Valuing Cybersecurity Research Datasets

Tyler Moore*, Erin Kenneally+, Michael Collett* and Prakash Thapa*

* Tandy School of Computer Science, The University of Tulsa

+ International Computer Science Institute, Berkeley and
Office of Science & Technology, Department of Homeland Security

Workshop on the Economics of Information Security
June 3, 2019
Harvard University

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
- Cybersecurity *research* datasets often a public good
  - Costs of developing and sharing data is substantial
  - Yet the datasets are typically given away for free
  - But this makes it hard to quantify the value of research data

How much data are researchers sharing?

- We examined 965 leading cyber research publications 2012-16 *[Zheng et al. CSET 2018]*
  - Researchers use public data as input to their research 74% of time
  - When they create data, **make it public only 18% of the time**

Outline

- Economics of supporting cybersecurity research datasets
- Valuing cybersecurity datasets: the case of IMPACT

Outline

- **Economics of supporting cybersecurity research datasets**
- Valuing cybersecurity datasets: the case of IMPACT

Beneficial outcomes of data for cybersecurity research

1. Advancing scientific understanding
2. Cybersecurity-enabling infrastructure
3. Parity
4. Cybersecurity operational support

Barriers to supporting research datasets

- **Legal and ethical risk**
  - Privacy and confidentiality sensitivities often volunteered, and steps to mitigate this (e.g., NDAs, anonymizing datasets) reduce data value
  - Laws designed to mitigate this risk for operational sharing (e.g., CISA) have not brought about wide sharing of data with researchers
- **Direct costs**: Significant but data often expected to be shared for free
- **Value uncertainty**: risks are readily apparent but benefits often aren’t
- **Value asymmetry**: benefits of receiving data greater than receiving
- **Value mismatch:** between data provider wants to collect and what a requester would like to receive

Incentives to supporting research datasets

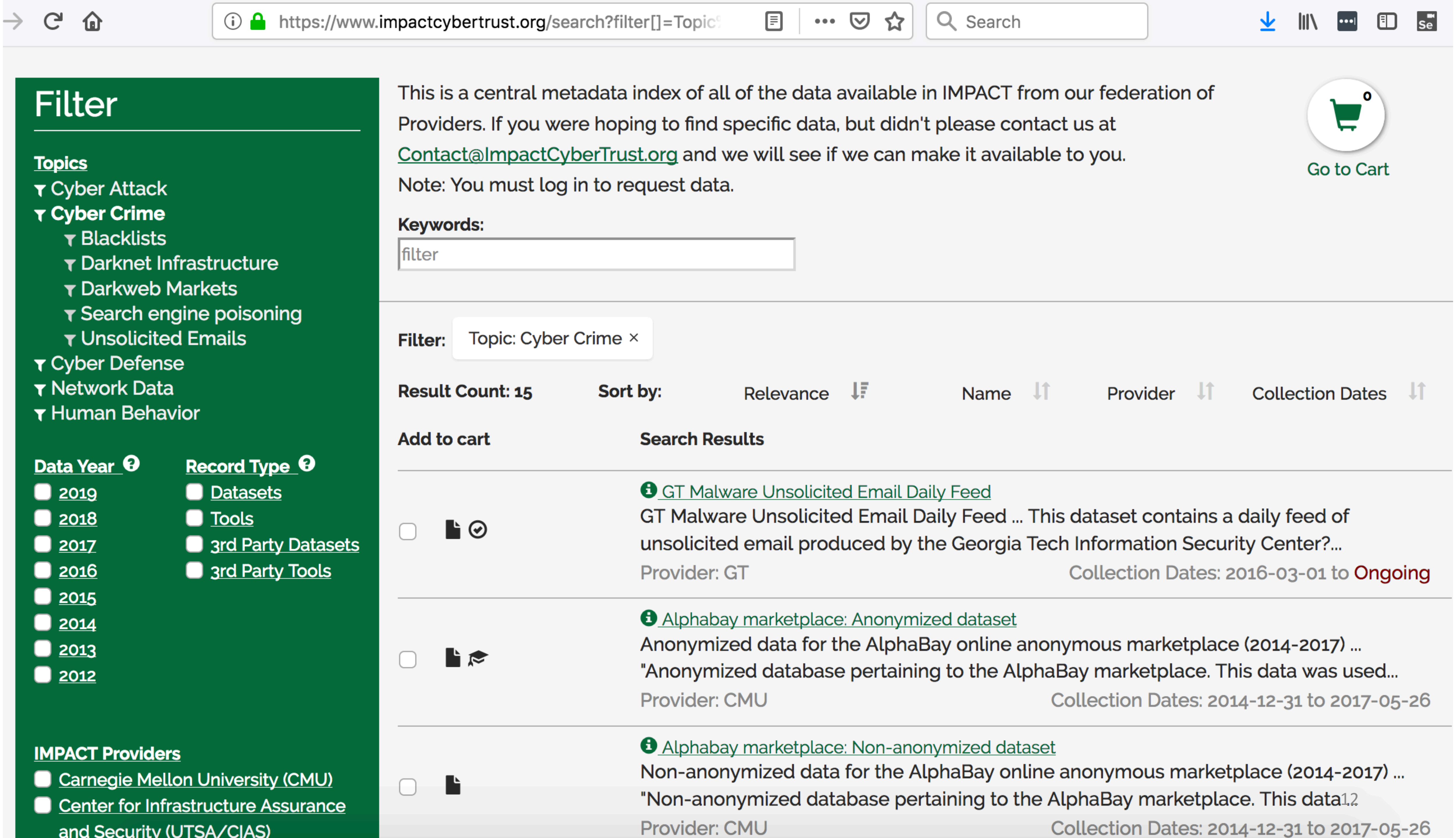
- **Fame and glory**: Researchers who make datasets public get cited more often [Zheng. et al. CSET 2018]
- **Direct compensation:** If governments value cybersecurity research data as a public good, they can pay researchers to create and share it
- **Liability safe harbors**: these exist for operational data (e.g., CISA)
- **Reducing costs to share**: the IMPACT program creates a two-sided platform to make it easier to find datasets, adopt standardized legal agreements, provide basic vetting of researchers

