assigned categories. We used the corpus to measure the relevance of each category to

the documents. We identified the duplicate terms semantically or syntactically in the corpus and labelled them as term variants. For instance, "infected computer" is a variant of "infected machine," just as "sensitive data" is a variant of "sensitive information." We found that security solutions or attacks, especially those related to sensitive information (*information privacy*, *security software/service*, and *security update/vulnerability*) have the largest corpora of terms. In contrast, political or nationwide threats (*election security* and *national security*) contain fewest terms.

4.2.4 Metrics and Analysis. Instead of using the topic probability computed by LDA, we define the category relevance based on our identified security categories. We obtained a set of term corpora for K categories as $\boldsymbol{C} = \{C_1, C_2, \ldots, C_K\}$.

Category relevance. The category relevance of a document measures the proportion of terms in each category corpus that occur in the document. More specifically, the relevance of a document to each category is computed as

$$\gamma(d_i, C_k) = \frac{|c_k|}{|C_k|}, 1 \leq k \leq K, \tag{1}$$

where $|C_k|$ is the number of terms in a category corpus, and c denotes the subset of C whose terms occur in the document d_i. A term is counted once whether it or its variants occur in the document.

Dominant categories. As explained in Reference [77], we define the dominant categories of each document as

$$dc(d_i) = \{C_k\}, if \gamma(d_i, C_k) > \theta(C_k), 1 \leq k \leq K, \tag{2}$$

where $\theta(C_k)$ is the threshold to determine whether a category is dominant or not. We selected a representative sample and manually labelled the dominant categories as ground truth. For each category, the threshold is the value where the accuracy archives the highest. Each document can have multiple dominant categories. The concept of dominant categories enables us to classify the documents based on the values of relevance.

Category popularity. We applied the measures of popularity and impact defined in Reference [77] on the categories. We applied these two metrics to measure the interest of different categories in the security articles and the temporal trends of the categories. We define the popularity for the category C_k within the dataset D as

$$popularity(D, C_k) = \frac{|\{d_i\}|}{|D|}, d_i \in D, C_k \in dc(d_i). \tag{3}$$

The popularity of a category measures the proportion of documents with the given category as dominant.

Category impact. The absolute and relative impact of the category C_k is defined as

$$impact_{absolute}(D(month), C_k) = \sum_{d_i \in D(month)} \gamma(d_i, C_k), \tag{4}$$

$$impact_{relative}(D(month), C_k) = \frac{impact(D(month), C_k)}{|D(month)|}, \tag{5}$$

where $D(ts)$ represents a collection of documents posted in a month. The absolute impact of a category measures the cumulative relevance to the category of the posted documents over a month. The absolute impact is influenced by the number of posts and their category relevance. The relative impact is not affected by the number of posts. It measures the average relevance to the category of the posted documents during a month.

Table 4. Three χ^2 Tests on the Proportions of Articles with Each Category as the Most Relevant across the LDA-generated Topics, Datasets, and Sources Separately

Category	Across LDA topics		Across datasets		Across sources	
	χ^2	p	χ^2	p	χ^2	p
CycmnAc	7020	<0.01	978	<0.01	768	<0.01
CysePrg	8715	<0.01	1528	<0.01	1222	<0.01
ElecSe	6696	<0.01	418	<0.01	366	<0.01
FMClm	72679	<0.01	27129	<0.01	7796	<0.01
IdtFncFrd	12393	<0.01	886	<0.01	677	<0.01
InfPry	7216	<0.01	566	<0.01	554	<0.01
IoTThr	14526	<0.01	3241	<0.01	3131	<0.01
MalVr	32297	<0.01	3230	<0.01	2758	<0.01
MbAppSe	14383	<0.01	1434	<0.01	1409	<0.01
NatSe	26250	<0.01	2688	<0.01	2335	<0.01
NetAtk	14529	<0.01	1659	<0.01	1253	<0.01
PwdEnc	17847	<0.01	2452	<0.01	2060	<0.01
SeSwServ	1075	<0.01	43	<0.01	25	<0.01
SeUdVnb	42059	<0.01	1992	<0.01	1773	<0.01
SpmPh	19368	<0.01	1105	<0.01	263	<0.01
WbAtk	13433	<0.01	3158	<0.01	1359	<0.01

5 RESULTS

We exhibit our results following our methodology in this section. Through data analysis, we try to answer our three research questions.

5.1 RQ1. What Are the Security Issues Reported in Cybersecurity Texts?

We first generated 68 topics from our datasets by using LDA; 18 from news, 15 from security blogs, and 35 from the three web datasets. We then manually checked each topic and removed one that was hard to infer any specific topic. However, the LDA-generated topics were not good representatives and were hard to distinguish, too. Therefore, we further studied the terms from the topics and manually found 16 security categories that can represent the articles, as demonstrated in Table 3. The table explains each security category with examples in detail. The categories were identified based on different perspectives on cybersecurity, including attack types (e.g., network, web, IoT, or mobile/application attacks), security techniques (e.g., encryption and security services as well as updates) and recently emerged security issues (e.g., election and national security).

Category validation. We validated the effectiveness of our 16 security categories. We applied the chi-squared (χ^2) test on the consistency of each category prevalence. More specifically, we tested if the proportions of the articles were similar with a given category as the most relevant category. The most relevant category of a document is the category where it achieves the highest relevance (Equation (1)). Table 4 shows the results of χ^2 tests for each category across 67 LDA-generated topics in the five datasets from the three sources. p values are corrected with the Holm–Bonferroni correction [7]. As all the p values are smaller than 0.01, the prevalence of our classified categories is significantly varying across the LDA-generated topics, datasets and sources. It indicates our identified categories are representative and effective, because the differences between them are statistically significant.

Category co-occurrences. We explored the relationships between different categories by calculating their co-occurrences in each document. The co-occurrences show the associations between

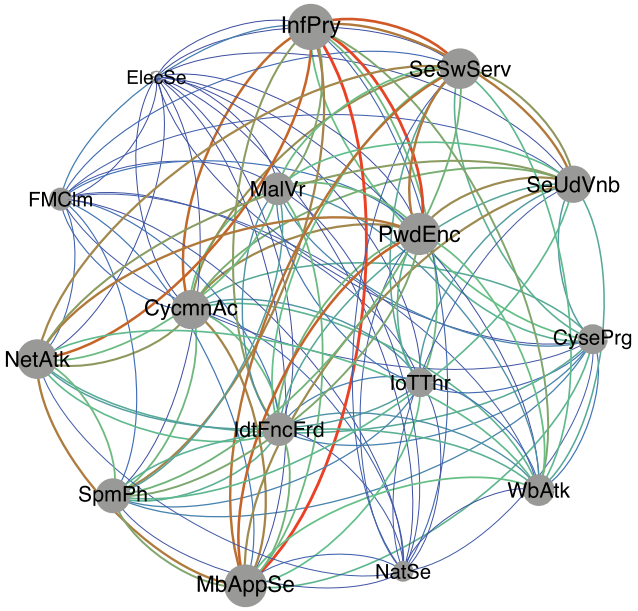

Fig. 3. The network graph of co-occurrences between different categories within an article. The node size represents the relative occurrences of each category. Red (thick) lines mean strong relationships (high co-occurrences) between two nodes, while blue (thin) represent weak relationships.

different security issues to pinpoint the challenges faced by researchers and practitioners. Figure 3 presents the network graph of the co-occurrences, where larger nodes indicate more frequently occurred categories and red (thicker) lines show strong relationships. We find that *information privacy*, *password/encryption*, *mobile/application security*, and *network attack* have the largest co-occurrences compared to other categories. These four categories have also had high numbers of occurrences in our dataset. *cybercriminal activity* exhibits strong relationship with *information privacy*. This indicates that information privacy still remains a dominant topic in the last decade and is largely due to criminal offence, including password attack (e.g., brute force attack), mobile application attack (e.g., malicious code injection exposure), and network attack (e.g., DDoS attack). Yet, usable authentication methods, mobile security solutions, and network protection are still challenges in safeguarding the sensitive data (e.g., credentials) of users and enterprises. In addition, the strong co-occurrences between *spam/phishing* and both *network attack* and *mobile/application security* denote that spam and phishing messages including malicious links are still spreading rampantly in the internet through email, SMS or other communications.

