Forging New Paths in Cybersecurity Doctoral Research with Open Datasets and Synthetic Data Generation

Michelle Liu
School of Technology and Innovation
Marymount University
Arlington VA, USA
xliu@marymount.edu

Nathan Green
School of Technology and Innovation
Marymount University
Arlington VA, USA
ngreen@marymount.edu

Diane Murphy
School of Technology and Innovation
Marymount University
Arlington VA, USA
dmurphy@marymount.edu

Donna Schaeffer
School of Technology and Innovation
Marymount University
Arlington VA, USA
dschaeff@marymount.edu

***Abstract*—- This research-to-practice full paper addresses the important need for relevant and comprehensive datasets to advance cybersecurity research by proposing methods for curating open datasets and generating synthetic datasets. Cybersecurity research is a rapidly evolving scientific field, making robust datasets crucial for empirical analysis. Unfortunately, current doctoral research is hindered by the scarcity, limited accessibility, and outdated or irrelevant nature of existing open-source datasets. This paper tackles these challenges by focusing on two main initiatives: (1) curating a pilot collection of open datasets aligned with the National Initiative for Cybersecurity Education (NICE) Cybersecurity Workforce Framework, and (2) using Generative Adversarial Networks (GANs) to generate synthetic datasets. Our research highlights the obstacles faced by doctoral students due to fragmented, outdated data and underscores the importance of accessible datasets for rigorous scientific inquiry. We also demonstrate how synthetic data can ease privacy concerns while still offering researchers realistic data. By incorporating these approaches into doctoral curricula, we aim to equip future cybersecurity researchers with the skills resources for impactful research. The authors will continue to expand their dataset curation efforts and study how discoverable, high-quality datasets can influence doctoral research, particularly empirical studies and their outcomes.**

***Index Terms*—cybersecurity, open data, synthetic data, GAN, doctoral curriculum**

I. Introduction

The field of cybersecurity continues to rapidly evolve, reflecting a shift towards a more rigorous scientific approach that emphasizes empirical research and data-driven analysis. For example, a bibliometric study by Furstena et al. [1] maps out two decades of cybersecurity research, revealing an expanding scope of themes from intrusion detection to complex issues like privacy and smart grids. The study highlights the growing reliance on quantitative methodologies and the increasing sophistication of cybersecurity research, underscoring the field's evolution from practical countermeasures to a structured scientific discipline. Complementing this perspective, [2] introduces the concept of "cybersecurity dynamics", further establishing the field's foundation by advocating for a systemic and scientific approach to understanding and modeling cybersecurity phenomena. The framework by Xu [2] reinforces the necessity for a scientific discipline that can adapt to and anticipate evolving cybersecurity challenges. Adding to this foundation, [3] highlight the critical role of mathematical approaches in elevating cybersecurity to a scientific discipline, arguing that these methodologies provide the precision and replicability needed to transform cybersecurity from a protoscience to a fully developed science. This stream of literature corroborates the pressing needs of transformation of cybersecurity from practical, ad hoc solutions to a more structured and empirical field. However, the current research landscape is marred by outdated and fragmented datasets that fail to capture the evolving dynamics of cyber threats, limiting the scope and depth of their potential use in research and their applicability to research on modern cybersecurity challenges [4].

Our previous work highlighted that the field of cybersecurity research, particularly at the academic doctoral level, faces a significant challenge posed by the scarcity of accessible and comprehensive datasets [5]. This gap hinders the ability of academic doctoral students to conduct thorough and scientifically robust research. Furthermore, this scarcity and inaccessibility of quality data are not just academic issues but also practical concerns that affect the development of effective cybersecurity measures. The authors called for collective action to make more accessible datasets freely available, promoting transparency and strengthening the scientific credibility and trustworthiness of the cybersecurity discipline [5].

This paper aims to address these challenges through two

primary initiatives: the curation of a pilot collection of open datasets aligned with the National Initiative for Cybersecurity Education (NICE) Cybersecurity Workforce Framework [6], and the innovative generation of synthetic datasets using Generative Adversarial Networks (GANs).

For the first initiative, a diverse array of cybersecurity-related datasets is being compiled and categorized according to the distinct areas of cybersecurity functions defined by the commonly used NICE Framework [6]. The goal is to enable researchers to easily locate and utilize datasets relevant to their specific area of interest within cybersecurity. For the second initiative, we showcase the application of GANs to generate synthetic datasets as a promising avenue to augment the limited available data. This approach not only addresses the gaps in data availability but also can mitigate privacy and security concerns associated with the use of real data, thus supporting a broader range of research activities while adhering to ethical standards. Together, these efforts are expected to underpin more rigorous scientific inquiries and enable more sophisticated investigations than currently possible.

