Aug 10th, 12:00 AM

A Review of Cyberattack Research using Latent Dirichlet Allocation

Ming Xiao
UNC Greensboro, m_xiao@uncg.edu

Gurpreet Dhillon
IT and Decision Sciences, gurpreet.dhillon@unt.edu

Kane J. Smith
University of North Carolina at Greensboro, kjsmith9@uncg.edu

A Review of Cyberattack Research using Latent Dirichlet Allocation

Emergent Research Forum (ERF)

Ming Xiao
UNC Greensboro
m_xiao@uncg.edu

Gurpreet Dhillon
University of North Texas
gurpreet.dhillon@unt.edu

Kane J Smith
UNC Greensboro
Kjsmith9@uncg.edu

Abstract

Cyber Attacks have been on an increase in the past decade. Dominant academic research has considered a range of theoretical approaches to the study of cyberattacks on organizations and society at large. However, there is a paucity of research that systematically reviews the current literature to define current emphasis and propose future research directions. In this paper we use the Latent Dirichlet Allocation approach to review cyberattack research in the Chartered Association of Business Schools level 3 and above journals over the last two decades. In a final synthesis we propose a future research agenda.

Keywords

Information systems security, cyberattacks, topic modeling, Latent Dirichlet Allocation

Introduction

Information security incidents create negative impacts on both enterprises and individuals. Therefore, issues related to information systems security research are of great interest to IS researchers, practitioners, as well as policy makers. The IS discipline is uniquely positioned to advance our understanding of information systems security and thus, has been contributing to this field of research for more than thirty years (Totty et al. 2020). Since the early 1990s, IS security research has continued to evolve from practical research focused on the individual level to a more theory-driven field that also addresses organizational-level phenomena (Totty et al. 2020). The works of Baskerville (1993), Dhillon and Backhouse (2001), and Siponen (2005) have provided comprehensive and cumulative reviews and assessments for the literature in each stage, as well as prophesy directions of research in the future (Dhillon et al. 2021). The most recent holistic review by Dhillon et al. (2021) indicates that IS Security Behaviors, Privacy Concerns, and Security Policy Compliance are dominant topics in IS literature, yet their Delphi study shows that practitioners maintain a dominant focus on IS security attack issues. The panelists of Chief Information Security Officers (CISO) consider the management of IS security attacks to be the most important issue, as opposed to the current topics that are dominant in IS academic research.

There is a common debate amongst academics and practitioners that Information System research can, at times, have limited relevance between the two groups. The academic community is frequently discussing a perceived lack of relevance of information systems (IS) research to practitioners (Baskerville and Wood-Harper 1996; Rosemann and Vessey 2008; von Bary et al. 2018). Several scholars argue that IS research is falling behind in the dynamic, fast developing world with which IS practitioners are confronted (Dhillon et al., 2021). This could lead to a disconnect between the needs of practitioners and the research focus of academic publications (Ali et al. 2020). The so-called disconnect is potentially driven by the missing interaction between academics and practitioners, perhaps limiting the academic exploration of topics relevant to practitioners. (Lee et al. 1995; Rosemann and Vessey 2008; von Bary et al. 2018).

The purpose of this study is to explore this potential concern and examine exactly how IS security researchers have addressed the issue of cyberattacks in the field. To achieve this goal, we collected articles

from academia and analyzed them by using topic modeling to provide concrete evidence of where IS security researchers have already contributed to understanding and solving this problem, providing a comprehensive analysis of literature. Within our review of the academic literature, we find the topic related to cyberattacks has also been well developed by Computer Science, where the focal point has been centered on algorithmic aspects. Meanwhile, different specific terminologies have been used to describe the various types of cyberattacks, such as penetration, intrusion, hack, phishing, etc., instead of using the word cyberattack as their topic. In this case, we refer to the proposition of Dhillon et al. (2021), taking all cyberattacks within IS security related works into consideration and filtering out irrelevant articles. By applying data science methods and techniques on the results from the academic literature related to cyberattack cases, more detailed analyses are illustrated by the visualized topic modeling output as well as a comprehensive analysis in the following sections. To begin, we present the methodology used, and then discuss the preliminary results of emergent themes and the current content of academic cyberattack research in the field of IS.

