Cybersecurity Research Datasets: Taxonomy and Empirical Analysis

Muwei Zheng *Hannah Robbins* *Zimo Chai* *Prakash Thapa*
Tyler Moore

Tandy School of Computer Science
The University of Tulsa, USA
{muz856,her131,zic337,prt4832,tyler-moore}@utulsa.edu

Abstract

We inspect 965 cybersecurity research papers published between 2012 and 2016 in order to understand better how datasets are used, produced and shared. We construct a taxonomy of the types of data created and shared, informed and validated by the examined papers. We then analyze the gathered data on datasets. Three quarters of existing datasets used as input to research are publicly available, but less than one fifth of datasets created by researchers are publicly shared. Using a series of linear regressions, we demonstrate that those researchers who do make public the datasets they create are rewarded with more citations to the associated papers. Hence, we conclude that an under-appreciated incentive exists for researchers to share their created datasets with the broader research community.

1 Introduction and Background

Cybersecurity research and practice is becoming increasingly data-driven. Cybercrime indicators can be used to better quantify risks and inform proactive defenses based on what has been targeted previously. Researchers have been using cybersecurity datasets as input to their own work and producing it as an output of research for years. Unfortunately, such data is not always shared with the broader research community, which makes replicating results difficult and developing new innovations using existing data infeasible.

Despite its importance, a number of challenges can stymie the sharing of cybersecurity data. Legal and privacy issues are frequently volunteered as an impediment to sharing. A more basic barrier is that sharing can be costly (in terms of time and dollars), while the benefits of such sharing largely accrue to others. In economics terms, sharing data creates positive externalities, which tends to result in less of the good (in this case, cybersecurity datasets) being produced than is socially optimal. Finally, the incentives to share data are not always strong. Competitive concerns can inhibit firms who offer security products or services from sharing data with others. Researchers invest substantial effort in amassing reliable datasets and may prefer to mine the data themselves, fearing that others might take credit for discoveries based on data they have produced.

Pioneering efforts have encouraged data sharing by researchers, principally DHS's PREDICT program and its successor IMPACT [19]. These have reduced some barriers to sharing, notably legal ones. Despite these efforts, it is clear that these resources remain underutilized.

Unlike many forms of data, security datasets are often seen as especially sensitive for several reasons. First, the data collected by one party could reflect poorly on others, e.g., if it indicated that a particular network had poor operational security. Second, the data might reveal to adversaries what is known about their activities, inadvertently assisting them in their criminal activities. Third, concerns over inadvertent sharing of private data preclude some organizations from sharing widely.

An economic perspective on cybersecurity offers many insights into the behavior of attackers and defenders [2]. Misaligned incentives on the part of defenders are often to blame when security mechanisms fail and attacks succeed [3]. Incentives also play a big role in influencing when security data is shared. Researchers have constructed game theory models to explain why firms would choose to share security data via Information Sharing and Analysis Centers (ISACs). Gordon et al. found that sharing can enable firms to invest optimally in security at lower cost [12]. But they also found that the incentives to share data were lacking, and absent coordination, firms would elect to free ride off firms that did share instead. Subsequently, Gal-Or and Ghose constructed a model which found that sharing works best in competitive industries where product substitutability is higher [11]. On the empirical side, Moore and Clayton found evidence that competition among security services firms can preclude data sharing, to the detriment of security overall [17]. They presented evidence that website takedown companies who remove phishing content from websites do not share data with competitors, which contributes to long delays in remediating the affected websites. More recently, Jhaveri et al. examined how abuse data such as malware URL blacklists and records of botnet activity are shared among Internet infrastructure operators [15]. They found that the incentives to collect and share data by individual operators play a significant role in determining whether data is collected and acted upon. The existing literature has primarily been concerned with data sharing among operators. By contrast, the focus of this project is on cybersecurity data sharing among researchers. Many concerns overlap (notably concerns over privacy, difficulty of sharing), but others manifest differently (e.g., competition is less of a concern).

In this paper, we study data usage and creation in nearly 1,000 top academic information security publications using a methodology outlined in Section 2. We iteratively construct a taxonomy of dataset categories informed by the papers inspected, described in Section 3. We then analyze this data in Section 4. Key findings include quantifying how often created datasets are made public, and how this varies by dataset type and over time. We also provide evidence that papers that share a created dataset publicly are correlated with higher citation rates.

