Valuing Cybersecurity Research Datasets

Tyler Moore*[1], Erin Kenneally†[2], Michael Collett[1], and Prakash Thapa[1]

[1]Tandy School of Computer Science, The University of Tulsa
[2]International Computer Science Institute, Berkeley and
Office of Science & Technology, Department of Homeland Security

Abstract

Cybersecurity research datasets are incredibly valuable, yet efforts to broaden their availability have had limited success. This paper investigates why and advances understanding of paths forward using empirical data from a successful sharing platform. We start by articulating the benefits of collecting and sharing research datasets, followed by discussing key barriers that inhibit such efforts. Using extensive data on IMPACT, a long-running cybersecurity research data sharing platform, we identify factors that affect the popularity of datasets. We examine over 2,000 written explanations of intended use to identify patterns in how the datasets are used. Finally, we derive a quantitative estimate of the financial value of sharing on the platform based on the costs of collection avoided by requesters.

1 Introduction and background

Data is an essential input to cybersecurity research. It takes many forms, from reports of compromised websites to network topologies, and from geolocations of backbone routers to traces of anonymous marketplaces peddling illegal goods. Whereas historically, the development of security-enabling technologies such as cryptography could be designed from mathematical foundations alone, today's security controls usually require data as input to the technology's design and to evaluate its effectiveness. Ultimately, to improve cybersecurity in the marketplace with scientific backing [2], empirical data must be more democratized.

Researchers have made considerable progress in advancing our scientific understanding of cybersecurity. For example, we know a great deal more about the supply chains underpinning cybercrime [18, 6, 17]. New forms of attacks have been uncovered by researchers, such as malware command-and-control domain infrastructure identified by inspecting passive DNS traces [3] and DDoS amplification attacks [36] . Retrospective analysis of antivirus

*`tyler-moore@utulsa.edu`

†`erink@icsi.berkeley.edu`. The views expressed are those of the author and not that of the Department of Homeland Security Office of S&T or the U.S. Government.

Workshop on the Economics of Information Security (WEIS), Cambridge, MA, June 3–4, 2019.

telemetry data has identified zero-day vulnerabilities and pinpointed the time of exploitation [4]. We also know more about the effectiveness of countermeasures, from the time required to remove phishing websites [26] to time lags in updating compromised certificates from high-profile vulnerabilities [11] to how well notifications sent to webmasters hosting compromised sites work [19]. Researchers have even begun to explore the link between security levels and susceptibility to compromise. For example, researchers have found that network misconfigurations may be predictive of security breach [22].

An analysis of top security publications from 2012 to 2016 has found that around half of inspected papers either used existing datasets as input to their research or created data as a byproduct [45]. However, we note that in most cases, data is collected in an ad hoc, one-off fashion, requiring special arrangements with source companies. The resulting datasets are not further shared. This makes reproduction or replication of results somewhere between difficult and impossible, hindering scientific advances. The practice is inefficient, as efforts are duplicated. Assessments of long-term trends and progress are infeasible because researchers are unable to conduct longitudinal studies. Finally, a dearth of data publication and sharing means that research is either chilled or researchers chase insignificant cybersecurity problems [33]. The aforementioned study of research papers also found that 76% of existing datasets used in papers were public, but only 15% of created datasets were made available. This signals significant structural asymmetries in cybersecurity research data supply and demand. It also underscores the opportunity to assist an underserved market.

This paper sets out to investigate the economics of provisioning cybersecurity research datasets. We enumerate the benefits to wider availability, outline the barriers to achieving that (since the community has been trying for many years with limited success), and identify incentives to change this trajectory. We then empirically examine an exemplar of research data sharing, the IMPACT Program. Using regressions, we identify factors that affect the demand for research datasets. We also investigate how to value the sharing of research data: first, by examining the data request purposes; and second, by quantifying value as costs avoided by the requesters.

Note that there has been considerable attention paid to information sharing among operators through organizations such as ISACs [14, 13, 24, 16]. In contrast, we examine data provisioning done primarily for research purposes. Cybersecurity data resides on a use spectrum – some research data is relevant for operations and vice versa. Yet, as difficult as it can be to make the case for data sharing among operators, its even harder for researchers. Data sharing for research is generally not deemed as important as for operations. Outcomes are not immediately quantifiable.Bridging the gap between operators and researchers, rather than between operators alone, is further wrought with coordination and value challenges. Finally, research data is often a public good, which means it will likely be undervalued by the parties involved. Overcoming this problem requires benefactors whose remit or motivation is to support and protect the collective good (e.g., governments). But benefactors vary in their support for applied research and advanced development and its enabling data infrastructure. All too often, then, support for research data provisioning resides in a purgatory between essential operations and fundamental research.

2 The economics of supporting cybersecurity research datasets

2.1 Beneficial outcomes of data for cybersecurity research

Cybersecurity research data yields many benefits, but they are not monolithic. Instead, they accrue along several, sometimes overlapping, dimensions. Value can vary by stakeholder, be it academic researchers, government or commercial organizations, or society as a whole. Data can provide direct benefits to individual stakeholders. But it can also accrue value to society through its ongoing availability to a broader set of stakeholders [25, 41]. Lastly, there can be derivative beneficial outcomes when the direct outputs from using data are used as input to higher-order challenges, such enterprise cyber risk management or cyber insurance underwriting.

The overarching benefits from expanded access to research data can be summarized as advancing scientific understanding, enabling cybersecurity infrastructure, enhancing parity, and improving operational cybersecurity. We describe each benefit category in turn.

Advancing Scientific Understanding Trust is a benefit that does not easily lend itself to concrete formulae or universal specifications, but scientific methodology has long been one of society's principal proxies for trust since it is predicated on transparency and falsifiable observation, measurement and testing to reach accepted knowledge. Cybersecurity has long lacked reliable metrics and measurements, from quantifying risks to evaluating the effectiveness of countermeasures. Science is the quintessential process by which society can achieve progress, as well as assure objectivity and foster trust. Data is the raw material upon which science subsists. Without data, there can be no systematic advancement of cybersecurity as a computational, engineering, and social science discipline. Without scientific underpinning, we are left with a cybersecurity market built on opinion, conjecture, hyperbole and faith.

In addition, data science[1] and analytics[2] are increasingly generating automated and augmented decisions and actions related to cyber risk management, and are critical to cybersecurity capabilities in a dynamic threat and interconnected world. Cyber risk management demands a more integrated, holistic understanding of the cyber-physical environment. It involves multidimensional data, complex association and fusion of data, and high context presentation. Cybersecurity decisions require abstraction of the low-level knowledge and labor-intensive tasks needed to augment, aggregate, and enrich data. Such tasks are costly to undertake and essential to advancing scientific understanding. Trust in the fairness and reliability of data science and analytics starts with provenance and integrity of the data upon which they are built.

[1] Viz: umbrella set of different techniques to obtain answers that incorporates computer science, predictive analytics, statistics, and machine learning to parse through massive data sets in an effort to establish solutions to problems that haven't been thought of yet.

[2] Viz: subset of data science that focuses on realizing actionable insights that can be applied immediately based on existing queries.

Cybersecurity-enabling Infrastructure Scalable and sustainable availability of data are critical to R&D capabilities. Researchers can get access to certain data at times, but such access is often ad hoc, expensive, and/or dependent on opportunistic relationships with individuals at data-rich companies. Although not always recognized as such, data is itself research- and operations-enabling infrastructure. While the "Big Data" era may in fact spawn a proverbial growth of data on trees relative to the past, extracting value from data in a scalable and sustainable manner demands an infrastructure to pick, sort, truck, process, store, bottle and ship data. Data as enabling infrastructure for research reduces duplication of costs and effort to find, curate, and use that data. Data as infrastructure lowers the barrier to entry to engage innovative research and makes investments in cybersecurity more efficient. A research-enabling data infrastructure reduces the time and cost associated with stewarding data in a manner that is mindful of the associated operational, legal and ethical risks. A sustainable and scalable data infrastructure counteracts the narrow mindset that has defined cybersecurity data sharing heretofore. Information sharing tends to focus on immediate concerns such as cyberattacks and imminent threats; sharing for research addresses longer-term trends, illuminates evolving attacker strategies, and provides a foothold for improvements in defensive technologies. Finally, sharing for research also affects broader facets of cybersecurity – education and training, workforce, controls acquisition, laws, long-term challenges like building security into the design of hardware and software, changing incentives, and developing wider scoping needs and requirements.