Among the 16 categories, *election security* and *national security* occur the least frequently and have the weakest correlation with the rest of the categories. The articles in these two categories mainly present nationwide attacks and espionage at high levels, with a focus on the infrastructure and attack consequences. They hardly analyse the related techniques in detail. Since the targets of these threats, such as governments, are harder to compromise compared to regular users, the attacks are not frequent. However, they are to be taken seriously, since they can cause significant losses such as political or military information leakage. Compared to these two categories, *false/misleading claim*, *IoT threat*, *web attack*, *cybersecurity program* happen more regularly but have weaker connections to other categories. Specific attacks such as *web attack* and *IoT threat* are partly related to a few categories. For instance, criminals can leverage cross-site scripting (web)

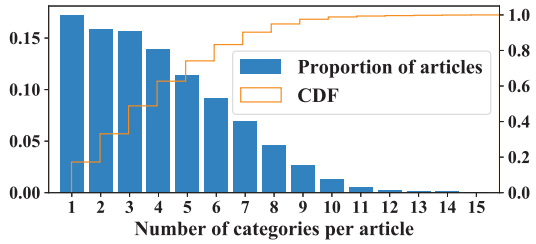

Fig. 4. Proportion of articles with different dominant categories and the CDF vs. the number of dominant categories (per article).

attacks to inject malicious codes into web applications (*mobile/application security*) and access sensitive information (*information privacy*).

Categories per article. We empirically found the threshold per category (Equation (2)) to determine the dominant categories for each article in our dataset. In Figure 4, we have plotted the probability distribution of the number of dominant categories for the articles along with its **Cumulative Distribution Function (CDF)**. With the increase in the number of dominant categories, the number of articles gradually decreases. The results show that 83% of the articles have six dominant categories or less. This percentage reaches 90% with seven categories. The results are aligned with our observations. In practice, different from other fields, these articles generally include multiple topics. For example, when an article introduces cyber attacks and prevention methods, it always explains the techniques and the related effects in detail. For example, a security update addresses an exploitable vulnerability through which remote code execution by hackers is possible. Hackers use this to gain admin access and run malware on infected computers. In this scenario, *network attack*, *password/encryption*, *malware/virus*, and *security update/vulnerability* are discussed. It also indicates that a security article generally discusses multiple security issues. And also, the informal online sources cover sufficient cybersecurity stories of varied contents to deliver the knowledge to home computer users.

- We identified 16 categories for the security articles from news, blogs and webs.
- Information privacy still remains a dominant topic in the last decade and is largely due to criminal offence, including password attack, mobile application attack, and network attack.
- Most of the articles (83%) have six dominant categories or less.

5.2 RQ2. How Have the Security Categories Varied and Evolved over the Last Decade?

5.2.1 Category Popularity. We compared the popularity of different categories. We empirically found the threshold to separate dominant categories (Equation (2)) and calculated the category popularity, too (Equation (3)). Figure 5 plots the popularity of the security categories. *cybercriminal activity* marks the most substantial category amongst all (with 65% popularity). This category contains the terms indicating cyber attacks such as "hack." The three categories *information privacy*, *security software/service*, and *cybersecurity program* share similar popularities at around 40%. In contrast, other categories are discussed less popularly, such as articles introducing specific threats (e.g., spam/phishing, malware/virus).

5.2.2 Category Absolute Impact. We calculated each category absolute impact (Equation (4)) to analyse the trends. Figure 6 demonstrates the comparison between the trends of absolute impact and relative impact per category (see the comparison between categories in Figure 7). Overall, there is an upward trend in the absolute impact for almost all categories since 2009 starting from

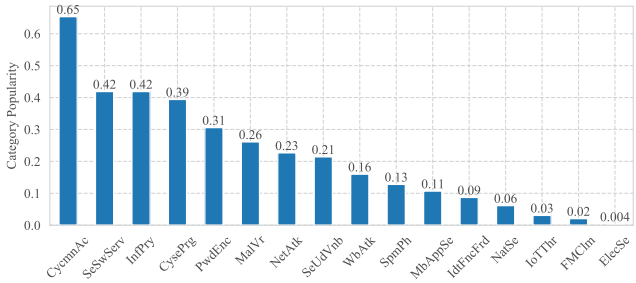

Fig. 5. The popularity of our security categories.

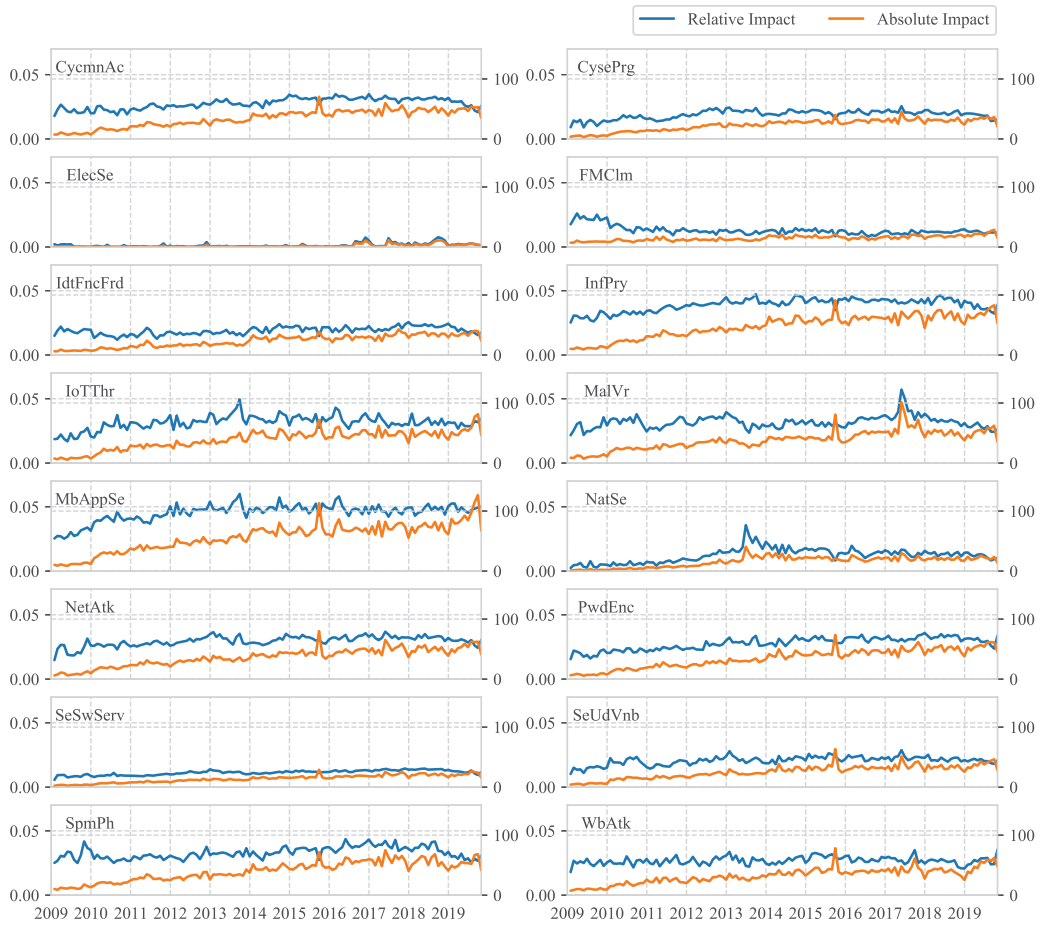

Fig. 6. The separate absolute impact and relative impact of 16 security categories over the last decade.

nearly zero. The increase indicates a considerable evolution of security incidences in both amount and sophistication. The explosion of ransomware in 2017 brings the impact of *malware/virus* to its peak, especially with the worldwide break out of WannaCry, which infected 200,000 computers across 150 countries [5].

