Cybersecurity Research Datasets: Taxonomy and Empirical Analysis

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

The University of Tulsa

USENIX Workshop on Cyber Security Experimentation and Test (CSET)

August 13, 2018

Federal Cybersecurity Research and Development Strategic Plan (2016)

"Sound science in cybersecurity research must have a basis in controlled and well-executed experiments with operational relevance and realism. That requires tools and test environments that **provide access to datasets** at the right scale and fidelity, ensure integrity of the experimental process, and support a broad range of interactions, analysis, and validation methods. **The Federal Government should encourage the sharing of high-fidelity data sets for research**"

But there is a problem

- Incentives for sharing research data are conflicted
  - Sharing often framed as community service or duty
  - Sharing can be time-consuming, costly, erode competitive advantage
  - Benefits perceived to accrue to others
- One potential benefit to sharing: fame and glory!
  - AKA increased citations

Our contributions

1. Empirical analysis of 965 papers for data use, creation, sharing
2. Development of a taxonomy of cybersecurity datasets
3. Measure the rate of public dataset sharing
4. Regression models demonstrate that papers that create public datasets are cited more often

Outline

1. Data Collection Methodology
2. Taxonomy of Cybersecurity Research Datasets
3. Empirical Analysis of Research Datasets

Outline

1. **Data Collection Methodology**
2. Taxonomy of Cybersecurity Research Datasets
3. Empirical Analysis of Research Datasets

Methodology: Data sources

- Sampled from top conferences and specialist workshops from 2012-2016
  - ACM Conference on Computer and Communications Security (CCS)
  - USENIX Security Symposium (USENIX)
  - IEEE Symposium on Security and Privacy (S&P)
  - Network and Distributed System Security Symposium (NDSS)
  - Internet Measurement Conference (IMC)
  - International Conference on Financial Cryptography and Data Security (FC)
  - Workshop on the Economics of Information Security (WEIS)
  - AI & Security Workshop at CCS
  - Cyber Security Experimentation and Test (CSET) Workshop at USENIX Security
  - Workshop on Bitcoin and Blockchain Research at FC (BITCOIN)
- Inspected 965 papers out of 2,037 total

Methodology: Dataset classifier

- Constructed binary classifier
  - **Dataset-related**: use or create at least one dataset
  - **Non-dataset-related**: otherwise
  - Manually labeled 391 papers (209 dataset-related)
  - Random forest using features based on TF-IDF wordlists
  - Used classifier to identify predicted dataset vs. non-dataset papers; all analyzed data was manually verified