Parity Improving availability of data creates several benefits. When data sharing is pervasive, data sources provision and exchange data that might otherwise be left on the cutting room floor. Parity lowers barriers for academic and industrial researchers, cybersecurity technology developers, and decision makers to access ground truth to inform their own work. An ecosystem that relies on data to develop, test and evaluate theories, techniques, products and services works better when there is not a stark gap between the data rich and data poor. Large technology platforms own access to stockpiles of user behavior and infrastructure data which is critical to cybersecurity. They can leverage this information advantage to study evolving attacker strategies and develop more effective countermeasures than smaller rivals. Meanwhile, academic researchers can be severely disadvantaged if not completely shut off from obtaining ground truths about threats, vulnerabilities and assets. The interconnected and interdependent nature of cybersecurity means that cooperation through data sharing is necessary for defenses to be effective.

Data parity diminishes information rent-seeking, thwarts anti-competitive behavior, unencumbers innovation by reducing costs to cybersecurity startups and individual experts, and increases the quality and effectiveness of products and services that are engendered by competition. Higher quality data for research can help correct the negative externalities that arise from organizations' reluctance to share data. Data parity also impacts the efficiency dividends that traditionally define value for organizations: having access to data which is a core substrate to cybersecurity products and services can reduce costs, increase profitability, and possibly introduce new sources of revenue.

Cybersecurity Operational Support What is the difference between the benefits that accrue from data for cybersecurity research and those for operations, and are they mutually exclusive? There has long been a tacit bias, borne out in legislative efforts to encourage data sharing [12], that relegates 'research data' less important than 'operational data' when it comes to prioritizing investments in and support for cybersecurity data sharing. The juxtaposition, however, tees up a false choice. Prioritizing data sharing for operations over research can be likened to expending health care budgets on clinical and emergency room medicine while forgoing preventative medicine. Like the former, data sharing for operations is used for acute, tactical and incident-driven cybersecurity needs. Often it takes the form of indicators of compromise (IOCs) such as IP addresses, URLs, file hashes, domain names, and TTPs. Data for research has typically comprised more longitudinal and broader scale data, such as blackhole address space, BGP routing, honeypot data, IP geolocation mapping, Internet infrastructure data, Internet topology, and traffic flows [30]. The presumptive differences between research and ops data, however, blurs against a canvas of APTs, perimeterless organizations, and advanced analytics. In each case, data needs to be representative of contemporary dynamic threats, traffic and communication patterns, and correlated risks to inform new, effective ways to protect critical information systems and assets. IOC-centric data addresses only part of the picture.

Data for cybersecurity research is increasingly needed to meet the growing needs of owners, operators and protectors of cyber infrastructures for dynamic and responsible operational support. These needs include situational awareness, decision support and optimization, risk modeling and simulation, economic analysis, statistical analysis and scoring, and incident response [38]. These capability needs can be met with research infrastructure that is responsive to the data and analytic requirements that support cyber security operations in a reusable and repeatable manner.

There are many beneficial outcomes for cybersecurity operations that stem from broader availability of research data [39]. Examples of operational benefits include:

- Traffic analysis, network forensic investigation, and real-time network event identification and monitoring (e.g., Internet outage detection, network hijacks) via on-demand query and measurement of streaming data;
- Event reconstruction and threat assessment by correlating data across multiple different sources and timeframes to offer insights and responses to suspected events;
- Tactical and strategic resource allocation for cyber resilience by assessing security and stability properties such as hygiene, robustness, and economic sustainability;
- Cyber risk management at various level by understanding cyber dependencies, risk aggregation, and cascading harm using integrated data (perimeter data like packet capture and firewall logs, internal data like DNS and DHCP logs, and cyber environment data);
- Threat detection by conducting time series analyses over coalesced signals/observed patterns;
- Investments in cybersecurity controls based on benchmark and efficacy measurements.

2.2 Incentives and disincentives to support datasets for research

Appreciating the positive outcomes from sharing data is critical to its broader availability. But achieving that desired end state requires understanding why data sharing for cybersecurity continues to conjure up "Groundhog Day" sentiments despite several decades of dialog extolling its virtues. We therefore turn to the barriers that hinder broader provisioning of data, followed by a discussion of available incentives that can enable more noticeable progress.

2.2.1 Barriers

We characterize barriers as legal and ethical, operational, and value impediments to the availability of data for cybersecurity research.

Legal and Ethical Risk Legal barriers to sharing data invariably top the list of obstructions and are both colloquially and formally recognized as such (see e.g., [21, 40, 5]). In general they comprise privacy and proprietary rights and interests, private contracts, intellectual property rights, data protection laws, and antitrust liability. Federal and state regulations and laws around personal data and communications privacy, consumer protection, and data protection create legal obligations on organizations who collect, use and disclose information that may otherwise be useful for cybersecurity research. Note that these sources of liability are not aimed to prohibit data sharing, per se, but by not carving out exceptions for allowable research they can functionally serve to disincent otherwise lawful data sharing.

Legal liability may also spawn from contracts between and among individuals and organizations which prescribe or proscribe behavior relating to shared data, in which cases clauses related to warranties, terms of service, limitation of liability, and indemnification for harms/damage/loss, and license terms can impede data sharing. While antitrust barriers have been undermined by official policy statements not to mention a paucity of precedent, there nevertheless are some unresolved legal questions about the nexus between sharing cybersecurity information and anti-competition law [34]. Antitrust risk has heretofore arisen in the context of business-to-business sharing of data for tactical cybersecurity operations, not for research purposes. In fact, if companies were to share data for research purposes, this could mitigate against antitrust concerns since presumably the scientific knowledge that is produced would inure to the benefit of consumer welfare and against information asymmetries that characterize and favor anticompetitive behavior.

Privacy and confidentiality sensitivities are a frequently cited disincentive to sharing data. At least for privacy this is owing to a confluence of legitimate privacy risk, evolving applications of privacy law to new technologies, legal conservatism, and/or the opportunistic use of uncertain legal liability as a foil for other motivations not to share. Progress has been made in disentangling privacy-sensitive data from what is needed for cybersecurity, i.e., sharing machine-to-machine data that does not contain first-order personal identifying information. Nonetheless, sensitive data risk resurfaces in the wake of advanced analytic capabilities such as machine learning and other AI-techniques that enable re-identification of pseudonymized data, or that spawn new risks of harm that stem from poorly understood privacy and confidentiality sensitivities created by these analytics [12].

The ability to realize value from shared data can be impeded by techniques or policies that attempt to prevent or mitigate data sensitivity risks. Technically obfuscating sensitive data or invoking data use limitations or NDAs can negatively impact utility of the shared data. For example, anonymizing IP addresses in network traces can hinder the ability to reassemble attack traffic data needed to test and improve new IDS technology. Prohibiting the probing of those IP addresses in a data use agreement may preclude research efforts to detect Internet outages.

Organizational sensitivities surrounding data sharing anchor on the potential exposure of confidential data, such as network configurations, system architectures, security controls, passwords and identifiers, trade secrets, customer or partner relationships, other proprietary financial and business information, and intellectual property (patent, copyright, trade secret). Improper release of this data may raise concerns about shareholder liability, loss of revenue, exposure of vulnerabilities and victimization, or otherwise induce competitive advantages for fellow market contenders. A related, albeit less quantifiable, risk of sharing cybersecurity-relevant data is reputation harm. The archetypal example is borne out in organizations' data breach reporting strategies, where legal mandates to report supersede notions of voluntarily sharing in support of collective defense. Here, organizations regularly weigh the costs of compliance with breach laws versus the impact of notification on revenue, sales and stock prices.

In addition to the plethora of legal risks related to proscriptions on sharing certain data, few laws actually encourage data sharing by neutralizing those liability concerns, and even then the focus is not on data for cybersecurity research purposes (e.g., [42, 27]). Industry does not usually share real, high fidelity data with researchers. There are exceptional cases where sensitive data is made available by organizations to specific researchers. However, these one-off, ad hoc situations do little to advance trusted, collective use of data. Besides the limited availability, there is no opportunity to peer review, hold results to account, or leverage the data to improve upon similarly situated efforts. These situations fail to establish sharing precedent that would help lower the risk perceptions and realities of data sharing, and mitigate some of the barriers [38, pp. 35–36].