Existing models for supporting research datasets

- **Research student internships**
- **Enclaves**: companies making some data available to vetted external researchers
- **Trade organizations**: Industry orgs (e.g., FS-ISAC, APWG) collect data for operational security but will share with researchers
- **Commecial DaaS providers**: firms that sell threat intelligence feeds
- **Researcher self-publishing**
- **Collaborative platforms for sharing research data**: PREDICT, WOMBAT, Cambridge Cybercrime Centre, IMPACT

Outline

- Economics of supporting cybersecurity research datasets
- **Valuing cybersecurity datasets: the case of IMPACT**



Filter

Topics

- Cyber Attack
- **Cyber Crime**
  - Blacklists
  - Darknet Infrastructure
  - Darkweb Markets
  - Search engine poisoning
  - Unsolicited Emails
- Cyber Defense
- Network Data
- Human Behavior

Data Year

- □ 2019
- □ 2018
- □ 2017
- □ 2016
- □ 2015
- □ 2014
- □ 2013
- □ 2012

Record Type

- □ Datasets
- □ Tools
- □ 3rd Party Datasets
- □ 3rd Party Tools

IMPACT Providers

- □ Carnegie Mellon University (CMU)
- □ Center for Infrastructure Assurance and Security (UTSA/CIAS)

This is a central metadata index of all of the data available in IMPACT from our federation of Providers. If you were hoping to find specific data, but didn't please contact us at Contact@ImpactCyberTrust.org and we will see if we can make it available to you.

Note: You must log in to request data.

Keywords: filter

Filter: Topic: Cyber Crime ×

Result Count: 15 **Sort by:** Relevance Name Provider Collection Dates

Add to cart **Search Results**

□ GT Malware Unsolicited Email Daily Feed
GT Malware Unsolicited Email Daily Feed ... This dataset contains a daily feed of unsolicited email produced by the Georgia Tech Information Security Center?...
Provider: GT Collection Dates: 2016-03-01 to Ongoing

□ Alphabay marketplace: Anonymized dataset
Anonymized data for the AlphaBay online anonymous marketplace (2014-2017) ... "Anonymized database pertaining to the AlphaBay marketplace. This data was used...
Provider: CMU Collection Dates: 2014-12-31 to 2017-05-26

□ Alphabay marketplace: Non-anonymized dataset
Non-anonymized data for the AlphaBay online anonymous marketplace (2014-2017) ... "Non-anonymized database pertaining to the AlphaBay marketplace. This data
Provider: CMU Collection Dates: 2014-12-31 to 2017-05-26

IMPACT

- DHS S&T has operated platform for sharing research data since 2006
  - DHS funds researchers to collect, create, clean and provision datasets
  - Data is free to requesters, who must be researchers from approved countries
  - Researchers create standardized data use agreements, governing IP, usage restrictions, liability limitations, security requirements for data storage, etc.
- For this project, IMPACT program operators shared:
  - All requests for data on the platform since inception
  - Time when datasets are made available
  - *Purpose requests* indicating how the requester intends to use the data
  - Attributes of the dataset (provider, use restrictions, time period)
  - 14 providers made 209 distinct datasets; 2,276 distinct access requests