2 Methodology

Data Sources In order to study how researchers use and produce datasets, we first selected suitable publication venues to examine. We started by selecting the top four computer security research conferences, ACM Conference on Computer and Communications Security (CCS), USENIX Security Symposium (USENIX), IEEE Symposium on Security and Privacy (S&P), and Network and Distributed System Security Symposium (NDSS). We complemented these with outlets that regularly publish data-intensive research: Internet Measurement Conference (IMC), International Conference on Financial Cryptography and Data Security (FC), and the Workshop on the Economics of Information Security (WEIS). Finally, we included relevant workshops associated with top conferences: the AI & Security Workshop at CCS, Cyber Security Experimentation and Test (CSET) Workshop at USENIX Security, and the Workshop on Bitcoin and Blockchain Research at FC (BITCOIN). We collected papers from these conferences from 2012 to 2016, inclusively. Focusing on these conferences makes our research efforts tractable, but this necessarily limits the scope of our findings.

We first downloaded all of the papers and collected their citation information. We used DBLP [8] to get information on the papers and their linked URLs, and then crawled 2,037 papers from their corresponding websites. We then obtained citation information for all papers from Google Scholar.

Dataset Classifier We constructed a binary classifier to distinguish dataset-related papers and non-dataset related papers. Dataset papers are defined as those with at least one dataset that was used or created during the research. Non-dataset papers are papers that do not include a dataset as defined above. To build the machine learning model, we manually classified 391 papers into data (209) and non-data (182) papers. These 391 papers are randomly selected from the set of 2,037 papers, while ensuring sufficient coverage of all venues and years from 2012 to 2016. To construct features, we first extracted a list of "base form" (i.e., case and tense insensitive) words for each paper using the textblob Python package. We also filtered all stop words using NLTK's build-in list [5]. From each paper's final word list, we built a word vocabulary from all papers and computed a TF-IDF vector for each paper. We then constructed several models using the sklearn Python package including Multinomial Naive Bayes, Bernoulli Naive Bayes, Gaussian Naive Bayes, SVC Confusion Matrix, and Random Forest. We used 10-fold cross validation to evaluate the models. Random Forest was the most accurate, but it still has 21.1% false positive rate and 17.2% false negative rate.

We used this model to classify all papers, 1,129 of which were predicted to include data. We inspected a random sample of 356 predicted data papers, confirming that 308 had data and 48 did not. Additionally, we inspected a sample of 218 predicted non-data papers that did not include references to dataset names identified in the training set. In total, we confirmed the presence or absence of data in 965 papers. 517 papers have data, with 209 coming from the training set and 308 confirmed from the test set. 448 papers do not have data, with 182 from training set and 266 confirmed from the test set. This labeled dataset, along with the R scripts to analyze the data, are available at `doi:10.7910/DVN/4EPUIA`.

3 Taxonomy of Cybersecurity Research Datasets

We were unable to find an existing taxonomy of cybersecurity datasets suited to our needs. Some databases categorize their datasets based on protocols or other technical attributes of the data, such as BGP or DNS data [19]. While this is a clear way to categorize data, there are drawbacks when viewing datasets from a cybersecurity perspective. For example, both BGP hijacking reports and BGP route announcements could be considered BGP data, but they share little in common beyond the proto-

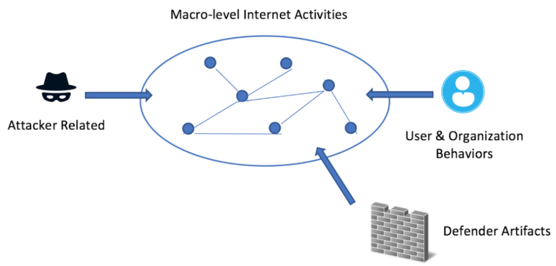
Macro-level Internet Activities
Attacker Related
User & Organization Behaviors
Defender Artifacts

Figure 1: Dataset categories.

col itself. BGP hijacking reports would be of interest to researchers studying Internet outages or attacks, while BGP route announcements are of interest to those studying Internet topology.

Grajeda et al. come closest to meeting our needs. They examined 715 papers to identify the types of data used and produced in digital forensics research [13]. They made similar distinctions between when data was used as input to or an output of research. Their subject matter, while related, is more narrowly focused on a particular aspect of cybersecurity that is inherently data-intensive.