Following this introduction, the next section delves into the current state of cybersecurity research and the challenges associated with data accessibility. We then discuss the scientific methodologies being employed to address these challenges, with a focus on the creation of open dataset repositories and the application of GANs for synthetic data generation. The subsequent section will provide a discussion of how these methodologies can be integrated into the curriculum of our doctoral research courses. Finally, the paper concludes with a summary of our study's contributions to the scientific community and suggestions for future research directions.

II. Impact of Data Challenges on Doctoral Research in Cybersecurity

Recent national initiatives including the NSF's Cyberinfrastructure for Public Access and Open Science (CI PAOS) (National Science Foundation, 2024) and the OSTP's memorandum on public access to federally funded research [7] highlight the government's efforts to create research data infrastructure ecosystems across scientific disciplines and domains, emphasizing the importance of "accessibility, discoverability, reliability, reproducibility, sustainability, and utility" of data products [8]. We believe these initiatives are particularly pertinent to the emerging field of cybersecurity, as it evolves into a recognized scientific discipline. The demands for rigorous and verifiable data to support empirical research is becoming apparent [9], [10]. For doctoral students, this evolution requires a dynamic approach to education, where curriculum content continuously integrates the latest methodologies and technological advancements in data analysis. This constant updating is critical to ensure that doctoral research outcomes are scientifically robust and practically relevant [11].

A. Current Data Accessibility Issues

The landscape of cybersecurity research is complicated by data accessibility challenges that impact doctoral research. This section discusses and summarizes several key obstacles faced by our doctoral students, which impede their capacity to conduct advanced and timely research.

A primary issue is the rapid evolution of cybersecurity threats, which outpaces the data collection and results in datasets that quickly become outdated and less relevant for current security challenges [12]. This dynamic nature of threats requires datasets that are not only comprehensive but also timely, an aspect often missing from available open-source data sources. Second, legal and ethical barriers present substantial hurdles. Data privacy laws, such as the General Data Protection Regulation (GDPR) in Europe and privacy regulations in various states in the US, impose stringent restrictions on data collection and sharing, particularly with data that involve personally identifiable information (PII) [13]. These regulations complicate the process for doctoral students who may benefit from access to sensitive data for their research. The proprietary nature of many cybersecurity datasets further exacerbates this issue, as companies and institutions are reluctant to share data that may reveal vulnerabilities or business insights [14]. Third, doctoral students often face technological and resource constraints that limit their ability to gather and analyze large-scale cybersecurity data. The high costs associated with data storage, processing, and analysis tools can be prohibitive, restricting the scope of doctoral projects to smaller, less impactful studies. This is particularly challenging in cybersecurity, where large datasets are crucial for developing effective security measures and algorithms [15]. Fourth, the lack of robust platforms for data sharing and collaboration among researchers hinders doctoral students' ability to access diverse datasets and collaborate with peers. While the value of shared datasets for advancing cybersecurity research is increasingly recognized, effective data sharing and collaboration platforms are still lacking, limiting the quality and scope of research that can be conducted [10]. Finally, addressing these challenges requires concerted efforts to develop more open and collaborative research environments. Initiatives that promote open data platforms and partnerships between academia, government, and industry are crucial as they provide essential resources for doctoral students. These platforms should enable data sharing, encourage standardization, facilitate replication, and support the creation of comprehensive datasets that reflect current and emerging cybersecurity threats [16].

B. Impact on Doctoral Research

The above challenges have shaped the scope and quality of doctoral research in cybersecurity, particularly affecting the choice of research methodologies. Many doctoral students find themselves constrained to using qualitative methods like surveys and interviews due to the unavailability of comprehensive cybersecurity datasets. While these methods provide insights into behavioral and policy-related aspects of cybersecurity, they lack the depth required for analyzing complex technical systems and cyber-attack patterns. This reliance on limited data types can skew research towards certain areas, potentially

neglecting others that are crucial for a holistic understanding of emerging cybersecurity threats [17].