Ethical risk may arise from the nature of the collection, use or disclosure of shared data. Ethical risk can spur legal liability when ethical obligations have been codified into law, as in the U.S. with the Common Rule and 45 C.F.R. 47, which requires any researcher receiving federal funds to abide by protections it establishes for research involving human subjects. A major challenge in cybersecurity research is whether it involves humans and triggers ethical oversight, or as is often argued, non-human machine research that is exempt from oversight. ICT research ethics challenges and guidelines are well-documented in the seminal Menlo Report [9]. Even if cybersecurity research is technically exempt under a strict interpretation of human subjects research, nevertheless ethical risk arises when research involves potentially human-harming activity such as interactions with malware that controls compromised user devices, embedded medical devices controlling biological functions, or process controllers for critical infrastructure.

Direct Costs Engaging data for research can have nontrivial direct financial costs. With the exception of data that is shared on a one-off, acute basis, technical infrastructure costs can impede research data collection and sharing. These can accrue to both data providers and user-recipients.

At a fundamental level data is not cost-free, and all sharing barriers can be boiled down to economic consequences. Even most data sharing efforts focused on tactical operations come at a cost, be it the price for direct data acquisition, membership in ISAOs or ISACs (e.g., $10,000 to $100,000 according to [20]), threat feed subscriptions, personnel to administer the data, and/or infrastructure to appropriately use the data. Certainly, advances in technology have created unprecedented amounts of data raw materials which in theory should lessen the need for data sharing. Yet there are undoubtedly resource requirements in dealing with real world data sets: finding, collecting, generating, preparing, storing, understanding and using the data. These include data storage and computation, semantically effective data searches, curation and annotation of noisy data, and cross-validation of data with limited provenance. Qualitative and quantitative data for effective cybersecurity demands infrastructure to make it actionable. As with our terrestrial roads, bridges, and waterways, digital infrastructure does not exist via assumed affordances, rather, deliberate resource expenditures. While this may not be revelatory, the research that often demands larger-scale, longer-term empirical data requires the equivalent in investment.

The problem is that cybersecurity research data is a club good, and often provisioned as a public good. Data is inherently non-rival. By design, in order to promote parity and advance scientific understanding, it is also often made non-excludable. Many research datasets are given away for free. When this happens, research data becomes undervalued and under provisioned, unless an entity is willing to underwrite the cost to society's benefit. In the absence of a benefactor, one could restrict access to those who are willing to pay for it. But this is problematic, since most researchers work in academic or other non-profit settings.

Value Uncertainty, Asymmetry and Misalignment While the benefits of data sharing to support tactical operations is often readily apparent, the benefits of sharing for research can be latent, indirect and correlative. When faced with situations where the risks and cost of sharing are direct, foreseeable, and causal (e.g., legal liability), behavioral economics tells us that people will do what is less uncertain. Here that means erring on the side of not sharing data when the countervailing benefits are not articulated or persuasive relative to costs [35, pp. 9–10].

The difficulty in realizing benefits from sharing data may also dissuade efforts. Effectuating value and avoiding harm from shared data is a contextual endeavor which involves understanding the utility profile for the shared data. Consider the following dimensions of data that can affect how to value sharing outcomes: duration (e.g., multi-year timescale attack traffic are needed for trend analysis but irrelevant for near real time incident response); timeliness (e.g., delayed sharing may be unhelpful, real time may not be actionable); detail (e.g., different users have different needs from broad policies and events, to incidents and IOCs, noting that even IOCs without context may have lower value); sensitivity (e.g., whether data is classified, confidential, proprietary, or personal will impact its availability); purpose (e.g., stakeholders have varying needs from situational awareness, specific defensive actions/measures, planning, and capacity building; noting that even threat signatures for attacks on specific networks or assets will not necessarily transfer to others); processing maturity (e.g., whether the data needs additional curation and processing to be valuable, such as raw data versus derivative dataset); and audience (e.g., public researchers will have different needs and disclosure controls than industry consortia) [7]. In other words, articulating value up front is rarely enough. The task's complexity often inevitably introduces knowledge and administrative friction that can be a barrier to sharing.

Just as stating value up front can be hard, so is articulating the harm caused by not sharing in advance. Proving a negative – that the sharing will not cause undue harm – can be impossible. Regrettably, it becomes that much easier to conclude that the cost of sharing likely outweighs the benefits.

Even when research benefits are palpable, they often accrue asymmetrically between data providers and seekers, thereby disincentivizing sharing. Some entities find that the benefits of receiving outweigh the benefits of providing data. This "free riding" can be a barrier to sharing and is not uncommon for social goods. [3]

Value mismatches may arise between the type of data researchers produce and the needs of recipients. As previously mentioned, most sharing occurs with tactical or breach-certain information between and among private companies and the government. There is very little sustainable sharing done with individual researchers or non-commercial research institutions for research purposes. "[R]esearch in cybersecurity requires realistic experimental data which emulates insider threat, external adversary activities, and defensive behavior, in terms of both technological systems and human decision making." [33]. The relevance and quality of shared data can be a barrier. Simply put, there can be a mismatch between what a data provider can and wants to generate and what a requester needs.

Even when the data is of interest, the collection, curation and/or provisioning process and workflow might not align with the requester's consumption capabilities. For example, resilience or outage detection research may need to be accessed dynamically via API rather than be downloaded in raw form from static repositories [39]. Similarly, network attack traffic needs to be labeled when it is provisioned to make it useful for researchers applying artificial intelligence techniques. High volume data may be difficult for the recipient to receive and process, or data may need to be transformed or combined prior to analysis. Sometimes the mismatch arises from a lack of agreeable legal and technical standards– both semantic (e.g., ontologies) and syntactic (e.g., schemas or APIs) [28, 31].

2.2.2 Incentives

Incentives to share data for research are those that lower the barrier to entry for cybersecurity R&D and address the operational, legal, and administrative costs that otherwise impede the scalable and sustainable data sharing needed to enable higher quality cybersecurity innovation in a responsible manner. We challenge assumptions that incentives to share both research and operational data are sufficient, and that organizations will embrace data sharing in light of general acknowledgement that it is critically lacking. The incentives to sharing largely mirror the barriers discussed above. Fundamentally, there is a need to align incentives between producers, seekers and beneficiaries of shared data for research. Sharing for

[3] See, for example, [43, 44]. Regarding the data breaches federal employees' information revealed in June 2015 by the Office of Personnel Management, it is not clear that specific information about the threat or even defensive measures would have resulted in effective defense against the attacks.

operational cybersecurity suffers from misaligned incentives, so support for research is more attenuated given its different value dividends with cybersecurity research spend, including expending resources to support sharing. In the operational realm, for example, companies that suffer a cybersecurity breach such as the theft of credit card information do not pay the full cost of the breach. As well, software companies are primarily driven by time-to-market pressures which come at the expense of cybersecurity needs to immediately fix security and other bugs.

On the data supply side, the most obvious yet arguably difficult incentive to effectuate is direct economic investment in large-scale, long-term and freely available data. Described more fully in the next section, the IMPACT Program provides a unique example of how funding to support data infrastructure addresses global cybersecurity research data needs. Regarding incentives on the demand side, the monetary investment in data sharing organizations (e.g., less than $100K [20]) can be much more cost effective than purchasing MSSP services. It is worth noting that the cost of providing information, including joining a specialized sharing organization, is likely to be less than $100,000[4].

Currently law and regulation does not create data sharing incentives. Few laws or regulations directly encourage data sharing. Nevertheless, calls by industry for liability safe harbors are manifest (e.g., those provided by the Cyber Information Sharing Act), thus supporting the claims that offering protections would help assuage anxiety about legal risk with data sharing. While often viewed as a stick rather than a carrot, regulations such as data breach notification laws and the SEC's requirement to disclose "material information" on cyber risks serves as a forcing function to engender publish and share data.

Lacking hard enforcement to share data, levers to encourage data sharing anchor on reciprocity, reputation, and retribution. There are few rewards for organizations who share data, but positive public relations and attribution in publications that cite shared data can cultivate reputations as good corporate citizens or achieving corporate responsibility. The equivalent on the research side are the reputational benefits that come from increased citations if shared data is referenced in derivative papers that use that data [45]. Data providers are incentivized to continue doing so if they likewise receive some benefit such as feedback on the utility of the data or perhaps getting access to data that would otherwise be unavailable without recipient stakeholders' recognition that reciprocity creates network effects. The threat of retribution might also encourage multilateral sharing. Examples include negative publicity or "peer shaming" when terms of shared data are violated or data sharing is otherwise exploited.

