Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

SAFEGUARDING INFRASTRUCTURE FROM CYBER THREATS WITH NLP-BASED INFORMATION RETRIEVAL

Christin J. Salley
Neda Mohammadi
John E. Taylor

School of Civil and Environmental Engineering
Georgia Institute of Technology
756 W Peachtree Street Northwest
Atlanta, GA 30308, USA

ABSTRACT

Natural disasters disrupt systems, leading to critical infrastructure vulnerabilities prone to cyber-attacks. The MITRE ATT&CK Enterprise Matrix is a knowledge base for threat analyses in the cybersecurity community. Existing processes to derive possible attack methodologies from this Matrix are largely manual and time-consuming. It is essential to automate the information retrieval process to reduce human errors, improve efficiency, and free up resources for identifying unrevealed cyber-attacks. We propose a framework that incorporates Natural Language Processing (NLP) and Text Mining to automatically generate sets of attack paths from the technique descriptions in the Matrix. The framework generates similarity between techniques based on their descriptions and creates an output showing potential pathways an adversary can take to infiltrate a system. The outputs are compared against an annotated approach and attack report. The results of this study provide an approach to more quickly and effectively assess potential cyber-attacks towards protecting critical infrastructure.

1 INTRODUCTION

Community resilience has become a national priority, and a key component of community resilience is protecting critical infrastructure. A disruption to critical infrastructure can have severe consequences for communities including loss of life, economic disruption, and social instability. Furthermore, protecting critical infrastructure that helps provide essential needs such as electricity and water, which can make communities less vulnerable to disruptions, is a national priority (U.S. Government Accountability Office 2023). With the digital world progressively growing (e.g., internet usage increasing) cyber-attacks will continue to be an imminent risk society faces, particularly during natural disasters and hazards. In fact, it is reported that during natural disasters and hazards attackers increase their strikes by up to 50% (GoldSky Security 2022). During such crisis events, attackers can take advantage of an already chaotic situation and leverage the vulnerable state of an area to infiltrate critical infrastructure systems such as power, water, transportation, emergency response, and healthcare; all of which can disrupt daily life activities and are components integrated in emergency management. Computer networks have been called the “central nervous system” of infrastructure (Walker 2012), therefore it is important to continuously investigate new approaches to keep these systems safe.

Over ten years ago, an executive order was declared by former President Barack Obama which declared that recurring cyber-attacks on critical infrastructure requires better cybersecurity approaches, and that “the cyber threat to critical infrastructure continues to grow and represents one of the most serious national security challenges we must confront” (Office of the Press Secretary 2013). This led to the National Institute of Standards and Technology’s (NIST) Cybersecurity Framework. Widely adopted, the NIST

979-8-3503-6966-3/23/$31.00 ©2023 IEEE

Cybersecurity Framework is designed to help mitigate risks associated with cyber threats, however, it is not a one-size-fits-all approach for organizations (NIST 2023). The framework provides a cyclical process to managing cyber risks through functions categorized as Identify, Protect, Detect, Respond, and Recover (NIST 2023). Since the creation of this framework, there has been more demand in research to discover new ways to manage cyber risks, such as automated techniques to analyze cybersecurity text and uncover useful understanding of cyber threats (Trong et al. 2020). This in turn can aid in better detection of cyber-attacks, through understanding and predicting potential scenarios.

Methods such as Natural Language Processing (NLP) have begun to be used in the cybersecurity field for tasks such as analyzing cyber-related documents (Georgescu 2019). One of the most prominent knowledge bases with cyber-related texts is MITRE ATT&CK. Being able to understand this information and patterns to connect tactics, techniques and procedures (TTPs) would lead to augmented mitigation of such malicious behaviors and threats. The use of MITRE ATT&CK textual information about techniques an adversary can use provides key personnel with the additional insight and situational awareness needed for defense against attackers. Taking action to prevent and anticipate potential cyber-attacks is crucial to protecting critical infrastructure and its associated systems.

2 RELATED LITERATURE

The protection of critical infrastructure, such as emergency management systems, can aid in assisting the response to and mitigation of crisis events. Effective emergency management systems can help to reduce the impacts of disasters and ensure that critical infrastructure can continue to function during and after an emergency event. Literature reveals many organizations are relying on telecommunications and computational systems to support their emergency response efforts as data collection and real-time information exchange occurs with disaster management (Seba et al. 2019). Preparing for cybercrimes and intrusions of these systems will help organizations defend against potential disruptions and should be integrated into emergency response plans (Janczewski and Colarik 2008). When it comes to emergency management, it is important to keep these systems safe not only for communication and information exchange of organizations, but also to keep the private identifiable details of citizens uncompromised (Sutedi et al. 2021). In healthcare, research has been done to show data breaches have steadily increased since 2010 (Ponemon 2016), leaving patients' personal information in hospital systems at risk. If an adversary successfully attacks this system, the intruder can get information such as name, date of birth, address, insurance provider, etc. and the damage can range from fraud to psychosocial harm (Argaw et al. 2020). Studies on critical infrastructure and cybersecurity suggest ways to bridge gaps between emergency management and cybersecurity personnel. These include improving communication about cyber crises (Bolton 2013), establishing information science-based data systems to equip communities to better manage natural disasters (Li et al. 2014), and expressing concerns for a lack of cyber situational awareness during natural disasters (Walker et al. 2010) more effectively.