Regression analysis of dataset requests

- Response variables: (1) total # requests, (2) # approved requests
- Only consider requests on or after Jan 1, 2016
- Explanatory variables
  - **Time available for requests**: # years dataset is available to researchers
  - **Dataset age**: how old, in years, dataset is (time since collection started)
  - **Commercial allowed:** whether dataset permits commercial use (T/F)
  - **Restriction type**: Unrestricted, Quasi-restricted, Restricted
  - **Ongoing collection**: whether data is constantly collected and shared (T/F)
  - **Dataset category**: using taxonomy from Zheng et al. CSET 2018

Analysis of IMPACT Dataset Requests

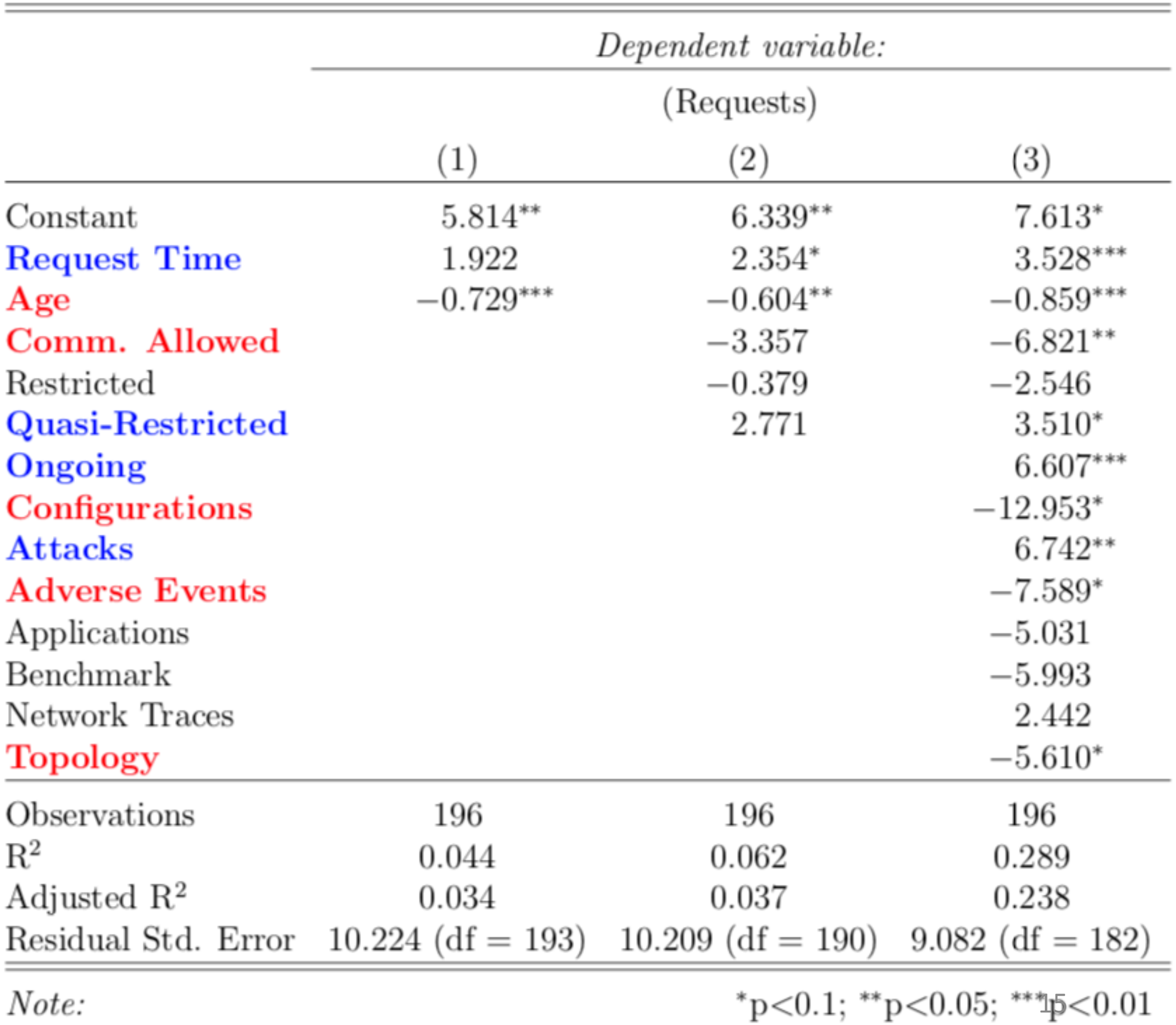
(+): Time available to download

(-): Dataset freshness

Access restrictions

(+): Longitudinal data collection

Dataset type



	Dependent variable:		
	(Requests)		
	(1)	(2)	(3)
Constant	5.814**	6.339**	7.613*
Request Time	1.922	2.354*	3.528***
Age	−0.729***	−0.604**	−0.859***
Comm. Allowed		−3.357	−6.821**
Restricted		−0.379	−2.546
Quasi-Restricted		2.771	3.510*
