COMMITTEE ON DATA
CODATA
INTERNATIONAL SCIENCE COUNCIL

Data Sharing and Use in Cybersecurity Research

RESEARCH PAPER

]u[ubiquity press

INNA KOUPER

STACY STONE

*Author affiliations can be found in the back matter of this article

ABSTRACT

Data sharing is crucial for strengthening research integrity and outcomes and for addressing complex problems. In cybersecurity research, data sharing can enable the development of new security measures, prediction of malicious attacks, and increased privacy. Understanding the landscape of data sharing and use in cybersecurity research can help to improve both the existing practices of data management and use and the outcomes of cybersecurity research. To this end, this study used methods of qualitative analysis and descriptive statistics to analyze 171 papers published between 2015 and 2019, their authors' characteristics, such as gender and professional title, and datasets' attributes, including their origin and public availability. The study found that more than half of the datasets in the sample (58%) and an even larger percentage of code in the papers (89%) were not publicly available. By offering an updated in-depth perspective on data practices in cybersecurity, including the role of authors, research methods, data sharing, and code availability, this study calls for the improvement of data management in cybersecurity research and for further collaboration in addressing the issues of cyberinfrastructure, policies, and citation and attribution standards in order to advance the quality and availability of data in this field.

CORRESPONDING AUTHOR:
Inna Kouper
Luddy School of Informatics, Computing, and Engineering, Indiana University Bloomington, USA
inkouper@indiana.edu

KEYWORDS:
data sharing; cybersecurity research; open data; cyberinfrastructure

TO CITE THIS ARTICLE:
Kouper, I and Stone, S. 2024. Data Sharing and Use in Cybersecurity Research. *Data Science Journal*, 23: 3, pp. 1–19. DOI: https://doi.org/10.5334/dsj-2024-003

INTRODUCTION

Cybersecurity focuses on safeguarding cyberspace from unauthorized access, malicious damages, and disruptions (Cybersecurity 2009; Kemmerer 2003). Cybersecurity research is an interdisciplinary domain that, in addition to developing safeguarding technologies, explores security and privacy-related events and human-oriented processes (Cavelty 2018; Craigen et al. 2014). Recently, cybersecurity research also expanded its repertoire of data sources to include network and application traces, database and information system activities, and user activities (Sarker et al. 2020). Many government, commercial, and non-profit organizations now collect cybersecurity-related information that can be used for research (Choucri et al. 2018).

Data plays a critical role in cybersecurity research. As threats continue to evolve, becoming more sophisticated and harder to detect, researchers need access to a wider range of data to stay ahead of potential risks and find solutions that work in the ever-changing landscape of data and technologies (Linger et al. 2017; Walton et al. 2021). Mitigating new forms of malware, ransomware, and phishing attacks requires a proactive collaborative approach to cybersecurity that involves prompt sharing of knowledge, including sharing of data and techniques (Sebastian 2022). However, organizations and research groups often operate in isolation when it comes to cybersecurity efforts. The reluctance to share sensitive information, even for the purpose of enhancing security, limits the scope and effectiveness of research (Camp et al. 2009).

Given these factors, an understanding of the existing landscape of data sharing and use in cybersecurity research can help to identify barriers to effective data sharing, contribute to the development of a more robust cybersecurity infrastructure, and encourage a more collaborative approach, thereby enhancing overall digital security. Apart from several studies and reports, cybersecurity data sharing is an unexplored area (Balenson et al. 2020; Sauerwein et al. 2019; Serrano et al. 2014). Considering that research data sharing in other domains has already been shown to be crucial for strengthening research integrity and outcomes (Berman et al. 2014; Douglass et al. 2013; Maxson Jones et al. 2018), this paper aims to fill a gap on this topic in cybersecurity research and stimulate the discussion about broader data sharing and re-use.

BACKGROUND

Cybersecurity research depends on the availability and quality of data (Sarker et al. 2020). If shared, many types of data, including network data, malware samples, website crawling results, social media, and human user event data, could help to advance research (Camp et al. 2009; Sun et al. 2019). And yet, researchers have consistently reported lack of quality datasets, particularly datasets that are dynamic and reflect the changing nature of security-related behaviors (Brown et al. 2009; Shiravi et al. 2012; Sommer & Paxson 2010). Difficulties obtaining operational security data also slow down newer forms of research that rely on data science and machine learning (Li & Oprea 2016).

Recent calls for cybersecurity data to become more available to broader academic audiences received a slow response. Common barriers include lack of incentives, data sensitivity, and fear of getting scooped (Nelson 2009; Tenopir et al. 2011). Additionally, cybersecurity research lacks consistent frameworks that can help consolidate the domain's views on what to share and how to maintain adequate levels of quality (Baker & Millerand 2012; Serrano et al. 2014). Privacy, security, proprietary restrictions, and legal concerns make security data gathering and dissemination challenging for all stakeholders, including infrastructure and data owners as well as data collectors, producers, and distributors (Balenson et al. 2015; Mathew & Cheshire 2018; Serrano et al 2014). However, the domain is engaged in ongoing discussions about the appropriate models for data collection, storage, and sharing (Atapour-Abarghouei et al. 2020; Fisk et al. 2015; Scheper et al. 2011; Shou 2012).

One barrier is that preparing data for sharing is costly and time-consuming, even though cybersecurity papers that share data were shown to receive more citations (Zheng et al. 2018). To alleviate the cost of sharing for individual researchers, shared environments have been established with support from federal, academic, or commercial organizations, which resulted in the creation of several valuable public resources of cybersecurity data (Blackfire Technology, Inc. 2019; InfraGuard 2018; MIT Lincoln Laboratory 2016; San Diego Supercomputer Center 2020; US Department of Homeland Security 2022). Cybersecurity exercises and competitions publicized via websites and academic papers also have generated several public data sources (Abbott et al. 2015; Brynielsson et al. 2016; Munaiah et al. 2019; Shou 2012; Sommestad & Hallberg 2012).

Despite their acknowledged value, those data sharing environments and websites face difficulties in long-term maintenance and often provide limited functionality or cease to exist after a short period of operation (see, for example, Dumitraş 2018; Dumitraş & Shou 2011; ISCX 2007; University of California Irvine 1999). Some data sources have been criticized for their high levels of anonymization and dubious authenticity (Maciá-Fernández et al. 2018; Mahoney & Chan 2003; McHugh 2000; Sun et al. 2019). As larger data-sharing platforms in cybersecurity remain a desirable goal of the future, researchers are left to navigate the current fragmented landscape of publicly available data, collecting or synthesizing their own data and metadata and making ad-hoc decisions about how to share them (Brown et al. 2009; Fontugne et al. 2010; Moustafa & Slay 2015; Sperotto et al. 2009; Tavallaee et al. 2009).

This brief overview shows that researchers in cybersecurity rely on a limited range of existing data sources, and they tend to not share their own data. In 2018, Zheng et al. examined 965 cybersecurity papers between 2012 and 2016 to understand the patterns of data production, sharing, and use. The authors found that papers that used data were split between using the existing data and creating new data, and over the years only 15–19% of the created datasets were made public, although the trend was rising, closer to 30% in 2016. Another exploratory study of public cybersecurity data sources found that those sources had a strong focus on vulnerability and a low degree of standardization (Sauerwein et al. 2019).

The present study contributes to the discussions about models of data sharing and use in cybersecurity research. It complements Zheng et al.'s (2018) study by providing an updated view on the data landscape in cybersecurity research. Additionally, this study examined a broader set of questions and conducted a more detailed analysis of the patterns of sharing, including the authors, research methods, data, and code. These findings help make the case for more nuanced approaches to open data sharing and to build better support for diverse forms of collective sharing of research objects, including data and code.

METHODS

The study aimed to examine the nature, use, availability, and modes of sharing of cybersecurity data for research. It draws on the concept of research objects and incorporates code, or analytic techniques, in its examination of sharing resources in support of research objectives and claims (Bechhofer et al. 2010). It addressed the following research questions:

- (RQ1) Who contributes to cybersecurity research and its sharing?
- (RQ2) What methods do researchers use and how are those methods related to data availability?
- (RQ3) What is the availability of cybersecurity data and software tools?

To address these questions, papers published between January 2015 and September 2019 were collected using two search strategies: a localized and an expanded search. For the localized search, we reviewed websites of eight highly ranked universities, focusing on the US Midwest and Western regions, and identified researchers who described themselves as working in cybersecurity. Using Google Scholar, Web of Science, the ACM Digital Library, and the IEEE Digital Library, we compiled a list of publications authored by those researchers. Seventy-seven papers were collected using this approach.