Data scarcity also hinders innovation, particularly in areas requiring extensive datasets such as artificial intelligence (AI) and machine learning (ML) applications in cybersecurity [18]. The lack of robust, real-time data impedes doctoral students' ability to conduct groundbreaking research that could lead to significant advancements in the field. This limitation not only slows the progress of developing more sophisticated security solutions but also restricts doctoral researchers' contributions to cutting-edge innovations. Doctoral students who wish to engage in experimental or simulation-based research encounter additional barriers. Simulations and modeling are essential for testing cybersecurity solutions in controlled environments, but they require accurate and extensive data to be effective [18]. Without access to high-quality datasets, the reliability of simulations and the conclusions drawn from them may be compromised, reducing the impact of the research and its applicability to real-world scenarios [19], [20].

Unlike disciplines such as bioinformatics or environmental science, where data repositories are rich and collaborations across institutions are common [21], cybersecurity research often suffers from data isolation. This isolation, partly due to the sensitive nature of cybersecurity data, complicates efforts to establish comprehensive data-sharing mechanisms. Although initiatives like the IMPACT project are making strides in addressing these gaps [22], much more needs to be done to foster a collaborative and open research environment that enhances the scale and relevance of doctoral research in cybersecurity.

Addressing these challenges involves not only institutional and policy changes but also a cultural shift towards more openness in data sharing within the cybersecurity research community. Enhancing data accessibility for doctoral students is essential for nurturing a new generation of cybersecurity professionals equipped to tackle future challenges with innovative solutions and robust research methodologies.

III. Methodology for Addressing Data Scarcity

A. Curating Open Datasets

There is currently no one specific repository for cybersecurity datasets. That is not to say that there are none available, dispersed across multiple sources, but their "discoverability" is lacking, requiring considerable effort on behalf of any researcher looking for specific types of cybersecurity data. In many cases, there is limited information available on the origins of the data and its provenance, leading to concerns about the "reliability" of the dataset. In May 2022, the Executive Office of the President issued a document giving "desirable characteristics" for data repositories from federally funded research [23]. Many of these recommendations can, however, be applied to any dataset.

In our experiment, three students in our DSc in Cybersecurity program set out to discover available cybersecurity datasets, with data produced in the last 10 years, primarily using the Internet and their knowledge of the field. Their searches included many different search terms, reflecting the evolution of computer security terminology. They first set out to find repositories specific to cybersecurity, for example the Canadian Institute for Cybersecurity from the University of New Brunswick (https://www.unb.ca/cic/datasets/index.html). This resource lists around 30 datasets classified as Ground Truth, IoT, Dark Web, IDS, SCX, Malware and Operational Technology (OT). Of these only 3 categories had recent data (Ground Truth, IOT, and OT), probably reflecting the current research interests at the university.

The students next extended their searches beyond cybersecurity resources, to generally available dataset repositories such as GitHub and Kaggle, as well as vendor open-source data repositories, such as the Registry of Open Data Amazon Web Services (AWS) and Google Large Public Data Directory, and government resources on data.gov. Students found a few cybersecurity datasets on these sites, although again they were limited, many not current, and most with limited information on the provenance of the data. For example, AWS had only 2 cybersecurity data sources listed (https://registry.opendata.aws/).

Our list was growing slowly, with many datasets found dated earlier than 2010 or not dated, and so not eligible for inclusion. In addition, the same dataset was often mentioned on multiple repositories or had links that no longer worked. We now turned our attention to "utility" and the classification of datasets into categories that were in general use. As noted above, we settled on the NIST/NICE Workforce Framework with its seven major functions [24].

At the end of the experiment, the doctoral students had identified and classified some 40 datasets, primarily in the Analysis and Investigate categories of the NIST framework. The research work demonstrated the difficulty in "discovering" open-source cybersecurity-related datasets, and ensuring that cybersecurity datasets that are found are reliable and current.

B. Synthetic Data Generation Using GANs

Generative Adversarial Networks (GANs) have been a major advancement in AI in recent years with end products such as ChatGPT [25], DALL-E [26], and Co-pilot [27]. These tools were trained on large datasets to achieve their particular success defined in their domain. An underused use of GANs has been in the generation of synthetic data where large datasets do not exist, a use case that is increasingly being used in cybersecurity. GANs are typically composed of two neural networks, a generator and a discriminator. These neural networks are trained together, to compete with each other. The generator's goal is to create data that is indistinguishable from real data, and the discriminator's goal is to create a model that can evaluate whether the data is real or synthetic. The discriminator provides this feedback to the generator to improve the output of the generator [28].

The use of GANs for synthetic data generation offers significant advantages to cybersecurity, a field with particularly sensitive data issues. Synthetic data can sidestep typical research concerns around the use of real-world datasets that

often contain PII while still providing a similar distribution of data. This approach is more important than ever given the increase in data privacy laws such as the GDPR that come with restrictions on data usage [13].