Ongoing			6.607***
Configurations			−12.953*
Attacks			6.742**
Adverse Events			−7.589*
Applications			−5.031
Benchmark			−5.993
Network Traces			2.442
Topology			−5.610*
Observations	196	196	196
R^2	0.044	0.062	0.289
Adjusted R^2	0.034	0.037	0.238
Residual Std. Error	10.224 (df = 193)	10.209 (df = 190)	9.082 (df = 182)
Note:			*p<0.1; **p<0.05; ***p<0.01

Option 1 for establishing value: how data is used

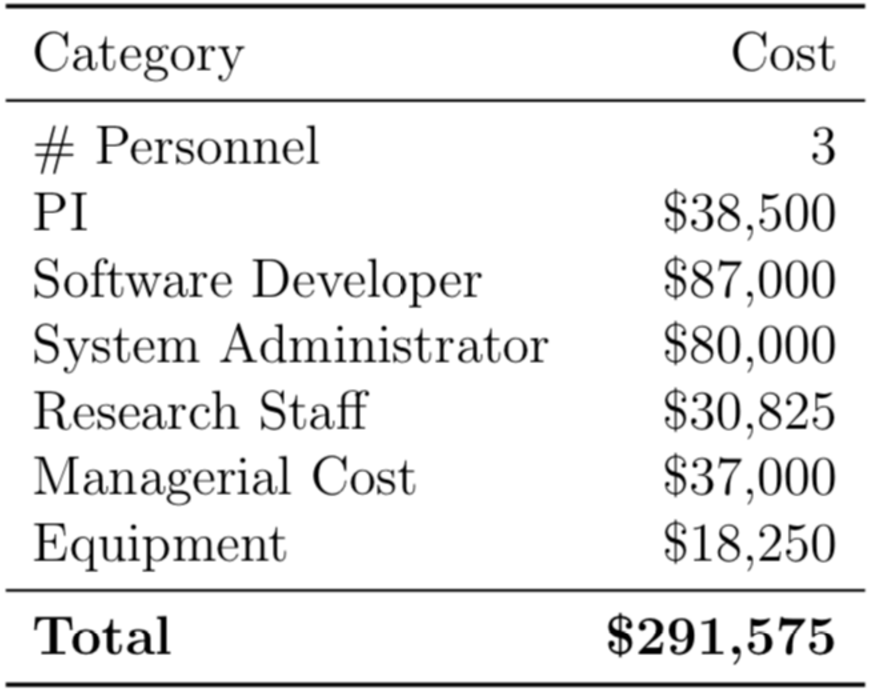
- Technology Evaluation (28%): “evaluate if our new DDoS detection in-line analytical module in NetFlow optimizer can detect this attack”
- Technology Development (28%): “Devise an automated process of detecting and controlling malicious insiders to mitigate risks to org.”
- Data Analysis (31%): “analyze how DDoS affects OS production systems”
- Operational Defense (6%): "provide intelligence on passive DNS malware that can be used to block it from entering my network”
- Education (3%): “develop exercises for an introductory stats and data science course that emphasizes cybersecurity awareness for state of Virginia”