Economic and collective security objectives can incentivize data sharing. Fostering a longer-term secure infrastructure and economic growth is not antithetical to the notion that maximizing shareholder value means employing any means to increase stock price. On the contrary, if the value that flows from sharing data for cybersecurity (see Section 2.1) lowers operational, financial, reputational, or public relations costs or increases revenues, there is a strong argument that public organizations are fulfilling their obligations to shareholders by spending on cybersecurity viz data sharing.

At the operational process and legal level, the IMPACT Program serves as a good example of how some barriers can be overcome. This model enables data providers to leverage

[4] See, for example, Financial Services ISAC, Membership Benefits at `https://www.fsisac.com/join`.

standardized data use agreements that allow for customized additional data restrictions by the Provider. Common features of its data use agreements include:

- IP rights protections for providers; purpose limitations for use of data, and duration limitations;
- balanced liability limitations;
- strong privacy and security requirements for data storage, including use of encryption;
- requirements for the destruction of data at the conclusion of the research;
- ownership and control of data resides with providers, who host and provision their own data.

Furthermore, balancing utility and data sensitivity is achieved via technical and policy controls. Providers can engage disclosure control-as-a-service for very sensitive data that allows analysis without the recipients seeing the sensitive raw data (e.g., SGX enclaves, multiparty computation) . Furthermore, oversight and accountability measures such as vetting the legitimacy of the sharing participants and data provenance helps establish trust that is often needed to enable sharing. In short, models that have successfully operationalized data sharing for research can incentivize replication and further investment. While IMPACT does succeed in reducing these barriers, its approach has been to treat cybersecurity research data as a public good in which the U.S. government subsidizes its creation by funding data providers and offering the data to users for free.

3 Existing models for supporting research datasets

A number of models have been attempted to support cybersecurity research datasets, each with their own advantages and drawbacks. We briefly review several of them here, in light of the preceding discussion on the value, barriers and incentives associated with sharing.

Research student internships Perhaps the most tried-and-true method for sharing data between industry and researchers is to temporarily hire research staff at the firm who has raw data available for collection and analysis. Ph.D. students regularly spend months working at companies so that they might work on a project of mutual interest to the company and researcher. Becoming an employee sidesteps thorny issues such as seeking legal permission to share and quantifying values and risks that are more often necessary when working with outsiders. The downside of course is that the data itself is typically not shared and cannot be used beyond the project for which it was originally collected.

Enclaves Some companies have made portions of their data available to vetted external researchers on request. Perhaps the best-known example is the Symantec WINE program [10], which made antivirus telemetry data available to run experiments. Unfortunately, these programs have struggled to meet the demand from users and are often short-lived.

Trade organizations Most industry organizations, such as ISACs, that collect and share operational data only do so between industry members. A few, however, also make their data available to researchers. For example, the Anti-Phishing Working Group has regularly shared its phishing URL blacklists with researchers who request access. Similarly, the Shadowserver Foundation and Spamhaus regularly share abuse data with vetted researchers.

Commercial DaaS providers Industry data providers such as Farsight sell threat intelligence feeds to private customers. They also share data with researchers, who often elect to share operational data from their organizations back to the commercial operators in appreciation.

Information sharing and analysis centers Significant data sharing takes place at sector-specific information sharing and analysis centers and organizations (ISACs and ISAOs). However, to date, these organizations have focused on data sharing between operators within the same sector, as opposed to sharing with outside researchers.

Open data The Open Data model primarily concerns access to certain government data and is premised on transparent and free availability of some data for use and republication by anyone without intellectual property or other control restrictions. This model faces technical barriers such as data processing difficulties, API deficiencies, lack of machine readable formats, sophistication needed to link and fuse data, and a lack of integrated tool sets to combine data from different data providers [46]. Other infrastructure challenges include access administration, storage, integration, and data analysis [15].

Researcher self-publishing Some self-motivated researchers elect to publish datasets on their own, either by self-hosting on websites or by partnering with organizations such as the Harvard Dataverse. Such activity is comparatively rare because only public data can be shared and because norms to share data have not taken root in the academic cybersecurity research community. Even when it does occur, such publishing is often short-lived and typically does not support ongoing data publication.

Government-facilitated sharing Governments can support data sharing beyond the unilateral Open Data publication model. In addition to fostering cybersecurity data sharing by directly funding the IMPACT R&D-enabling infrastructure described above, DHS champions multilateral operational sharing between and among civil society and governments [8]. Two notable models are the Cyber Information Sharing and Collaboration Program (CISCP) and the Automated Indicator Sharing (AIS) program. CISCP involves private sector participant organizations voluntarily submitting cybersecurity data that is subsequently analyzed and context-enhanced to provide recipients with more appropriate threat assessment and response. In contrast to CISCP's low-volume, deliberate curation approach, AIS tries to commoditize cyber threat indicator sharing using more automated processes to facilitate quantitatively broader sharing. AIS participants include Federal departments and agencies, state, local, tribal, and territorial governments, private sector entities, information sharing and analysis centers and organizations, foreign governments and companies.

Critical analysis of these models is beyond the scope of this paper because they are not research-focused. It is instructive, however, to consider how publicized shortcomings of these approaches might be attenuated by a complementary cybersecurity research data sharing regime. Sharing threat intelligence with the private sector at the DHS is hamstrung by prioritizing automated ingestion and speed of release over qualitative context-enhancement, and because there's a failure to integrate relevant databases [32]. Furthermore, only six non-federal entities share data with DHS via AIS, for example [23]. The result is an incomplete picture of risk exposure and insufficient details to be actionable.

Collaborative platforms for sharing research data Over the past 10–15 years, a few attempts have been made to collect and disseminate cybersecurity research data by establishing a dedicated platform to do so. The first attempt was PREDICT (the predecessor to IMPACT), an effort launched in 2006 [37]. PREDICT sought to reduce legal and technical barriers to sharing data by establishing unified agreements and serving as a clearinghouse of disparate datasets. Additional efforts have been funded as research projects by governments to collect relevant cybersecurity datasets and make the collected data more broadly available (e.g., the WOMBAT project [1], the Cambridge Cybercrime Centre [29]). Because these programs are in effect providing public goods free of charge, their continued operation requires support from a benefactor, typically a government research program.

4 Valuing cybersecurity research datasets: The case of IMPACT

We now investigate more closely IMPACT, a notable platform that disseminates cybersecurity research datasets and which has been supported for over a decade by the Department of Homeland Security, Science & Technology Directorate. Cybersecurity data provisioning can be thought of as a two-sided market that must satisfy incentives for both the producers of relevant datasets and consumers of such datasets. The IMPACT Program has funded cybersecurity researchers to undertake the significant steps of collecting or creating, cleaning, and finally making cybersecurity-related data available for free to qualified researchers. The programs federated technical distribution model achieves scalable and sustainable sharing via normalized legal agreements and centralized administrative processes, including vetting prospective researchers, datasets and providers.

The operators of the IMPACT program have shared with us information on dataset requests, namely:

1. all requests for data made to the platform, from its inception in 2006 through September 30, 2018;

2. time when datasets are made available;

3. *purpose requests* in which the requester outlines its intended use in free-form text;

4. attributes of the dataset (e.g., provider, restrictions on use, time period of collection).

Table 1: Linear regression tables for all requests (left) and approved requests (right)

	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

	Dependent variable:		
	(Approved)		
	(1)	(2)	(3)
Constant	4.640**	5.584**	5.915*
Request Time	1.524	1.929*	3.002***
Age	−0.653***	−0.535**	−0.748***
Comm. Allowed		−2.385	−4.885**
Restricted		−3.204**	−5.269***
Quasi-Restricted		1.832	2.369
Ongoing			5.054***
Configurations			−9.424
Attacks			6.203**
Adverse Events			−6.538*
Applications			−2.698
Benchmark			−4.253
Network Traces			2.615
Topology			−4.536*
Observations	196	196	196
R^2	0.053	0.107	0.342
Adjusted R^2	0.043	0.083	0.295
Residual Std. Error	8.505 (df = 193)	8.325 (df = 190)	7.302 (df = 182)
Note:			*p<0.1; **p<0.05; ***p<0.01

In total, 14 providers have made available 209 distinct datasets. 2,276 distinct requests for these datasets have been made.[5]

Additionally, the IMPACT team shared the results of email inquiries sent to all requesters in the summer of 2018 asking about whether and how the data was used. Furthermore, several data providers shared information on their costs. More details on these datasets are provided in the following subsections.

4.1 Regression analysis of dataset requests

The first way in which we evaluate the value of cybersecurity datasets provided by IMPACT is to examine the factors that affect how frequently they are used. Many variables could influence a dataset's popularity among researchers, from the restrictions placed on its commercial use to the type of data being shared. We empirically examine multiple factors using regressions described below.