Synthetic data in cybersecurity has demonstrated potential in two main areas. First, synthetic datasets can be used to train machine learning models. This approach limits the attacker's ability to expose sensitive information through machine learning attacks. This also helps to provide large-scale data and diverse data that mirror the real world but do not contain the same ethical or legal risks that traditional cybersecurity machine learning carries that might be fined under policies such as GDPR [4].

Secondly, often in cybersecurity, datasets are quickly outdated as the underlying technology has changed or a researcher or company has only been able to release a fragment of a dataset. In many areas of cybersecurity, cybersecurity-related datasets have been determined to be outdated. This reduces the researcher's ability to generalize research results to current real-world scenarios [12]. By solving these problems and creating synthetic data that reflect current cybersecurity landscapes, GANs allow researchers and security measures to be developed that are effective against cyber threats.

Although GANs are a growing trend in research, they do have limitations. The output quality of the synthetic data depends on the quality of the initial datasets used to train the models. Any bias in the original data will likely continue in synthetic data as it tries to follow similar distributions. Additionally, if the original data are not representative of the real-world conditions, the synthetic data is unlikely to match the real-world conditions as well. For many cybersecurity tasks, this could allow for flawed security measures or false confidence in the system as a whole.

Despite these limitations, as GANs mature, we expect them to play an important role in overcoming traditional data collection limitations. Ultimately, the promise of GANs in cybersecurity should give industry and researchers a bridge to connect while allowing privacy and security of data [16].

In one of our doctoral student's dissertation research, various hypotheses relating to the impact of synthetic data on the utility of machine learning models were tested. The student generated synthetic data to preserve privacy, while additionally testing that machine learning models could learn accurately from datasets that did not directly contain sensitive data. This research was driven by the objective of maintaining ethical standards along with utility in data-driven research, a growing domain in our doctoral program. The experiments generated synthetic data based on the European Credit Card Fraud Dataset, available on Kaggle (https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud), while integrating differential privacy techniques.

C. Limitations of Using Synthetic Data in Academic Research

Doctoral students and other academic cybersecurity researchers may increasingly view synthetic data as the viable alternative to the collection of real data or as a valuable methodology for the augmentation of real data where collecting data, particularly from human subjects, is time-consuming, less than comprehensive, subject to privacy regulations, or not free from bias.

While the use of GANS to develop synthetic data, as discussed above, is improving continuously, it is also important for cybersecurity researchers to consider its limitations, most importantly, its "responsible use" in research. Unlike in human subject research, there is currently a lack of legal or ethical frameworks in place to ensure effective use of synthetic data whether, for example, to increase the size of a dataset for machine learning training, to ensure data privacy, to generate missing data, or to deliberately diversify the data.

However, general guidance on ethical uses of AI is becoming more prevalent, with a leading proponent being the European Union with the recent passing and implementation of its Artificial Intelligence (AI) Act as well the pending guidelines from the United States Artificial Intelligence(AI) Safety institute, within the National Institute of Standards and Technology (NIST), with its initial focus on the priorities assigned under President Biden's Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. NIST AI 100-4 has a draft, available on public comment, that addresses techniques to reduce the risks posed by a broad range of synthetic content [29]. Included in their report are the need to ensure the providence of the synthetic data is identified, content is authenticated, synthetic data is detected, and that ownership of the data is known. NIST also identifies a series of harms and risks from synthetic content, including discrimination based on gender, race and ethnicity, and other factors. They also acknowledge that further scientific research is necessary.

Based on this draft guidance, cybersecurity researchers, including doctoral students, must take responsibility for any synthetic data they generate and use in their research, including clearly identifying the data and models used to create the data.

It is noted that the positive impact of synthetic data in cybersecurity research may be greater than the negative effects, particularly in doctoral research settings where obtaining real data is often time-consuming and difficult, particularly given time constraints on dissertation completion. However, the risks of synthetic data must be understood, generation and use of synthetic data must be fully detailed in the dissertation, and the evolving standards and guidelines being developed need to be monitored.