Survey of IMPACT Users

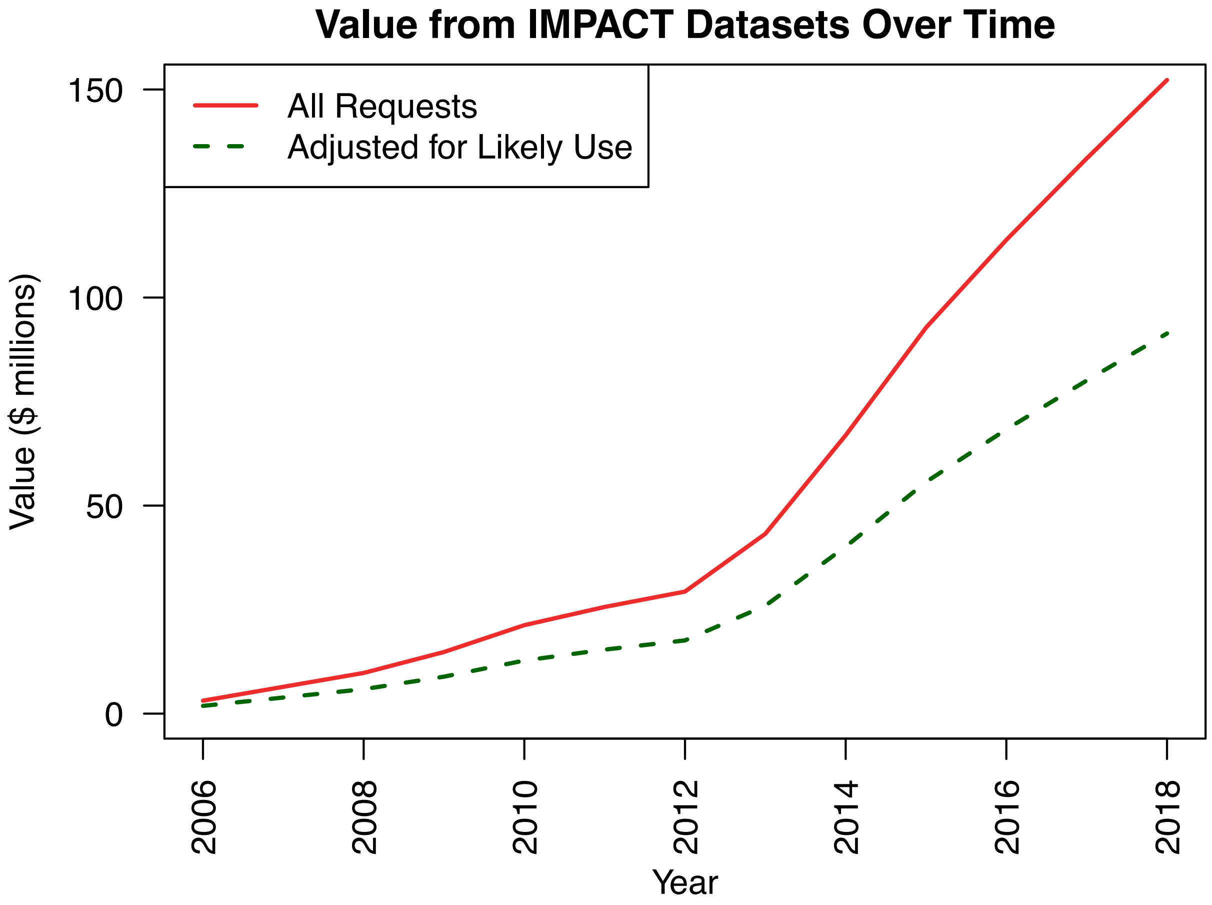
- In Summer 2018, sent email to every requester, referencing the dataset(s) they requested, asked a few questions on use
  - 114 responded. Of these:
  - 60% said they used the dataset requested
  - 91% reported using the data in the same way they requested
  - 72% said they would not have collected the data themselves if it wasn’t available in IMPACT

Option 2 for establishing value: quantifying avoided cost

Median annual cost of providing datasets (8 providers)