Consequently, we set out to develop a taxonomy that is more focused on describing datasets in the context of cybersecurity research writ large. We iteratively constructed the categories by placing datasets identified in papers into prospective categories, modifying the groupings when uncertainties arose over which category the dataset should belong. Once the categories and subcategories stabilized, each dataset was manually labeled with a category, subcategory, whether the dataset already existed or was created by the authors, and whether the data was made publicly available. We now discuss full details of how these values are defined.

We divided data into four major categories: attack-related dataset; defender artifacts; end users and organization characteristics; and macro-level Internet characteristics. These categories are visualized in Figure 1.

Attacker-Related Attackers exploit vulnerabilities to launch attacks. So we classified any data that is already deemed malicious (e.g., scams, malware), or is used by attackers (vulnerabilities, cybercrime infrastructures) as *attacker-related*. There are four subcategories. **Attacks** contain information on attempts to harm digital assets perpetrated intentionally by malicious actors. For example, Park and McCoy [20] extracted 29K rental scam postings from Craigslist website. **Vulnerabilities** contain information on weaknesses in digital assets that can be exploited by an attacker. For example, Bilge and Dumitras used Open Source Vulnerability Database as a dataset [4]. **Exploits** contain information on how attacks may be perpetrated, but not when a particular system has been targeted by a malicious actor. For example, Sabottke and Suciu used exploits from Microsoft security advisories [21]. Finally, **cybercrime activities** describe unlawful activities distinct from attacks, as well as information on the infrastructure and operations used by malicious actors to perpetrate attacks. A typical example is the crawl of the Silk Road anonymous marketplace [7].

Defender Artifacts People and organizations construct defenses such as firewalls or secure configurations to block or prevent attacks. Frequently, as a consequence of constructing these defenses, data is generated (e.g., logs of blocked connection requests). These *defender artifacts* include configurations and alerts. **Alerts** contain outputs of defender artifacts, such as firewall logs or blackhole traffic [9]. **Configurations** contain information about how defender artifacts are set up and configured (e.g., SSL certificate configurations [10]).

User & Organization Characteristics Many datasets study the behaviors of individuals or organizations. **User activities** contain information about users or organizations online behavior, such as tweets [18]. **User attitudes** contain information about opinions or attitudes towards an issue, often gleaned through surveys [14]. **User attributes** contain information about the characteristics of users or organizations themselves (e.g., user profiles).

Macro-level Internet Characteristics The Internet does not only consist of human activities but also many technical protocols and traffic. Datasets that are focused on studying network characteristics, as opposed to user or organizational characteristics, fit in here. **Applications** contain information about Internet end products and services such as websites, Android apps, bitcoin, extensions, or code. A typical example of Applications is the Alexa list of top websites [1]. **Network traces** are usually network traffic dumps that not only contain information regarding the application level, but also information about lower layers. Data usually comes from a benign resource, like an organization's internal network, but malicious traffic might be included. For example, Wang and Dyer used "packet-level traces for Tor Pluggable Transport traffic collected in controlled environments" [22]. **Topology** datasets contain information about relationships between Internet components. A typical example in this category is CAIDA's AS relationship database [6]. **Benchmarks** contain information about measurements of Internet performance, such as upload/download speed or end-to-end network reliability. For example, Jiang and Wang constructed a dataset that

	Not Public		Public	
Dataset Type	#	%	#	%
Created Deriv.	89	85	16	15
Created Prim.	213	81	50	19
Existing	129	24	398	76

Table 1: Number and percentage of datasets made public, split by whether or not the dataset was created by researchers or already existed. Note: seven datasets could not be categorized as public or not based on the dataset description alone, so they are excluded from the table.

measured 3G/4G network performance in the US and Korea [16]. Finally, **adverse events** contain information on events that harm digital assets where malicious intent has not been established (e.g., outages caused by routing misconfigurations).

Additional dataset characteristics We not only log datasets' subject matter, we also track how they are used and shared. If a dataset already existed before the study undertaken by the research paper, we label this dataset as an **existing** dataset. In other cases, the dataset is **created** by the researchers. There are two possibilities here. If it is generated from some other datasets, like extracting all malicious websites from a large website list, the paper has created a derivative dataset. If the dataset is generated entirely by the authors without using other datasets as input, such as collecting and analyzing campus network traffic, we say the dataset has been created primarily by the authors.