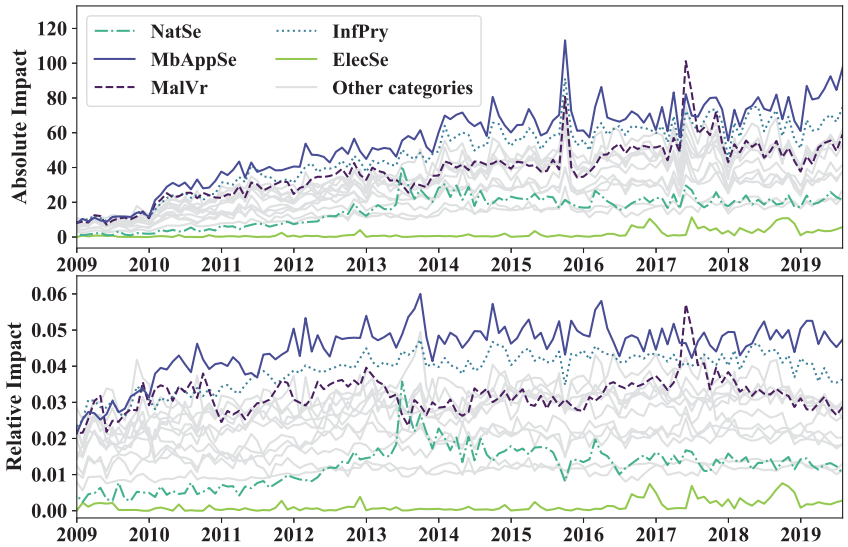

Fig. 7. The absolute impact and relative impact of 16 security categories over the last decade. Some categories are plotted in light grey to avoiding mixing the lines up.

We aggregated the overall absolute impact for all the categories and compared it to the monetary damage caused by cybercrimes in the 2010s (data from Reference [39]). We used Spearman correlation coefficient to measure the correlation between the overall absolute impact of security articles and the amount of financial loss caused by recorded cybercrimes. The result showed a strong correlation ($corr = 0.85, p = 0.0037$). The increasing impact of security categories reflects exponential economic loss caused by reported cyber crime to the IC3, from $485 million in 2011 to $3.5 billion in 2019.

We observe that there was a sharp jump in the absolute impact at the end of 2015 for most categories, followed by another steady growth in 2017. Interestingly, almost all the categories had decreasing trends in absolute impact in 2018 but climbed to the highest point in 2019. Different from other categories, *election security* had a significant increase in the absolute impact in 2016, while it was around zero before that. This increase coincides with the Russian interference in 2016 U.S. presidential election [46]. *national cybersecurity* became popular earlier, with the absolute impact gradually going up from 2009, before a sudden rise in 2013. National security, including national cyber attacks and cyber-espionage, was first considered to be more harmful than other threats (e.g., terrorism) by U.S. officials in 2013 [41]. It is worth noting that absolute impact and relative impact almost overlap for both *election security* and *national security*, as shown in Figure 6.

5.2.3 Category Relative Impact. We additionally computed category relative impacts (Equation (5)) for the sake of comparison. Relative impact reflects the average impact that each article has on the security categories during a month. The results of studying the articles from 2009 to now are depicted in Figure 6 and the lower subplot of Figure 7. Among the 16 categories, *mobile/application security* and *information privacy* have had the largest relative impacts over time as well as the largest absolute impacts. Meanwhile, *election security* shows the smallest relative impact and absolute impact.

We computed the Pearson correlation coefficient of all the pairs relative impacts from the security categories. We found the trends of the relative impacts for seven categories (i.e., *cybercriminal activity*, *cybersecurity program*, *information privacy*, *network attack*, *password/encryption*, *security*

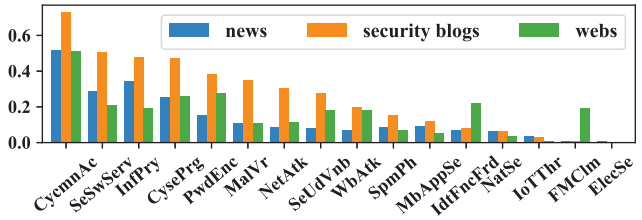

Fig. 8. The popularity of our security categories in different sources about cybersecurity.

software/service, and *spam/phishing*) are similar to each other ($corr > 0.7, p < 0.01$). The relative impacts of them have progressively risen from 2009 to 2015, and fluctuated around the peak afterwards. We further applied the Mann–Kendall trend test [36] to statistically measure the trends of relative impacts for the categories. The results suggest that 15 of our 16 categories have statistically increasing trends ($p < 0.05$). Only one category (*false/misleading claim*) experiences a downtrend.

- *cybercriminal activity* has been the most popular and was discussed in most security articles (65%), followed by *information privacy*, *security software/service*, and *cybersecurtiy program*, with similar popularities at 40%.
- Almost all the categories show upward trends in both absolute impact and relative impact over the last decade.
- Security issues in mobile/application and information privacy gained the largest absolute/relative impact over time.
- The absolute impacts from cybersecurity texts strongly correlate with the monetary loss caused by cybercrimes.

5.3 RQ3. How Have the Security Categories Varied and Evolved across Different Sources on Cybersecurity over the Last Decade?

We compared the security categories in terms of their popularity and impact across different sources of cybersecurity articles, i.e., news, security blogs and websites. This provides insights on how categories become popular on different platforms. In addition, the comparisons between different sources could assist users in source selection based on their preference such as interest in a specific topic or the latest cybersecurity technique.

5.3.1 Category Popularity. Figure 8 demonstrates the category popularity of security articles across the three sources. We find that almost all the categories are popularly present within all the sources except *election security* and *false/misleading claim*. *election security* only presents its prevalence in news at a significantly lower popularity compared to other categories. *false/misleading claim* refers to fake or deceptive online advertisements designed to mislead customers. This category is mainly active in web pages, but has shallow popularity in news. Among the 16 categories, *cybercriminal activity* has the highest popularity in all the sources. Among the three sources, security blogs stand popular in the majority of categories. This is because security blogs are more domain-specific and contain more detailed security knowledge in the content. Interestingly, only five categories (i.e., *identity theft/financial fraud*, *password/encryption*, *security update/vulnerability*, and *web attack*), and *false/misleading claim* show higher popularity in web sites than in news.

5.3.2 Category Absolute Impact. The absolute impact for the security categories at different sources are depicted in Figure 9(a) and the upper plot of Figure 10. Overall, there is an increasing trend in the absolute impacts for all the three sources. We further used the Mann–Kendall trend