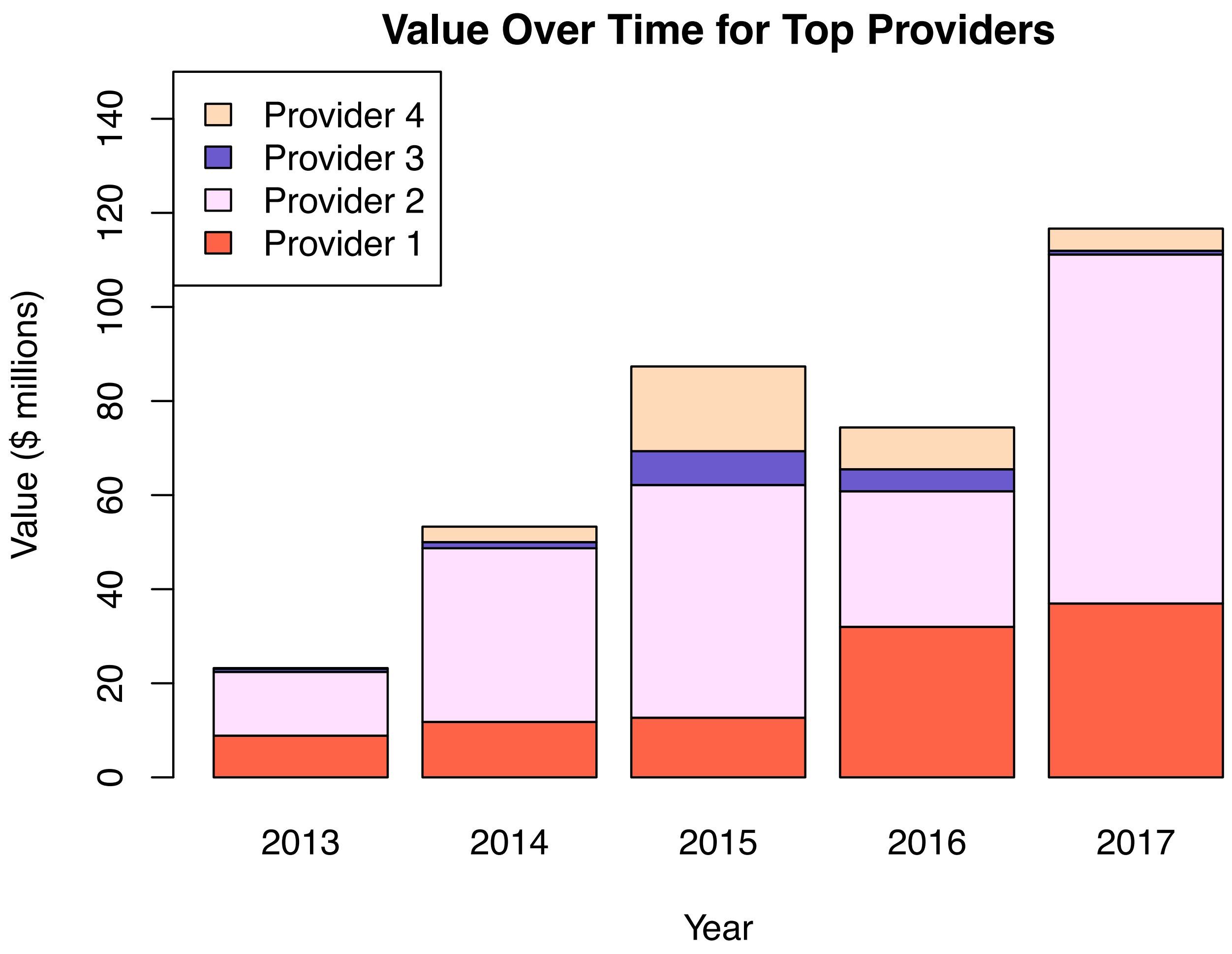
Category	Cost
# Personnel	3
PI	$38,500
Software Developer	$87,000
System Administrator	$80,000
Research Staff	$30,825
Managerial Cost	$37,000
Equipment	$18,250
Total	**$291,575**