NLP is known for being able to emulate human language. NLP can execute tasks such as extracting information from unstructured data, such as text, and is used in applications of cybersecurity (Ukwen and Karabatak 2021). To meet the challenge of identifying cyber threats in documents or reports, studies have been completed to create both word and phrase embeddings for the cybersecurity domain to use in modeling (Purba et al. 2020; Ranade et al. 2021). Regarding NLP integration into cybersecurity, studies have also examined event detection (Trong et al. 2020), analyzing sentiments from cybersecurity reports (Phandi et al. 2018), and development of a semi-supervised NLP model for security entity extraction (Jones et al. 2015). The issue with most approaches, current and past, is that they are manual (i.e., time consuming) and costly (Kuhl and Sudit 2007). Additionally, models can assume that all attack steps can be performed instantly, not including multiple paths or relationships between attack techniques (Xiong et al. 2022). One study that used NLP to address this was Haque et al. (2023), who used NLP to advance attack graphs by mapping them to MITRE ATT&CK Enterprise Matrix techniques using term frequency-inverse document frequency (TF-IDF) and cosine similarity to validate adversarial actions. However, while this approach maps components of the MITRE ATT&CK Enterprise Matrix, it lacks connecting the technique definitions within the cyber-context.

The context of cybersecurity terms can differ from plain English to the cybersecurity field. For instance, the term "actor" in the dictionary would be related to a performer or theatrics, while the definition of actor in the cybersecurity space is "an individual or a group posing a threat" (NIST 2019). When performing NLP, research demonstrates this could produce very different outcomes when comparing to similar texts, as one word can have many meanings (Gaikwad et al. 2014). The need for integration of domain knowledge into the corpus is evident (Talib et al. 2016) to possibly increase the quality of the paths created and produce a higher number of generated pairings in comparison to manual processes. There is also a continued need for automated approaches to streamline processes, reduce human errors, improve efficiency through enabling the forecasting of potential cyber-attacks more quickly, and identify potential paths that the annotated process would produce. With technology continuing to advance and the complex digital world growing, the demand for automated systems is increasing. There is a gap in research incorporating more NLP-based systems into cybersecurity and digital technology that are automated (Ukwen and Karabatak 2021). Furthermore, there is a limited number of NLP studies trained with cybersecurity-based context to protect critical systems, like emergency management systems. Therefore, in this study we generate hypothetical cyber-attack scenarios (i.e., potential attack paths) and present an NLP-based approach with cyber-based trained text that automates this process. The attack paths generated show relationships between every technique, thus the length of the attack paths span from shortest to longest path and represent a range of attacker skill levels from script kiddies to top-tier nation states. This process is a step towards further protecting critical infrastructure, and can be applied to scenarios such as natural disasters when a community is most vulnerable and resilience against such attacks is low.

3 METHODOLOGY

For decades, digital text such as social media posts, emails, reports, etc. have become more prominent in society (Humphreys and Wang 2018). Text mining is a common multidisciplinary tool used to analyze such data and is at the intersection of computation linguistics, artificial intelligence/machine learning, statistics, and information science. NLP has been known to decrease issues that occur with text mining (Talib et al. 2016). For instance, NLP techniques such as removing stop words or tokenization can be used for pre-processing the text data before text mining occurs. In determining the best text mining technique for generating potential attack paths in this study, we analyzed previous studies that have compared various text mining techniques and their usefulness (Dang and Ahmad 2014). In assessing the advantages and disadvantages of alternative approaches such as categorization, summarization, etc., we had to consider the importance of extracting relevant information and finding patterns within the given set of words (Talib et al. 2016). Techniques found within Information Retrieval (IR) were the most promising to achieve the goal of identifying patterns within words and establishing connections between text descriptions. Leveraging the benefits of NLP pre-processing and a text mining technique, we employed IR in the form of Term-Frequency Inverse Document Frequency (TF-IDF) to assess the relevance of words in each text description to one another, facilitating the extraction of valuable information. TF-IDF will be defined in more detail in the next section. The high-level steps taken to execute this methodology were: collecting the data, transforming the data from unstructured to structured, numerically representing the text, extracting valuable information, and discovering patterns. Figure 1 below shows the proposed framework to use an NLP-based approach to automatically generate potential attack paths.

3.1 Data Sets

To execute the objective of our study, we used data in the form of the pre-trained Word2Vec Cyber-Phrase model (Purba et al. 2020), and textual data from the MITRE ATT&CK Enterprise Matrix. Below we will further describe each data set in more detail.