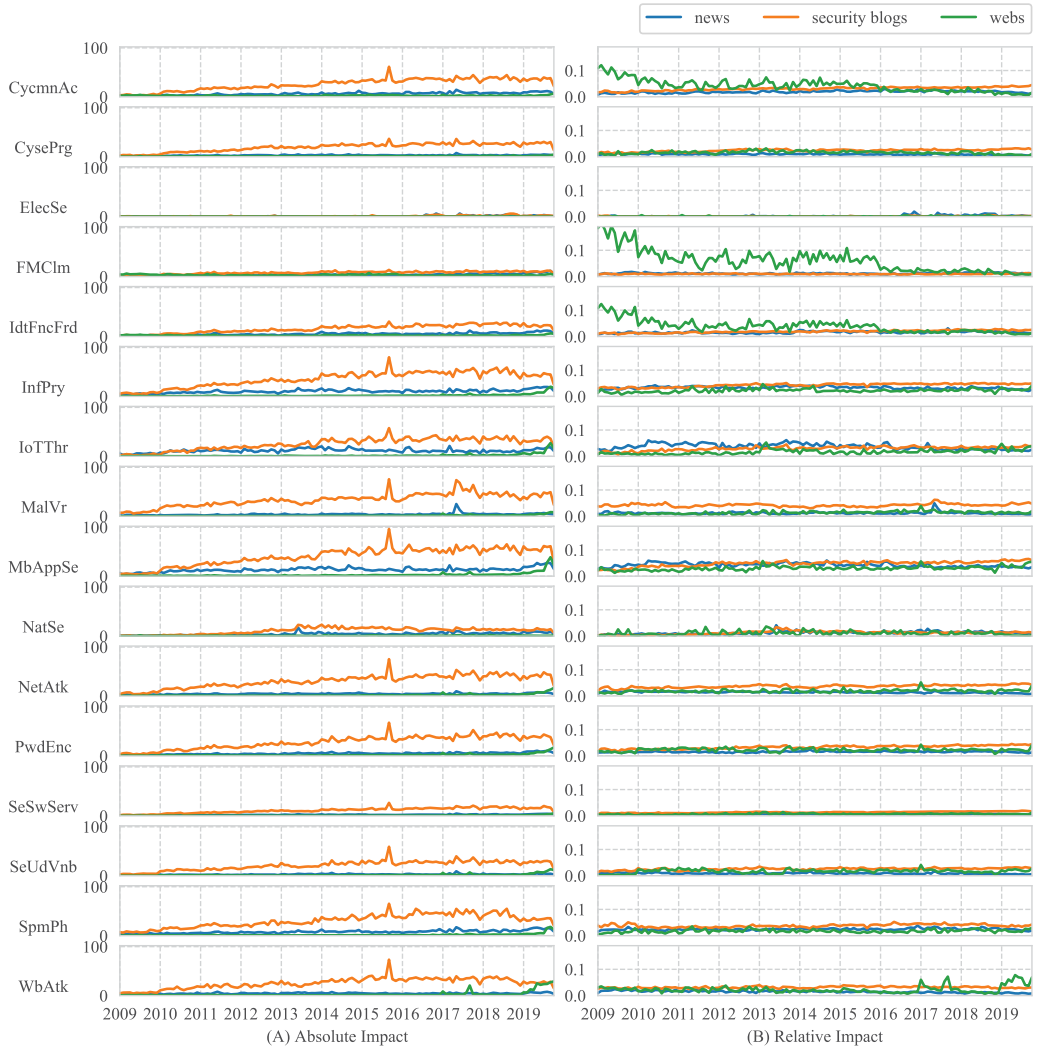

Fig. 9. The (a) absolute impacts and (b) relative impacts of our 16 security categories across various sources (news, security blogs, webs) over the last decade.

test [36] to check whether the trend is statistically significant or not. The results show that the increasing trends of absolute impacts for all the categories in news and security blogs are significant ($p < 0.05$). In webs, two categories (i.e., *cybercriminal activity* and *identity theft/financial fraud*) do not have any significant trends ($p = 0.06, 0.2$), whether increasing or decreasing. Only *false/misleading claim* experiences a significant downtrend ($p = 0.008$) in absolute impact during the 2010s.

From Figure 10, we observe that the distinction in the absolute impacts of different categories is less significant in webs than in the other two sources over time. Overall, security blogs have the largest absolute impact among the three sources. The high value of absolute impact indicates that security blogs have been the dominant source of delivering security knowledge in the last 10 years. However, security blogs generally contain technical jargon, which harms the readability. Our

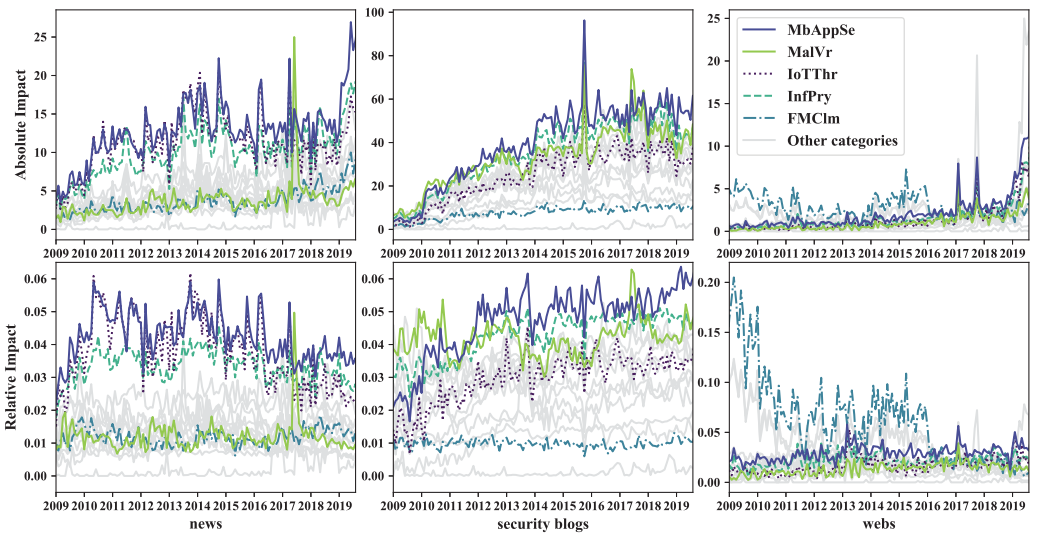

Fig. 10. Absolute and relative impacts of our security categories for different sources of cybersecurity from 2010 to date. Some categories are plotted in light grey to avoiding mixing the lines up. Figure 12 in Appendix A shows the version where the absolute and relative impacts are set to the same interval, separately.

previous work [80] studied users' understanding of security blogs and found a real-time glossary of tech terms could help. Home computer users are encouraged to leverage a blog reading assistant to learn more technical details in addition to reading on news and web pages. Meanwhile, security blogs can serve as a reliable source for more cybersecurity text-mining tasks. Compared to the other sources, web sources have had low absolute impacts ever since 2009; however, the trend has ended with a dramatic increase in 2019. Among the security categories, *mobile/application security* has gained the highest absolute impact, especially in news and security blogs. Besides, *information privacy* has achieved the second highest absolute impact across all the sources at almost all times. In news, its absolute impact exceeded *IoT threat* and moved up to the second after 2016. In webs, it surpassed *mobile/application security* and reached the highest in 2019.

Figure 9(a) plots the value of absolute impact for each category separately. In security blogs, we observe that most categories experience a rapid rise in the absolute impact in 2015. Except for *election security* and *national security*, the trends of the remaining categories are similar; with a steady increase at different paces. Compared to security blogs, news and web pages gain considerably lower absolute impacts, except for *election security*. Moreover, news absolute impact is slightly higher than that of web pages.

5.3.3 Category Relative Impact. Figure 9(b) and Figure 10 demonstrate the relative impacts of the security categories across the three sources. They show that *mobile/application security* has had the largest relative impact in news and security blogs during the study period. This is similar to its absolute impact. It is worth noting that *IoT threat* almost mirrored the absolute impact and the relative impact of *mobile/application security* in news. Interestingly, *malware/virus* rarely made a higher relative impact than *mobile/application security* in security blogs. One can also see that *false/misleading claim*, *identity theft/financial fraud*, and *cybercriminal activity* have had the highest relative impacts in webs before 2016.