IV. Case Study

Marymount university's Doctor of Science (DSc) in Cybersecurity began in 2018 and is primarily focused on knowledge transfer to and from experienced cybersecurity professionals. These professionals generally come in with one or more master's degrees in the field and a desire to obtain an applied research terminal degree, leading to promotion or a career change on retirement, say into teaching. Many come to the

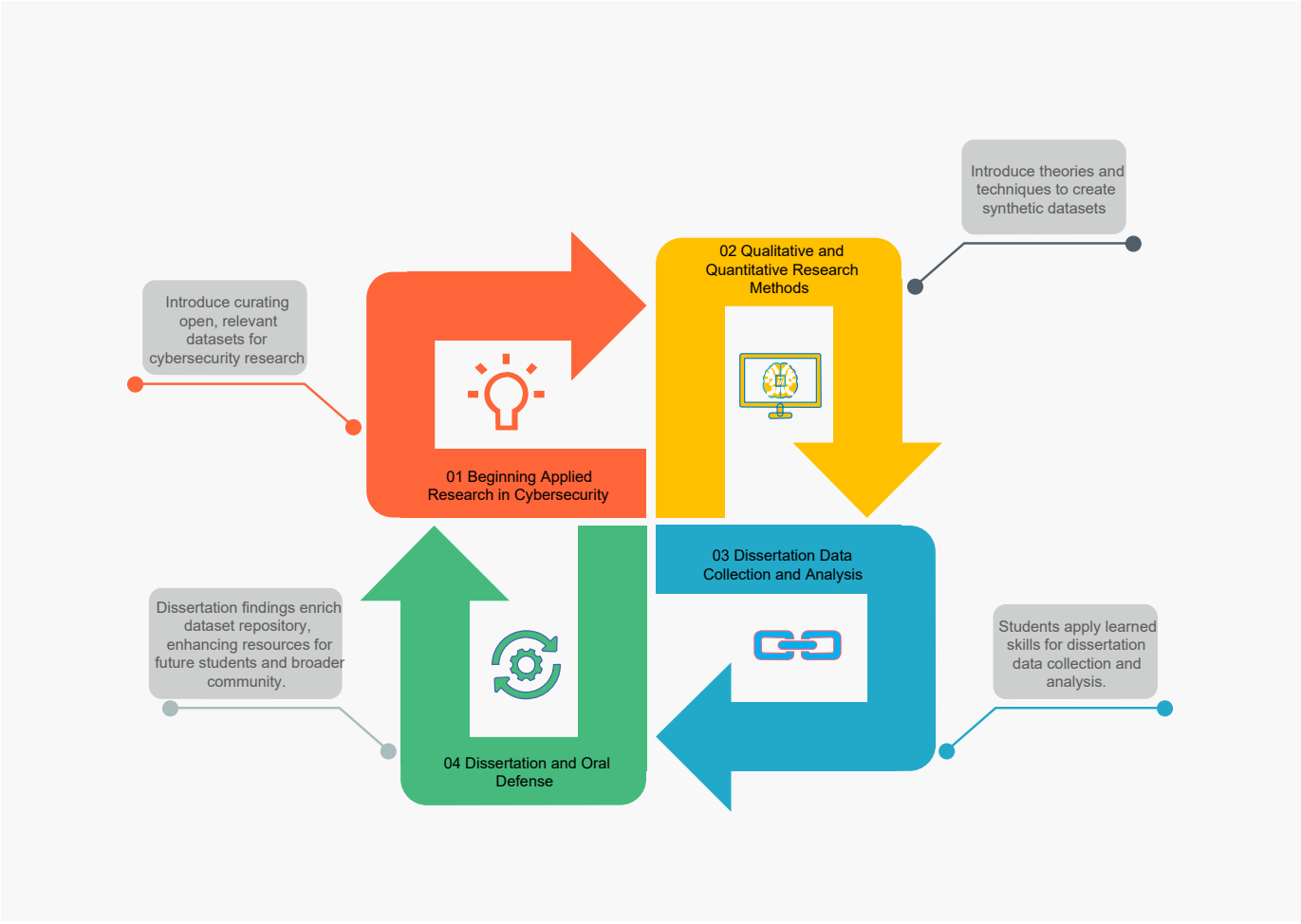

Fig. 1. Proposed data integration in our research course progression

program with their own research topic based on their experiences in the field. In the application process, we talk with the applicants about their research focus and ensure that we have the faculty to provide guidance in their research area as well as our ability to teach them new aspects of cybersecurity and how to ethically conduct academic research.

The post-master's program has two parallel components: new knowledge acquisition and research leading to a dissertation. For knowledge acquisition, doctoral students select six 700-level courses to supplement their existing knowledge base obtained during their master's program, which might now be dated, and the practical knowledge gained from their specific work experience in the military, government agenesis, consulting organizations, and commercial companies. Courses offered include topics such as cyber threat intelligence, machine learning, advanced cybersecurity, global and national cybersecurity policy, and advanced malware analysis, to name a few. The doctoral students chose the areas where they want to gain new knowledge and collaborate in making presentations and preparing publications, learning from each other's experiences. Several datasets of relevance to the topic are used in these courses but are generally provided by the faculty members based on their research.