Regression Analysis We constructed a series of linear regressions with two distinct response variables: (1) the total number of requests a dataset receives and (2) the total number of approved requests. For these models, we only considered requests from 2016 onward because IMPACT utilization was relatively stable during this period.

Explanatory variables include:

1. **Time Available for Requests:** This variable indicates how long, in years, dataset is made available to researchers since January 1, 2016. We anticipate that the longer a dataset is available, the more requests it receives.

[5] IMPACT lets providers choose whether to treat datasets that are collected over time as a single, ongoing dataset or as distinct datasets collected at different points in time. We consolidated datasets with the same basic structure collected at different points in time into a single dataset. This also affects how we count requests. We counted multiple requests for the same type of data collected at different intervals as a single request.

2. **Dataset Age:** This variable indicates how old, in years, the dataset is. Age is determined by the time that has passed since the start of data collection. We expect that the older a dataset is, the less likely it is to be requested.

3. **Commercial Allowed:** IMPACT allows data providers to choose whether to permit commercial use or to restrict use to academic or government purposes. We hypothesize that this variable may affect the number of requests either by allowing more people to request it or only allowing commercial organizations to access less crucial datasets.

4. **Restriction Type:** We hypothesize that as access to datasets are made less restrictive, they will be requested more often. The three restriction types in IMPACT are *Unrestricted*, *Quasi-Restricted* and *Restricted*. These categories designate the potential sensitivity of the data, the ease with which the request can be processed, and the policy controls in the associated legal agreement. For example, *Unrestricted* data is low risk and can be requested by a click-through agreement that has fewer user obligations. This is compared to *Restricted* data that has privacy or confidentiality risk and requires a signed MOA, authorization by the provider, and more use encumbrances. *Unrestricted* is used as the baseline in the regressions.

5. **Ongoing Collection:** Some datasets encompass a snapshot of time, while others are being constantly collected and publicized in IMPACT. We expect that datasets with ongoing collection will be requested more often.

6. **Dataset Category:** We expect that characteristics of a dataset will influence the number of requests it gets. We do not presume to know which categories will be requested more often, but we do anticipate that the type of data within a dataset will affect request totals. We note that the data appearing in IMPACT reflects the interests of the data providers, not necessarily what requesters actually want. This categorical variable uses *Alerts* as a baseline.

The tables in Table 1 present the results of the linear regressions. Surprisingly, the baseline model does not find the amount of time a dataset is available to researchers to significantly affect the number of requests it receives, though the overall age of the dataset is negatively correlated with requests. Adding in variables that cover access restrictions (model 2) yields more surprises. On their own, these variables have limited effect. None of the variables are significant for the regression measuring requests. Restricted datasets do receive fewer approved requests than do unrestricted datasets, however, and that difference is statistically significant. Furthermore, in Model 2, permitting commercial access does not affect utilization. However, the variables become significant and *negative* once additional explanatory variables are added in Model 3. In other words, permitting commercial use is associated with a reduction in requests. Additionally, quasi-restricted datasets are requested more often than unrestricted datasets, statistically significant at the 10% level. One possible explanation is that the more attractive datasets place more restrictions on access.

Model 2 alone explains roughly 3.7% and 8.3% of the variance in total requests and approved requests respectively. Adding in whether collection is ongoing and the dataset category (model 3) helps explain a lot more of the variance: 24% and 30% respectively.

Category	Data Analysis	Tech. Eval.	Tech. Dev.	Op. Def.	Education
% of Requests	31.0	28.2	27.9	5.62	3.12

Table 2: Incidence of request categories in purpose requests.

Ongoing collection corresponds to six more dataset requests. Topology and adverse event datasets are requested less often than alerts, while attacks are requested more often. In the request regression, configurations are also weakly underrepresented.

4.2 Empirical analysis of value

We have just examined how the number of requests a dataset receives can vary by the terms on which it is shared, as well as the type of data involved. We now investigate the value created by utilizing datasets in IMPACT. Valuing information goods such as cybersecurity datasets is fraught with difficulty. The most obvious approach is to assign a value corresponding to the amount others are willing to pay to obtain it. This is not an option for public goods like IMPACT datasets that are given away for free, not to mention that there is no objective pricing of somewhat-similar data that is “sold” by data brokers or as part of fee-based data sharing consortium. An alternative is to investigate how others use the data, thereby creating value. This is a worthwhile approach because it can shed light on the outputs or outcomes that result from data use. The challenges with this approach is that is hard to aggregate the myriad uses into a single dollar estimate of value. We defer until the next section a discussion of a method to provide a dollar estimate of IMPACT datasets.

Whenever a researcher requests a dataset offered by IMPACT, the person is required to explain how he or she intends to use the dataset in a free-form text response. Data providers review these requests in order to assess whether the request is legitimate[6]. We examined all 2,276 of these reasons and developed a taxonomy to encompass the various types of purposes researchers have for requesting this data. There are six distinct categories and any individual reason may be classified into one or more of these categories. No reason was ever classified into more than three categories. These categories are described below.

Technology Evaluation Requests are categorized as Technology Evaluation when requested for evaluating the effectiveness of some technology. This may be an algorithm, framework, model, application, theory or any other form of technology that the requester wishes to test. Datasets used for ML are not considered to be Technology Evaluation unless they are exclusively used to evaluate a model. In other words, datasets used for ML training and testing are only considered Technology Development.
Example request: “Need to evaluate if our new DDoS detection in-line analytical module in NetFlow Optimizer can detect this attack.”

[6] The guidance given to requesters states the criteria: “The things we are looking for are some statement about what is novel about what you need to do (”new spectral analyis”), some statement on how you'll do it (”spectral analysis to identify DDoS in aggregate traffic”), and some statement of the context of the work (”for PhD-thesis research”)”

Example request: “Evaluation of the risk methodology presented in the paper, as it applies to current USG network communications.”

Technology Development: These are requests for assisting with the development of some technology. The requester may wish to extract features from the dataset that aid them in developing a technology (which we consider different from Data Analysis). Datasets that are used to train machine learning applications are also considered technology development.
Example request: “We are designing an anomaly detection system (on the victim side) for NIST. This dataset will be analyzed to capture the uniform attack behavior for our research.”
Example request: “Incorporate the attack scenarios to devise an automated process of detecting and controlling malicious insiders to mitigate risks to the organization.”

Data Analysis The requester wishes to analyze the data for its own sake. Data analytics, data visualization, and characteristic extraction all fall under data analysis. Again, datasets that are used for feature extraction as a means of technology development are not labeled as data analysis.
Example request: “The data will be used to analyze how DDoS affects the open source production systems.”
Example request: “Government funded research to benefit humanitarian aid and disaster relief community. Looking to see if we can correlate changes in BGP routing data with loss of power/communications infrastructure.”

Operational Defense The data is requested in order to help protect some critical resource of the requester’s organization. Requesters may want to see if the data has any specifics about their organization or if the data can help strengthen their defenses. Improving a defense resource to be used as a *product* is not considered operational defense.
Example request: “My objective is to protect Marine Corps data. This database can provide intelligence on passive DNS malware that can be used to block it from entering my network.”
Example request: “We intend to use this information to make our institutions’ IT related programs and computers as secure as possible. The ultimate goal is to ensure that our customer data is safe from malware attacks by keeping informed of recent trends and software that may require patches.”

Education Data is requested for education purposes such as use in courses or clubs in a school setting such as a University or High School.
Example request: “I’d like to develop exercises for an introductory stats and data science course that emphasizes cybersecurity awareness for the state of Virginia.”
Example request: “Mentoring project course for cadets at the Air Force Academy. Using data to develop new heuristics for anomaly detection.”

Unspecified The request reason was either too vague or we were unable to determine/understand what their request was for. They may have specified what their research is, but we couldn’t discern/easily assume what part of their research the data is being used for.

	Attacks			Topology			Network Traces		
% of Requests	58%			21%			21%		
Request Cat.	#	%	Sig.	#	%	Sig.	#	%	Sig.
Data Analysis	338	25		183	**37**	(+)	112	22	
Tech. Eval.	334	25		92	**18**	(−)	148	**30**	(+)
Tech. Dev.	344	25		84	**17**	(−)	157	**32**	(+)
Op. Defense	98	**7**	(+)	24	5		4	**1**	(−)
Education	39	3		9	2		14	3	
Unspecified	199	15		109	**22**	(+)	63	13	

Table 3: Three largest dataset categories split by request categories. Statistically significant under- and over-representations are indicated in bold with a (+/-).