From Figure 9(b), we find that the relative impact of web pages fluctuates dramatically. While security blogs still took the dominant place in relative impact for most the categories during the

Table 5. Significance Test on the Trends of Relative Impact, Including Increasing (↑), Decreasing (↓), and Stable(→) Trends ($p < 0.05$)

Category	News	Security blogs	Webs
CycmnAc	↑	↑	↓
CysePrg	↓	↑	↓
ElecSe	↑	↑	↑
FMClm	→	→	↓
IdtFncFrd	↑	↑	↓
InfPry	↓	↑	↑
IoTThr	↓	↑	↑
MalVr	→	↑	↑
MbAppSe	↓	↑	↑
NatSe	↑	↑	→
NetAtk	↓	↑	↑
PwdEnc	↑	↑	→
SeSwServ	→	↑	↑
SeUdVnb	↓	↑	→
SpmPh	↑	↑	↑
WbAtk	↓	↑	↓

study period, it was taken over by news and web pages sometimes. Webs maintained the highest relative impact in *cybercriminal activity*, *false/misleading claim*, and *identity theft/financial fraud* until 2016. Moreover, news showed to have a relative impact comparable to security blogs in some categories, namely *information privacy*, *national security*, and *spam/phishing*. This source took over security blogs in *election security*, *IoT threat*, and *mobile/application security* occasionally. In our collected data, the proportion of web articles with publication dates is significantly smaller than that of news or security blogs, with percentages around 66.6% compared to at least 98% for the latter two. This leads to low absolute impacts in webs, in contrast to noticeably higher relative impacts compared to the other two sources.

We further applied the Mann–Kendall trend test [36] to see whether the trend of relative impact has been statistically significant over time or not. The results are reported in Table 5. Only security blogs showed increasing trends ($p < 0.05$) in the relative impact for nearly all the categories, except *false/misleading claim*, which was the only category that remained stable. This category showed the same behaviour in the other two sources, too. News and webs had a few decreasing trends among their categories. The results also suggest that half of the security categories experienced statistically-significant increasing trends ($p < 0.01$) in relative impact across the three sources (i.e., *election security*, *information privacy*, *IoT threat*, *malware/virus*, *mobile/application security*, *network attack*, *security software/service*, and *spam/phishing*).

- For most categories, security blogs have been the most popular and impactful among the sources in the 2010s.
- Security issues in mobile/application have been the most impactful in news and security blogs over time.
- *IoT threat* almost mirrored the absolute impact value of *mobile/application security* in news over time.
- Only security blogs experienced statistically increasing trends in relative impact for nearly all the categories.

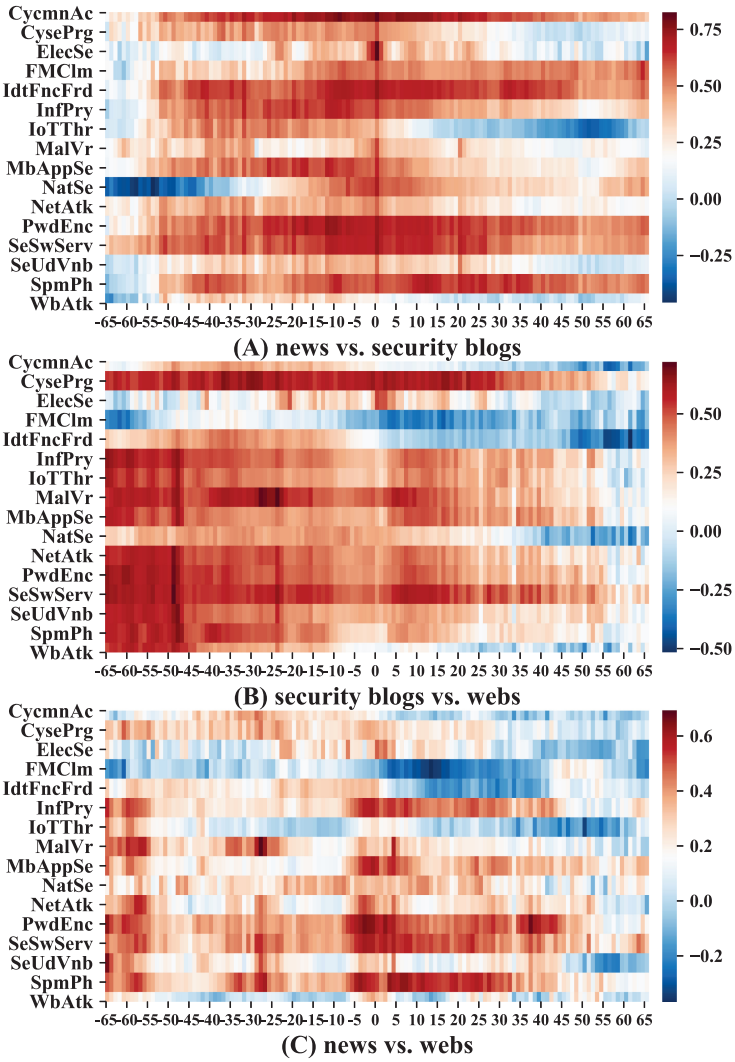

Fig. 11. The TLCC [11] in the absolute impact between (a) news vs. security blogs, (b) security blogs vs. webs and (c) news vs. webs. It plots the correlation between each pair of sources with shifting the second source (after "vs.") backwards (−) or forwards (+) in months. Darker red colour represents positively stronger correlation, and the peak correlation (dark red) indicates the two sources are most synchronised at that time.

Timeliness of different sources. We further measured the timeliness of different sources in reporting security incidents. We compared the sources temporally to measure the difference in information delivery delay. We applied **time-lagged cross correlations (TLCC)** [11] to calculate the correlation progressively with shifting one time series incrementally. TLCC identifies any temporal (leader-follower) relationship between two time series. The comparisons between each pair existed in the sources in the absolute impact per category are plotted in Figure 11. Each subplot depicts the dynamic correlation when we pull the second source backward (negative: −) or forward (positive: +). Darker colours indicate stronger correlation, with red representing positives and blue representing negatives in the spectrum. The peak correlation (dark red) shows where

the two sources are most synchronised in time, either when the first source leads (−) or when the second source leads (+).

From Figure 11(b), we clearly observe that 'security blogs' drove "webs" in the study period. A strong correlations is seen if one moves webs backwards for at least 10 months. Web pages show a few month delay compared to news in security information delivery, as shown in Figure 11(c). Overall, webs show very weak correlation with news, in contrast to the other two pairs. This indicates that websites do not focus on the timeliness when publishing cybersecurity texts, which is aligned with their low proportion of articles having publication dates. Figure 11(a) suggests that news and security blogs report security events firsthand and at almost similar speeds in most categories. News led in *information privacy*, *IoT threat*, *mobile/application security*, *password/encryption*. Only *spam/phishing* was an exception and became influential in security blogs earlier.

- News and security blogs report security events firsthand at similar speeds in most categories.
- Websites deliver security information without caring about timeliness much, where 30% articles do not specify the date and the rest have a time lag in posting emerging security issues.

6 DISCUSSION

6.1 Implications

Security education. Home computer users are struggling to resist the ever increasing cyber threats. While formal security education designed by certified experts is essential, it is still challenging to standardise the training as users might need security knowledge at different levels. Moreover, different users may have different backgrounds in dealing with cyber attacks. That is why the grey literature have become a major platform for users to learn security advice from. Understanding cybersecurity texts and the difference between sources can help users with the identification of useful information by themselves. Our analysis can additionally help to improve current security information sharing systems by capturing the trending topics with time. In this article, we only studied three sources (news, security blogs and websites). There are more online information sources in the real world, such as technical reports, which are not covered here. However, the three sources we picked are the best representatives in terms of prevalence, authority and users' click rate. They also broadly cover the security information reported by other sources.

Cyber attack prediction. Criminals are leveraging advanced technologies to perform sophisticated hackings such as cryptojacking (cryptomining attacks) based on rapidly grown cryptocurrencies (e.g., blockchain). By studying the topic patterns and following the tendencies in existing security incidents, future study can be conducted to predict the security categories that might be exploited by hackers and additionally, infer the potential technologies to be used. Since cybersecurity texts discuss security issues at different levels of technicality, it is unlikely that one can create a globally accepted standard set of security topics. Traditional classifications are either too abstract (architecture-based classification, e.g., application layer, endpoint layer) or too specific (common cyber attack types). In contrast to these, our manual classification that that uses card sorting [73], provides a comprehensive set of security categories that cover considerably different levels of security issues. Each article is associated with some categories, which is in line with the observation of real-world data and makes comparison of articles discussing even similar attacks, possible.