Doctoral students are not required to take all the 700-level courses before they begin the research process since they have generally acquired the foundational cybersecurity knowledge from their master's programs and through working in the field. Hence, doctoral students may elect to start the applied research component (the so-called 800-level courses) at any time and so can tailor their 700-level courses to support their planned research, for example, by taking a machine learning course. The 800-level courses are designed to scaffold the doctoral students through the academic research process, firstly in a collaborative environment, again learning from each other. Only when the proposal is fully defined and evaluated, do they complete their research individually under the guidance of their committee. In the initial three courses including IT 800 Beginning Applied Research in Cybersecurity, IT 810 Applied Research Topics and Methods in Cybersecurity, and IT 820 Applied Research: Qualitative and Quantitative Research, faculty expose the students to the scientific research process, the importance of establishing the research output of others basis for their own research, ethical research practices including IRB requirements, and a variety of research methodologies.

The availability of datasets to further the research initiatives may come up in any of these courses, and the doctoral students are advised by faculty on potential sources of the data. Several have relied on obtaining data from their organizations, but many of them faced legal obstacles around the release of the data, despite promises to protect the privacy of the individuals and the organization. A few did manage to obtain technical datasets such as logfiles and network incident data from their employers with permission, but these datasets tended to be small and not generalizable.

Those researchers who used open datasets generally used data from other related disciplines, such as criminal justice or financial services. In the synthetic data generation example above, the source of the data was from credit card transactions made by European cardholders each year where the data had been anonymized to protect the cardholders' identities. The primary objective of releasing this dataset through Kaggle was to facilitate the development of fraud detection algorithms to identify potentially fraudulent transactions. In another example, a doctoral student based his research on cyber threats on older adults by using a series of crime datasets, the primary one being from the AARP Fraud Watch System which provided raw victim sentiment data. However, this was not an open dataset and required permission to access and use it.

Web scraping was another method used by doctoral students to generate appropriate datasets. For example, a doctoral student used an initial dataset from Reddit to research the use of social media in influence operations. He then used web scraping to obtain more data on the entities identified in that dataset. Another student built a web scraping tool for the dark web and surface web to assist in deanonymization tasks. Often on these datasets, the students add additional information in the form of annotation that can be other released to researchers in the field.

However, given the lack of discoverability of available open datasets by doctoral students, many researchers settled on doing surveys or interviews to collect data to generate new knowledge. As noted above, the results of these studies provided information on important areas such as user behavior and motivation, but they often lacked technical depth, and because of the constrained data collection were often not generalizable.

Based on our experiences to date, we believe our program needs to address the availability of reliable and open datasets to provide students with more in-depth research options, as discussed below.

V. Discussion

To prepare our doctoral students for the continuously evolving challenges in the cybersecurity research field, we propose integrating curating open, accessible datasets and synthetic data generation methodologies into our D.Sc. program in Cybersecurity as specialized topic modules within various courses. These modules will provide doctoral students with advanced analytical, technical, and ethical competencies, enhancing their cybersecurity research capabilities.

As shown in Figure 1, the integration of topic modules into 800-level courses follows a structured approach. Students begin with applied research in cybersecurity in IT 800 and IT 810, focusing on curating relevant, open datasets for cybersecurity research. They then learn qualitative and quantitative research methods in IT 820, in which we can add the topic modules covering how to use GANs to generate synthetic data. In the dissertation research phase in IT830 The Dissertation Proposal, IT840 Data Collection and Analysis, and IT850 Dissertation and Oral Defense Designs, they apply these skills to collect and analyze data, ensuring their research is rigorous and impactful. Finally, their research findings contribute to a more comprehensive dataset repository, enriching resources and adding valuable insights for future students and the broader research and practitioner community.

For example, for the research course focusing on exploring cybersecurity research topics, we propose introducing a topic module focused on curating current, relevant, and open datasets. This module would involve training doctoral students in identifying, sourcing, and evaluating high-quality open datasets relevant to their research topic in cybersecurity. As part of this module, students would engage in practical exercises that require them to access various cybersecurity data repositories, evaluate the datasets for completeness, accuracy, and relevance, and then curate a collection of datasets that could be used in their individual doctoral research projects. Students would be tasked with creating a curated dataset that they present at the end of the course, which could also contribute to a larger, shared repository accessible to other researchers and students within the university.