We are especially interested in whether a dataset is **publicly available** or not. We note whether the paper explicitly claims that the dataset is publicly available. We include public repositories that place restrictions on who can download the data (e.g., IMPACT) as still being public. When the dataset is publicly available, we check to see if it is in fact available visiting the link directly to determine if it is still valid or not.

4 Empirical Analysis of Research Datasets

We now examine the datasets identified in research papers. We first consider how datasets are created and used in Section 4.1. In Section 4.2, we report on how often datasets in different categories are created, used, and made public. Finally, in Section 4.3, we show that dataset creators who share publicly are rewarded with modestly higher citation rates compared to other papers.

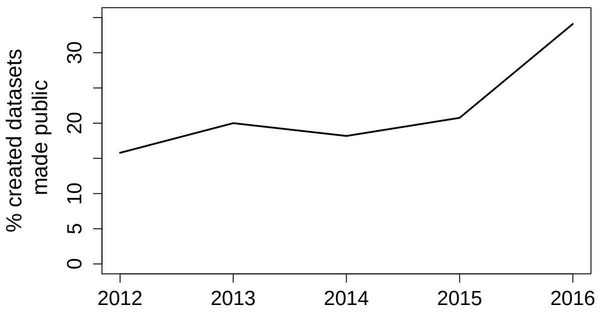

Figure 2: Percentage of papers that create datasets which are made public.

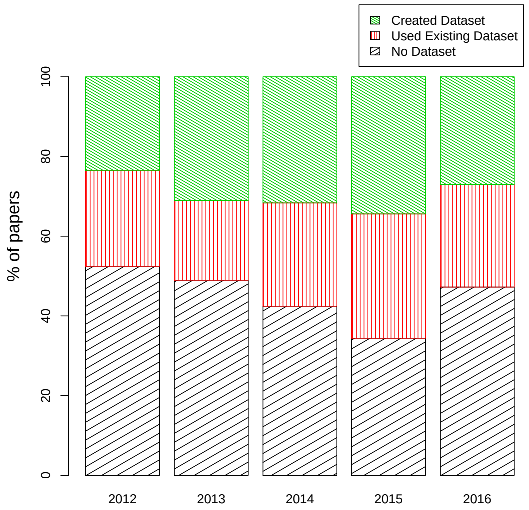

Figure 3: Datasets in papers over time.

4.1 Dataset creation and usage

Of the 965 research papers we labeled, 517 (55%) included data in some form. 229 papers (24% of the total) did not create a new dataset, but used an existing dataset as input to their research. Meanwhile 288 papers created datasets, either from scratch or by using other data as input. Just 61 papers, or 6% of those examined, created a dataset and made it publicly available.

Table 1 breaks down the datasets used in the papers, according to what was done and whether the underlying data was publicly available. Note that a single paper can use or create multiple datasets. In total, 902 datasets were created or used in the 517 papers. For created data, the distinction is made between primary and derivative data. Primary data is created solely by the authors, while derivative data leverages at least one existing dataset in order to make a new dataset. We can see that 76% of existing datasets used by researchers as input to their work

	% Datasets	% Created		% Public	
Attacks	13	**30**	(–)	53	
Vulnerabilities	5	**71**	(+)	39	
Exploits	3	29		**75**	(+)
Cybercrime Inf.	1	56		44	
Alerts	3	30		**74**	(+)
Configurations	5	55		48	
Applications	24	36		**62**	(+)
Network Traces	9	**60**	(+)	**22**	(–)
Topology	9	**22**	(–)	**67**	(+)
Benchmarks	3	**81**	(+)	34	
Adverse Events	2	**67**	(+)	33	
User Activities	12	38		**41**	(+)
User Attitudes	1	**90**	(+)	**10**	(+)
User Attributes	10	**26**	(–)	**66**	(+)

Table 2: Incidence of datasets split by subcategory, plus proportion of datasets in each subcategory that are created and made public. Statistically significant under- and over-representations are indicated in bold with a (**+**/**-**).

are publicly available. This makes sense because public datasets are easier to access. Unfortunately, once the researchers create datasets, they are much less likely to return the favor by publishing their own datasets. 81–85% of created datasets described in research papers are not made publicly available. While not surprising, this stark difference highlights the opportunities missed by researchers failing to reciprocate by making their own data publicly available.