Example request: "I'm doing some research on cyber situation awareness and feel this data would be beneficial to this work."
Example request: "Need for Research".

We manually categorized each request according to the taxonomy described above. Table 2 breaks down the incidence of requests that matched each category. Requests could correspond to more than one category, or to no category at all. Data analysis was most common (31%), followed by 28% each for technology evaluation and development.

We further investigated a question of whether or not the intended *use* for the data varied by the *type* of data being requested. Using the dataset categorization from [45], we analyzed the three most requested dataset categories split by request categories using a χ^2 test. Table 3 presents the results. Operational defense is overrepresented in the datasets describing attacks: 7% of the requests for attack data state operational defense as the intended use, compared to 5.6% overall. Data analysis is overrepresented in the requests for topology datasets, with 37% of all requests for topology datasets listing it as the reason for use. By contrast, technology evaluation and development are both underrepresented in the topology requests. For network traces, the trend is the opposite: both are overrepresented, while operational defense is rarely given as the reason for requesting network trace data.

We additionally sought to understand not only what the dataset was requested for, but also what it was ultimately used for. DHS surveyed all IMPACT requesters whose requests had been approved. Each survey response was associated with a certain dataset, or multiple, that the respondent specified. In total, 114 requesters responded, a few of which were the same requester responding for different datasets.

When asked whether or not they actually used the dataset they had requested, 60.4% of respondents said they had. To better understand what those requesters actually used the dataset for, we asked them to categorize their request reason and to categorize what their actual use was using the request taxonomy described above. 90.8% of requesters reported that they used the datasets in the same manner that they originally requested. This suggests that the preceding analysis on intended use accurately reflects actual use.

Furthermore, we asked the requesters who used the dataset whether or not they would have collected the themselves had IMPACT not provided the dataset. 72% answered that

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**

Table 4: Median reported annual cost of providing datasets to IMPACT split by category for eight data providers.

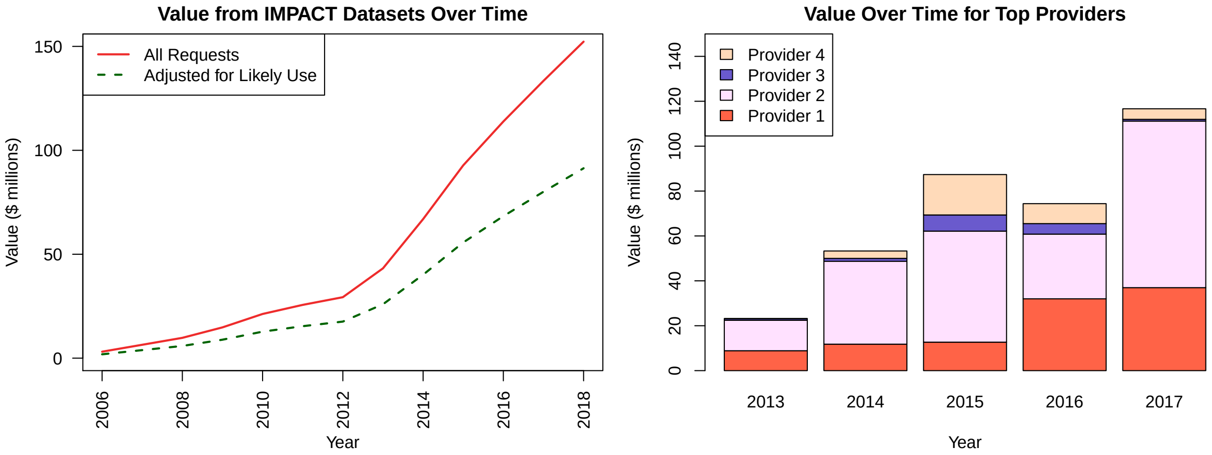


Figure 1: Value of data shared by IMPACT since inception using avoided cost definition (left). Value of data shared by top 4 providers on IMPACT using their costs reported for the year in which data was requested (right).

they would *not* have collected the data themselves. For those that wouldn't have collected the data themselves, their research may not have continued. For the 28% that would have collected the data themselves, they would have been replicating costly data collection and wasting time or resources that could be spent elsewhere. Motivated by this finding, in the next section we construct a quantitative model of value based on the avoided cost of data collection.

4.3 Quantifying value through avoided cost

While it would be preferable to value cybersecurity datasets by quantifying the benefits that accrue when individuals and organizations use the data, this is typically infeasible. Even if it were possible to reach every user of a dataset, translating the many uses into a dollar benefit is usually not possible even for the consumer of the dataset. One alternative method for quantifying the value of datasets that can be aggregated is to think of value as the cost

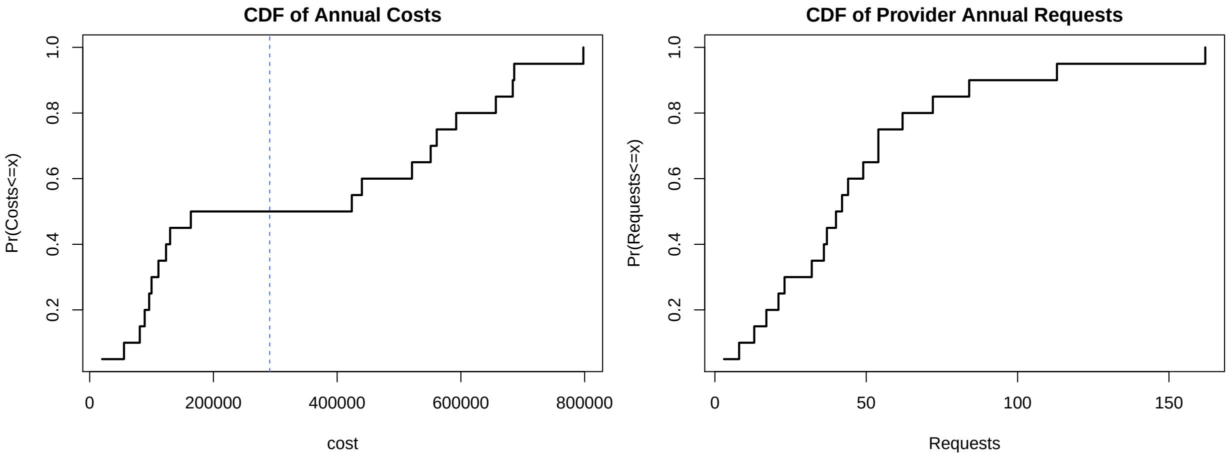


Figure 2: Cumulative distribution functions for the annual costs of providing data to IMPACT (left) and the number of annual requests providers receive for all shared datasets.

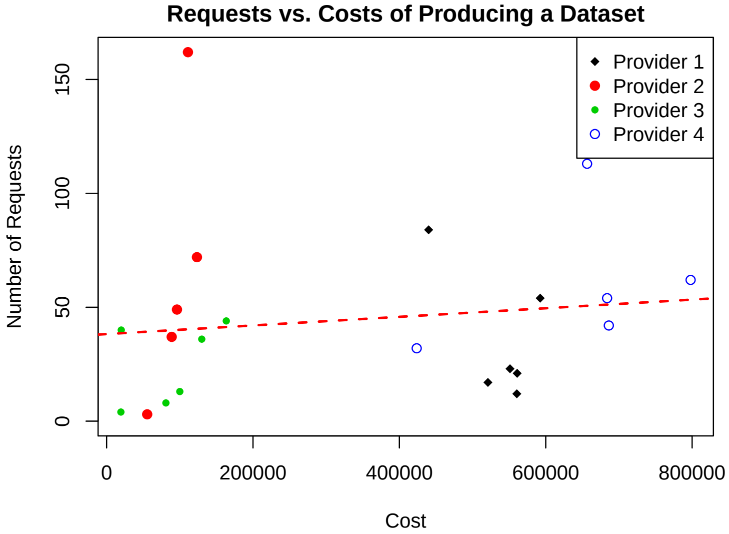
avoided by data consumers not having to collect the data themselves. Fortunately, such data is readily available, as the IMPACT program pays data providers to share their data with requesters.

Eight IMPACT performers shared detailed cost estimates for a number of categories such as personnel and equipment. Annual figures from 2012–17 were provided. Table 4 reports the median cost figures for each category, along with the total of $291K. Given that IMPACT has shared data with 2,276 requesters, the total value created as measured by this metric since the program's inception in 2006 is $663 million.