Methodology

Literature Search

To explore the thematic structure of IS security articles related to cyberattacks, we applied a two-phase process. First, a keyword search was performed by using "security attack" and "cyberattack" and related terminologies on the article's title, abstract, and keywords from the Scopus database. This keyword search was limited to articles published between January 2003 to December 2020, in journals of Information Systems discipline with 3 and higher rankings in the academic journal guide 2021 published by the Chartered Association of Business Schools (CABS). Moreover, we also took Information and Computer Security journal and Computer & Security journal into consideration, which are CABS 1 in the latest version, but both are of few information security specific journals. Journal samples are restricted to CABS ranked and IS security-specific since our objective is to analyze how the IS scholar community has contributed to IS security attack research with the consideration of "major contributions are likely to be in the leading journals" (Webster and Watson 2002, p. xvi). The selected journals are internationally recognized as top-tier IS publication outlets so we believe our approach can reveal the mainstream research on cyberattack-related studies in the journals searched, as well as within the IS discipline as a whole. Then, we conducted an intensive reading and screening of the collected literature. References without abstracts or keywords, such as editorials, introductions to special issues, or review of papers, were excluded from these observations. Further, the authors of this study individually browsed all collected abstracts and screened out articles deemed irrelevant to the topic at hand. After discussion and consensus, our samples resulted in 162 references.

Thematic Analysis Approach Using Topic Modeling

To carry out an automatic classification of literature, we proceeded to apply topic modeling on the titles, abstracts, and keywords. We made this decision because article titles or author-provided keywords are often limited to a low, specified number of words/characters and full papers would bring a tremendous amount of noise to data. The topic modeling approach uses the distances among terms on the collection of objects as criteria to cluster them into smaller numbers of groups, which could be a useful source for categorization. Latent Dirichlet Allocation (LDA) is one of the common topic modeling means in analogous research. LDA was proposed by Blei et al. (2003), which uses a statistical generative model that assumes a document is composed of a set of words, and there is no sequential relationship between those words. In this logic, a document is supposed to be generated in two steps. First, each topic is randomly selected based on the topic distribution in the document. Second, each word is randomly drawn based on the word distribution in the topic. This process is applied iteratively until the observed words in the document find the best set describing the topic and word distribution (Blei et al. 2003; Huang et al. 2018). Additionally, LDA is a typical bag-of-words model, by which a document could contain multiple topics at the same time, as well as each word in the document being generated by one of the topics. Thus, by using LDA, the topic of each document in the sample set can be given in the form of a probability distribution; the words of each topic in the sample set will be clustered as a form of distribution as well. After extracting the word distribution

and topic distribution by analyzing the sample set, topic clustering and document classification can be performed.

In line with the literature we collected, we applied text parsing to remove irrelevant data noise. Preprocessing steps include converting text to lower case, removing punctuation, prepositions, numbers, and common stop words (e.g., "the", "a", "are"), and lemmatization to reduce the inflectional form of a word to a base dictionary form. We then conducted the topic modeling with the LDA algorithm to identify latent themes in those articles' title, keywords, and abstracts (Blei et al. 2003). Although topic modeling can automatically classify articles based on text information, the number of identified topics often requires researchers to determine. The LDA algorithm also requires the number of topics to be prespecified as an input parameter before execution. This is not easy to make an effortless decision since a high value could lead to meaningless topics, while a low value could eliminate important information. Since there is no straightforward way to use mathematical analysis to determine the "optimal" number of topics (Mortenson and Vidgen 2016), we derived the optimum number of topics by using four algorithms: two minimization algorithms proposed by Cao et al. (2009) and Arun et al. (2010); as well as two maximization algorithms proposed by Griffiths and Steyvers (2004) and Deveaud et al. (2014). To overcome the bias of individuals, we ran analyses specifying the program to create fifty topics in the first round and observed output figures. The polyline of the topic numbers almost completely converges at eight in the minimization algorithm, while these eight topics cover more than ninety-five percent of possible topics in the maximization algorithm (see Figure 1). Therefore, we extracted eight topics for preliminary result analysis using LDA implementation through R 'topicmodels' package and 'ldatuning' package. The authors then independently read the output of topic clusters and word distributions, then discussed whether the result was reliable based on the criteria of how the terms distributed in each topic are distinguished from those in other clusters.