Cyber-Phrase Model: To obtain cybersecurity context, we used the data found in the pre-trained model by Purba et al. (2020), which is trained on cybersecurity related material. This is a Word2Vec based model that produces phrase vectorization on text, and demonstrated it outperformed word vectorization of the popular, competing models by the IBM funded UMBC model and Google's model (Purba et al. 2020). These were initially considered prior to discovering the performance of the phrase model. The data itself used to train the model include "CTI reports (Fireeye, Talos, Symantec, APTnotes, Sans and others), Common Weakness Enumeration (CWE), National Vulnerability Database (NVD), Common Vulnerabilities and Exposures (CVE), MSDN documents, security books, and security papers" (Purba et al. 2020). Using this model, this embedded data is what is also included in our study for cyber-related information since we employ this model in the analysis.

MITRE ATT&CK Enterprise Matrix: The relationships between the various tactics and techniques can be seen in this matrix. This matrix in particular represents the most traditional platforms and technologies, including those used in infrastructure systems such as hospitals. The descriptions provided by each technique connect to the tactics and strategies used by an adversary for an attack. Tactics are the "why" of a technique, whereas the technique itself is the "how" an adversary achieves its attack (The MITRE Corporation 2023). As this study is most interested in how an attacker can execute a cyber intrusion, and is looking at critical infrastructure that represents and uses more traditional platforms, the Enterprise Matrix techniques were the proper matrix and information to study. There are 196 techniques in the Enterprise Matrix; each with paragraph descriptions entailing actions that lead to successful execution. This data is what was used as the techniques in potential attack paths produced.

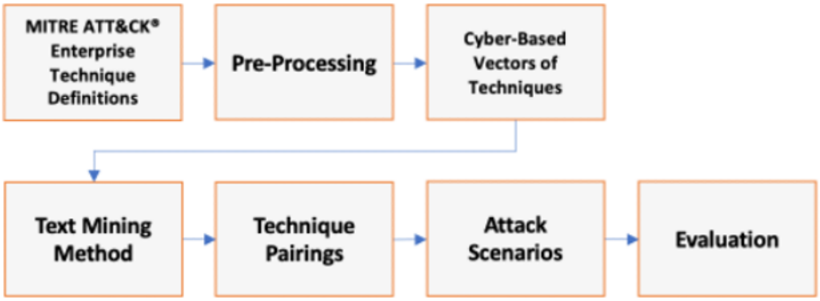
MITRE ATT&CK® Enterprise Technique Definitions
Pre-Processing
Cyber-Based Vectors of Techniques
Text Mining Method
Technique Pairings
Attack Scenarios
Evaluation

Figure 1: NLP-based framework to automatically generate sets of attack paths from the MITRE ATT&CK Enterprise Matrix.

3.2 NLP-Based Framework

The workflow begins with a web-scraping the Enterprise Matrix definitions to obtain a tabular database with its technique descriptions found on the MITRE ATT&CK website. As mentioned previously, textual data is unstructured therefore the use of NLP is performed via pre-processing. Standard techniques such as tokenization, stop words removal, stemming, and lemmatization were performed, as done in previous studies (Salley et al. 2021), to pre-process the data and transform it so it is prepared to be integrated into a machine learning model. Words with less than three letters were also removed, as well as filtering out words that were not nouns, verbs, adjectives or non-English. When the descriptive text was cleaned, it was then a new corpus of only words that fit this criterion. The corpus of pre-processed words from the Enterprise Matrix definitions are then ran through the downloaded, pre-trained cyber-phrase model to begin to create embeddings and analyze similarity of the techniques based on cyber-based context.

A vectorizer was constructed to assign numerical values to each word or term in the corpus (e.g., embeddings) and build a numerical TF-IDF matrix based on an anagram analyzer. TF-IDF is a common Information Retrieval (IR) approach used to determine how relevant the words in each text description are to one another based on the TF-IDF scores and compares every data entry point to one another. The words that appear more frequent, appearing the most in the corpus as a whole, are considered more relevant. A pairwise matrix using cosine similarity was also performed to measure the semantics of each technique description to one another, to further develop relationships between the techniques. The pairwise matrix with cosine similarity values in the model had a threshold of 0.5. Despite this threshold seeming relatively conservative it produces satisfactory results (Zhai et al. 2011), and we wanted to fully maximize the number of pathways possible. Once similarity and relevancy were determined through the TF-IDF model, the techniques were grouped based on their likeness as determined by the TF-IDF similarity scores to improve the effectiveness of the information retrieval through assembling the techniques into their respective groupings. This produced 57 unique groups, the number of techniques in each group being at least two. When each group was further broken down and split to create pairings of starting and resulting techniques for illustrative purposes, mapping the techniques to one another generated 643 combinations of starting and resulting technique pairs that were matched based on their pairwise similarity. The techniques were paired in starting and resulting techniques to visualize their connections and potential pathways (Xiong et. al 2022). The techniques and the potential paths they create based on textual similarity were then illustrated with a network graph to show the interlinked entities for further analysis.