We recognize that the metric's validity rests on a number of assumptions that may not hold in each circumstance. We assume each request is independent. We assume that the researchers experience no other sunk costs or utilize any existing resources when provisioning data. We assume that outside researchers would not have to expend resources gaining a sufficient technical understanding of the data collection requirements. We also assume that outside researchers would exercise the same level of care in collecting the data that the IMPACT performers do. Even if these assumptions do not hold universally across all requesters, the metric nonetheless provides a valuable estimate of what the "true" value might be.

Figure 1 (left) plots the annual value created for all requests (solid red line), as well as a more conservative measure that normalizes for intended use (green dashed line). A normalization factor of 60% is used since that is the proportion of surveyed recipients who reported using the dataset they requested. Figure 1 (right) splits the value created among the top four IMPACT providers who have shared annual costs. We can see considerable variation, which is a consequence of highly variable costs of data production and dataset popularity.

Figure 2 (left) plots a cumulative distribution function for the annual provider costs. The plot reveals a barbell-like distribution in which half of the providers have low costs around $100K and half have higher costs around $600K. The vertical blue line shows the median value of $291K presented in Table 4. Meanwhile, Figure 2 (right) plots a CDF of the number



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

Figure 3: Scatter plot of provider annual requests compared against the cost of providing the data (top); cost per request for top 4 providers (bottom).

of annual requests providers have received since 2013. While the median number of annual requests during this period is around 50, some providers receive much fewer requests while other receive many more.

In fact, there is little to no relationship between the number of requests a dataset receives and what it costs to produce. Figure 3 plots the annual provider cost against the number of requests received that year for the top 4 providers. The best-fit line indicates a very slight positive correlation between cost and requests, but it is clear that many other latent factors besides cost of production affect a dataset's popularity. Finally, the table in Figure 3 lists the cost per dataset request for each of the top providers. We can see that the cost per request varies by an order of magnitude.

On one level, it is not surprising that the relationship between the cost of data production and the resulting demand for it is weak at best. What drives researcher interest is how the data can be leveraged, not the person-hours required to collect the data in the first place. Nonetheless, the implications for funding cybersecurity research data production are significant. Ideally, program managers should (and assuredly do) consider the potential demand for a dataset when deciding whether to support an effort financially. But perhaps more weight should be given to the anticipated requests per unit cost in order to maximize the impact of limited budgetary resources. To do so would also require more work estimating the demand of datasets in advance. To an extent, the regressions in Section 4.1 can help identify dataset categories that are in higher demand, but more work is needed to test whether such retrospective analysis is predictive of future demand.

5 Discussion and concluding remarks

In this paper we have undertaken two interconnected objectives. First, we articulated the benefits of making data broadly available to cybersecurity research: advancing scientific understanding, providing infrastructure to enable research, improving parity by lowering access costs and broadening availability, and bolstering operational support. Despite these benefits, access to data remains a nontrivial impediment to cybersecurity research. Therefore, we discussed barriers that inhibit broader access: legal and ethical risks, costs of operating infrastructure, and uncertainties, asymmetries and mismatches related to the value such data can provide. We also considered available incentives to promote data sharing, finding them to be lacking at present. We reviewed existing models for supporting research datasets, from student internships to government-facilitated sharing and suggested that the economics of sharing data for research requires appropriate investment not unlike that of other social goods. It is hoped that the readers are left with a better understanding of the value of cybersecurity data in research, how it works today, and what needs to change in order to improve the situation moving forward.

Our second objective has been to empirically investigate the sharing that has taken place on IMPACT, a long-running platform that has uniquely facilitated free access to cybersecurity research data. Controlling for the time available on IMPACT, we have found that the dataset's age is negatively correlated with requests. This makes sense given that researchers may prefer more recent data for their efforts. We also found that the restrictions placed on access to data affect how often they are requested, but in unexpected ways. For example, permitting commercial use of the data is negatively correlated with utilization, and quasi-restricted datasets are requested more often than unrestricted ones. These may reflect either a perception (or the reality) that datasets placing modest restrictions are more likely to be useful. Note that when we do move to the restricted category that introduces significant additional costs and verification, approved requests fall.

We also find that datasets that are made available on an ongoing basis are requested more often. Ongoing availability can be thought of as a proxy for current relevance and longitudinal cohesiveness, two properties valued by researchers. Additionally, ongoing datasets are more likely to be relevant to operational defense, which comprises around 6% of IMPACT requests.

We also find that there is considerable variation among the types of datasets. Twenty percent of the variance in requests can be explained by the type of data offered and whether or not it is made available on an ongoing basis. Difficult to collect, topically relevant, and potentially sensitive data such as attacks are requested more often, while more general and less sensitive data such as network topology are requested less often.

We also investigated the value created by data shared on IMPACT in two ways. First, we looked at what the requesters themselves said they intended to do with the data. We identified five categories of use: technology evaluation, technology development, data analysis, operational defense, and education. Data analysis was the most common intended use, followed by technology development and evaluation. Strikingly, when asked, 60% of requesters said they used the data requested and 90% of those said they used it in the way they originally intended. This suggests that the IMPACT users are highly sophisticated in their understanding of their research data needs. Most significantly, 72% of surveyed requesters stated that they would not have collected the datasets themselves if they could not have obtained it through IMPACT. This highlights the value of investing in research data infrastructure and underscores how much research may not be conducted when data access is limited.

This motivates the second approach to valuing data shared on IMPACT, by quantifying value in terms of the costs avoided by data recipients. We obtained annual provisioning costs from data providers. Matching this to requests, we estimate that the value created since program inception in 2006 is $663 million. Digging deeper into the costs uncovers two surprising insights. First, the normalized cost per request varies widely, by one order of magnitude. Second, there is little if any relationship between the cost of data provisioning and its resulting demand.

How do the findings for the case of IMPACT compare to the benefits, barriers and incentives identified? IMPACT has realized each of the benefits described, from enabling scientific advances to understanding to improving data access (at least among eligible participants). Under IMPACT, standardized legal agreements have been accepted by providers, and experience has shown little difficulty in sharing restricted datasets. Furthermore, requesters have seldom objected to the terms outlined in agreements. So it seems that for the data shared, legal barriers can be overcome. Of course, we cannot say much about the datasets *not* shared on IMPACT due to perceived legal issues. The direct financial costs can absolutely be a barrier, but these costs have been addressed by government funding for data providers. The fact that 72% of those asked said they would not have directly collected the data themselves if not for IMPACT suggests that direct financial costs are in fact a significant barrier.

Discrepancies in dataset popularity reflects challenges due to uncertainty over dataset value, as well as value asymmetries between data provider and requester. Simply put, researchers do not always create and share data that requesters want. IMPACT is a platform serving a two-sided market of data consumers and producers. Each makes independent decisions, and so it is inevitable that there can be mismatches. This also is indicative of the lack of collective dialog and agreement about cybersecurity data needs.

On the one hand, we should be encouraged by the success of the IMPACT Program: thousands of users, year-over-year increases in account and user requests, and hundreds of technical papers published using data hosted by the platform. On the other hand, there are reasons to be concerned: the lack of a comparable data sharing platform for cybersecurity research, as well as the present market immaturity in valuing data. It is reasonable to conclude that investment in research data infrastructure is an essential requirement for assuring the availability of data for cybersecurity R&D. Failure to support data as a social good will exacerbate an existing cybersecurity challenge: the individual and collective risk and harms that can cascade from shared and interdependent systems whose exposure is only knowable when individual stakeholders collaborate.

Acknowledgements 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] Worldwide observatory of malicious behaviors and threats, 2011. http://www.wombat-project.eu.

[2] National Security Agency. Science of security, 2019. https://www.nsa.gov/what-we-do/research/science-of-security/.

[3] Manos Antonakakis, Roberto Perdisci, Yacin Nadji, Nikolaos Vasiloglou, Saeed Abu-Nimeh, Wenke Lee, and David Dagon. From throw-away traffic to bots: detecting the rise of dga-based malware. In *Presented as part of the 21st {USENIX} Security Symposium ({USENIX} Security 12)*, pages 491–506, 2012.

[4] Leyla Bilge and Tudor Dumitraş. Before we knew it: An empirical study of zero-day attacks in the real world. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, CCS '12, pages 833–844, New York, NY, USA, 2012. ACM.

[5] Aaron J. Burstein. Amending the ECPA to enable a culture of cybersecurity research. *Harvard Journal of Law and Technology*, 22(167), 2008.