Has the publication of datasets changed over time? Figure 2 plots the percentage of papers that create datasets and subsequently publish the data. The trend is variable but slightly increasing.

Finally, Figure 3 examines how data usage in research papers have changed over time. In 2012, just over half of the papers studied did not include use or create a dataset. This steadily declined to one third in 2015, before jumping to 48% in 2016. The fraction using existing datasets as input to research without creating their own dataset fluctuated between 18-31%. Papers creating datasets rose from 23% in 2012 to a peak of 34% in 2015, before falling to 27% in 2016.

4.2 Dataset categories

We now study the prevalence of different types of data used in cybersecurity research, according to the taxonomy outlined in Section 3. The leftmost numerical column in Table 2 shows the percentage breakdown of datasets across categories. Macro-level Internet characteristics comprise 47% of the total datasets encountered, with another 23% for user and organizational characteristics. Datasets related to attack (22%) and defense (8%) fill out the remainder. The next column in Table 2 looks at how datasets in each subcategory are used. It reports the percentage of datasets in each subcategory that are created by the research, as opposed to re-use of existing data. Differences in proportion that are statistically under- and over-represented (according to a χ^2 test) are indicated. For example, we note that 71% of datasets describing vulnerabilities are created, compared to just 30% of datasets of attacks. This may indicate that attack datasets are particularly valuable inputs to research, or that vulnerabilities are more likely to be identified than subsequently used by others. Similarly, network traces, benchmarks and adverse events are disproportionately likely to be created rather than used. Note that low levels of reuse could also reflect difficulty sharing data, as is likely for network traces and user attitudes.

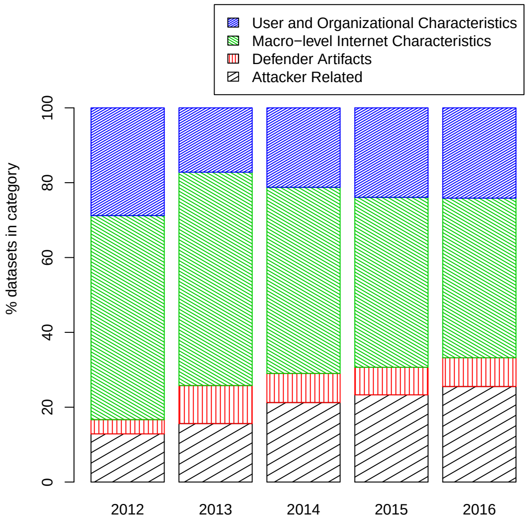

Figure 4: Dataset categories over time.

The last column in the table reports the proportion of datasets (created or existing) that are public. Exploits, application, topology and user attribute datasets are more likely to be public, while alerts, network traces, user activities and user attributes are less so.

Have the types of data appearing in research papers changed over time? Figure 4 shows that the proportion of attacker-related datasets in papers has increased, from 14% to 25% overall. While less common, defender datasets have also increased.

4.3 Could citations incentivize publishing datasets?

The preceding analysis has identified that while datasets are frequently created and used by cybersecurity re-

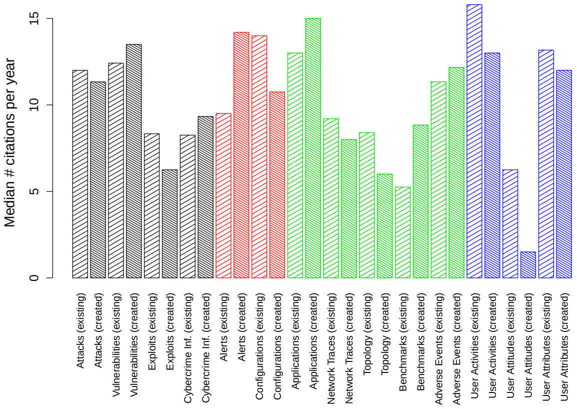

Figure 5: Median annual citations per subcategory and whether or not the data was created or already existed.

searchers, it is uncommon for these datasets to be shared by the researchers who gather the data. There are many reasonable (and not so reasonable) explanations, including privacy considerations, restrictions on sharing by partners, and competitive concerns. On top of all that, it can be expensive and time-consuming to prepare the data for sharing and to accommodate requests.