For the expanded search, we reviewed proceedings of four national cybersecurity conferences: IEEE Symposium on Security and Privacy, Computer and Communications Security (CCS), USENIX Security Symposium, and Networked and Distributed Security Symposium (NDSS). Papers that used data and focused on cybersecurity of computers and networks were included in the sample. The combined data from both searches was reviewed for duplicates and empirical focus, that is, the publications had to use data and report research based on observations or experimentation. The final dataset included 171 publications (see Table 1).

PUBLICATION VENUE	COUNT	PERCENT
ACM Computer and Communications Security Conference (CCS)	40	23%
USENIX Security Symposium	40	23%
IEEE Symposium on Security and Privacy (SP)	21	12%
Network and Distributed System Security Symposium (NDSS)	17	10%
ArXiv	8	5%
Other (journals, conferences)	45	26%
Total	**171**	**100%**

Table 1 Venues of Sampled Publications.

The analysis involved close reading of the papers and subsequent coding of text segments. Upon detailed examination of each publication and its metadata, we extracted relevant information into a spreadsheet, including publication title, year, and venue. We also extracted information about authors and datasets. For authors, we examined the information available in the papers and performed Internet searches to record their names, positions, gender, institution, and research focus. For datasets, we recorded the dataset name, its origin, and availability of both data and analytical tools. Any tools mentioned in the publications were recorded in the spreadsheet for further aggregation and analysis (see the 'Results' section below). If there was a URL for either dataset or analysis software, it was included in the coding spreadsheet. To avoid duplication of datasets, we gave the datasets consistent names, descriptions, and URL links (when available).

Additional characterizations of how the datasets were used in each publication were documented in a separate column labeled 'Methods of analysis.' The codes for methods of analysis in the publications emerged bottom-up as the authors read the papers and recorded types of analysis performed in the papers with free 2–3-word labels. The coding was later aggregated and standardized into several categories (see codebook in Appendix A). To ensure consistency in coding, we examined the first ten papers together and discussed coding and interpretations. After reaching an agreement on clear interpretations of each code, the rest of the coding was split into equal shares with both authors reviewing the final analyses and discussing any questions and potential disagreements.

Datasets were also coded using two taxonomies developed in previous studies that described cybersecurity data sources (Sauerwein et al. 2019; Zheng et al. 2018). A simplified version of both taxonomies was used in the coding; namely, we took the main categories and did not use any additional facets and subcategories. Our coding was guided by the descriptions provided in the original papers. In case of disagreements, we aimed for internal consistency within our own study rather than consistency across our study and the studies by Sauerwein et al. and Zheng et al. because our data set differed from theirs. From Zheng et al. (2018) the following categories were used: 1) attacker-related, defined as any data that is already deemed malicious or is used by attackers, including scams, malware, and vulnerabilities, 2) defender artifacts, such as firewalls or secure configurations, 3) user and organization characteristics, defined as information about users and organizations online behavior, and 4) internet characteristics, defined as network characteristics, including applications, traffic and traces, and various adverse events.

From Sauerwein et al. (2019) we used the following main categories: 1) vulnerability, defined as weaknesses that might be exploited by a threat, 2) threat, defined as potential causes of unwanted incidents, 3) countermeasure, defined as any administrative, managerial, technical or legal control that is used to counteract an information security risk, 4) attack, defined as information regarding any unauthorized attempt to access, alter or destroy an asset, 5) risk, defined as the consequences of a potential event, such as an attack, and 6) asset, defined as any object or characteristic that has value to an organization. The codebook used in this study is provided in Appendix A.

RESULTS

PUBLICATION AUTHORS AND METHODS

Overall, 823 individuals contributed to cybersecurity research in our sample between 2015 and 2019. The number of authors per paper ranged between 2 and 12, with the average of about five authors per publication (see Table 2).

Table 2 Number of Authors in Publications.

TOTAL NUMBER OF AUTHORS	NUMBER OF PAPERS	PERCENT
2	20	12%
3	39	23%
4	33	19%
5	21	12%
6	22	13%
7	13	8%
8	12	7%
9	4	2%
10	4	2%
11	2	1%
12	1	1%
Mean # authors per paper	**4.81**	
Standard deviation	**2.21**	

To better understand the range and nature of authors' contributions, we coded and analyzed the professional profiles of the publications' first authors. The majority of published research came from academic institutions in the US, but there were also commercial and government organizations such as Microsoft Research, Lawrence Livermore National Laboratory, and Symantec Research Labs. The position titles of first authors at the time of publication included mostly traditional academic positions and a few research- and practice-oriented positions, including engineers, computer scientists, and software developers. The majority of the first authors (75%) were doctoral students (see Table 3).

Table 3 Positions of First Authors in Publications.

FIRST AUTHOR POSITION	NUMBER OF PAPERS	PERCENT
Graduate student (PhD)	128	75%
Faculty	18	11%
Postdoctoral researcher	8	5%
Graduate student (MS)	8	5%
Other	9	5%
Total	**171**	**100%**

Cybersecurity research remains a male-dominated field, at least in terms of primary authorship recognition. Out of 171 first authors, only 24 of them (14%) were females. The relative proportion of female students was slightly smaller than the proportion of females in faculty, but these numbers are difficult to evaluate due to a very small number of faculty first authors (Table 4). Three faculty women who were first authors in our dataset were full professors. Male faculty first authors were in various ranks, including assistant, associate, full, and research professors. The positions of male first authors were also more diverse as they included research scientists, undergraduate students, and engineers (category 'Other' in Table 4).

FIRST AUTHOR POSITION GENDER	FEMALE	MALE
Graduate student (MS or PhD)	18 (13%)	118 (87%)
Faculty	3 (17%)	15 (83%)
Postdoc	3 (37%)	5 (63%)
Other	0	9 (100%)

Table 4 First Author Positions by Gender.

Next, we examined methodologies and analytical approaches that were used in publications. In describing the types of analyses, we focused on whether the authors developed their own system (prototype and evaluation) or an algorithm (algorithm development and testing); used machine learning in a specific domain (machine learning application); examined vulnerabilities in a system (vulnerability analysis); or emphasized conceptual development (conceptual model) or statistical analysis. When there was an overlap between methodologies, the paper would be categorized first based on the primary goal and then, a secondary (and if necessary, a tertiary) category would be assigned to the paper. For example, if authors developed a prototype that included a novel algorithm to identify cyberattacks, the paper would be first categorized as 'Prototype and evaluation' and then as 'Algorithm development and testing.' The results for primary types of analysis are presented in Table 5 below.

TYPE OF ANALYSIS	FREQUENCY	PERCENT
Prototype and evaluation	81	47%
Algorithm development and testing	42	25%
Vulnerability analysis	13	8%
Conceptual model	12	7%
Machine learning application	8	5%
Statistical analysis	8	5%
Other	7	4%
Total	**171**	**100%**

Table 5 Primary Types of Analysis in Publications.

Almost half of the papers (47%) used prototyping and evaluation as their main type of analysis. The second largest category was algorithm development and testing (25%), with the remaining categories representing less than 10% of total papers. The methods that were included in the 'Other' category were network scanning and surveys of the domain or stakeholders.

About one-third of the publications (58 or 34%) used more than one type of analysis. Thus, more than a quarter of publications (29 out of 81) that used prototype development and evaluation, also used algorithm development as their methodology. Prototype development was also used in conjunction with vulnerability analysis, machine learning applications, and network scanning. Prototypes and algorithms were developed for a wide range of uses and applications, including phishing and malware detection, network and data monitoring, privacy protection and data anonymization, and threat intelligence.

Publications relied on a large variety of computing tools, including operating systems, major programming languages, data science tools, benchmarking platforms, and data sources and platforms, such as VirusTotal. Linux was among the most popular operating systems (mentioned 45 times), followed by Windows (mentioned 19 times). Mac OS was mentioned only three times along with other rare operating systems such as Graphene and Redox/Rust.

The large variety of tools and software mentioned in the publications made the creation of standard categories difficult. Overall, we counted over 450 technological tools and their variations mentioned in the publications. Python programming language and its various packages were mentioned about 60 times. WEKA, a free collection of machine learning algorithms, and some other machine learning packages were also used in algorithm development and ML applications. Less common languages and scripting tools included C/ C++, Java, R, and shell scripting. Virtualization and cloud computing tools included Qemu/KVM, AWS/Amazon, VM, Docker, and VirtualBox, with AWS/Amazon being the most common one (mentioned in at least eight publications). Almost one-fifth of the publications (33) did not mention any software or technologies. Some of them focused on mathematical proof and conceptual analysis, while others engaged in data analysis, prototype evaluation, or algorithm development without providing specifics about which technologies they used.