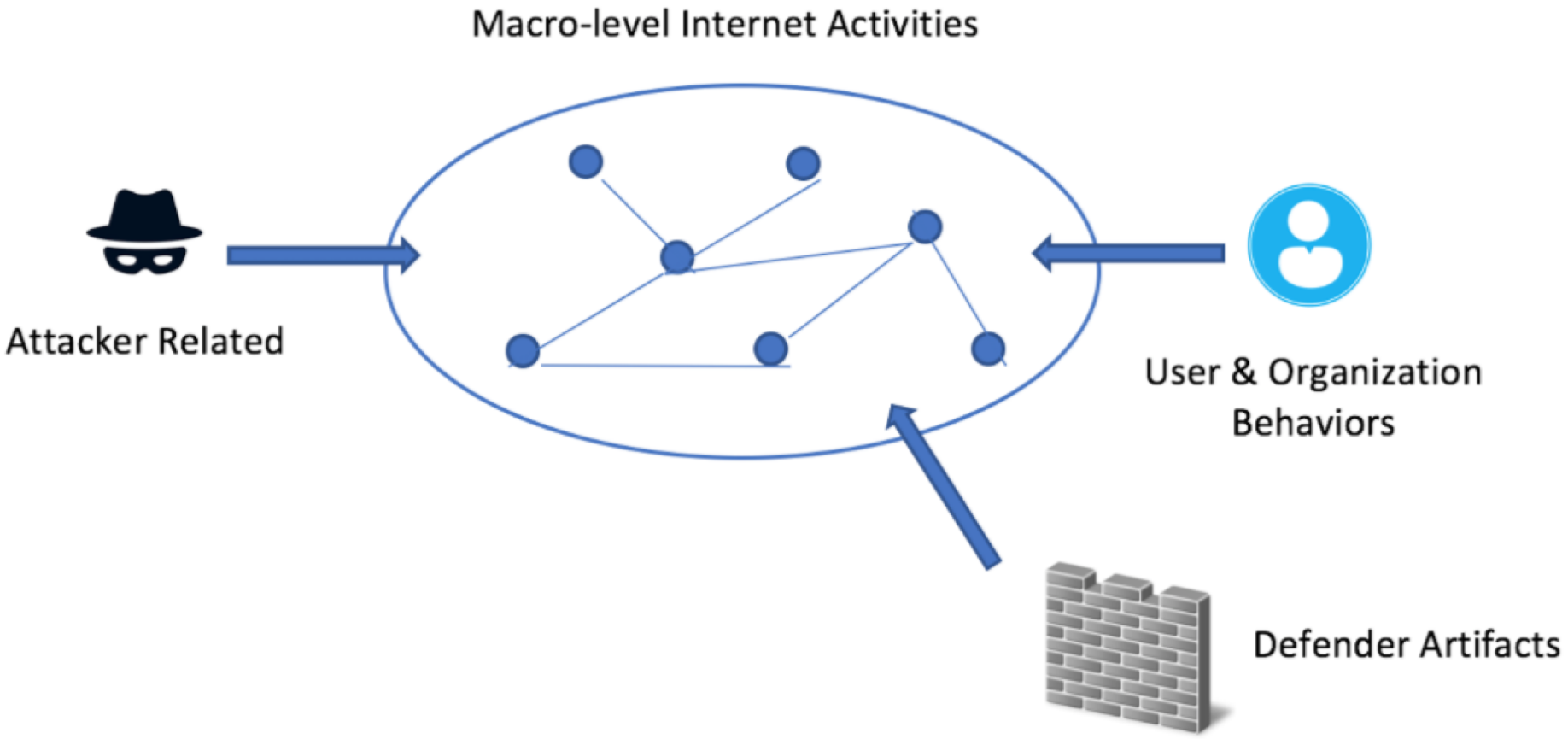
Methodology: Definitions

- **Existing datasets**: already existed before the study undertaken by the research paper
- **Created datasets:** otherwise
  - **Create primary**: generated entirely by the authors without using other datasets as input
  - **Create derivative**: generated from some other datasets
- Papers can involve multiple datasets, both existing and created
- **Public datasets**: paper must explicitly claim that the datasets is publicly available

Outline

1. Data Collection Methodology
2. **Taxonomy of Cybersecurity Research Datasets**
3. Empirical Analysis of Research Datasets

Taxonomy of cybersecurity research datasets: Categories



Taxonomy of cybersecurity research datasets: Subcategories

- **Attacker-Related**
  - Attacks
  - Vulnerabilities
  - Exploits
  - Cybercrime Infrastructure
- **Defender Artifacts**
  - Configurations
  - Alerts
- **Macro-level Internet Characteristics**
  - Applications
  - Network Traces
  - Topology
  - Benchmarks
  - Adverse Events
- **User & Organizational Characteristics**
  - User Activities
  - User Attitudes
  - User Attributes