Despite these downsides, there are definitely benefits to publishing datasets that accrue to the researcher. We explore one possible such benefit that is highly prized by academic researchers, namely, citations to the paper that created the dataset. We hypothesize that papers describing publicly available datasets will be cited more often, since other researchers can apply the data in their own follow-up work.

The summary statistics are encouraging. Papers that do not involve data or only use existing datasets in their work are cited 10 times per year (median), compared to 9.3 citations per year for papers that create datasets but don't publish them. By contrast, papers that do publish their data receive a median of 14.2 citations per year.

Figure 5 looks at the median annual citations split by data type, as well as whether the data was used or created. There is considerable variation in citation rates by data type, with papers creating data describing user attitudes cited roughly once to user activities receiving 13–16 citations annually. Comparing the differences within subcategories based on whether the data was used or created reveals some interesting trends. For macro-level Internet and attacker-related characteristics, the citation rates are broadly similar for existing and created datasets. For alerts, papers that created datasets were cited more often than those that used them. For user and organizational characteristics, the trend is reversed: papers that leveraged an existing dataset were cited more frequently than those that created datasets.

Regression Analysis While the above summary statistics about how citations vary suggest interesting trends, these effects are interrelated. To disentangle their effects, we constructed several linear regressions using the number of citations as the response variable. The explanatory variables include:

1. **# years since published**: We expect that the passage of time will lead to more citations.

2. **Publication venue**: The reputation and visibility of the publication outlet doubtless influences how often the paper is likely to be cited. In our regressions, this is represented as a categorical variable, using ACM CCS as the baseline.

3. **Created public dataset**: We hypothesize that creating a dataset and making it public will yield more citations than keeping it private.

4. **Dataset category**: We expect that for papers that create datasets, the type of data created will influence its citation frequency. While we do not presume which categories will be cited more often, we do anticipate that differences among data types affect citation frequency. This is represented in the regressions as a categorical variable, with the *Attacks* subcategory as baseline.

Table 3 presents the results of four linear regressions that incrementally incorporate the above explanatory variables. These regressions include only those papers that create datasets, in order to evaluate our key hypothesis that publishing datasets will be associated with higher citations. The baseline model (1) finds that, as expected, time since publication affects citation rates. Each additional year since publication corresponds to 23 more citations. Approximately 10% of the variance in citation rates can be explained by time since publication alone. Adding in publication venue (model 2) explains a further 6.3% of the variance in citation rates. Papers creating datasets and published in the CSET, AISEC, and BITCOIN workshops are less likely to be cited than those in CCS, while those appearing in IEEE S&P are considerably more likely to be cited than papers from CCS. Citations for other outlets (FC, IMC, NDSS, USENIX Security, and WEIS) were indistinguishable from CCS.

Model 3 adds in a Boolean variable for whether the created dataset was made public. It is positive and statistically significant. The coefficient can be interpreted as papers that publish their datasets receive a boost of around 31 citations.

Model 4 adds in dataset subcategories. Relative to attack datasets, papers that create datasets of vulnerabilities, alerts, network traces, topology and benchmarks are cited less often. Note that the number of observations is

Table 3: Linear regression tables for papers that create datasets.

	Dependent variable:			
	citeNum			
	(1)	(2)	(3)	(4)
Years Published	23.059^{***}	24.957^{***}	25.619^{***}	24.779^{***}
FC		−26.982	−26.848	−24.712
IMC		−17.616	−23.730	−20.464
NDSS		−11.401	−15.367	−11.330
IEEE S&P		60.211^{***}	55.741^{**}	29.723^{**}
USENIX Security		4.586	−0.717	−3.582
WEIS		−25.607	−27.932	−30.750
Workshops		-46.998^{**}	-48.271^{**}	-54.410^{***}
Created Public			30.718^{**}	24.651^{**}
Vulnerabilities				-33.029^{*}
Exploits				−29.843
Cybercrime Inf.				−2.050
Alerts				-51.072^{*}
Configurations				−22.363
Applications				−12.232
Network Traces				-30.925^{*}
Topology				-37.760^{*}
Benchmarks				-36.534^{*}
Adverse Events				−36.323
User Activities				−10.679
User Attitudes				−26.017
User Attributes				−14.081
Constant	−16.172	−16.412	−21.488	2.895
Observations	288	288	288	453
R^2	0.099	0.162	0.176	0.192
Adjusted R^2	0.096	0.138	0.149	0.151
Note:				$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 4: Linear regression tables for all papers.