DATA AND CODE

Origin and Availability

We identified 438 datasets in our sample. Eight publications used no datasets as they relied on mathematical proof and software development, or their data sources could not be identified. Some publications used the same datasets or sampled from the same sources with varying characteristics (e.g., different date ranges or selected variables). Twenty-eight datasets were used more than once across all publications. After those duplications were identified and removed from the sample, the resulting set consisted of 387 unique datasets total. This deduplicated sample was used for subsequent analysis. The number of datasets per paper ranged between one and 12 (see Table 6).

Table 6 Number of Datasets in Each Paper ($N_{datasets} = 387$).

DATASETS IN EACH PAPER	NUMBER OF PAPERS	PERCENT
1	61	16%
2	54	14%
3	93	24%
4	36	9%
5	55	14%
6 or more	88	23%
Mean	**2.6**	

Most of the publications used 1–3 datasets, with an average use of 2.6 datasets per paper. However, several publications relied on larger data gathering efforts. For example, one paper used data from nine sources, eight of which were the existing sources with varying levels of public availability. One paper used 12 datasets. This paper developed a model of an offline password cracker; it tested the model on the data from recent massive password breaches of such companies as Yahoo!, Dropbox, LastPass, dating service site Ashley Madison, and others. Three of these datasets (000webhost, Ashley Madison, and Yahoo! passwords) are publicly available, while all others are not available.

For each dataset, we coded its *origin*, that is, whether the data was drawn from the existing sources, collected, simulated, or synthesized; and its *availability*, that is, whether it was made publicly available or not. Data was considered simulated when it was collected from an experimental setup or a simulated environment, while synthetic data was the data generated to reproduce certain characteristics of the existing real-world data. Additionally, we coded the public availability of processing software (analytics).

In terms of the data origin, more than half of the datasets (55%) used in the publications were *existing* datasets, that is, datasets that were previously collected by others (see Table 7). The second largest group (27%) was data collected by publication authors themselves.

Table 7 Origin of the Datasets Used in Cybersecurity Research.

DATA ORIGIN	NUMBER OF DATASETS	PERCENT
Existing	211	55%
Collected	105	27%
Simulated	17	4%
Synthetic	10	3%
Other	44	11%
Total	**387**	**100%**

The category 'Other' was applied to datasets that were compiled from multiple sources or when there was not enough detail to determine the contents and origin of the dataset. Below is an example of how compilations of multiple datasets were described:

> We use a set of reputation blacklists to measure the level of malicious activities in a network. This set further breaks down into three types: (1) those capturing spam activities, … (2) those capturing phishing and malware activities, … and (3) those capturing scanning activities, including the Darknet scanners list …

Figure 1 below illustrates public availability of datasets used in cybersecurity research publications.

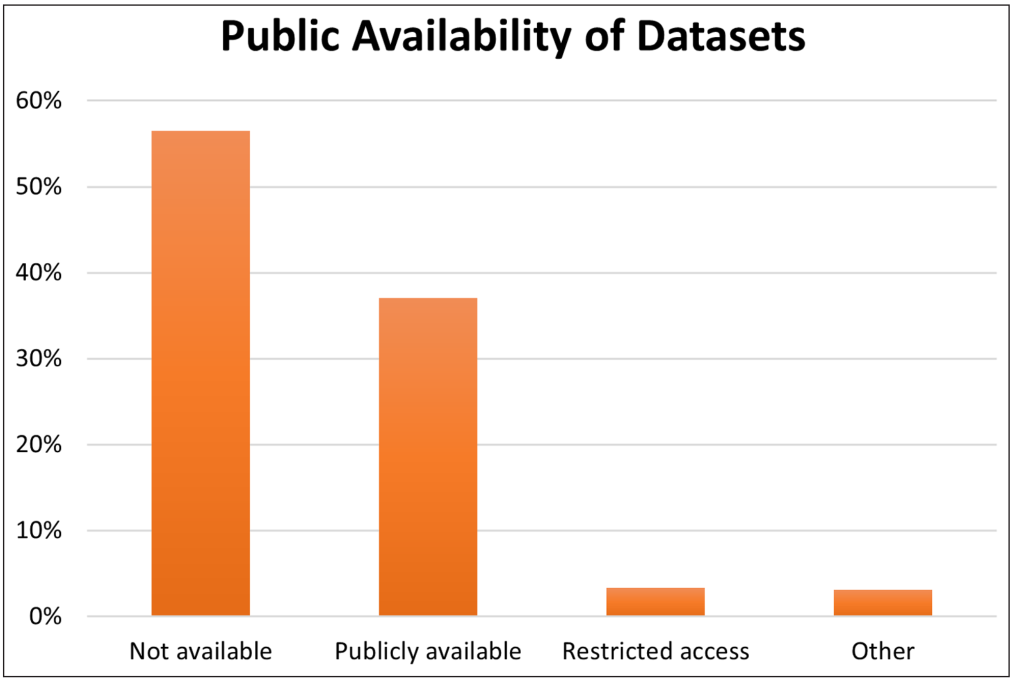

Figure 1 Public Availability of Datasets in The Sample.

As evidenced in the figure above, a larger majority of the datasets (58%) were not publicly available. At the same time, a good share of the datasets (37%) was publicly available. Another two small categories (3% each) were restricted availability or other sharing arrangements, such as partial availability, availability upon request, and sampling or assemblages from multiple existing datasets that were not clearly defined or were not reproducible with the details available in the paper.

Analytical tools used to process and analyze datasets had a different availability pattern (see Figure 2). Only a small fraction of code and analytical tools was made publicly available (11%), with two more items partially available or available upon request. The rest (89%) was not publicly available.

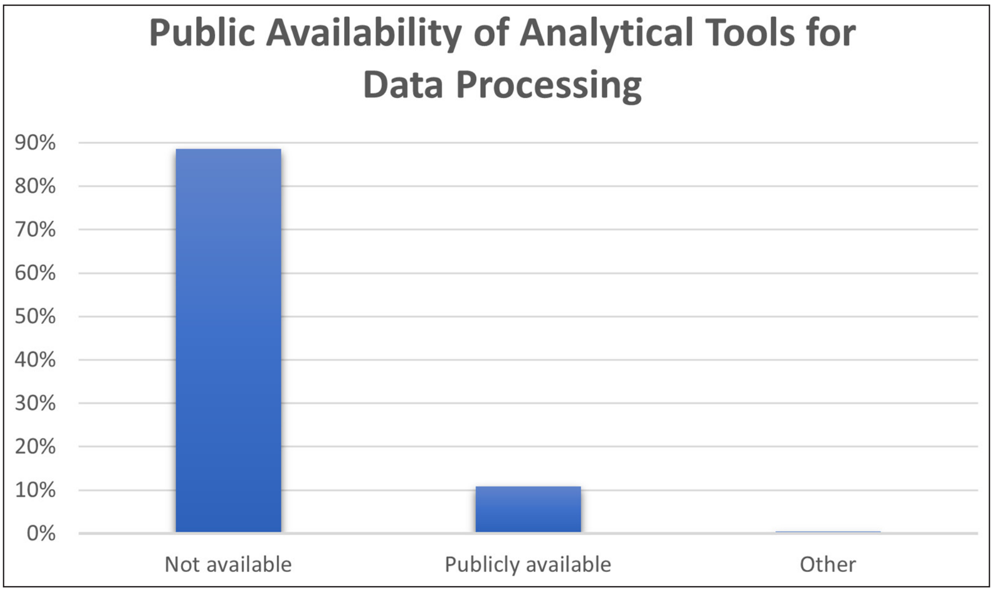

Figure 2 Public Availability of Computing and Analytical Tools for Data Processing.

All publicly available code except for one publication used GitHub for sharing. Some papers used one repository to share code for processing more than one dataset. There were only 25 instances where both the dataset and the code were made publicly available, those instances came from 14 publications (8% of the sample). Three more papers had data available upon request or partially available, and one other paper used data from a restricted source that is currently described as a 'past project' on the website with no means of accessing the data.[1]

Upon further examination, the notion of 'public availability of data' turned out to be complicated. When coding for public availability, we considered datasets publicly available when some information was provided in the publication to assist others in locating the datasets. However, when we tried to find the data using the provided sources, the ease of discovery varied significantly among publicly available datasets. To understand this variability better, we coded for the types of availability (see Table 8 below).

Table 8 Types of Data Availability in Publications.

