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

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

<table>
<thead>
<tr>
<th>Dataset Type</th>
<th>Not Public</th>
<th>Public</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
<td>%</td>
</tr>
<tr>
<td>Created Deriv.</td>
<td>89</td>
<td>85</td>
</tr>
<tr>
<td>Created Prim.</td>
<td>213</td>
<td>81</td>
</tr>
<tr>
<td>Existing</td>
<td>129</td>
<td>24</td>
</tr>
</tbody>
</table>

The research community is *not* getting much better at publishing datasets.
Empirical analysis: Dataset categories

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

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Years Published</td>
<td>23.059***</td>
<td>24.957***</td>
<td>25.619***</td>
<td>24.779***</td>
</tr>
<tr>
<td>IMC</td>
<td>-17.616</td>
<td>-23.730</td>
<td>-20.464</td>
<td></td>
</tr>
<tr>
<td>IEEE S&amp;P</td>
<td>60.211***</td>
<td>55.741**</td>
<td>29.723**</td>
<td></td>
</tr>
<tr>
<td>USENIX Security</td>
<td>4.586</td>
<td>-0.717</td>
<td>-3.582</td>
<td></td>
</tr>
<tr>
<td>WEIS</td>
<td>-25.607</td>
<td>-27.932</td>
<td>-30.750</td>
<td></td>
</tr>
<tr>
<td>Workshops</td>
<td>-46.998**</td>
<td>-48.271**</td>
<td>-54.410***</td>
<td></td>
</tr>
</tbody>
</table>

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
Observations: 288 288 288 453
R²: 0.099 0.162 0.176 0.192
Adjusted R²: 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

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>citeNum</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years Published</td>
<td>16.552***</td>
<td>16.515***</td>
<td>16.774***</td>
<td>19.774***</td>
</tr>
<tr>
<td>FC</td>
<td>-19.837**</td>
<td>-19.661**</td>
<td>-20.705*</td>
<td></td>
</tr>
<tr>
<td>IMC</td>
<td>-12.393</td>
<td>-16.079*</td>
<td>-13.840</td>
<td></td>
</tr>
<tr>
<td>NDSS</td>
<td>-0.142</td>
<td>-1.269</td>
<td>1.149</td>
<td></td>
</tr>
<tr>
<td>IEEE S&amp;P</td>
<td>46.035***</td>
<td>44.973***</td>
<td>27.904***</td>
<td></td>
</tr>
<tr>
<td>USENIX Security</td>
<td>13.805*</td>
<td>11.958*</td>
<td>-3.012</td>
<td></td>
</tr>
<tr>
<td>WEIS</td>
<td>-28.671**</td>
<td>-29.128***</td>
<td>-35.974**</td>
<td></td>
</tr>
<tr>
<td>Workshops</td>
<td>-38.876***</td>
<td>-39.534***</td>
<td>-46.114***</td>
<td></td>
</tr>
<tr>
<td>Created Not Public</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Created Public</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Only Existing Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vulnerabilities</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exploits</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cybercrime Inf.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alerts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Configurations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Applications</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Network Traces</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topology</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benchmarks</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adverse Events</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>User Activities</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>User Attributes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>-0.029</td>
<td>0.271</td>
<td>-0.998</td>
<td>8.460</td>
</tr>
<tr>
<td>Observations</td>
<td>957</td>
<td>957</td>
<td>957</td>
<td>702</td>
</tr>
<tr>
<td>R²</td>
<td>0.099</td>
<td>0.186</td>
<td>0.194</td>
<td>0.193</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>0.098</td>
<td>0.179</td>
<td>0.184</td>
<td>0.166</td>
</tr>
</tbody>
</table>

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