# Data for Cybersecurity Research: Process and "Wish List"

Jean Camp, Lorrie Cranor, Nick Feamster, Joan Feigenbaum,
Stephanie Forrest, Dave Kotz, Wenke Lee, Patrick Lincoln, Vern Paxson, Mike Reiter,
Ron Rivest, William Sanders, Stefan Savage, Sean Smith, Eugene Spafford, Sal Stolfo

June 10, 2009

This document identifies data needs of the security research community. This document is in response to a request for a "data wish list". Because specific data needs will evolve in conjunction with evolving threats and research problems, we augment the wish list with commentary about some of the broader issues for data usage.

We divide this document into two parts. Section 1 provides background on data collection as often practiced today and a few of its uses. Section 2 identifies the need for a process for ongoing data sharing with the research community, and then provides the wish list itself.

## 1 Background on Empirical Computer Security Research

The types of data and the means by which the data might be delivered is difficult to fully enumerate. Hence, we start with a general discussion about how the security research community currently *uses* data, both for assessing monitoring threats and for developing new defenses. We discuss how data is currently exchanged between industry and researchers and the various shortcomings of these modalities and delivery mechanisms. Finally, we survey current legal issues with operational security data.

### 1.1 Example Uses for Data

**Monitoring and mitigation of global threats** Network traffic data can help researchers track the activity and evolution of global threats, such as spam, botnets, phishing attacks, and scams. This data can also provide insights into how cybercriminals use the network to mount attacks (e.g., to what extent multiple attacks are coordinated). Network trace data can also help researchers determine the extent to which a proposed mitigation technique (e.g., a new spam filtering or sender reputation technique, a phishing detection scheme) might work in practice before it is widely deployed.

**Threat model validation** One important reason to obtain as complete a trace as possible of a network, even over a limited period, is validation of traffic and threat models. The networking community depends on sound modeling, but testing and validating models and their assumptions is generally a difficult task, exacerbated by the difficulty of obtaining appropriate data. The lack of a measurement infrastructure in the Internet limits our ability to improve our models of network behavior and user behavior.

**Insider threat mitigation** Insider threat mitigation requires the evaluation of methods and models for detecting malicious insider activity and behavior on a computer system. Malicious insiders refers to either authorized users who are violating policy and intentionally triggering malicious activities on a system or external users who have appropriated a valid user's credentials and are masquerading as the valid user for nefarious purposes. Today there are no data sources for insider threat studies that include "true masquerade data". What forensic data may be available for studying true masquerader events are usually highly confidential and beyond the reach of researchers.

## 1.2 Modalities and Delivery Mechanisms

**Modalities of data collection today** It is quite difficult for researchers to obtain access to operational data. Generally speaking, some level of cooperation and reciprocity can help initiate data-sharing efforts. Service providers are often unlikely to deliver data to researchers (particularly in light of privacy and proprietary concerns) without first seeing some upside or added value to their own enterprise. After such value is demonstrated, researchers may gain access to data, but this outcome is rare.

Currently, when researchers obtain data access in some form, it happens in one of the following ways:

1. A bilateral exchange absent strong contracts (e.g., non-disclosure agreements), typically via operators, but at best engaging middle management, and inevitably with the understanding that the data will not be attributed to the company nor used to embarrass them.
2. Access to the data as a "black box": data purveyor offers to run processing scripts on the data and return the output to the researcher, without ever letting the researcher see the raw data.
3. A temporary employee who is sent as a "data envoy"—the employment contract thus acts as the contractual instrument (regardless of how carefully the data is actually stewarded).

The research community has made some progress using these methods of data acquisition, but it is safe to say that more empirical research efforts fail more at this juncture than at any other. Each such relationship to gain access to data typically must be cultivated in an *ad hoc* fashion, often requiring significant effort and "face time", if it can be cultivated at all. Moreover, where such relationships can be cultivated by individual researchers, that accessibility almost never extends to other researchers, posing an impediment to experimental repeatability—a basic tenet of the scientific process. Such customized arrangements also result in variable quality of data stewardship and privacy oversight; even the "black box" model can divulge sensitive data in the outputs of computations. The "data envoy" model introduces additional problems, in the form of impediments when an envoy's employment contract ends (e.g., when a student leaves the company after the end of an internship, but before a project is completed).