TYPE OF DATA AVAILABILITY	NUMBER OF DATASETS	PERCENT
URL to a repository	47	28%
Citation, no URL	40	24%
No URL or citation	40	24%
URL to dataset	21	13%
Broken link	16	10%
DOI	1	1%
Total	**165**	**100%**

Out of 165 datasets with public or other availability, 47 datasets (28%) provided a URL to a repository rather than to the dataset itself. Additional browsing or searching was needed to identify the specific data mentioned in the publications. About one-fourth of available data provided a citation, but no URL. For example, one paper cited the source of their data in the reference section as follows: 'A. Asuncion and D. Newman. UCI machine learning repository, 2007.' While this repository can be easily found via Internet search, it contains hundreds of datasets that were deposited at various times.

Almost one-fourth of the available data (24%) provided neither URLs nor citations to their data. Some of those publications used software as an input, so they simply listed that software in the text of the paper. Others referred to other publications that presumably had information about the datasets or described their data in a non-specific way, for example, 'We use a collection of nine live blacklist feeds, summarized in Table I, to label relevant entries….' Thirteen percent of available datasets provided a URL to the dataset, and an additional 10% resulted in a broken link at the time of our analysis. Only one dataset had a persistent identifier (DOI).

One particular publication clearly illustrates the complexities of data sharing and use. We identified seven datasets in that publication. One dataset came from the Center for Applied Internet Data Analysis (CAIDA) repository. It had a direct URL to the dataset, but the data was restricted as access to it requires registration and approval. The remaining six datasets came from various sources and used several tools to collect or acquire network traffic data, including a campus network, browser extensions, external websites, and so on. None of those six datasets or their combination (a merged dataset was also reported in the paper) were available for re-use, and only one of them had a link to a website that resulted in an error at the time of our analysis.

Another publication used exploit kit samples (an existing dataset) and provided a link to a repository that contained those kits. However, the dataset was classified as unavailable in our analysis because it was not possible to determine which kit samples were used in the paper given the information provided. The same paper mentioned other datasets, which could be found via internet search; however, the paper itself did not provide a citation or a URL.

[1] https://www.feinberg.northwestern.edu/sites/chip/our-projects/past-projects/healthlnk.html.

Some publications included URLs to GitHub repositories, but a significant effort was needed to find the data that was used for analysis. URLs in several other publications turned out to be broken links at the time of our analysis. Several publications used the Alexa Top Sites service, which was retired on May 1, 2022. Until that point, the repository had been available as a subscription service; to access the data, users would need to pay for subscription and then reconstruct the dataset with parameters described in the publication. Because the service was constantly updating the data, the availability of historical data was unclear.

Existing and Collected Data

As noted above, many publications in our sample relied on the datasets previously collected by others, that is, on the existing data. Such data included network traces, files or software excerpts, various statistics, and text information collected from various websites, forums, and newsgroups. Very few existing datasets used simulated datasets, such as simulated network traffic data: 'The ISCX-IDS-2012 dataset was gathered by simulating real normal network traffic along with multi-staged attacks in a testbed environment.'

Out of the 211 existing datasets used in the publications, slightly more than half (118 or 56%) were determined to be publicly available. The rest of them were either not available (36%), were restricted (6%) or were collected from multiple sources (2%, see all percentages in Table 9).

Table 9 Availability of the Previously Existing Datasets.

EXISTING DATASETS	DATASETS IN PUBLICATIONS	PERCENT
Public	119	56%
Not available	75	36%
Restricted access	13	6%
Other	4	2%
Total	211	100%

Restrictions could include registration, payment, or both. For example, the Alexa Top Sites web service mentioned above used to provide lists of web sites ranked by traffic. To access data through this service, one had to create an account and pay $0.0025 per URL returned.[2] CAIDA at the San Diego Supercomputer Center at University of California San Diego required a data use agreement and full registration with details about the user(s), their affiliation, and their project.

The existing datasets that were not available publicly or had restrictions included data collected from commercial organizations, such as Uber, Cisco, Symantec, and others. Some papers used breached or leaked data found on the dark web and did not share or link to the data to avoid wider publicity. For example, one paper developed and tested a password similarity model using billions of username-password pairs that were compiled from major data breaches around 2017 and shared on the dark web. Malware sample datasets were also not available. Another example of the use of existing datasets not available for re-use included intrusion alerts collected from the 2017 National Collegiate Penetration Testing Competition (CPTC), where teams worked to identify and capture vulnerabilities of the same infrastructure using Suricata software. This dataset is an example of a considerable data collection and curation effort that the researchers undertook to test their models. The dataset was coded as 'existing' even though no information was provided regarding who collected the data as illustrated by the quote below:

> This paper demonstrates ASSERT's capability using the intrusion alerts collected from the 2017 National Collegiate Penetration Testing Competition (CPTC) …, where approximately 60 people from 10 teams attempting to penetrate into the same computing infrastructure to find as many vulnerabilities as possible. Suricata was installed to capture malicious activities over approximately a 9-h period, and the Suricata alerts were used as inputs for the experiments shown in this paper ….

However, in additional search for the sources of this dataset, we found another paper published by the co-authors of this publication that described a related dataset from CPTC from another year and provided a link to datasets from competitions in 2018 and 2019 (Munaiah et al. 2019).

[2] https://aws.amazon.com/alexa-top-sites/faqs/.

This second paper illustrates that the effort to collect and organize the data was considered substantial enough to merit a separate publication. At the same time, lack of standards in describing and publishing the data creates situations where important details can be missing, as it is still not clear from either publication whether the dataset from year 2017 have been made available.

To better understand the nature of the existing datasets used in publications, we used data classifications from Zheng et al. (2018) and Sauerwein et al. (2019) described above. According to Zheng et al.'s classification, the existing datasets were split across three categories: user and organizational characteristics, attacker-related data, and Internet characteristics (see Table 10).

Table 10 Nature of the Existing Cybersecurity Datasets per Zheng et al. (2018) Classification.

CATEGORY	EXAMPLES	NUMBER OF DATASETS	PERCENT
User and organization characteristics	Patient or financial records, social media, reviews	57	27%
Attacker-related	Malware, vulnerability data, security certificates	49	23%
Internet characteristics	Network traces, IP packets, access logs	49	23%
Defender artifacts	Security alerts, non-leaked password databases	13	6%
Other	Images, citation data, web pages	43	20%
Total		**211**	**100%**

One-fifth of our datasets could not be categorized within this classification: particularly, the datasets that were used for algorithm development and testing or machine learning applications, such as samples from the ImageNet database that contains image data[3] and the MNIST database containing images of handwritten digits.[4] These machine learning datasets are not derived from or related to user data, attacker footprint, or Internet traffic, but they are important in developing or optimizing methods that protect machine learning approaches from unintended consequences and uses. Below is an example of how the use of images in cybersecurity research was justified:

> Deep learning algorithms have shown exceptionally good performance in speech recognition, natural language processing, and image classification. However, there is growing concern about the robustness of the deep neural networks (DNN) against adversarial attacks. … For image classifiers, it has been shown that adding small perturbations to the original input image (known as 'adversarial examples') can force an image classifier to make mistakes, which can yield practical risks.

In comparison to Zheng et al. (2018), the Sauerwein et al. (2019) classification method was even harder to apply. Its categories were action-oriented (e.g., attack versus countermeasure) and many datasets could have been described as related to those actions, but not necessarily representing the actions themselves. Therefore, a large majority of the existing datasets (71%) were coded as 'Asset,' that is, any object or characteristic that has value to an organization. The second largest category (15% of the datasets) was 'Threat,' that is, a potential cause of unwanted incidents (Table 11).

Table 11 Nature of the Existing Cybersecurity Datasets per Sauerwein et al. (2019) Classification.

CATEGORY	EXAMPLES	NUMBER OF DATASETS	PERCENT
Asset	Whitelists, network traffic, emails, images	150	71%
Threat	Security alerts, data breaches	32	15%
Countermeasure	Spam samples, VirusTotal samples, security certificates	13	6%
Attack	DDoS attack data	7	4%
Vulnerability	Vulnerability data	8	4%
Risk	Market transactions	1	
Total		**211**	**100%**

[3] https://image-net.org/.

[4] http://yann.lecun.com/exdb/mnist/.

Since these classifications were developed to address research questions that were different from ours, they were difficult to apply consistently and therefore, must be interpreted with caution. Nevertheless, understanding the nature of the datasets is important for further promotion of their sharing and for building necessary infrastructure to support the practice of sharing. As our findings show, data in cybersecurity research varies. Therefore, the data management and sharing infrastructure will need to support this variety. In the future, it may be beneficial to develop a more robust and expandable data classification approach that covers a larger variety of data used in cybersecurity research and has guidelines for consistent application.

Among the datasets we analyzed, a majority (55%) were pre-existing. Datasets that were collected by the authors of the publications comprised 27% of the overall number of datasets. Most of these datasets (87%) were not available either publicly or by request (see Table 12).