	Dependent variable:			
	citeNum			
	(1)	(2)	(3)	(4)
Years Published	16.552^{***}	16.515^{***}	16.774^{***}	19.774^{***}
FC		-19.837^{**}	-19.661^{**}	-20.705^{*}
IMC		−12.393	-16.079^{*}	−13.840
NDSS		−0.142	−1.269	1.149
IEEE S&P		46.035^{***}	44.973^{***}	27.904^{***}
USENIX Security		13.805^{*}	11.958^{*}	−3.012
WEIS		-28.671^{**}	-29.128^{***}	-35.974^{**}
Workshops		-38.876^{***}	-39.534^{***}	-46.114^{***}
Created Not Public			−1.267	
Created Public			27.587^{***}	22.105^{**}
Only Existing Data			0.505	3.147
Vulnerabilities				−17.684
Exploits				−28.089
Cybercrime Inf.				−3.910
Alerts				−28.798
Configurations				−19.208
Applications				−7.488
Network Traces				-26.889^{**}
Topology				-32.887^{**}
Benchmarks				-29.999^{*}
Adverse Events				−20.551
User Activities				−5.372
User Attitudes				−17.022
User Attributes				−13.265
Constant	−0.029	0.271	−0.998	8.460
Observations	957	957	957	702
R^2	0.099	0.186	0.194	0.193
Adjusted R^2	0.098	0.179	0.184	0.166
Note:				$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

higher for model 4 because the unit of analysis is datasets in order to compare citations by dataset category.

We note that the boost to R^2 from models 3 and 4 is modest. Consequently, while we can conclude that making datasets public and the type of dataset created do meaningfully affect citation rates, there is much unexplained variance in citation rates beyond what is captured by these simple regression models. This is to be expected, since there are many characteristics of papers that affect their likelihood to be cited that we do not consider (e.g, topical relevance, author reputation and influence, media attention).

Table 4 presents four more regressions, the only difference being that we now consider all papers, not only those that create datasets. We construct a categorical variable about dataset usage in papers with four values: no data (baseline), created not public, created public, or only using existing data. The findings for the other variables is consistent with the first regression, with a few more publication venue variables gaining significance.

Papers that create and make datasets public are once again more likely to be cited, but this time it is relative to papers with no datasets. Also, the citations for papers that create datasets but do not make them public and papers that only use existing datasets but do not create their own are indistinguishable from papers without datasets.

Of course, citing a paper and using a dataset in research are not the same thing. We defer to future work ascertaining whether or not the rise in citation reflects authors reusing public datasets in their own research. Unfortunately, it is currently difficult to automatically determine whether the dataset has been used directly since the cybersecurity research community has not established norms to cite datasets rather than papers. More widespread use of DOIs for datasets, as is common in some social sciences, could one day make this possible.

5 Conclusion

While datasets are recognized as valuable to cybersecurity research, there has been little work that examines what datasets are created, how they are used, and how to encourage more sharing among researchers. This paper makes contributions towards improving our understanding of each of these points. By examining nearly 1,000 papers, we have taken a data-driven approach to constructing a taxonomy of cybersecurity research datasets. We then apply that taxonomy to the labeled datasets, shedding light on which types of data are being created, used and shared with the broader community.

Some findings underscore the disincentives to make datasets publicly available to others. While 76% of existing datasets used by researchers are publicly available, just 18% of created datasets return the favor. Despite increasing attention being paid to cybersecurity research involving data and exhortations to share data publicly, the proportion of datasets that are shared publicly has remained consistently low.

So what can be done to disrupt the status quo? Paying attention to the incentives to share by eliminating barriers and rewarding publication is key. This paper has made a start in that direction by identifying that citation rates are higher for papers that make created datasets publicly available. To the extent that this finding and subsequent research can shift the narrative about data sharing away from community service towards it being individually rational, we believe more researchers will elect to publish datasets and the science of security can be advanced.

Acknowledgements We thank the anonymous reviewers and shepherd Fanny Lalonde Lévesque for their constructive feedback. We are also grateful to the assistance of Michael Collett in reviewing the data created for this paper.

This material is based on research sponsored by DHS Office of S&T under agreement number FA8750-17-2-0148. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DHS Office of S&T or the U.S. Government.