Scientific research generally relies on shared access to data, to allow for repeatable experimentation and evaluation. Current modalities thus leave much to be desired. Instead, we desire processes for sharing data so that researchers can compare experimental results and compare competing solutions to emerging security threats and problems.

**Delivery mechanisms and data stewardship** Data can be delivered as a one-time archive trace, periodically, or as a continuous feed. The appropriate mechanism for delivering data depends on the nature of the research effort. For example, research efforts that focus on monitoring the network for global threats may require continuous data streams to provide useful information to both researchers and operators. Examples of such frameworks include those that monitor the Internet routing infrastructure for stolen routing information ("hijacks"), such as the Internet Alert Registry, and those that update their detection heuristics in response to the continually changing behavior of spam, botnets, and phishing sites. In these cases, the process of providing a data stream should involve a dialog of precisely what information is required, and at what granularity, so that no more data than necessary is revealed.

The other common mode of data delivery is via archived traces. Such traces may range from application-level traces (e.g., HTTP clickstreams) to network-level traffic information (e.g., statistics about the flows at the IP layer) to user activity data. In these cases, the data can be useful for assessing the extent of a threat or for evaluating a proposed solution offline, although the data itself may have a limited shelf life as threats evolve. Data in which the activities of malicious parties, if any, is identified ("labeled data") is most useful. It is often the case that this type of data is subject to anonymization, to protect what is (or may be perceived as) proprietary or private information. When such data is anonymized, the method that is used for anonymization should be clearly documented so that researchers can reason about and compare results. Stewardship of these traces is also important; ultimately, it may make sense to have a sanctioned broker mediate the type and extent of access given to certain traces. The PREDICT project (`http://www.predict.org/`), sponsored by the Department of Homeland Security, offers one such possible stewardship model.

**Metadata** Information about how, where, and when data is collected—as well as additional information, such as whether it contains known attacks—is critically important for scientific research. Such metadata helps researchers

distinguish meaningful conclusions from artifacts (e.g., side effects of data collection). The particular format of this metadata will differ depending on the actual type of data. In cases where metadata is incomplete, researchers should be able to communicate with data purveyors to clarify ambiguities about the data. CAIDA's DatCat (`http://www.datcat.org/`) provides one possible model for collecting and publishing metadata regarding specific network traces.

### 1.3 Legal and Privacy Concerns

Data analysis today is mired by a host of legal and legitimate privacy concerns that, if not adequately addressed, often stymie research. For example, in 2006, AOL planned to make anonymized search data available to academic researchers. The lawsuits are still ongoing.

Anonymizing some of the data sets needed in computer security research is non-trivial. Simply replacing personally identifiable information with pseudonymous identifiers may not be sufficient, as profiles can be built based on the pseudonymous identifiers and context may be used to re-identify some users. This issue is paramount when the data includes content (e.g., search terms, movie preferences), location information, detailed demographics or job information, or social network information. In some cases, there are legal restrictions that will not permit the release of this data. Companies may also have issues due to proprietary concerns or due to the commitments they have made in their privacy policies.

## 2 A Wish List

We briefly outline a process for data sharing and provide a wish list for data based on current needs. Due to the fact that threats and research problems continually evolve, we view that establishing a process by which researchers and operators can exchange data on an ongoing basis is as important as the wish list itself.

### 2.1 A Process for Data Sharing

The data needs of the security community are continually evolving with the nature of threats, and as new projects and questions emerge. While the data described in the list in Section 2.2 would be useful to advance a segment of today's computer security research, the usefulness of this particular data to the research community will decay over time, owing to the ever-changing landscape of computer and network usage, computer security threats, and research ideas. This evolution regularly requires more current or even new types of data in order to perform evaluations. *We cannot emphasize enough the need to establish a process by which researchers can engage with operators and vendors of relevant technologies to obtain data for research purposes on an ongoing basis.* In particular, this process would ideally allow for the collection of data that is customized to a particular experiment, since data collected in the absence of an experimental method is typically insufficient to conduct sound experiments.