Table 12 Availability of the Collected Datasets.

CATEGORY	NUMBER OF DATASETS	PERCENT
Not available	91	87%
Public	12	12%
Available upon request	2	1%
Total	**104**	**100%**

Using Zheng et al.’s taxonomy, the majority of the collected datasets could be described as ‘Internet characteristics’ (54%). The researchers collected logs from various computer configurations, network traffic data, software samples and baseline data. Another large category of collected data was attacker-related data (31%). Only four out of 32 datasets in this category were publicly available. They included data on honeypot attacks and specific types of vulnerabilities. The dataset that was available upon request included 30 exploit kits. The rest of the datasets included honeypot data, various emails collections, malware samples, and vulnerability scanning results. None of those datasets were publicly available.

DISCUSSION

This study provides an in-depth look into the practices of data sharing and use in cybersecurity research in 2015–2019. It contributes to a larger body of literature that calls for broader sharing in cybersecurity, including the sharing of vulnerability information, threat intelligence, incidents reports, research findings, and best practices (Pala & Zhuang 2019). Focusing on the sharing of research data as a narrower aspect of information sharing, our study reveals several gaps in the data practices, and points to the need of creating a more robust data access and sharing ecosystem in cybersecurity research. Figure 3 below provides a synthesis of the main themes of this study and ties them to a broader set of factors that help to promote public access to data (Arzberger et al. 2006; Chawinga & Zinn 2019).

Our study points to tensions or contradictions that are depicted in the shades of green and orange in the figure above. On one hand, cybersecurity research is innovative and collaborative, and it prioritizes team effort and student work; it also uses multiple tools and data sources, while engaging in active technological experimentation and data re-use. On the other hand, it lacks gender diversity, performs most of the experimentation with free open-source technology—which in some ways limits applicability of its solutions—and does not share much of its tools or data.

Gender disparities in cybersecurity follow persistent global patterns (Larivière et al. 2013; Peacock & Irons 2017; Ross et al. 2022). Being underrepresented in first authorship, females are less likely to lead studies and contribute to the advanced research design practices. They are also less likely to serve as role models and encourage other women to go into cybersecurity research, thereby limiting the diversity of perspective in the field, its data, and its code practices. Limited gender diversity in cybersecurity can negatively impact data practices as a more homogenous group is likely to approach solutions based on a limited set of experiences, potentially overlooking sources of data, user insights, and innovation. This can leave systems vulnerable to unforeseen threats or provide solutions for a limited set of cyber environments (Azhar et al. 2019; Tuma & Van Der Lee 2022).

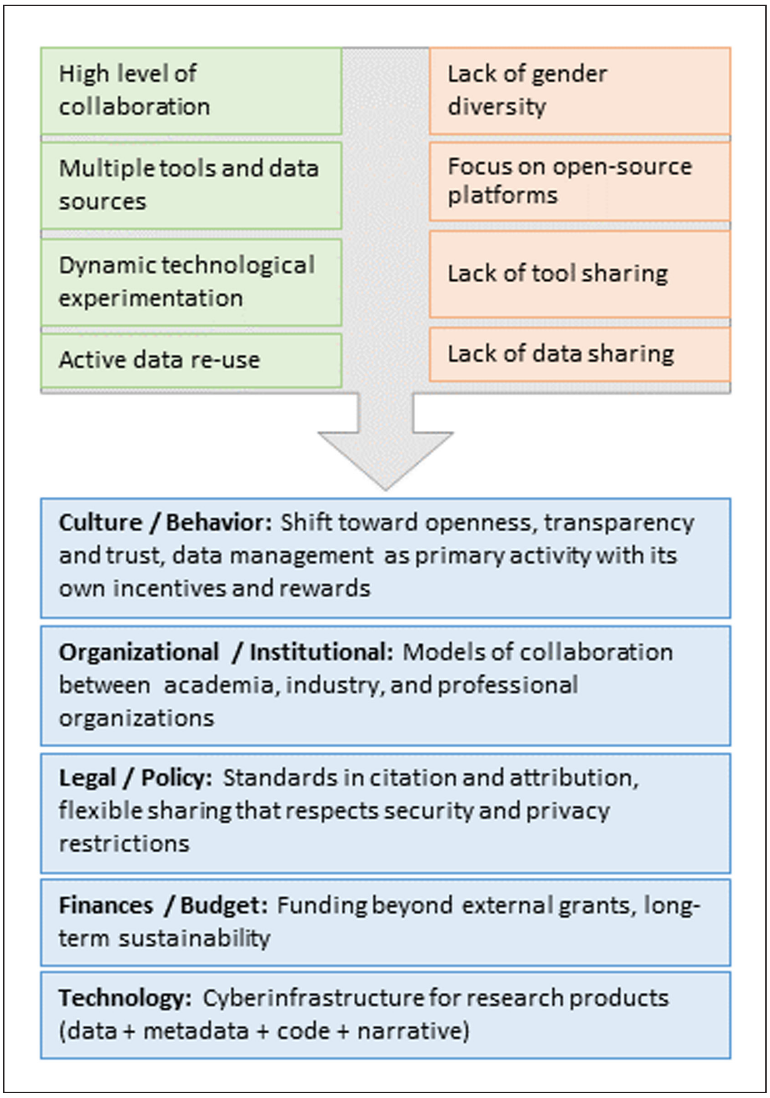
High level of collaboration
Multiple tools and data sources
Dynamic technological experimentation
Active data re-use
Lack of gender diversity
Focus on open-source platforms
Lack of tool sharing
Lack of data sharing
Culture / Behavior: Shift toward openness, transparency and trust, data management as primary activity with its own incentives and rewards
Organizational / Institutional: Models of collaboration between academia, industry, and professional organizations
Legal / Policy: Standards in citation and attribution, flexible sharing that respects security and privacy restrictions
Finances / Budget: Funding beyond external grants, long-term sustainability
Technology: Cyberinfrastructure for research products (data + metadata + code + narrative)

Figure 3 Data and Cybersecurity Research, from Present to Future.

Cybersecurity research relies on a variety of methods, with prototyping and algorithm development being the primary methods in our sample, pointing to dynamic technological experimentation as one of the features of this field. As a necessary condition and a consequence of extensive experimentation, cybersecurity uses a large variety of advanced technologies and data sources. At the same time, it appears that software licensing fees may be a barrier in cybersecurity research as most papers favored Linux and Python as their tools of choice. On one hand, open-source free software is beneficial as it promotes a wider use and availability of tools. On the other hand, proprietary environments, such as Windows or Mac OS environments, could also benefit from cutting-edge prototyping and evaluation as well as from algorithm development, but it is not clear how much of that is part of the ongoing cybersecurity research.

More than half of the papers in our sample relied on the existing datasets; researchers often used more than one dataset in their studies. Despite active re-use of data, there was a notable lack of tool and data sharing. Coupled with concerns over lack of data sharing mentioned in the background section, this study reaffirms the need for robust, standardized, and quality-controlled data sharing frameworks in cybersecurity research.

While there are many valid concerns in sharing cybersecurity data, including possible harm from sharing dark web data or malicious vulnerabilities data, wider availability with appropriate precautions will benefit the field (Plale et al. 2019). Moreover, arguments of harm or negative impact of data sharing should be confirmed with evidence from practice; otherwise, proprietary practices will engender obscurity rather knowledge and stall the advancement of the field (Atapour-Abarghouei et al. 2020). Testing prototypes and algorithms in real-world scenarios that use large amounts of data can provide more robust, applicable results. The availability of data for testing, evaluation, and development increases its value in cybersecurity research and saves expense for those who re-use the existing data (Moore et al. 2019). Finding nuanced solutions to the challenges of sharing data in cybersecurity research requires addressing a combination of factors, including cultural/behavioral factors, organizational/institutional factors, legal/policy factors as well as financial and technological factors (see Figure 3). These factors are briefly discussed below.

Cultural factors appear to be the largest, most challenging to address, as has been pointed out in the data sharing literature (Gormley & Gormley 2012; Poirier & Costelloe-Kuhn 2019). Addressing culture involves fostering a shift toward openness, transparency, and trust; challenging existing stereotypes; modifying education and training practices; as well as establishing incentives for both data collection and management activities. Diversity efforts could benefit from more general strategies of increasing equity and inclusivity—such as better work environments, inclusive job advertisements, and work-life balance (Su et al. 2015)—and from strategies tailored to this field, such as addressing the 'hacker' and 'protector' stereotypes, acknowledging women's contributions, and inviting broader expertise to cybersecurity (Shumba et al. 2013; Poster 2018). Considering the high participation of graduate students in research, changes in their training toward prioritizing open science and data work could help change the existing patterns of data use toward more sharing, while building the necessary infrastructure for it (Campbell et al. 2019; Hrynaszkiewicz et al. 2021).