3.3 Results

The visualization of the network graph serves as the generation of attack scenarios step in the framework. A circle graph was used, as shown in Figure 2 below, to represent the relationships between the techniques (i.e., how close the nodes are). Each node represents one of the techniques in the MITRE ATT&CK Enterprise Matrix that generated a relationship with another technique within the 0.5 similarity threshold. The edges represent that similarity connection between nodes. The total number of nodes in the network graph created is 78, and the total number of edges is 140. The nodes for "Search Open Technical Databases" and "Gather Victim Host Information" had the highest degrees (15), indicating that they have the most relationships with the other techniques in the Enterprise Matrix. The graph in Figure 2 depicts the various paths an adversary can take based on the textual definitions of the techniques, showing realistic portrayals of some nodes circling back to themselves in attempts to try techniques again and that the paths to each technique in this network is not too far from one another.

3.4 Evaluation

To determine if the attack paths created with this NLP approach were indeed adequate in size (i.e., could generate a majority of attack paths predicted by manual processes) and faster in completion (i.e., not taking several hours or days to annotate to determine sequences), we compared the quantity and time of our execution of potential cyber-attacks with a previous study that manually generated attack steps with the MITRE ATT&CK Enterprise Matrix and starting and resulting techniques as well. Xiong et. al (2022) proposed enterpriseLang based on meta attack language. Through parsing the main enterpriseLang file (i.e., not considering single attacks) there were 1009 pairs of techniques. Manual annotation not only can take significant time to do and requires domain expertise, but there are also challenges with disagreements between annotators (Trong et al. 2020) which can call into question consistency. Our approach was able to capture nearly 64% of the same number of potential attack paths as a manual process and complete the task in under 20 seconds. The manual process would take hours to days to complete with a team of researchers; completing the task in under 20 seconds substantially outpaces the manual processing speed.

To confirm the quality of the pairings, we conducted testing to see if our technique pairs can model a real-life attack scenario. Due to availability, the publicly available cyber-attack report we were able to utilize was on the Ukraine Cyber Attack to demonstrate the potential accuracy of the pairings. The attack occurred in 2015 and caused over 3 hours of power outages, affecting nearly 225,000 people (CISA 2021). According to the SANS Industrial Control Systems Library's ICS Defense Use Case report on the incident, there were nine technical components, that were consolidated into six, that were used by the adversaries and can be translated to MITRE ATT&CK techniques: 1) Spear phishing, 2) Credential Theft, 3) VPN access, 4) Workstation remote, 5) Control and Operate, 6) Tools and Tech (Lee et al. 2016). While our data