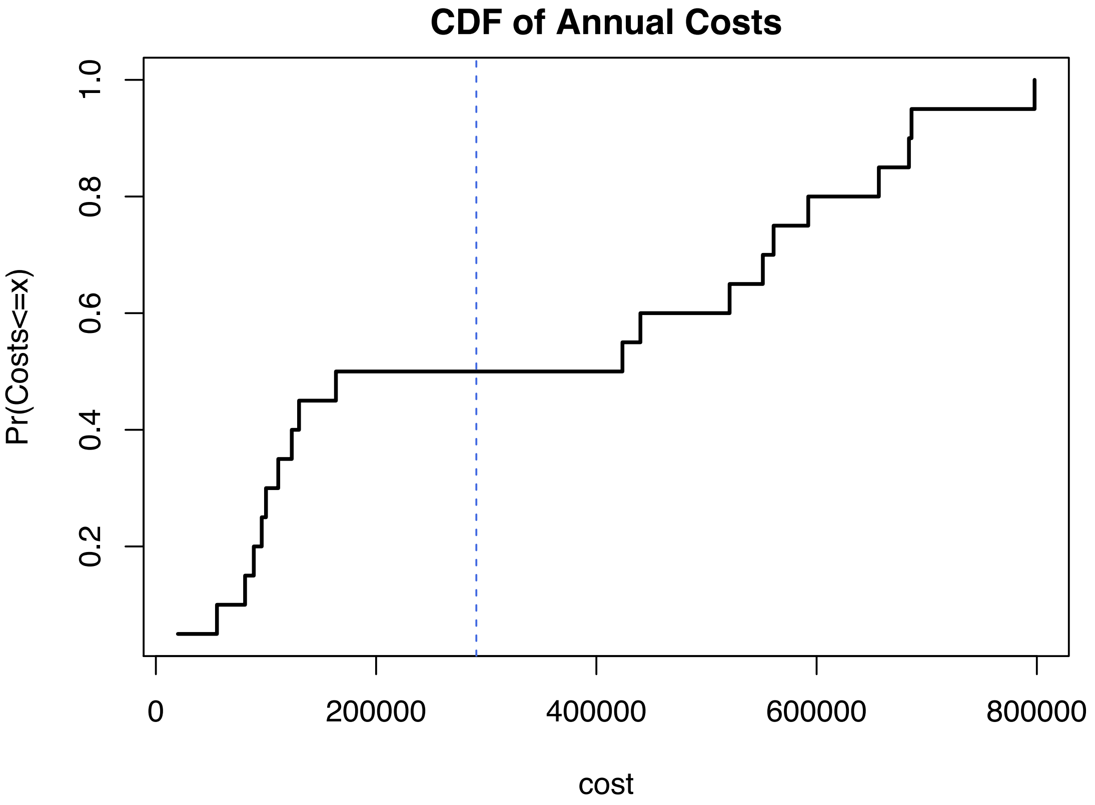
Total value since 2006: $663 million

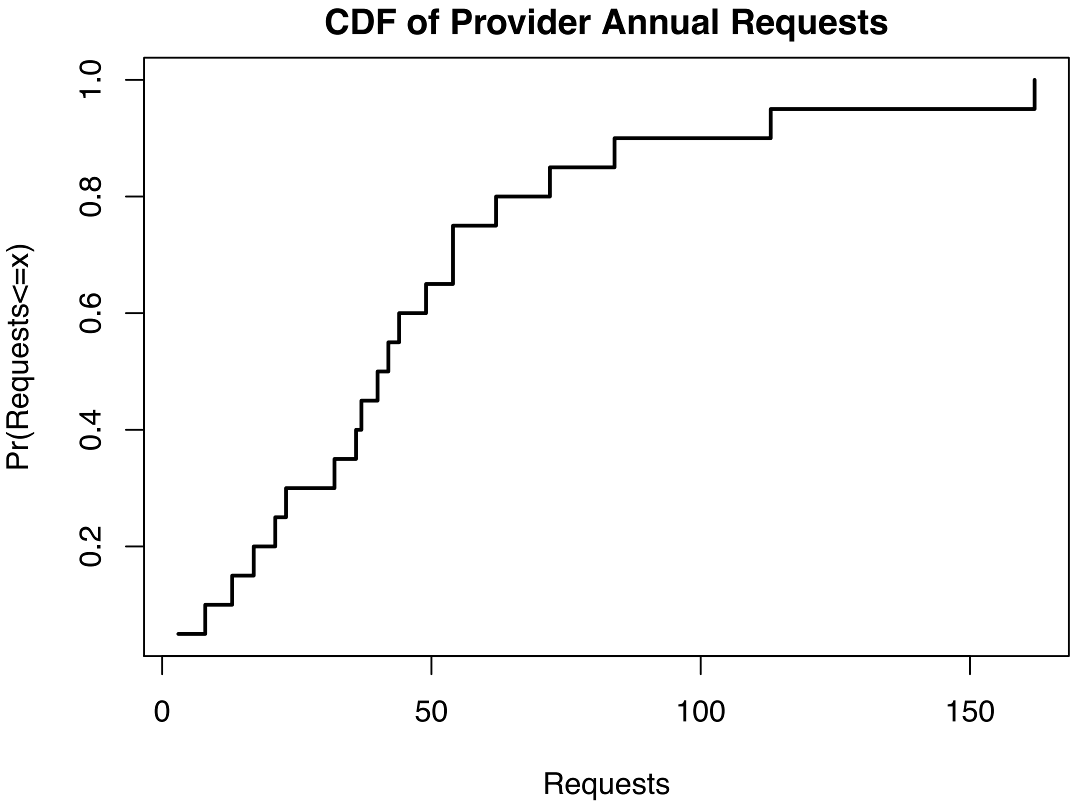
Option 2 for establishing value: quantifying avoided cost

For 4 large providers that provided costs, we can tally their value compared to actual costs