[6] Juan Caballero, Chris Grier, Christian Kreibich, and Vern Paxson. Measuring pay-per-install: the commoditization of malware distribution. In *Usenix security symposium*, pages 13–13, 2011.

[7] Scott Coull and Erin Kenneally. Toward a comprehensive disclosure control framework for shared data. In *IEEE International Conference on Technologies for Homeland Security*, 2013.

[8] Department of Homeland Security. Protected Critical Infrastructure Information Program (PCII), 2019. https://www.dhs.gov/cisa/information-sharing.

[9] David Dittrich and Erin Keneally, 2012. https://www.impactcybertrust.org/link_docs/Menlo-Report.pdf; companion https://www.impactcybertrust.org/link_docs/Menlo-Report-Companion.pdf.

[10] Tudor Dumitraş and Darren Shou. 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*, BADGERS '11, pages 89–96, New York, NY, USA, 2011. ACM.

[11] Zakir Durumeric, Frank Li, James Kasten, Johanna Amann, Jethro Beekman, Mathias Payer, Nicolas Weaver, David Adrian, Vern Paxson, Michael Bailey, and J. Alex Halderman. The matter of heartbleed. In *Proceedings of the 2014 Conference on Internet Measurement Conference*, IMC '14, pages 475–488, New York, NY, USA, 2014. ACM.

[12] Eric A. Fischer. Cybersecurity and information sharing: Comparison of H.R. 1560 (PCNA and NCPAA) and S. 754. Technical Report R44069, Congressional Research Service, November 2015.

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

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

[15] Thorhildur Jetzek, Michel Avital, and Niels Bjørn-Andersen. Generating value from open government data. In *International Conference on Information Systems (ICIS)*, 2013.

[16] Stefan Laube and Rainer Böhme. Strategic aspects of cyber risk information sharing. *ACM Comput. Surv.*, 50(5):77:1–77:36, November 2017.

[17] Nektarios Leontiadis, Tyler Moore, and Nicolas Christin. A nearly four-year longitudinal study of search-engine poisoning. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, pages 930–941. ACM, 2014.

[18] K. Levchenko, N. Chachra, B. Enright, M. Felegyhazi, C. Grier, T. Halvorson, C. Kanich, C. Kreibich, H. Liu, D. McCoy, A. Pitsillidis, N. Weaver, V. Paxson, G. Voelker, and S. Savage. Click trajectories: End-to-end analysis of the spam value chain. In *IEEE Symposium on Security and Privacy*, pages 431–446, Oakland, CA, May 2011.

[19] Frank Li, Grant Ho, Eric Kuan, Yuan Niu, Lucas Ballard, Kurt Thomas, Elie Bursztein, and Vern Paxson. Remedying web hijacking: Notification effectiveness and webmaster comprehension. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 1009–1019, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

[20] Martin C. Libicki. Prepared testimony of Martin C. Libicki, Senior Management Scientist at The RAND Corporation, before the House Committee on Homeland Security, Subcommittee on Cybersecurity, Infrastructure Protection, and Security Technologies, 2015. `http://docs.house.gov/meetings/HM/HM08/20150304/103055/HHRG-114-HM08-Wstate-LibickiM-20150304.pdf`.

[21] Edward C. Liu, Gina Stevens, Kathleen Ann Ruane, Alissa M. Dolan, Richard M. Thompson III, and Andrew Nolan. Cybersecurity: Selected legal issues. Technical Report R42409, Congressional Research Service, April 2013.

[22] Yang Liu, Armin Sarabi, Jing Zhang, Parinaz Naghizadeh, Manish Karir, Michael Bailey, and Mingyan Liu. Cloudy with a chance of breach: Forecasting cyber security incidents. In *24th USENIX Security Symposium (USENIX Security 15)*, pages 1009–1024, Washington, D.C., 2015. USENIX Association.

[23] Joseph Marks. Only 6 non-federal groups share cyber threat info with Homeland Security. *NextGov*, 2018. `https://www.nextgov.com/cybersecurity/2018/06/only-6-non-federal-groups-share-cyber-threat-info-homeland-security/149343/`.

[24] Alain Mermoud, Marcus Matthias Keupp, Kévin Huguenin, Maximilian Palmié, and Dimitri Percia David. Incentives for human agents to share security information: a model and an empirical test. In *Workshop on the Economics of Information Security (WEIS)*, 2018.

[25] Microsoft News Center. Adobe, Microsoft and SAP announce the Open Data Initiative to empower a new generation of customer experiences, 2018. `https://news.microsoft.com/2018/09/24/adobe-microsoft-and-sap-announce-the-open-data-initiative-to-empower-a-new-generation-of-customer-experiences/`.

[26] T. Moore and R. Clayton. Examining the impact of website take-down on phishing. In *Second APWG eCrime Researcher's Summit*, Pittsburgh, PA, October 2007.

[27] NASA. Aviation safety reporting system agreement. https://asrs.arc.nasa.gov.

[28] European Network and Information Security Agency. Standards and tools for exchange and processing of actionable information, November 2014. https://www.enisa.europa.eu/activities/cert/support/actionable-information/standards-and-tools-for-exchange-and-processing-of-actionable-information.

[29] University of Cambridge. Cambridge cybercrime centre, 2019. http://www.cambridgecybercrime.uk.

[30] Department of Homeland Security. Information marketplace for policy and analysis of cyber-risk and trust. https://www.impactcybertrust.org. Last accessed February 14, 2019.

[31] Department of Homeland Security. Information sharing specifications for cybersecurity, 2015. https://www.us-cert.gov/Information-Sharing-Specifications-Cybersecurity.

[32] Department of Homeland Security. Biennial report on DHS' implementation of the Cybersecurity Act of 2015 OIG-18-10, 2017. https://www.oig.dhs.gov/sites/default/files/assets/2017-11/OIG-18-10-Nov17_0.pdf.

[33] Department of Homeland Security. Cyber risk economics capability gaps research strategy, 2018. https://www.dhs.gov/publication/cyrie-capability-gaps-research-strategy.

[34] Department of Justice and Federal Trade Commission. Antitrust policy statement on sharing of cybersecurity information, 2014. http://www.justice.gov/sites/default/files/atr/legacy/2014/04/10/305027.pdf.

[35] Government Accountability Office. Critical infrastructure protection: Improving information sharing with infrastructure sectors, July 2004. http://www.gao.gov/products/GAO-04-780.

[36] Christian Rossow. Amplification hell: Revisiting network protocols for DDoS abuse. In *Network and Distributed Security Symposium (NDSS)*, 2014.

[37] Charlotte Scheper, Susanna Cantor, and Renee Karlsen. Trusted distributed repository of internet usage data for use in cyber security research. pages 83 – 88, 04 2009.

[38] National Science and Technology Council. Federal cybersecurity research and development strategic plan: Ensuring prosperity and national security, FEB 2016.

[39] United States. Broad agency announcement solicitation/call: HSHQDC-17-R-00030 Project: Information Marketplace for Policy and Analysis of Cyber-risk & Trust (IMPACT) Research and Development (R&D), 2017. https://www.fbo.gov/utils/view?id=1f18dfa7debc01e90fbc8b61a85bfb2b.

[40] Kurt Thomas, Chris Grier, and David M. Nicol. Barriers to Security and Privacy Research in the Web Era. In *Proceedings of the Workshop on Ethics in Computer Security Research*, 2010.

[41] United States Congress. OPEN Government Data Act (S. 760 / H.R. 1770), 2019.

[42] US-CERT. Cybersecurity information sharing act - frequently asked questions, 2016. https://www.us-cert.gov/sites/default/files/ais_files/CISA_FAQs.pdf.

[43] N. Eric Weiss. Legislation to facilitate cybersecurity information sharing: Economic analysis. Technical Report R43821, Congressional Research Service, June 2015.

[44] Denise Zheng and James Lewis. Cyber threat information sharing: Recommendations for congress and the administration, 2015.

[45] Muwei Zheng, Hannah Robbins, Zimo Chai, Prakash Thapa, and Tyler Moore. Cybersecurity research datasets: Taxonomy and empirical analysis. In *11th USENIX Workshop on Cyber Security Experimentation and Test (CSET 18)*, Baltimore, MD, 2018. USENIX Association.

[46] Anneke Zuiderwijk, Natalie Helbig, J. Ramón Gil-García, and Marijn Janssen. Special issue on innovation through open data: A review of the state-of-the-art and an emerging research agenda: Guest editors' introduction. *J. Theor. Appl. Electron. Commer. Res.*, 9(2):i–xiii, May 2014.