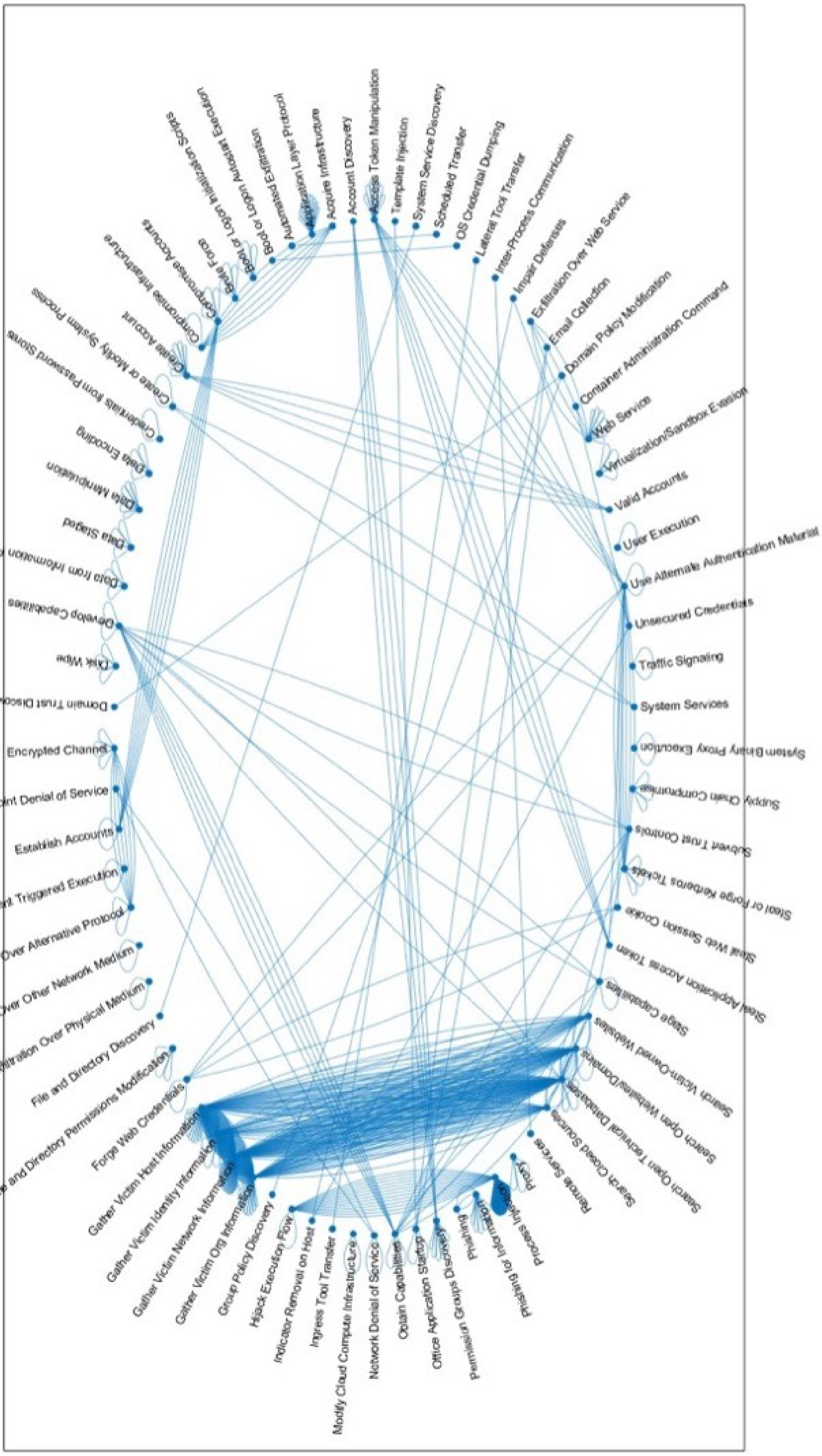
Account Discovery
Access Token Manipulation
Template Injection
System Service Discovery
Scheduled Transfer
OS Credential Dumping
Lateral Tool Transfer
Inter-Process Communication
Impair Defenses
Exfiltration Over Web Service
Email Collection
Domain Policy Modification
Container Administration Command
Web Service
Virtualization/Sandbox Evasion
Valid Accounts
User Execution
Use Alternate Authentication Material
Unsecured Credentials
Traffic Signaling
System Services
System Binary Proxy Execution
Supply Chain Compromise
Subvert Trust Controls
Steal or Forge Kerberos Tickets
Steal Web Session Cookie
Steal Application Access Token
Stage Capabilities
Search Victim-Owned Websites
Search Open Websites/Domains
Search Open Technical Databases
Search Closed Sources
Remote Services
Proxy
Process Injection
Phishing for Information
Phishing
Permission Groups Discovery
Office Application Startup
Obtain Capabilities
Network Denial of Service
Modify Cloud Compute Infrastructure
Ingress Tool Transfer
Indicator Removal on Host
Hijack Execution Flow
Group Policy Discovery
Gather Victim Org Information
Gather Victim Network Information
Gather Victim Identity Information
Gather Victim Host Information
Forge Web Credentials
File and Directory Permissions Modification
File and Directory Discovery
Exfiltration Over Physical Medium
Exfiltration Over Other Network Medium
Exfiltration Over Alternative Protocol
Event Triggered Execution
Establish Accounts
Endpoint Denial of Service
Encrypted Channel
Domain Trust Discovery
Disk Wipe
Develop Capabilities
Data from Information Repositories
Data Staged
Data Manipulation
Data Encoding
Credentials from Password Stores
Create or Modify System Process
Create Account
Compromise Infrastructure
Compromise Accounts
Brute Force
Boot or Logon Initialization Scripts
Boot or Logon Autostart Execution
Automated Exfiltration
Application Layer Protocol
Acquire Infrastructure

Figure 2: Network graph illustrating technique pairings based on textual similarity, with each node representing a different technique. This output highlights the potential paths an adversary can take to execute an attack.

set has techniques that match these components, they do not generate paths exactly in the same order. This could be due to the nature of the definitions, pairing with others that are more similar to predict a natural progression of attack, and sometimes cyber- attacks are unpredictable. Our ongoing research aims to expand the data set to include more attack paths to better model real-life attack scenarios; this is further discussed in the next section. Overall, however, the NLP-based approach was a success in the area of faster computation and achieving nearly two thirds the quantity of attack paths in comparison to traditional manual annotations.

4 DISCUSSION

State and local government officials need to implement cybersecurity plans, drills, or workshops for emergency preparedness, so they can properly respond to crisis events and help those in need without the disruption of a cyber-attack in their networks generating more issues to an already heightened event. For instance, in 2021 the state of Indiana's federally funded Multi-State Information Sharing and Analysis Center (which helps protect against cyber threats) participated in exercises to assist them with how to deal with the potential impacts of both a natural disaster and a cyber-attack at the same time, and in 2018 the city of Houston and the U.S. Army Cyber Institute hosted a three-day drill on how to handle cyber-attacks during a hurricane (Bergal 2021). Efficient training, communication, and regularly updated policies are pertinent to plan for cyber-attacks, as well as state-of-the-art systems being in place that can mitigate cyber threats promptly after a major incident occurs that weakens critical infrastructure needed to sustain communities. While an approach such as ours can be generalized to use across multiple sectors, we emphasize the urgency of using protective measures against cyber-attacks during natural disasters as it is a rising concern.