A key idea for addressing **organizational/institutional factors** includes developing models of collaboration between academia, industry, and professional organizations. Each of these entities has resources to contribute. However, the synergy among them is often hindered by the lack of structures and frameworks of collaboration that address risks on all sides, while encouraging more openness and transparency (Hui 2010; Kashef 2023; Yanakiev 2020). In some ways, these models are also connected to **legal/policy** and **financial factors** as certain organizations may promote or hinder data sharing and determine how resources are allocated. Considering the high variability in approaches to data and code citations, there is also a need for standardization of data and software policies, which requires coordination among various organizations. Consensus-building activities across academic and commercial data providers could help cybersecurity researchers develop common tools, guidelines, and policies for sharing data and analytical tools. The funding models also need to address the need of working with commercial data and long-term sustainability of data sharing solutions.

Finally, **technological** factors in advancing data sharing in cybersecurity research include developing infrastructure that enables long-term sharing of all components of research products, including data, metadata, code, and narrative descriptions. Data sharing platforms such as CAIDA or Impact CyberTrust demonstrate how data availability can be increased to practice safe open science in cybersecurity, even if individual data downloads is not the ideal model for sharing and re-using large-scale data (Ives et al. 2008).

In this study we considered data as part of a research object and examined code availability as a key component of scientific evidence. Lack of code availability along with the differences in reporting technical experimentation raise questions about wider applicability and reproducibility of cybersecurity research. Unavailable code creates gaps in cybersecurity research methods, which rely on computing environment and tools to develop and test its hypotheses and create knowledge. While some publications were very specific in describing their technical environments and tools, many others omitted details that would allow other users to verify their approach without contacting the authors. Overall, our study revealed a lack of standardization in documenting/reporting experimentation and supporting technologies, which could be a sign of a still maturing field but is also a clear sign of challenges with infrastructure development and adoption.

Each of the factors discussed briefly above contains a multitude of avenues for future research directions and practical steps. Areas for further investigation may include the development of data sharing frameworks that facilitate sharing of commercial or potentially harmful data (e.g., malicious data), as well as the promotion of diversity in cybersecurity research, greater reproducibility through code availability, and infrastructure that better supports data repeated uses of existing data and code. In addition to resources for data sharing, the field needs resources for code sharing. A fuller systematic review that addresses the current state of the cultural, institutional, policy, and technological factors that affect data practices in cybersecurity research could also further advance the field and guide the research agenda.

This study has several limitations. First, our sampling technique, while covering top conferences in cybersecurity and supplementing it with a range of publications from highly ranked cybersecurity programs in the US, created a dataset that cannot be considered representative of the cybersecurity research because it does not cover a wide range of cybersecurity journals. It is also skewed toward the research in the US and does not fully represent the international perspectives. Such sampling could provide an incomplete representation of data practices in the field. Second, our analysis focused mostly on data and code availability, and it did not address the nature of the data in depth. Our attempt to use existing taxonomies has demonstrated their insufficiency for such a fast-developing field, but developing a taxonomy of data in cybersecurity was beyond the scope of this study. A deeper analysis of what types of data are shared and in what environments will help to create a fuller picture of data practices in cybersecurity and identify areas that information professionals can target for cyberinfrastructure development and training, outreach, and support services.

CONCLUSION

The findings of this study show that while the data and code in cybersecurity research are often not publicly available, the landscape of data sharing and use in cybersecurity is more complicated than a lack of incentives or unwillingness to share data. Many researchers rely on the existing data in their experimentation, but the nature of data they work with creates obstacles for accessing and re-using good data. Researchers often rely on more than one dataset in their studies, as they compile data from multiple sources over extended periods of time. Many generate code to process data, but as graduate students are often responsible for data collection, transformation, and analysis, it is not clear whether there is adequate training in data and software curation. The diversity of patterns of sharing and use found in this study indicates that individual researchers and teams may have their own idiosyncratic data and code management approaches that can benefit from standardization.

Sharing large-scale real-world data and code in security-related contexts needs a robust cyberinfrastructure that would support both large data producers (organizations and individuals) and data consumers (cybersecurity professionals and academic researchers). Building such infrastructure would benefit from broader diversity and inclusion strategies, consensus-building activities, and more graduate student training. The success of the data sharing ecosystem in cybersecurity depends on further standardization in data and software policies, including the policies of citation and attribution, and the mechanisms of persistent preservation and sharing of data collected in academic and commercial settings—all of which will pave the way for cyberinfrastructure, policies, and standards that advance the quality and availability of data in the field of cybersecurity research.

DATA ACCESSIBILITY STATEMENT

The data that supports the findings of this study, including the list of the analyzed publications, first author metadata, and coding of the datasets is available via the Figshare repository at https://doi.org/10.6084/m9.figshare.24639387.v1.

ADDITIONAL FILE

The additional file for this article can be found as follows:

- **Appendix A.** Codebook. DOI: https://doi.org/10.5334/dsj-2024-003.s1

ACKNOWLEDGEMENTS

The authors are grateful to Dalya Manatova at Indiana University, who assisted in gathering and annotating the literature for the background section.

FUNDING INFORMATION

This research was supported by the grant from the National Science Foundation ‘CICI: CSRC: Research Security Operations Center (ResearchSOC)’ (Award # 1840034).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Both authors have made substantial contributions to the paper. Kouper is responsible for the study funding, design, analysis, and interpretation of data, and drafting the paper. Stone is responsible for collecting, coding, and analyzing the data and revising the manuscript. The authors collaborated closely throughout this study and agreed to be on the author list.

AUTHOR AFFILIATIONS

Inna Kouper orcid.org/0000-0001-9801-2277
Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, USA

Stacy Stone
Center for Applied Cybersecurity Research, Indiana University, Bloomington, USA

REFERENCES

Abbott, RG, McClain, J, Anderson, B, Nauer, K, Silva, A and **Forsythe, C.** 2015. Log analysis of cyber security training exercises. *Procedia Manufacturing*, 3: 5088–5094. DOI: https://doi.org/10.1016/j.promfg.2015.07.523

Arzberger, P, Schroeder, P, Beaulieu, A, Bowker, GC, Casey, K, Laaksonen, L, Moorman, D, Uhlir, PF and **Wouters, P.** 2006. Promoting access to public research data for scientific, economic, and social development. *Data Science Journal*, 3: 135–152. DOI: https://doi.org/10.2481/dsj.3.135

Atapour-Abarghouei, A, McGough, AS and **Wall, DS.** 2020. Resolving the cybersecurity data sharing paradox to scale up cybersecurity via a co-production approach towards data sharing. In: *2020 IEEE International Conference on Big Data (Big Data)*. pp. 3867–3876. DOI: https://doi.org/10.1109/BigData50022.2020.9378014

Azhar, M, Bhatia, S, Gagne, G, Kari, C, Maguire, J, Mountrouidou, X, Tudor, L, Vosen, D and **Yuen, TT.** 2019. Securing the human: Broadening diversity in cybersecurity. In: *Proceedings of the 2019 ACM Conference on Innovation and Technology in Computer Science Education*. New York, NY, USA: ACM. pp. 251–252. DOI: https://doi.org/10.1145/3304221.3325537

Baker, K and **Millerand, F.** 2012. Infrastructuring ecology: Challenges in achieving data sharing. In: Parker, J, Vermeulen, N and Penders, B (eds.), *Collaboration in the New Life Sciences*. Ashgate. pp. 111–138.

Balenson, D, Tinnel, L and **Benzel, T.** 2015. Cybersecurity experimentation of the future (CEF): Catalyzing a new generation of experimental cybersecurity research. SRI International and USC Information Sciences Institute. Available at https://cef.cyberexperimentation.org/application/files/2616/2160/7871/CEF_Final_Report_Bound_20150922.pdf

Balenson, D, Tinnel, LS and **Kouper, I.** 2020. *Panel discussion and audience dialogue: Sharing artifacts and data for cybersecurity experimentation*. Available at https://www.usenix.org/conference/cset20/panel [Last accessed 11 August 2020].

Bechhofer, S, De Roure, D, Gamble, M, Goble, C and **Buchan, I.** 2010. Research objects: Towards exchange and reuse of digital knowledge. *Nature Precedings*, (ERIM Project Document erim1rep091103ab12). DOI: https://doi.org/10.1038/npre.2010.4626

Berman, F, Wilkinson, R and **Wood, J.** 2014. Building global infrastructure for data sharing and exchange through the research data alliance. *D-Lib Magazine*, 20(1/2): 1–4. DOI: https://doi.org/10.1045/january2014-berman

Blackfire Technology, Inc. 2019. Impact CyberTrust. Available at https://www.impactcybertrust.org/ [Last accessed 3 February 2022].