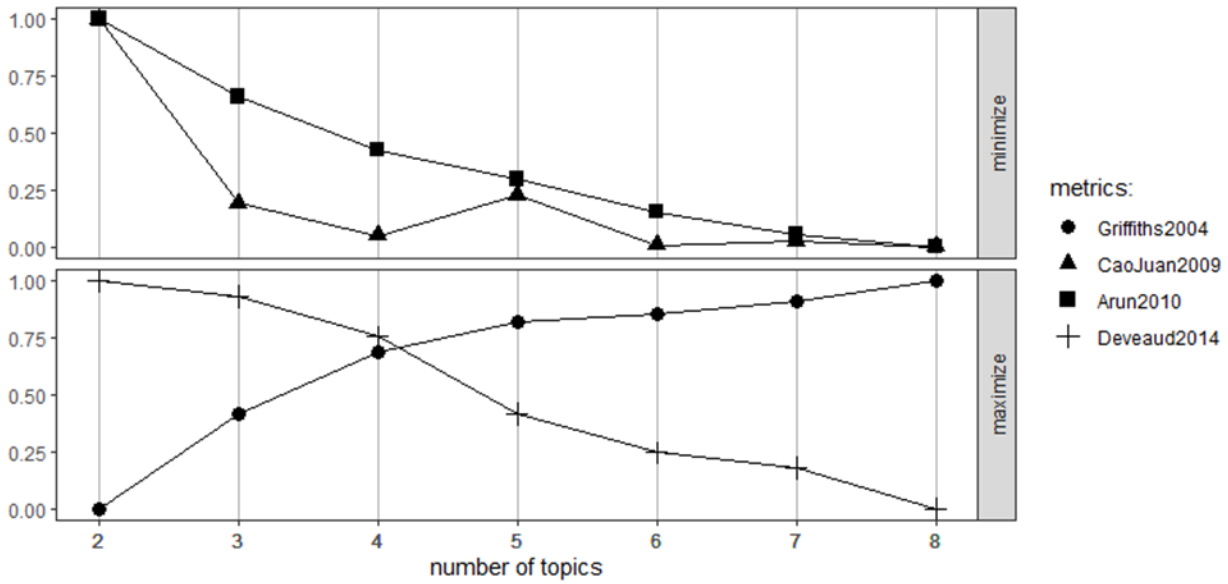

Figure 1. The Maximum-Minimum Topic Analysis

Results and Discussion

Preliminary Results

As discussed in the previous section, our topic modeling resulted in eight topic areas acknowledged in mainstream cyberattack research in the IS security domain. The authors agreed that eight topics was a reasonable result for our next phase of the analysis. By mapping the collected articles with each Topic ID, we get groupings of literature. Then, we compared the abstracts of the articles in each group, summarized the commonalities among them, and discussed the naming of each topic by the authors collaboratively.

We manually evaluated these eight topics and classified them into corresponding topic features for cyberattacks as presented in Table 1, namely: Cybercrime and Cyberwarfare, AI-enabled Cyberattacks, Security Forensics and Detection, Vulnerability Analysis, Critical Infrastructure Security, Network Security, Ubiquitous Security, and Deception Attacks. The table also presents the most frequent words for each topic (the order of topics is listed by quadrant in R output). Taking the topic of Network Security for instance, we identified twenty-two articles in this group and found that these articles are highly relevant to different types of cyber threat and countermeasures in different IS domains, whereas the commonality of these studies is identifying risks of information systems and work to enhance security. Hence, we discussed and named this topic as Network Security. We repeated the same process on the other seven topic areas and got the corresponding grouping themes. In the next section, we will detail the findings and discuss our future research agenda.

Topic ID	Most Frequent Keywords	Topic Theme
T1	Security, cyber, risk, management, threat, attacks, incident, cybersecurity, crime, forensics, computer, system	Security Forensics & Detection
T2	Detection, intrusion, system, security, network, systems, control, anomaly, industrial, attack, computing, cyber physical,	Network Security
T3	Data, learning, security, intelligence, internet, machine, theory, threat, information, cyber, networks	AI-enabled cyberattacks
T4	Security, information, attacks, software, cyber, vulnerability, analysis, game, web	Vulnerability Analysis
T5	Security, cyber, SCADA, analysis, risk, social, cybersecurity, detection, cybercrime, engineering, classification, intrusion	Critical Infrastructure Security
T6	Attack, security, things, privacy, control, IoT, device	Ubiquitous Security
T7	Learning, security, machine, data, clustering, log, unsupervised, Markov, phishing, deceptions, attack, hidden, sanitization	Deception Attacks
T8	Cyber, online, theory, warfare, cps, critical, malicious, fraud, models, malware, systems	Cybercrime & Cyberwarfare

Table 1. Keywords and Emergent Topics

Discussion & Conclusion

In this section, we analyze the research findings, with our review of literature providing a basis for the development of eight preliminary research questions. These questions are based on what was found in our literature review using content analysis, 8 distinct topics. A summary of the topics is found in table 1. The following eight preliminary research questions were developed based on the topic analysis of the academic information system security literature on cyberattacks and a detailed review of the papers that compose those eight topics. The goal is to provide potential future research directions that are both rooted in prior work, and extend research into new domains where meaningful academic and practical contributions can occur.