There are practical contributions of this study that can impact society. One, our study provides a framework to more quickly and efficiently assess potential cyber-attacks. It also aids in protecting critical infrastructure such as the electricity grid or emergency response telecommunications pertaining to natural disasters when implemented into cyber-attack prevention efforts. Also, this work can mitigate impacts of natural disasters by allowing efficient execution of response efforts without interference in pertinent networks. For instance, in the context of Emergency Medical Services (EMS), the dispatching system and sharing of patient information are vital components. If the EMS system were to be compromised, not only would private information be at risk of exposure, but the longer the system remains inoperable, the greater the potential loss of lives or increased risk of individuals not receiving the necessary medical attention needed during a crisis. To the domain of research related to protecting critical infrastructure, the study is a foundation for the need for more advanced NLP processes to generate attack scenarios (Ukwen and Karabatak 2021). It is an approach on how to ensure critical infrastructure is not impeded by cyber-attacks (Loukas et al. 2013). This study also is an automotive approach, that unlike others, uses cyber-based context to produce similarity pairings.

While this study demonstrated a quicker and effective way to generate potential cyber-attacks, there are some limitations to this study. One limitation is the scarcity of publicly available reports on actual cyber-attacks, primarily due to privacy concerns, which restricts access to impede learning opportunities for hackers. This in turn, however, makes researching actual cases or implementation of effective countermeasures difficult. Constant evolution and advancement of hackers pose another limitation to this study. As this system develops to produce more cyber-attack scenarios, if hackers got ahold of this information, they could then plan out how they might be stopped and mitigate around the preventative measures further. As NLP and text mining for cybersecurity continues to develop, there should be a larger discussion around what data, models, or code need to be more confidential. Cyber-attacks are also rarely identical in nature. Therefore, in order to use this framework new text and definitions should be added as a future work as time goes on.

Additionally, while the data set used was able to capture a good number of relationships between techniques, there are still more and this is not a comprehensive list of all possible attack paths. It is, though, a step toward the NLP-based automation of these time intensive efforts. Striking a balance between automated and manual processes that maximizes efficiency, accuracy, and cost-effectiveness, while also considering factors such as data volume, complexity, and criticality would best benefit cybersecurity analyses. Overall, the framework has versatile applications in various domains, including enhancing overall security awareness and simulating realistic attack paths and scenarios. It enables the implementation of proactive strategies to mitigate risks, supports security training programs, and facilitates the integration of countermeasures during the security architecture design phase.

5 CONCLUSION

Protecting critical infrastructure from cyber-attacks has become an important societal issue that affects various sectors from economics to public health. Vulnerability of cyber emergency response is not a new issue (Jennex 2007). However, it is very difficult to protect systems from cyber-attacks as they occur which is why it is important to be able to anticipate and predict what an adversary might do to be able to better mitigate on the front end and reduce such security events (Han et al. 2019). To minimize the threat of cyber-attacks, models have been generated to simulate potential cyber threats. However, the issue with most current approaches is that they are time-consuming and there can be inconsistencies with annotations. In this paper, we proposed a framework that incorporates NLP and Text Mining to automatically and systemically generate sets of attack paths from the technique descriptions in the Enterprise Matrix. The framework ingests attack related definitions, produces linkages between techniques, and creates an output graph that shows the relationship between techniques and the potential pathways an adversary can take to enact their desired consequence(s). While the application of this work focuses on using such generated scenarios to protect critical infrastructure during disasters, this work can complement various sectors. As it pertains to disasters, though, it can complement all parts in both the cyclical process of the four phases of emergency management (i.e., mitigation, preparedness, response, and recovery) and the five key functions of the NIST Cybersecurity Framework to enhance cybersecurity awareness and manage risk. The implementation of proper cybersecurity planning and mitigation tactics for emergency preparedness and response to crisis events can help those in need and reduce unnecessary interference in people's routines or lives with cyber-attacks. Protecting critical infrastructure from cyber threats directly aids community resilience efforts to better protect society. With stronger community resilience, society will be able to recover faster from crisis events such as natural hazards and disasters.

ACKNOWLEDGMENTS

This study was conducted in collaboration with Mike Nygaard, Deputy Associate Program Leader, Cyber Modeling & Simulation at Lawrence Livermore National Laboratory during an internship the first author had. This material is based upon work supported by the National Science Foundation under Grant No. 1837021. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Lawrence Livermore National Laboratory. The hypothetical scenarios can be shared upon reasonable request.