Brown, C, Cowperthwaite, A, Hijazi, A and **Somayaji, A.** 2009. Analysis of the 1999 DARPA/Lincoln Laboratory IDS evaluation data with NetADHICT. In: *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*. pp. 1–7. DOI: https://doi.org/10.1109/CISDA.2009.5356522

Brynielsson, J, Franke, U, Tariq, MA and **Varga, S.** 2016. Using cyber defense exercises to obtain additional data for attacker profiling. In: *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*. pp. 37–42. DOI: https://doi.org/10.1109/ISI.2016.7745440

Camp, LJ, Cranor, L, Feamster, N, Feigenbaum, J, Forrest, S, Kotz, D, Lee, W, Savage, S, Smith, S, Spafford, E and **Stolfo, S.** 2009. Data for cybersecurity research: Process and "Wish List." Available at https://www.researchgate.net/publication/255960171_Data_for_Cybersecurity_Research_Process_and_Wish_List.

Campbell, HA, Micheli-Campbell, MA and **Udyawer, V.** 2019. Early career researchers embrace data sharing. *Trends in Ecology & Evolution*, 34(2): 95–98. DOI: https://doi.org/10.1016/j.tree.2018.11.010

Cavelty, MD. 2018. Cybersecurity research meets science and technology studies. *Politics and Governance*, 6(2): 22–30. DOI: https://doi.org/10.17645/pag.v6i2.1385

Chawinga, WD and **Zinn, S.** 2019. Global perspectives of research data sharing: A systematic literature review. *Library & Information Science Research*, 41(2): 109–122. DOI: https://doi.org/10.1016/j.lisr.2019.04.004

Choucri, N, Madnick, S and **Koupke, P.** 2018. *Institutions for cybersecurity: International responses and data sharing initiatives*. The MIT Press. DOI: https://doi.org/10.7551/mitpress/11636.003.0003

Craigen, D, Diakun-Thibault, N and **Purse, R.** 2014. Defining cybersecurity. *Technology Innovation Management Review*, 4(10): 13–21. DOI: https://doi.org/10.22215/timreview/835

Cybersecurity. 2009. *Cybersecurity Glossary*. Available at https://niccs.cisa.gov/about-niccs/cybersecurity-glossary#C [Last accessed 22 July 2021].

Douglass, K, Allard, S, Tenopir, C, Wu, L and **Frame, M.** 2013. Managing scientific data as public assets: Data sharing practices and policies among full-time government employees. *Journal of the Association for Information Science and Technology*, 65(2): 215–429. DOI: https://doi.org/10.1002/asi.22988

Dumitraş, T. 2018. Worldwide Intelligence Network Environment (WINE). Available at http://users.umiacs.umd.edu/~tdumitra/blog/old/worldwide-intelligence-network-environment/ [Last accessed 29 June 2022].

Dumitraş, T and **Shou, D.** 2011. Toward a standard benchmark for computer security research: The worldwide intelligence network environment (WINE). In: *Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*. New York, NY, USA: Association for Computing Machinery. pp. 89–96. DOI: https://doi.org/10.1145/1978672.1978683

Fisk, G, Ardi, C, Pickett, N, Heidemann, J, Fisk, M and **Papadopoulos, C.** 2015. Privacy principles for sharing cyber security data. In: *2015 IEEE Security and Privacy Workshops*. pp. 193–197. DOI: https://doi.org/10.1109/SPW.2015.23

Fontugne, R, Borgnat, P, Abry, P and **Fukuda, K.** 2010. MAWILab: Combining diverse anomaly detectors for automated anomaly labeling and performance benchmarking. In: *Proceedings of the 6th International Conference*. New York, NY, USA: ACM. DOI: https://doi.org/10.1145/1921168.1921179

Gormley, CJ and **Gormley, SJ.** 2012. Data hoarding and information clutter: The impact on cost, life span of data, effectiveness, sharing, productivity, and knowledge management culture. *Issues in Information Systems*, 13(2): 90–95.

Hrynaszkiewicz, I, Harney, J and **Cadwallader, L.** 2021. A survey of researchers' needs and priorities for data sharing. *Data Science Journal*, 20(1): 31. DOI: https://doi.org/10.5334/dsj-2021-031

Hui, P, Bruce, J, Fink, G, Gregory, M, Best, D, McGrath, L and **Endert, A.** 2010. Towards efficient collaboration in cyber security. In: *2010 International Symposium on Collaborative Technologies and Systems*. IEEE Computer Society. pp. 489–498. DOI: https://doi.org/10.1109/CTS.2010.5478473

Information Centre of Excellence for Tech Innovation (ISCX). 2007. Datasets. Available at http://www.iscx.ca/datasets/ [Last accessed 3 February 2022].

InfraGuard. 2018. Available at https://www.infragard.org/ [Last accessed 3 February 2022].

Ives, ZG, Green, TJ, Karvounarakis, G, Taylor, NE, Tannen, V, Talukdar, PP, Jacob, M and **Pereira, F.** 2008. The ORCHESTRA Collaborative Data Sharing System. *ACM SIGMOD Record*, 37(3): 26–32. DOI: https://doi.org/10.1145/1462571.1462577

Kashef, R, Freunek, M, Schwartzentruber, J, Samavi, R, Bulgurcu, B, Khan, AJ and **Santos, M.** 2023. Bridging the bubbles: Connecting academia and industry in cybersecurity research. [Preprint]. DOI: https://doi.org/10.32920/24132645.v1

Kemmerer, RA. 2003. Cybersecurity. In: *Proceedings of the 25th International Conference on Software Engineering*. pp. 705–715. DOI: https://doi.org/10.1109/ICSE.2003.1201257

Larivière, V, Ni, C, Gingras, Y, Cronin, B and **Sugimoto, CR.** 2013. Bibliometrics: Global gender disparities in science. *Nature*, 504(7479): 211–213. DOI: https://doi.org/10.1038/504211a

Li, Z and **Oprea, A.** 2016. Operational security log analytics for enterprise breach detection. In: *2016 IEEE Cybersecurity Development (SecDev)*. pp. 15–22. DOI: https://doi.org/10.1109/SecDev.2016.015

Linger, R, Goldrich, L, Bishop, M and **Dark, M.** 2017. Agile research for cybersecurity: creating authoritative, actionable knowledge when speed matters. In: *Proceedings of the 50th Hawaii International Conference on System Sciences*. DOI: https://doi.org/10.24251/HICSS.2017.723

Maciá-Fernández, G, Camacho, J, Magán-Carrión, R, García-Teodoro, P and **Therón, R.** 2018. UGR'16: A new dataset for the evaluation of cyclostationarity-based network IDSs. *Computers and Security*, 73: 411–424. DOI: https://doi.org/10.1016/j.cose.2017.11.004

Mahoney, MV and **Chan, PK.** 2003. An analysis of the 1999 DARPA/Lincoln Laboratory evaluation data for network anomaly detection. In: Vigna Giovanni and Kruegel, C and Erland, J (eds.), *Recent Advances in Intrusion Detection*. Berlin, Heidelberg: Springer Berlin Heidelberg. pp. 220–237. DOI: https://doi.org/10.1007/978-3-540-45248-5_13

Mathew, AJ and **Cheshire, C.** 2018. *A fragmented whole: Cooperation and learning in the practice of information security*. UC Berkeley and Packet Clearing House. Available at https://www.pch.net/resources/Papers/A_Fragmented_Whole/.

Maxson Jones, K, Ankeny, RA and **Cook-Deegan, R.** 2018. The Bermuda Triangle: The pragmatics, policies, and principles for data sharing in the history of the Human Genome Project. *Journal of the History of Biology*, 51(4): 693–805. DOI: https://doi.org/10.1007/s10739-018-9538-7

McHugh, J. 2000. Testing intrusion detection systems: A critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory. *ACM Transactions on Information and System Security*, 3(4): 262–294. DOI: https://doi.org/10.1145/382912.382923

MIT Lincoln Laboratory. 2016. Cyber Grand Challenge – Datasets. Available at https://www.ll.mit.edu/r-d/datasets/cyber-grand-challenge-datasets [Last accessed 3 February 2022].

Moore, T, Kenneally, E, Collett, M and **Thapa, P.** 2019. Valuing cybersecurity research datasets (SSRN Scholarly Paper No. ID 3469364). Available at https://papers.ssrn.com/abstract=3469364 [Last accessed 22 July 2021].