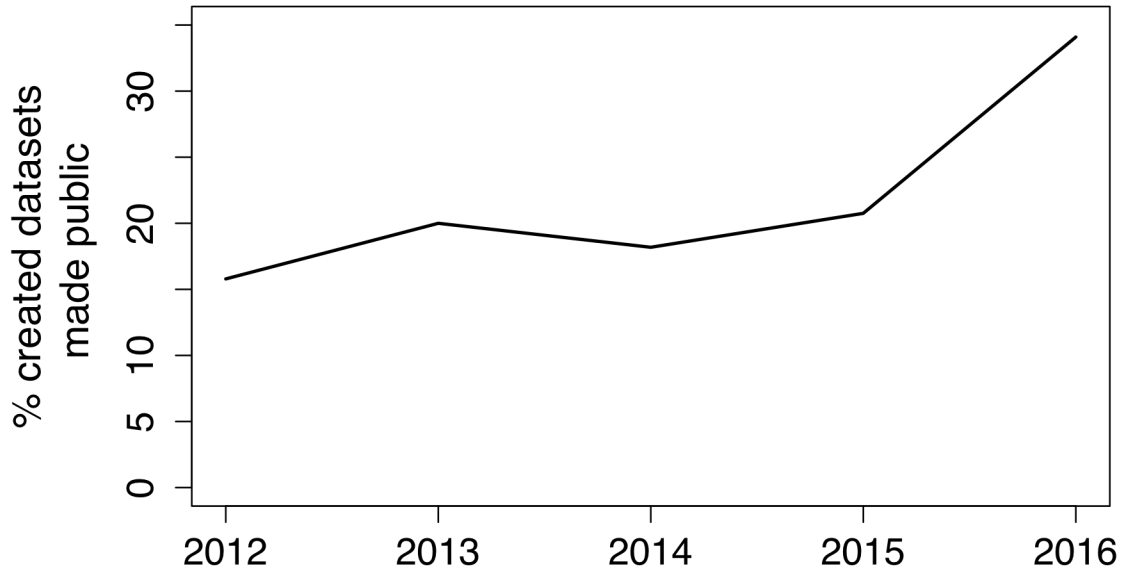
Outline

1. Data Collection Methodology
2. Taxonomy of Cybersecurity Research Datasets
3. **Empirical Analysis of Research Datasets**

Empirical analysis: Making research data public

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

Researchers use public data as input to their research, but don't reciprocate by making their own data public



The research community is *not* getting much better at publishing datasets

Empirical analysis: Dataset categories

	% 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**	(+)

- % Created
  - **Underrepresented (-):** more likely to be used than made
  - **Overrepresented (+):** more likely to be made than used
- % Public
  - **Underrepresented:** less likely to be public
  - **Overrepresented:** more likely to be public

Could citations incentivize publishing datasets?

- Summary statistics are encouraging
  - Papers that **do not involve data** or **only use existing datasets** are cited **10** times per year (median)
  - **9.3** citations per year for papers that **create datasets but don't publish** them
  - Papers that do **publish their data** receive **14.2** citations per year
- To disentangle other explanatory factors, we run regression models

Regression model

- Response variable: # citations
- Explanatory variables
  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 (baseline: ACM CCS)
  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 (baseline: Attacks)

Could citations incentivize publishing datasets?

What factors affect citation rates?

Passage of time

Reputation of publication venue

Creating a Public Dataset

Dataset type

	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

Could citations incentivize publishing datasets?

Making public a created dataset is associated with more citations

Papers with no data, only existing data, or created data kept private are indistinguishable

	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

Discussion

- The huge disconnect between existing (76%) and created (18%) datasets being public is staggering
- The community service narrative for publishing data is not working
- Our findings suggest that narrow self-interest might encourage researchers to publish datasets
- Limitations
  - A lot of unexplained variance in citation rates remains
  - Citing a paper and using created dataset not the same
  - We have demonstrated robust correlation, not causation

Concluding remarks

- We have taken a data-driven approach to building a taxonomy of data created by and used in cybersecurity research
- Researchers who create datasets and make them publicly available get cited more often
- Data and analysis scripts available at doi:10.7910/DVN/4EPUIA
- For more, see: https://tylermoore.utulsa.edu

Questions?

- Thanks
  - Anonymous reviewers and shepherd Fanny Lalonde Lévesque
  - Michael Collett in reviewing the data created for this paper
  - DHS S&T CSD Cyber Risk Economics (CyRIE) Program
    - 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

THE UNIVERSITY of TULSA