For the research seminar teaching quantitative and qualitative research methods, integrating a module on synthetic data generation using GANs would significantly enhance the research methodology training for doctoral students. This module would cover the theoretical foundations of GANs and their applications in creating synthetic datasets that are statistically like real-world data but do not carry the same privacy or security risks. Students would learn how to design GANs to generate datasets that could be used for testing hypotheses in environments where real data may be scarce or sensitive. A course on machine learning is one of the technical courses offered and this can provide the framework for this topic.

Integrating open datasets and synthetic data generation methodologies into our D.Sc. in Cybersecurity program offers both benefits and challenges. Curated open datasets of current data will provide structured, real-world resources that could improve research transparency, reproducibility, and valuable to the identification of emerging cyber threats. However, their curation requires careful attention to quality, legal, and ethical constraints, as well as evolving data management standards. Similarly, synthetic data generation through GANs helps tackle data scarcity and privacy concerns but requires a curriculum that balances theory with practice. These curriculum changes should ensure our doctoral students are better equipped to handle the technical, ethical, and methodological challenges

involved in cybersecurity research today and tomorrow.

VI. Conclusion and Future Directions

The government is working to improve the "accessibility, discoverability, reliability, reproducibility, sustainability, and utility" of data generated with their funding [8]. In addition, NIST is looking for public comments on standards and guidelines for the digital content transparency [29]. However, the authors believe strongly that similar initiatives should be applied to the many datasets identified or generated in academia, whether government-funded or not, to further develop the science of cybersecurity. While privacy and security remain major concerns, synthetic data generation provides a potential solution providing that limitations are identified and documented. In addition, the addition of relevant metadata during a structured curation process can significantly add to the quality and usability of the datasets. The authors will continue to expand their collection and curation of the datasets, through their own activities and the results of the doctoral student course work. They will also study the impact that the discoverability of quality current datasets will have on the types of research conducted by doctoral students, such as an increase in replications studies, and the quality of these outcomes.

In addition to the existing measures, the authors plan to further explore the impacts of the two modules added to the 800-level courses, which all students are required to take. This evaluation will not only assess their research methods, including the increased use of the open-source datasets and the generation of synthetic data, but also initiate a broader range of potential research projects specifically aimed at evaluating the learning outcomes associated with these methods. A survey will be conducted to examine students' selection of datasets and its effect on their research process, particularly focusing on the time and effort required for dissertation completion. This future work will aim to provide a deeper understanding of the educational and practical implications of our curriculum enhancements.

References

[1] L. Bertolin Furstenau, M. Sott, L. Kipper, A. Homrich, T. Cardoso, A. Abri, J. R. López-Robles, and M. Cobo, "20 years of scientific evolution of cyber security: a science mapping," 04 2020.

[2] S. Xu, *Cybersecurity Dynamics: A Foundation for the Science of Cybersecurity*, 05 2019, pp. 1–31.

[3] I. Trenchev, W. Dimitrov, G. Dimitrov, T. Ostrovska, and M. Trencheva, "Mathematical approaches transform cybersecurity from protoscience to science," *Applied Sciences*, vol. 13, no. 11, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/11/6508

[4] F. Cremer, B. Sheehan, M. Fortmann, A. N. Kia, M. Mullins, F. Murphy, and S. Materne, "Cyber risk and cybersecurity: a systematic review of data availability," *The Geneva Papers on Risk and Insurance - Issues and Practice*, vol. 47, no. 3, pp. 698–736, July 2022.

[5] M. Liu, D. Murphy, and N. Green, "The hunt for cybersecurity data: Exploring the availability of open datasets for cybersecurity scientific research," *J. Comput. Sci. Coll.*, vol. 39, no. 3, p. 171–181, dec 2023.

[6] N. I. of Standards and T. (NIST), "The workforce framework for cybersecurity (nice framework)," https://www.nist.gov/itl/applied-cybersecurity/nice/nice-framework-resource-center, 2024.

[7] O. of Science and T. Policy, "Memorandum on ensuring free, immediate, and equitable access to federally funded research," https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-access-Memo.pdf, 2022.

[8] N. S. Foundation, "Cyberinfrastructure for public access and open science (ci paos) program," https://new.nsf.gov/funding/opportunities/cyberinfrastructure-public-access-open-science-ci, 2024.

[9] J. Dykstra, *Essential Cybersecurity Science*. O'Reilly Media, Inc., 2015.