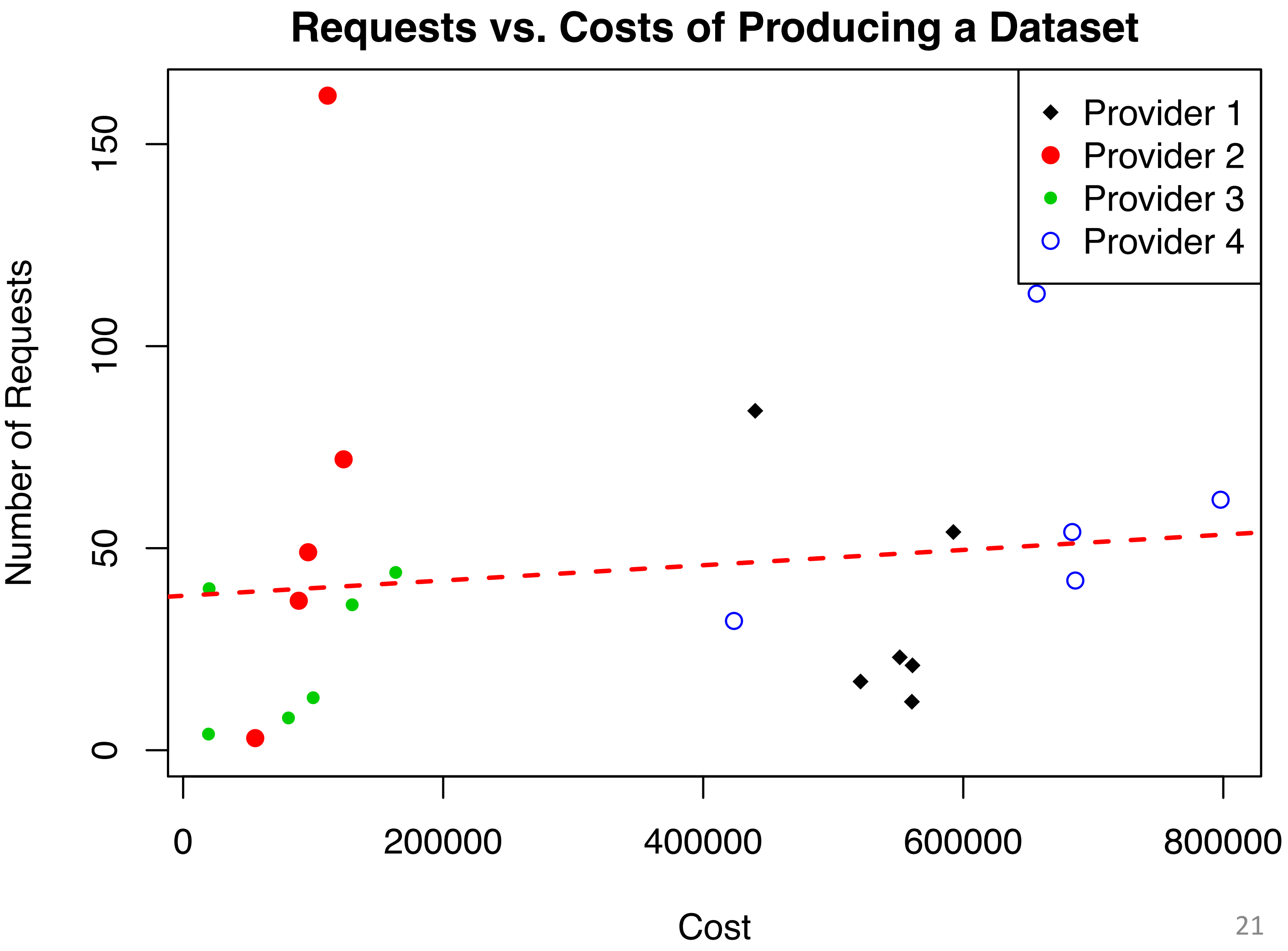
Dataset provisioning costs and demand vary widely





Cost per request varies greatly

Provider	Cost Per Request
P1	\$13 394
P2	\$10 056
P3	\$3 507
P4	\$1 567



Concluding remarks (1)

- While cybersecurity datasets are essential to research, their benefits are not always appreciated
- The paper sets out those benefits, along with barriers to sharing
- We also empirically investigated research dataset sharing on IMPACT
  - Dataset age is negatively correlated with requests
  - Permitting commercial use is negatively correlated with utilization
  - Ongoing collections are more popular
  - Hard to collect, topically relevant and potentially sensitive data more popular

Concluding remarks (2)

- Quantifying value of a public good like free sharing of cyber data is also hard, particularly in $ terms
  - Some progress can be made by treating value as data collection cost avoided
  - Using this method, IMPACT created $663 million in value since inception
  - Myriad uses for data do not easily translate into $
- Little relationship between data provision cost and customer demand
- To increase platform success, identify data users want AND providers can collect
- For more: https://tylermoore.utulsa.edu and https://www.impactcybertrust.org

THE UNIVERSITY of TULSA