6.2 Threats to Validity

Internal Validity. News is one source where we collect security texts. Due to the fact that news especially breaking can be republished by multiple news agencies, there might be duplicated stories in our dataset. To alleviate the threat, we ran a script to remove duplicated content on the whole

corpus. Also, during the manual sanitisation process when checking the relevance to cybersecurity, we further checked the content of the articles that have the same title to remove duplicates. We observed that some news reports similar stories but in varying ways. For example, a financial review that reports a cyber attack might raise some security issues for financial institutions. Therefore, different reports on the same news might have different values on our security categories. Further study can be conducted to investigate how the news varies across different agencies. In addition, the 16 security categories are manually identified by the authors, and there might be some bias. To mitigate the threat, we have referred to other literature and checked with security experts to ensure our categories cover the major security and privacy issues.

External Validity. Apart from online sources, there exist other ways for users to get advice and make security decisions. IT workers, especially those with qualified internet skills who process sensitive business data are likely to learn from negative experiences, too [56]. Home computer users also gain security knowledge from social learning, such as their family, friends and acquainted experts [20, 21, 55]. Regardless of the diversity in security learning methods, it is hard to collect the real-world information received from communications and convert it into a standard text format. Thus, in our study, we only considered the online articles with text content that could be easily accessed by end users. Our collected articles from security blogs and webs somehow contained social communications, too. Note that security experts are likely to share security stories and tips to safeguard both home computer users and businesses. People with negative experiences might also post their personal stories in discussion web forums to seek help from authorities as well as other online users.

7 CONCLUSION

In conclusion, we discovered the emerging topics in cybersecurity and preformed an empirical analysis based on our collected security texts from the sources of news, security blogs and websites. Since LDA cannot generate specific and distinguished topics for cybersecurity texts, we proposed a novel semi-automated classification method for this purpose. We applied the term extraction method based on the results generated by LDA. We then identified 16 security categories from the terms that could represent the articles statistically as a probabilistic distribution. We further analysed the evolution and variation of the collected articles across the security categories as well as sources over the last decade. We revealed several interesting findings, like the absolute impact of cybersecurity texts shows a strong correlation with the financial loss caused by cybercrimes, or websites (of authorised organisations), in contrast to news and security blogs, tend to publish general security articles without caring about their timeliness. Further research can be conducted to improve users' understanding of different sources of cybersecurity texts or predict cyber attacks based on our analyses. Further study can explore more resources such as discussion web forums and develop automated analysis tools.

APPENDIX

A SUPPLEMENTARY FIGURE

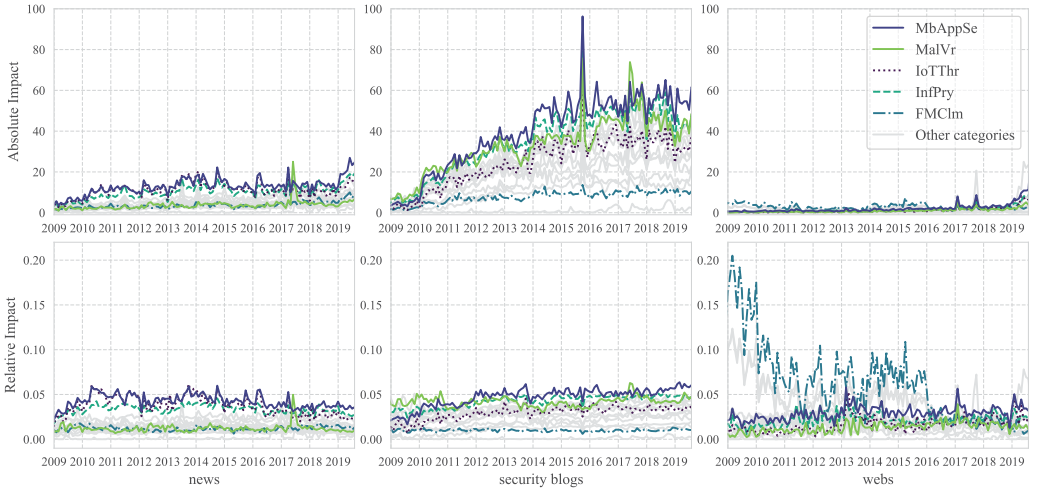

Fig. 12. Another version of Figure 10 where the absolute and relative impacts are set to the same interval, separately. Absolute and relative impacts of our security categories for different sources of cybersecurity from 2010 to date.

REFERENCES

[1] Md Sahrom Abu, Siti Rahayu Selamat, Aswami Ariffin, and Robiah Yusof. 2018. Cyber threat intelligence–issue and challenges. *Indones. J. Electr. Eng. Comput. Sci.* 10, 1 (2018), 371–379.

[2] Ruba Abu-Salma, Elissa M. Redmiles, Blase Ur, and Miranda Wei. 2018. Exploring user mental models of end-to-end encrypted communication tools. In *Proceedings of the 8th USENIX Workshop on Free and Open Communications on the Internet (FOCI'18)*.

[3] Alessandro Acquisti and Jens Grossklags. 2005. Privacy and rationality in individual decision making. *IEEE Secur. Priv.* 3, 1 (2005), 26–33.

[4] Devdatta Akhawe and Adrienne Porter Felt. 2013. Alice in warningland: A large-scale field study of browser security warning effectiveness. In *Proceedings of the 22nd USENIX Security Symposium (USENIX Security'13)*. 257–272.

[5] Henry Belot and ABC News Stephanie Borys. cited May 2020. Ransomware Attack Still Looms in Australia as Government Warns WannaCry Threat Not Over. Retrieved from https://www.abc.net.au/news/2017-05-15/ransomware-attack-to-hit-victims-in-australia-government-says/8526346.

[6] Noam Ben-Asher and Cleotilde Gonzalez. 2015. Effects of cyber security knowledge on attack detection. *Comput. Hum. Behav.* 48 (2015), 51–61.

[7] Ralf Bender and Stefan Lange. 2001. Adjusting for multiple testing—when and how? *J. Clin. Epidemiol.* 54, 4 (2001), 343–349.

[8] David M. Blei and John D. Lafferty. 2009. Topic models. In *Text Mining*. Chapman and Hall/CRC, 101–124.

[9] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, (Jan. 2003), 993–1022.

[10] John M. Blythe, Nissy Sombatruang, and Shane D. Johnson. 2019. What security features and crime prevention advice is communicated in consumer IoT device manuals and support pages? *J. Cybersecur.* 5, 1 (2019), tyz005.

[11] Steven M. Boker, Jennifer L. Rotondo, Minquan Xu, and Kadijah King. 2002. Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychol. Methods* 7, 3 (2002), 338.

[12] Eric W. Burger, Michael D. Goodman, Panos Kampanakis, and Kevin A. Zhu. 2014. Taxonomy model for cyber threat intelligence information exchange technologies. In *Proceedings of the 2014 ACM Workshop on Information Sharing & Collaborative Security*. 51–60.

[13] Nathanael Chambers, Ben Fry, and James McMasters. 2018. Detecting denial-of-service attacks from social media text: Applying nlp to computer security. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 1626–1635.