References

[1] Alexa. Top 500 websites. `https://www.alexa.com/topsites`. Last accessed May 10, 2018.

[2] R. Anderson. Why information security is hard – an economic perspective. In *Annual Computer Security Applications Conference*, Washington, DC, USA, 2001. IEEE Computer Society.

[3] R. Anderson and T. Moore. The economics of information security. *Science*, 314(5799):610–613, 2006.

[4] L. Bilge and T. Dumitraş. Before we knew it: An empirical study of zero-day attacks in the real world. In *ACM Conference on Computer and Communications Security (CCS)*, pp. 833–844, New York, NY, USA, 2012. ACM.

[5] S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition, 2009.

[6] CAIDA. Overview of datasets, monitors, and reports. `http://www.caida.org/data/overview/`. Last accessed May 10, 2018.

[7] N. Christin. Traveling the silk road: A measurement analysis of a large anonymous online marketplace. In *International Conference on the World Wide Web*, pp. 213–224, New York, NY, USA, 2013. ACM.

[8] DBLP. Computer science bibiliography. `dblp.uni-trier.de`. Last accessed May 10, 2018.

[9] Z. Durumeric, M. Bailey, and J. A. Halderman. An Internet-wide view of Internet-wide scanning. In *USENIX Security Symposium*, pp. 65–78, San Diego, CA, 2014. USENIX Association.

[10] Z. Durumeric, J. Kasten, M. Bailey, and J. A. Halderman. Analysis of the https certificate ecosystem. In *ACM Internet Measurement Conference (IMC)*, pp. 291–304, New York, NY, USA, 2013. ACM.

[11] E. Gal-Or and A. Ghose. The economic incentives for sharing security information. *Information Systems Research*, 16(2):186–208, 2005.

[12] L. Gordon, M. Loeb, and W. Lucyshyn. Sharing information on computer systems security: An economic analysis. *Journal of Accounting and Public Policy*, 22(6):461–485, 2003.

[13] C. Grajeda, F. Breitinger, and I. Baggili. Availability of datasets for digital forensics and what is missing. *Digit. Investig.*, 22(S):S94–S105, Aug. 2017.

[14] B. Henne, M. Koch, and M. Smith. On the awareness, control and privacy of shared photo metadata. In N. Christin and R. Safavi-Naini, editors, *Financial Cryptography and Data Security*, pp. 77–88, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg.

[15] M. H. Jhaveri, O. Cetin, C. Gañán, T. Moore, and M. V. Eeten. Abuse reporting and the fight against cybercrime. *ACM Computing Surveys (CSUR)*, 49(4):68, 2017.

[16] H. Jiang, Y. Wang, K. Lee, and I. Rhee. Tackling bufferbloat in 3G/4G networks. In *ACM IMC*, pp. 329–342, New York, NY, USA, 2012. ACM.

[17] T. Moore and R. Clayton. The consequence of non-cooperation in the fight against phishing. In *APWG eCrime Researchers Summit*, pp. 1–14. IEEE, 2008.

[18] A. Newell, R. Potharaju, L. Xiang, and C. Nita-Rotaru. On the practicality of integrity attacks on document-level sentiment analysis. In *ACM Workshop on Artificial Intelligent and Security Workshop*, pp. 83–93, New York, NY, USA, 2014. ACM.

[19] D. of Homeland Security. Information marketplace for policy and analysis of cyber-risk and trust. `https://www.impactcybertrust.org`. Last accessed May 10, 2018.

[20] Y. Park, D. McCoy, and E. Shi. Understanding craigslist rental scams. In J. Grossklags and B. Preneel, editors, *Financial Cryptography and Data Security*, pp. 3–21, Berlin, Heidelberg, 2017. Springer Berlin Heidelberg.

[21] C. Sabottke, O. Suciu, and T. Dumitras. Vulnerability disclosure in the age of social media: Exploiting twitter for predicting real-world exploits. In *USENIX Security Symposium*, pp. 1041–1056, Washington, D.C., 2015. USENIX Association.

[22] L. Wang, K. P. Dyer, A. Akella, T. Ristenpart, and T. Shrimpton. Seeing through network-protocol obfuscation. In *ACM CCS*, pp. 57–69, New York, NY, USA, 2015. ACM.