RQ1: Investigation of emergent patterns of cyber-criminal activity for predictive analysis

RQ2: Analysis of patterns of network activity to decipher criminal behavior

RQ3: The Role of Bots and AI-based Agents in perpetuating cyberattacks

RQ4: Investigation of Game Theoretic attacks and vulnerability assessment

RQ5: The role of nation states in disrupting critical IT-enabled infrastructure

RQ6: Investigating pervasive computing and cyber security challenges affecting people

RQ7: Investigating cyber-based deception and masquerading techniques with criminal intent

RQ8: How Denial of Service Attacks facilitate cyber-criminal activity

In this study, we conducted a rigorous and comprehensive review of the academic information systems security literature on cyberattacks, thus identifying key academic trends in research. This is an essential contribution as it represents the first comprehensive literature review of the academic literature in this area. In the next phase of our research, we will delve deeper into our findings, refine our research questions, and provide a more detailed analysis of each topic identified by our study. By doing so, we can provide a more nuanced understanding of where current academic research trends exist and where new opportunities exist for future researchers.

REFERENCES

Ali, S., Padmanabhan, V., and Dixon, J. 2014. "Why Cybersecurity Is a Strategic Issue Is Your Business One Hack Away from Disaster?", retrieved from https://www.bain.com/insights/why-cybersecurity-is-a-strategic-issue/.

Arun, R., Suresh, V., Veni Madhavan, C. E., and Narasimha Murthy, M. N. 2010. "On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations," *Advances in Knowledge Discovery and Data Mining* (6118), pp. 391–402.

Baskerville, R. 1993. "Information Systems Security Design Methods: Implications for Information Systems Development," *ACM Computing Surveys* (25:4), pp. 375–414.

Baskerville, R. L., and Wood-Harper, A. T. 1996. "A Critical Perspective on Action Research as a Method for Information Systems Research," *Journal of Information Technology* (11:3), pp. 235–246.

Blei, D., Edu, B., Ng, A., Jordan, M., and Edu, J. 2003. "Latent Dirichlet Allocation," *Journal of Machine Learning Research* (3:3), pp. 993–1022.

Cao, J., Xia, T., Li, J., Zhang, Y., and Tang, S. 2009. "A Density-Based Method for Adaptive LDA Model Selection," *Neurocomputing* (72:7), pp. 1775–1781.

Deveaud, R., SanJuan, E., and Bellot, P. 2014. "Accurate and Effective Latent Concept Modeling for Ad Hoc Information Retrieval," *Document Numérique* (17:1), pp. 61–84.

Dhillon, G., and Backhouse, J. 2001. "Current Directions in IS Security Research: Towards Socio-Organizational Perspectives," *Information Systems Journal* (11:2), pp. 127–153.

Dhillon, G., Smith, K., and Dissanayaka, I. 2021. "Information Systems Security Research Agenda: Exploring the Gap between Research and Practice," *The Journal of Strategic Information Systems* (30:4), p. 101693.

Griffiths, T. L., and Steyvers, M. 2004. "Finding Scientific Topics," *Proceedings of the National Academy of Sciences* (101: Supplement 1), pp. 5228–5235.

Huang, A. H., Lehavy, R., Zang, A. Y., and Zheng, R. 2018. "Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach," *Management Science* (64:6), pp. 2833–2855.

Lee, D. M. S., Trauth, E. M., and Farwell, D. 1995. "Critical Skills and Knowledge Requirements of IS Professionals: A Joint Academic/Industry Investigation," *MIS Quarterly* (19:3), p. 313.

Mortenson, M. J., and Vidgen, R. 2016. "A Computational Literature Review of the Technology Acceptance Model," *International Journal of Information Management* (36:6), pp. 1248–1259.

Rosemann, M., and Vessey, I. 2008. "Toward Improving the Relevance of Information Systems Research to Practice: The Role of Applicability Checks," *MIS Quarterly* (32:1), p. 1.

Siponen, M. T. 2005. "An Analysis of the Traditional IS Security Approaches: Implications for Research and Practice," *European Journal of Information Systems* (14:3), pp. 303–315.

Totty, S., Li, H., Janz, B., and Zhang, C. 2020. "Themes in Information Security Research in the Information Systems Discipline: A Topic Modeling Approach," *AMCIS 2020 Proceedings* (7).

von Bary, B., Westner, M., and Strahringer, S. 2018. "Do Researchers Investigate What Practitioners Deem Relevant? Gaps between Research and Practice in the Field of Information Systems Backsourcing," *IEEE Xplore*, July 1, pp. 40–49.

Webster, J., and Watson, R. T. 2002. "Analyzing the Past to Prepare for the Future: Writing a Literature Review," *MIS Quarterly* (26:2), pp. xiii–xxiii.