[10] S. Walton, P. Wheeler, Y. Zhang, and X. Zhao, "An integrative review and analysis of cybersecurity research: Current state and future directions," *Journal of Information Systems*, 04 2020.

[11] A. Hall, X. Liu, and D. Murphy, "Advancing cybersecurity through knowledge conversion: Industry-academia interchange in a doctoral program," in *2023 IEEE Frontiers in Education Conference (FIE)*. Los Alamitos, CA, USA: IEEE Computer Society, oct 2023, pp. 1–4. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/FIE58773.2023.10343286

[12] P. K. Mvula, P. Branco, G. V. Jourdan, and H. L. Viktor, "A systematic literature review of cyber-security data repositories and performance assessment metrics for semi-supervised learning," *Discover Data*, vol. 1, no. 1, p. 4, 2023. [Online]. Available: https://doi.org/10.1007/s44248-023-00003-x

[13] F. Bechara and S. Schuch, "Cybersecurity and global regulatory challenges," *Journal of Financial Crime*, vol. ahead-of-print, 11 2020.

[14] K. Logos, R. Brewer, C. Langos, and B. Westlake, "Establishing a framework for the ethical and legal use of web scrapers by cybercrime and cybersecurity researchers: learnings from a systematic review of australian research," *International Journal of Law and Information Technology*, vol. 31, pp. 1–27, 10 2023.

[15] D. B. Rawat, R. Doku, and M. Garuba, "Cybersecurity in big data era: From securing big data to data-driven security," *IEEE Transactions on Services Computing*, vol. 14, no. 6, pp. 2055–2072, 2021.

[16] I. Kouper and S. Stone, "Data sharing and use in cybersecurity research," *Data Science Journal*, vol. 23, p. 3, 01 2024.

[17] T. Moore, E. E. Kenneally, M. Collett, and P. Thapa, "Valuing cybersecurity research datasets," in *18th Workshop on the Economics of Information Security (WEIS)*, June 2019. [Online]. Available: https://ssrn.com/abstract=3469364

[18] H. Kavak, J. J. Padilla, D. Vernon-Bido, S. Y. Diallo, R. Gore, and S. S. Shetty, "Simulation for cybersecurity: state of the art and future directions," *J. Cybersecur.*, vol. 7, 2021.

[19] T. Takko, K. Bhattacharya, M. Lehto *et al.*, "Knowledge mining of unstructured information: application to cyber domain," *Scientific Reports*, vol. 13, p. 1714, 2023. [Online]. Available: https://doi.org/10.1038/s41598-023-28796-6

[20] X. Wu, W. Zheng, X. Xia, and D. Lo, "Data quality matters: A case study on data label correctness for security bug report prediction," *IEEE Trans. Software Eng.*, vol. 48, no. 7, pp. 2541–2556, 2022. [Online]. Available: https://doi.org/10.1109/TSE.2021.3063727

[21] C. Borgman, "The conundrum of sharing research data," *Journal of the American Society for Information Science and Technology*, vol. 63, 06 2011.

[22] M. Zheng, H. Robbins, Z. Chai, P. Thapa, and T. Moore, "Cybersecurity research datasets: Taxonomy and empirical analysis," in *11th USENIX Workshop on Cyber Security Experimentation and Test (CSET 18)*. Baltimore, MD: USENIX Association, Aug. 2018. [Online]. Available: https://www.usenix.org/conference/cset18/presentation/zheng

[23] T. N. Science and T. Council, "Desirable characteristics of data repositories for federally funded research," The National Science and Technology Council, Tech. Rep., 2022. [Online]. Available: https://doi.org/10.5479/10088/113528

[24] N. I. of Standards and T. (NIST), "Unveiling nice framework components v1.0.0," https://www.nist.gov/news-events/news/2024/03/unveiling-nice-framework-components-v100-explore-latest-updates-today, 2024.

[25] OpenAI, "Chatgpt: Optimizing language models for dialogue," OpenAI, 2024. [Online]. Available: https://www.openai.com/research/chatgpt

[26] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *ArXiv*, vol. abs/2204.06125, 2022.

[27] GitHub, "Github copilot: Your ai pair programmer," https://copilot.github.com/, 2021.

[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 3, 06 2014.

[29] National Institute of Standards and Technology (NIST), "Reducing risks posed by synthetic content: An overview of technical approaches to digital content transparency," National Institute of Standards and Technology, Tech. Rep. NIST AI 100-4, April 2024, draft for public comment. [Online]. Available: https://airc.nist.gov/docs/NIST.AI.100-4.SyntheticContent.ipd.pdf