REFERENCES

Argaw, S. T., J. R. Troncoso-Pastoriza, D. Lacey, M.V. Florin, F. Calcavecchia, D. Anderson, W. Burleson, J.M. Vogel, C. O'Leary, B. Eshaya-Chauvin, and A. Flahault. 2020. "Cybersecurity of Hospitals: Discussing the Challenges and Working Towards Mitigating the Risks." *BMC Medical Informatics and Decision Making* 20:1-10.

Bergal, J. 2021. Natural Disasters Can Set the Stage for Cyberattacks. https://stateline.org/2021/10/25/natural-disasters-can-set-the-stage-for-cyberattacks/, accessed 2nd March 2023.

Bolton, F. 2013. "Cybersecurity and Emergency Management: Encryption and the Inability to Communicate." *Journal of Homeland Security and Emergency Management* 10(1):379-385.

CISA. 2021. Cyber-Attack Against Ukrainian Critical Infrastructure. https://www.cisa.gov/news-events/ics-alerts/ir-alert-h-16-056-01, accessed 3rd March 2023.

Dang, S., and P.H. Ahmad. 2014. “Text Mining: Techniques and its Application.” *International Journal of Engineering & Technology Innovations* 1(4):22-25.

Gaikwad, S. V., A. Chaugule, and P. Patil. 2014. “Text Mining Methods and Techniques.” *International Journal of Computer Applications* 85(17):42-45.

Georgescu, T. M. 2019. “Machine Learning Based System for Semantic Indexing Documents Related to Cybersecurity.” *Economy Informatics* 19(1):5-13.

GoldSky Security. 2022. How To PROTECT Your Data Amid Natural Disasters. https://www.goldskysecurity.com/how-to-protect-your-data-amid-natural-disasters/, accessed 2nd March 2023.

Han, C. H., S.T. Park, and S.J. Lee. 2019. “The Enhanced Security Control Model for Critical Infrastructures with the Blocking Prioritization Process to Cyber Threats in Power System.” *International Journal of Critical Infrastructure Protection* 26:1-10.

Haque, M. A., S. Shetty, C.A. Kamhoua, and K. Gold. 2023. “Adversarial Technique Validation & Defense Selection Using Attack Graph & ATT&CK Matrix.” In *2023 International Conference on Computing, Networking and Communications (ICNC)*, 181-187.

Humphreys, A., and Wang, R. J. H. 2018. “Automated Text Analysis for Consumer Research.” *Journal of Consumer Research* 44(6):1274-1306.

Janczewski, L., and A. Colarik (Eds.). 2007. “Cyber Warfare and Cyber Terrorism.” IGI Global.

Jennex, M. E. 2007. “Modeling Emergency Response Systems.” In *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07),* 1-8.

Jones, C. L., R.A. Bridges, K.M. Huffer, and J.R. Goodall. 2015. “Towards a Relation Extraction Framework for Cyber-Security Concepts.” In *Proceedings of the 10th Annual Cyber and Information Security Research Conference*, 1-4.

Kuhl, M. E., M. Sudit, J. Kistner, and K. Costantini. 2007. “Cyber Attack Modeling and Simulation for Network Security Analysis.” In *Proceedings of the 2007 Winter Simulation Conference*, edited by S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, 1180-1188. IEEE.

Lee, R.M., M.J. Assante, and T. Conway. 2016. “Analysis of the Cyber Attack on the Ukrainian Power Grid.” Technical Report, Electricity Information Sharing and Analysis Center (E-ISAC), 388, 1-29.

Li, J., Q. Li, C. Liu, S.U. Khan, and N. Ghani. 2014. “Community-Based Collaborative Information System for Emergency Management.” *Computers & Operations Research* 42:116-124.

Loukas, G., D. Gan, and T. Vuong. 2013. “A Review of Cyber Threats and Defence Approaches in Emergency Management.” *Future Internet* 5(2):205-236.

NIST. 2019. Actor. https://csrc.nist.gov/glossary/term/actor, accessed 1st May 2023.

NIST. 2023. Quick Start Guide. https://www.nist.gov/cyberframework/getting-started/quick-start-guide, accessed 1st May 2023.

Office of the Press Secretary. 2013. ExecutiveOrder -- Improving Critical Infrastructure Cybersecurity. https://obamawhitehouse.archives.gov/the-press-office/2013/02/12/executive-order-improving-critical-infrastructure-cybersecurity, accessed 30th April 2023.

Phandi, P., A. Silva, and W. Lu. 2018. “SemEval-2018 Task 8: Semantic Extraction from CybersecUrity REports using Natural Language Processing (SecureNLP).” In *Proceedings of the 12th International Workshop on Semantic Evaluation,* 697-706.

Ponemon, I. 2016. “Sixth Annual Benchmark Study on Privacy & Security of Healthcare Data.” Technical Report, Ponemon Institute.

Purba, M. D., B. Chu, and E. Al-Shaer. 2020. “From Word Embedding to Cyber-Phrase Embedding: Comparison of Processing Cybersecurity Texts.” In 2*020 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 1-6.