[14] Tianying Chen, Jessica Hammer, and Laura Dabbish. 2019. Self-efficacy-based game design to encourage security behavior online. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–6.
[15] Hyo Shin Choi, Won Sang Lee, and So Young Sohn. 2017. Analyzing research trends in personal information privacy using topic modeling. *Comput. Secur.* 67 (2017), 244–253.
[16] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 1 (1960), 37–46.
[17] Damien Cram and Béatrice Daille. 2016. Terminology extraction with term variant detection. In *Proceedings of ACL'16 System Demonstrations*. 13–18.
[18] Robert E. Crossler and France Bélanger. 2017. The mobile privacy-security knowledge gap model: Understanding behaviors. In *Proceedings of the Hawaii International Conference on System Sciences*.
[19] CybSafe. Human error to blame for 9 in 10 UK cyber data breaches in 2019. Retrieved February 2020 from https://www.cybsafe.com/press-releases/human-error-to-blame-for-9-in-10-uk-cyber-data-breaches-in-2019/.
[20] Sauvik Das, Tiffany Hyun-Jin Kim, Laura A Dabbish, and Jason I. Hong. 2014. The effect of social influence on security sensitivity. In *Proceedings of the 10th Symposium On Usable Privacy and Security (SOUPS'14)*. 143–157.
[21] Sauvik Das, Adam D. I. Kramer, Laura A. Dabbish, and Jason I. Hong. 2014. Increasing security sensitivity with social proof: A large-scale experimental confirmation. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. 739–749.
[22] 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 CHI Conference on Human Factors in Computing Systems*. 1–12.
[23] Mark Evans, Leandros A. Maglaras, Ying He, and Helge Janicke. 2016. Human behaviour as an aspect of cybersecurity assurance. *Secur. Commun. Netw.* 9, 17 (2016), 4667–4679.
[24] Michael Fagan and Mohammad Maifi Hasan Khan. 2016. Why do they do what they do?: A study of what motivates users to (not) follow computer security advice. In *Proceedings of the 12th Symposium on Usable Privacy and Security (SOUPS'16)*. 59–75.
[25] Michael Fagan and Maifi Mohammad Hasan Khan. 2018. To follow or not to follow: A study of user motivations around cybersecurity advice. *IEEE Internet Comput.* 22, 5 (2018), 25–34.
[26] Chris Fennell and Rick Wash. 2019. Do stories help people adopt two-factor authentication? *Studies* 1, 2 (2019), 3.
[27] Alain Forget, Sarah Pearman, Jeremy Thomas, Alessandro Acquisti, Nicolas Christin, Lorrie Faith Cranor, Serge Egelman, Marian Harbach, and Rahul Telang. 2016. Do or do not, there is no try: User engagement may not improve security outcomes. In *Proceedings of the 12th Symposium on Usable Privacy and Security (SOUPS'16)*. 97–111.
[28] Cyber Security Research Center from Romania CCSIR. cited Feb 2020. The Impact of cybersecurity over the Last 5 Years. Retrieved from https://def.camp/impact-cybersecurity-five-years/.
[29] Masahiro Fujita, Mako Yamada, Shiori Arimura, Yuki Ikeya, and Masakatsu Nishigaki. 2015. An attempt to memorize strong passwords while playing games. In *Proceedings of the 18th International Conference on Network-Based Information Systems*. IEEE, 264–268.
[30] Maximilian Golla, Miranda Wei, Juliette Hainline, Lydia Filipe, Markus Dürmuth, Elissa Redmiles, and Blase Ur. 2018. "What was that site doing with my facebook password?" Designing password-reuse notifications. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. 1549–1566.
[31] Hui Guo, Özgür Kafalı, Anne-Liz Jeukeng, Laurie Williams, and Munindar P Singh. 2020. Çorba: Crowdsourcing to obtain requirements from regulations and breaches. *Empir. Softw. Eng.* 25, 1 (2020), 532–561.
[32] Lee Hadlington. 2017. Human factors in cybersecurity; examining the link between Internet addiction, impulsivity, attitudes towards cybersecurity, and risky cybersecurity behaviours. *Heliyon* 3, 7 (2017), e00346.
[33] Marlo Haering, Christoph Stanik, and Walid Maalej. 2021. Automatically matching bug reports with related app reviews. In *Proceedings of the IEEE/ACM 43rd International Conference on Software Engineering (ICSE'21)*. IEEE, 970–981.
[34] Hang Hu and Gang Wang. 2018. End-to-end measurements of email spoofing attacks. In *Proceedings of the 27th USENIX Security Symposium (USENIX Security'18)*. 1095–1112.
[35] Jiajun Hu, Xiaobing Sun, David Lo, and Bin Li. 2015. Modeling the evolution of development topics using dynamic topic models. In *Proceedings of the IEEE 22nd International Conference on Software Analysis, Evolution, and Reengineering (SANER'15)*. IEEE, 3–12.
[36] Md. Hussain and Ishtiak Mahmud. 2019. pyMannKendall: A python package for non parametric mann kendall family of trend tests. *J. Open Source Softw.* 4, 39 (25 7 2019), 1556. https://doi.org/10.21105/joss.01556
[37] Iulia Ion, Rob Reeder, and Sunny Consolvo. 2015. "... no one can hack my mind": Comparing expert and non-expert security practices. In *Proceedings of the 11th Symposium On Usable Privacy and Security (SOUPS'15)*. 327–346.
[38] Julian Jang-Jaccard and Surya Nepal. 2014. A survey of emerging threats in cybersecurity. *J. Comput. Syst. Sci.* 80, 5 (2014), 973–993.
[39] Joseph Johnson. IC3: Total Damage Caused by Reported Cyber Crime 2001–2020. Retrieved April 2020 from https://www.statista.com/statistics/267132/total-damage-caused-by-by-cyber-crime-in-the-us/.

[40] Özgür Kafalı, Jasmine Jones, Megan Petruso, Laurie Williams, and Munindar P Singh. 2017. How good is a security policy against real breaches? A HIPAA case study. In *Proceedings of the IEEE/ACM 39th International Conference on Software Engineering (ICSE'17)*. IEEE, 530–540.

[41] Ken Dilanian. *Los Angeles Times*. Cyber-attacks a Bigger Threat Than Al Qaeda, Officials Say. Retrieved April 2020 from https://www.latimes.com/world/la-xpm-2013-mar-12-la-fg-worldwide-threats-20130313-story.html.

[42] Katharina Krombholz, Wilfried Mayer, Martin Schmiedecker, and Edgar Weippl. 2017. "I have no idea what i'm doing"—On the usability of deploying HTTPS. In *Proceedings of the USENIX Security Symposium (USENIX Security'17)*. 1339–1356.

[43] Xiaozhou Li, Boyang Zhang, Zheying Zhang, and Kostas Stefanidis. 2020. A sentiment-statistical approach for identifying problematic mobile app updates based on user reviews. *Information* 11, 3 (2020), 152.

[44] Xiaojing Liao, Kan Yuan, XiaoFeng Wang, Zhou Li, Luyi Xing, and Raheem Beyah. 2016. Acing the IOC game: Toward automatic discovery and analysis of open-source cyber threat intelligence. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS'16)*. ACM, 755–766.

[45] Shike Mei and Xiaojin Zhu. 2015. The security of latent dirichlet allocation. In *Artificial Intelligence and Statistics*. 681–689.

[46] Ellen Nakashima, Missy Ryan, and Karen DeYoung. *The Washington Post*. Obama Administration Announces Measures to Punish Russia for 2016 Election Interference. Retrieved May 2020 from https://www.washingtonpost.com/world/national-security/obama-administration-announces-measures-to-punish-russia-for-2016-election-interference/2016/12/29/311db9d6-cdde-11e6-a87f-b917067331bb_story.html.

[47] Pradeep K. Murukannaiah, Chinmaya Dabral, Karthik Sheshadri, Esha Sharma, and Jessica Staddon. 2017. Learning a privacy incidents database. In *Proceedings of the Hot Topics in Science of Security: Symposium and Bootcamp*. 35–44.

[48] Anand Narasimhamurthy. 2005. Theoretical bounds of majority voting performance for a binary classification problem. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 12 (2005), 1988–1995.

[49] James Nicholson, Lynne Coventry, and Pamela Briggs. 2019. "If it's important it will be a headline": Cybersecurity information seeking in older adults. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–11.

[50] Norbert Nthala and Ivan Flechais. 2018. Informal support networks: An investigation into home data security practices. In *Proceedings of the 14th Symposium on Usable Privacy and Security (SOUPS'18)*. 63–82.