Moustafa, N and **Slay, J.** 2015. UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In: *2015 Military Communications and Information Systems Conference (MilCIS)*. pp. 1–6. DOI: https://doi.org/10.1109/MilCIS.2015.7348942

Munaiah, N, Pelletier, J, Su, S-H, Yang, S and **Meneely, A.** 2019. A cybersecurity dataset derived from the national collegiate penetration testing competition. In: 2019 Hawaii International Conference on System Sciences. Available at https://www.rit.edu/academicaffairs/facultyscholarship/submit/download_file.php?id=128537 [Last accessed 6 June 2023].

Nelson, B. 2009. Data sharing: Empty archives. *Nature*, 461(7261): 160–163. DOI: https://doi.org/10.1038/461160a

Pala, A and **Zhuang, J.** 2019. Information sharing in cybersecurity: A review. *Decision Analysis*, 16(3): 172–196. DOI: https://doi.org/10.1287/deca.2018.0387

Peacock, D and **Irons, A.** 2017. Gender inequality in cybersecurity: Exploring the gender gap in opportunities and progression. *International Journal of Gender, Science and Technology*, 9(1): 25–44.

Plale, BA, Dickson, E, Kouper, I, Liyanage, S, Ma, Y, McDonald, RH, Walsh, JA and **Withana, S.** 2019. Safe open science for restricted data. *Data and Information Management*, 3(1): 50–60. DOI: https://doi.org/10.2478/dim-2019-0005

Poirier, L and **Costelloe-Kuehn, B.** 2019. Data sharing at scale: A heuristic for affirming data cultures. *Data Science Journal*, 18(1): 48. DOI: https://doi.org/10.5334/dsj-2019-048

Poster, WR. 2018. Cybersecurity needs women. *Nature*, 555(7698): 577–580. DOI: https://doi.org/10.1038/d41586-018-03327-w

Ross, MB, Glennon, BM, Murciano-Goroff, R, Berkes, EG, Weinberg, BA and **Lane, JI.** 2022. Women are credited less in science than men. *Nature*, 608(7921): 135–145. DOI: https://doi.org/10.1038/s41586-022-04966-w

San Diego Supercomputer Center. 2020. Center for Applied Internet Data Analysis (CAIDA). *CAIDA*. Available at https://www.caida.org/ [Last accessed 29 June 2022].

Sarker, IH, Kayes, ASM, Badsha, S, Alqahtani, H, Watters, P and **Ng, A.** 2020. Cybersecurity data science: An overview from machine learning perspective. *Journal of Big Data*, 7(1): 41. DOI: https://doi.org/10.1186/s40537-020-00318-5

Sauerwein, C, Pekaric, I, Felderer, M and **Breu, R.** 2019. An analysis and classification of public information security data sources used in research and practice. *Computers & Security*, 82: 140–155. DOI: https://doi.org/10.1016/j.cose.2018.12.011

Scheper, C, Cantor, S and **Maughan, D.** 2011. PREDICT: A trusted framework for sharing data for cyber security research. In: *Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security BADGERS '11*. pp. 105–106. DOI: https://doi.org/10.1145/1978672.1978686

Sebastian, G. 2022. Could incorporating cybersecurity reporting into SOX have prevented most data breaches at U.S. publicly traded companies? An exploratory study. *International Cybersecurity Law Review*, 3(2): 367–383. DOI: https://doi.org/10.1365/s43439-022-00062-x

Serrano, O, Dandurand, L and **Brown, S.** 2014. On the design of a cyber security data sharing system. In: *Proceedings of the 2014 ACM Workshop on Information Sharing & Collaborative Security – WISCS '14*. pp. 61–69. DOI: https://doi.org/10.1145/2663876.2663882

Shiravi, A, Shiravi, H, Tavallaee, M and **Ghorbani, AA.** 2012. Toward developing a systematic approach to generate benchmark datasets for intrusion detection. *Computers & Security*, 31(3): 357–374. DOI: https://doi.org/10.1016/j.cose.2011.12.012

Shou, D. 2012. Ethical considerations of sharing data for cybersecurity research. In: Danezis, G, Dietrich, S and Sako, K (eds.), *Financial Cryptography and Data Security*. Berlin, Heidelberg: Springer Berlin Heidelberg. pp. 169–177. DOI: https://doi.org/10.1007/978-3-642-29889-9_15

Shumba, R, Ferguson-Boucher, K, Sweedyk, E, Taylor, C, Franklin, G, Turner, C, Sande, C, Acholonu, G, Bace, R and **Hall, L.** 2013. Cybersecurity, women, and minorities: Findings and recommendations from a preliminary investigation. In: *Proceedings of the ITiCSE Working Group Reports Conference on Innovation and Technology in Computer Science Education-Working Group Reports*. New York, NY, USA: ACM. pp. 1–14. DOI: https://doi.org/10.1145/2543882.2543883

Sommer, R and **Paxson, V.** 2010. Outside the closed world: On using machine learning for network intrusion detection. In: *2010 IEEE Symposium on Security and Privacy*. pp. 305–316. DOI: https://doi.org/10.1109/SP.2010.25

Sommestad, T and **Hallberg, J.** 2012. Cyber security exercises and competitions as a platform for cyber security experiments. In: Jøsang, A and Carlsson, B (eds.), *Secure IT Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg. pp. 47–60. DOI: https://doi.org/10.1007/978-3-642-34210-3_4

Sperotto, A, Sadre, R, van Vliet, F and **Pras, A.** 2009. A labeled data set for flow-based intrusion detection. In: Nunzi, G, Scoglio, C and Li, X (eds.), *IP Operations and Management*. Berlin, Heidelberg: Springer Berlin Heidelberg. pp. 39–50. DOI: https://doi.org/10.1007/978-3-642-04968-2_4

Su, X, Johnson, J and **Bozeman, B.** 2015. Gender diversity strategy in academic departments: exploring organizational determinants. *Higher Education*, 69(5): 839–858. DOI: https://doi.org/10.1007/s10734-014-9808-z

Sun, N, Zhang, J, Rimba, P, Gao, S, Zhang, LY and **Xiang, Y.** 2019. Data-driven cybersecurity incident prediction: A survey. *IEEE Communications Surveys & Tutorials*, 21(2): 1744–1772. DOI: https://doi.org/10.1109/COMST.2018.2885561

Tavallaee, M, Bagheri, E, Lu, W and **Ghorbani, AA.** 2009. A detailed analysis of the KDD CUP 99 data set. In: *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*. pp. 1–6. DOI: https://doi.org/10.1109/CISDA.2009.5356528

Tenopir, C, Allard, S, Douglass, K, Aydinoglu, A U, Wu, L, Read, E, Manoff, M and **Frame, M.** 2011. Data sharing by scientists: Practices and perceptions. *PLoS ONE*, 6(6). DOI: https://doi.org/10.1371/journal.pone.0021101

Tuma, K and **Van Der Lee, R.** 2022. The role of diversity in cybersecurity risk analysis: An experimental plan. In: *Proceedings of the Third Workshop on Gender Equality, Diversity, and Inclusion in Software Engineering*. New York, NY, USA: ACM. pp. 12–18. DOI: https://doi.org/10.1145/3524501.3527595

University of California Irvine. 1999. KDD Cup 1999 Data. Available at https://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html [Last accessed 3 February 2022].

US Department of Homeland Security. 2022. Cyber Information Sharing and Collaboration Program (CISCP). Available at https://www.cisa.gov/ciscp [Last accessed 3 February 2022].

Walton, S, Wheeler, PR, Zhang, Y (Ian) and **Zhao, X (Ray).** 2021. An integrative review and analysis of cybersecurity research: Current state and future directions. *Journal of Information Systems*, 35(1): 155–186. DOI: https://doi.org/10.2308/ISYS-19-033

Yanakiev, Y. 2020. A governance model of a collaborative networked organization for cybersecurity research. *Information & Security*, 46(1): 79–98. DOI: https://doi.org/10.11610/isij.4606

Zheng, M, Robbins, H, Chai, Z, Thapa, P and **Moore, T.** 2018. Cybersecurity research datasets: Taxonomy and empirical analysis. In: *11th USENIX Workshop on Cyber Security Experimentation and Test CSET-18*. Available at https://www.usenix.org/system/files/conference/cset18/cset18-paper-zheng.pdf [Last accessed 22 July 2021].

TO CITE THIS ARTICLE:
Kouper, I and Stone, S. 2024. Data Sharing and Use in Cybersecurity Research. *Data Science Journal*, 23: 3, pp. 1–19. DOI: https://doi.org/10.5334/dsj-2024-003

Submitted: 20 June 2023
Accepted: 15 December 2023
Published: 19 January 2024

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.