Ranade, P., A. Piplai, A. Joshi, and T. Finin. 2021. “CyBERT: Contextualized Embeddings for the Cybersecurity Domain.” In *2021 IEEE International Conference on Big Data (Big Dat*a), 3334-3342.

Salley, C., N. Mohammadi, and J.E. Taylor. 2021. “Semi-Supervised Machine Learning Framework for Fusing Georeferenced Data from Social Media and Community-Driven Applications.” In *Computing in Civil Engineering 2021*, 114-122.

Seba, A., N. Nouali-Taboudjemat, N. Badache, and H. Seba. 2019. “A Review on Security Challenges of Wireless Communications in Disaster Emergency Response and Crisis Management Situations.” *Journal of Network and Computer Applications* 126:150-161.

Strom, B. E., A. Applebaum, D.P. Miller, K.C. Nickels, A.G. Pennington, and C.B. Thomas. 2018. “MITRE ATT&CK: Design and Philosophy.” Technical Report, The MITRE Corporation.

Sutedi, A., E. Gunadhi, D. Heryanti, and R. Setiawan. 2021. “Data Privacy in Disaster Situation: A Review.” In *2021 International Conference on ICT for Smart Society (ICISS),* 1-4.

Talib, R., Hanif, M. K., Ayesha, S., and Fatima, F. 2016. “Text mining: techniques, applications and issues.” *International Journal of Advanced Computer Science and Applications* 7(11):414-418.

The MITRE Corporation. 2023. Enterprise Matrix. https://attack.mitre.org/matrices/enterprise/, accessed 13th July 2023.

Trong, H. M. D., D.T. Le, A.P.B. Veyseh, T. Nguyễn, and T.H. Nguyen. 2020. “Introducing a New Dataset for Event Detection in Cybersecurity Texts.” In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 5381-5390.

Ukwen, D. O., and M. Karabatak. 2021. “Review of NLP-based Systems in Digital Forensics and Cybersecurity.” In *2021 9th International Symposium on Digital Forensics and Security (ISDFS),* 1-9.

U.S. Government Accountability Office. 2023. Critical Infrastructure Protection: Time Frames to Complete CISA Efforts Would Help Sector Risk Management Agencies Implement Statutory Responsibilities. https://www.gao.gov/products/gao-23-106720, accessed 1st May 2023.

Walker, J. 2012. "Cyber Security Concerns for Emergency Management." In *Emergency Management*, edited by B. Eksioglu, 39-59. Rijeka, Croatia: InTech.

Walker, J., B.J. Williams, and G.W. Skelton. 2010. "Cyber Security for Emergency Management." In *2010 IEEE International Conference on Technologies for Homeland Security (HST),* 476-480.

Xiong, W., E. Legrand, O. Åberg, and R. Lagerström. 2022. "Cyber Security Threat Modeling Based on the MITRE Enterprise ATT&CK Matrix." *Software and Systems Modeling* 21(1):157-177.

Zhai, J., Y. Lou, and J. Gehrke. 2011. "ATLAS: A Probabilistic Algorithm for High Dimensional Similarity Search." In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, 997-1008.

AUTHOR BIOGRAPHIES

CHRISTIN J. SALLEY is a Ph.D. Candidate in the Network Dynamics Lab in the School of Civil and Environmental Engineering at the Georgia Institute of Technology. Her research interest revolves around mitigating natural hazards through emergency detection and communications and establishing equitable systems for the built environment. She utilizes methodologies such as information science and Machine Learning techniques like Natural Language Processing, classification, clustering, and regression for data retrieval and textual analyses. Her email address is csalley3@gatech.edu and she can be found on her lab's website at https://ndl.gatech.edu/.

NEDA MOHAMMADI is the Director of City Infrastructure Analytics in the Network Dynamics Lab in the School of Civil and Environmental Engineering at the Georgia Institute of Technology. Her primary research is in the area of city infrastructure systems with a focus on human-infrastructure system interactions. She primarily explores the complexities of evolving human-infrastructure interrelationships and their space-time fluctuations to quantify, explain, predict, and enhance urban dynamics towards sustainable urbanization. Her email address is nedam@gatech.edu and she can be found on her lab's website at https://ndl.gatech.edu/.

JOHN E. TAYLOR is the Frederick Law Olmsted Professor in the School of Civil and Environmental Engineering at the Georgia Institute of Technology, where he is Director of the Network Dynamics Lab. He specializes in investigating engineering network dynamics of industrial and societal importance. His current research focuses on: (1) achieving sustained energy conservation by coupling energy use with occupant networks and examining inter-building network phenomena in cities, and (2) understanding and improving response times by affected human networks during extreme events in urban areas. His research extends from developing virtual and augmented reality applications to collect and visualize urban scale data, to developing real-time interventions to improve urban sustainability and resilience. His email address is jet@gatech.edu and his lab's website is https://ndl.gatech.edu/.