[51] Sungrae Park, Wonsung Lee, and Il-Chul Moon. 2015. Efficient extraction of domain specific sentiment lexicon with active learning. *Pattern Recogn. Lett.* 56 (2015), 38–44.

[52] Maria Teresa Pazienza, Marco Pennacchiotti, and Fabio Massimo Zanzotto. 2005. Terminology extraction: An analysis of linguistic and statistical approaches. In *Knowledge Mining*. Springer, 255–279.

[53] Justin Petelka, Yixin Zou, and Florian Schaub. 2019. Put your warning where your link is: Improving and evaluating email phishing warnings. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–15.

[54] Emilee Rader and Rick Wash. 2015. Identifying patterns in informal sources of security information. *J. Cybersecur.* 1, 1 (2015), 121–144.

[55] Emilee Rader, Rick Wash, and Brandon Brooks. 2012. Stories as informal lessons about security. In *Proceedings of the 8th Symposium on Usable Privacy and Security*. 1–17.

[56] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. 2016. How i learned to be secure: A census-representative survey of security advice sources and behavior. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. 666–677.

[57] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. 2017. Where is the digital divide? A survey of security, privacy, and socioeconomics. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 931–936.

[58] Elissa M. Redmiles, Amelia R. Malone, and Michelle L. Mazurek. 2016. I think they're trying to tell me something: Advice sources and selection for digital security. In *Proceedings of the IEEE Symposium on Security and Privacy (SP)*. IEEE, 272–288.

[59] Elissa M. Redmiles, Michelle L. Mazurek, and John P. Dickerson. 2018. Dancing pigs or externalities? Measuring the rationality of security decisions. In *Proceedings of the ACM Conference on Economics and Computation*. 215–232.

[60] Elissa M. Redmiles, Miraida Morales, Lisa Maszkiewicz, Rock Stevens, Everest Liu, Dhruv Kuchhal, and Michelle L. Mazurek. 2018. First steps toward measuring the readability of security advice. In *Proceedings of the IEEE Security & Privacy Workshop on Technology and Consumer Protection (ConPro'18)*.

[61] Robert W. Reeder, Iulia Ion, and Sunny Consolvo. 2017. 152 simple steps to stay safe online: Security advice for non-tech-savvy users. *IEEE Secur. Priv.* 15, 5 (2017), 55–64.

[62] Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC Workshop on New Challenges for NLP Frameworks*. ELRA, 45–50. http://is.muni.cz/publication/884893/en.

[63] Leonard Richardson. 2007. Beautiful soup documentation. Dosegljivo: https://www.crummy.com/software/BeautifulSoup/bs4/doc/. [Dostopano: 7. 7. 2018] (2007).

[64] Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the WSDM'15*. ACM, 399–408.
[65] Hannah R. Rothstein and Sally Hopewell. 2009. Grey literature. *Handb. Res. Synth. Meta-anal.* 2 (2009), 103–125.
[66] Nader Sohrabi Safa and Rossouw Von Solms. 2016. An information security knowledge sharing model in organizations. *Comput. Hum. Behav.* 57 (2016), 442–451.
[67] Clemens Sauerwein, Irdin Pekaric, Michael Felderer, and Ruth Breu. 2019. An analysis and classification of public information security data sources used in research and practice. *Comput. Secur.* 82 (2019), 140–155.
[68] Stuart Schechter and Joseph Bonneau. 2015. Learning assigned secrets for unlocking mobile devices. In *Proceedings of the 11th Symposium On Usable Privacy and Security (SOUPS'15)*. 277–295.
[69] Tyler Schultz. If Your Employees Aren't Learning from Your Security Training, Are You Really Teaching? Retrieved February 2020 from https://www.infosecinstitute.com/blog/if-your-employees-arent-learning-from-your-security-training-are-you-really-teaching/.
[70] Karthik Sheshadri, Nirav Ajmeri, and Jessica Staddon. 2017. No (privacy) news is good news: An analysis of New York times and guardian privacy news from 2010–2016. In *Proceedings of the 15th Annual Conference on Privacy, Security and Trust (PST'17)*. IEEE, 159–15909.
[71] Ruth Shillair and Jingbo Meng. 2017. Multiple sources for security: Seeking online safety information and their influence on coping self-efficacy and protection behavior habits. In *Proceedings of the 50th Hawaii International Conference on System Sciences*.
[72] Florian Skopik, Giuseppe Settanni, and Roman Fiedler. 2016. A problem shared is a problem halved: A survey on the dimensions of collective cyber defense through security information sharing. *Comput. Secur.* 60 (2016), 154–176.
[73] Donna Spencer. 2009. *Card Sorting: Designing Usable Categories*. Rosenfeld Media.
[74] Rock Stevens, Daniel Votipka, Elissa M. Redmiles, Colin Ahern, Patrick Sweeney, and Michelle L. Mazurek. 2018. The battle for new york: A case study of applied digital threat modeling at the enterprise level. In *Proceedings of the 27th USENIX Security Symposium (USENIX Security'18)*. 621–637.
[75] Joshua Sunshine, Serge Egelman, Hazim Almuhimedi, Neha Atri, and Lorrie Faith Cranor. 2009. Crying wolf: An empirical study of ssl warning effectiveness. In *Proceedings of the USENIX Security Symposium*. 399–416.
[76] Wiem Tounsi and Helmi Rais. 2018. A survey on technical threat intelligence in the age of sophisticated cyber attacks. *Comput. Secur.* 72 (2018), 212–233.
[77] Zhiyuan Wan, Xin Xia, and Ahmed E Hassan. 2019. What is discussed about blockchain? A case study on the use of balanced LDA and the reference architecture of a domain to capture online discussions about blockchain platforms across the stack exchange communities. *IEEE Trans. Softw. Eng.* (2019).
[78] Rick Wash. 2010. Folk models of home computer security. In *Proceedings of the 6th Symposium on Usable Privacy and Security*. 1–16.
[79] Zikai Alex Wen, Zhiqiu Lin, Rowena Chen, and Erik Andersen. 2019. What. hack: engaging anti-phishing training through a role-playing phishing simulation game. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.
[80] Tingmin Wu, Rongjunchen Zhang, Wanlun Ma, Sheng Wen, Xin Xia, Cecile Paris, Surya Nepal, and Yang Xiang. 2020. What risk? I don't understand: An empirical study on users' understanding of the terms used in security texts. In *Proceedings of the 15th ACM Asia Conference on Computer and Communications Security (ASIA CCS'20)*. Association for Computing Machinery, New York, NY, 248–262.
[81] Xin Xia, Emad Shihab, Yasutaka Kamei, David Lo, and Xinyu Wang. 2016. Predicting crashing releases of mobile applications. In *Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*. 1–10.
[82] Maochao Xu, Kristin M. Schweitzer, Raymond M. Bateman, and Shouhuai Xu. 2018. Modeling and predicting cyber hacking breaches. *IEEE Trans. Inf. Forens. Secur.* 13, 11 (2018), 2856–2871.
[83] Xin-Li Yang, David Lo, Xin Xia, Zhi-Yuan Wan, and Jian-Ling Sun. 2016. What security questions do developers ask? A large-scale study of Stack Overflow posts. *J. Comput. Sci. Technol.* 31, 5 (2016), 910–924.
[84] Yixin Zou, Shawn Danino, Kaiwen Sun, and Florian Schaub. 2019. YouMight'be affected: An empirical analysis of readability and usability issues in data breach notifications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–14.
[85] Yixin Zou, Abraham H. Mhaidli, Austin McCall, and Florian Schaub. 2018. "I've got nothing to lose": Consumers' risk perceptions and protective actions after the equifax data breach. In *Proceedings of the 14th Symposium on Usable Privacy and Security (SOUPS'18)*. 197–216.
[86] Yixin Zou and Florian Schaub. 2019. Beyond mandatory: Making data breach notifications useful for consumers. *IEEE Secur. Priv.* 17, 2 (2019), 67–72.

Received February 2021; revised June 2021; accepted August 2021