As the Internet is global, various cultural and legal views of data privacy, retention, etc. must be taken into account in any such process. An advantage of having a structured process for data sharing is that it offers the opportunity for oversight on a more consistent basis, versus the *ad hoc* sharing conducted today (see Section 1.2). We envision that such a proper form of oversight would enable establishing an effective *legal safe harbor* for sharing data with academic researchers—this alone would greatly enhance data accessibility, we believe. For some types of data, this oversight might impose limitations on the data's use, retention, storage, and disclosure.

We hope that this document will spawn a discussion about the best ways to build relationships between researchers and operators for addressing critical security questions in the future. Despite the challenges for establishing such a process, we believe that it is imperative for coping with the continually evolving threats and research needs. The research community is eager to engage in this discussion.

### 2.2 Specific Data That Would be Useful Today

Below we will list examples of some of our current needs. For each of the example items below, we briefly describe the nature of the data that the operations community could provide, and how the data might be used.

**Labeled traffic traces** Annotated network traffic traces, such as network traffic from infected hosts, will allow researchers to develop and evaluate methods for detecting both infected hosts and coordinated global threats. *Traffic traces should be clearly labeled with their respective malicious or benign components, and they should include attacks that current detection systems have failed to detect.* Today, access to labeled traces typically occurs through the modalities outlined in Section 1.2, but, even in those cases, *labeled* traffic traces are particularly difficult to obtain, since labeling is often a manual process.

**Wireless and cellular network data** The CRAWDAD project at Dartmouth receives many requests for Wi-Fi network traces from large hotspot providers and enterprise networks. Researchers want data traffic, user mobility and handoff patterns, and call times and durations. A common security-related request is for labeled wireless-network traces that include either (a) Internet attacks (e.g., worms, botnets) or (b) MAC-layer attacks. As cellular networks become the Internet connection for a majority of the world's users, cellular traces will become increasingly important as well. Cellular network traces can help researchers better understand access and usage patterns, as well as vulnerabilities with the existing cellular network design and infrastructure for certain tasks (e.g., emergency alert systems).

**Information about network "agility"** Data sources that could provide additional information about the dynamics of infrastructure that is used to host scams and phishing attacks, as well as to send spam, could help researchers identify the invariants of such infrastructures. Although it is well-known that cybercriminals use various parts of the Internet infrastructure, including Domain Name System and the Internet routing infrastructure (e.g., BGP), to cloak their activities, data that helps identify which of these mechanisms is used to cloak various types of attacks could assist with both attribution and mitigation.

**URLs received in spam and legitimate email, instant messages, etc.** Many cyberattacks, such as phishing attacks, are mounted using pointers to Web sites (i.e., URLs) delivered in spam email messages. A critical part of monitoring and protecting the network infrastructure involves determining the nature of these attacks relative to legitimately hosted Web sites. A data feed that provided URLs delivered in legitimate email messages and spam messages could help researchers monitor threats and improve detection techniques for such attack sites. The state of the art in understanding how scam and phishing sites are hosted is relatively primitive: neither network operators nor researchers understand how scam sites move to evade detection; they also do not have sound methods for automatically detecting such sites. Better access to URL data sent via email and messaging services will help researchers better understand the nature of these threats and the effectiveness of proposed detection techniques.

**Statistical information about email usage** For example, a histogram of "the number of emails sent from a given IP address" for some time period (e.g., one month) could provide information about infection rates, and could help answer questions about whether it is possible to determine whether a machine is infected this way. This information could also provide information about the overall load a spam filtering system would face. For example, researchers are actively developing systems for email sender reputation, stamp-based email systems, etc.; information about the magnitude of email traffic that ISPs both transit and deliver to individual user mailboxes—as well as the rate of such traffic—could help researchers evaluate the technical and economic feasibility of various proposed solutions.

**Aggregate statistics about downtime and threats from ISPs** Improving Internet availability requires information about the most significant threats ISPs face. The causes of network downtime range from misconfiguration by human operators to physical failure to various types of attacks. Information about the various causes of network downtime, and the extent to which various threats cause network downtime, can help researchers develop tools and techniques that focus on the most serious threats.

**Example hostile workloads for a Web server** Internet services are continually subject to various types of attacks, including distributed denial of service (DDoS) attacks. These attacks can deny access to legitimate users, and, if they are distributed, they can sometimes be difficult to distinguish from legitimate requests. Researchers are continually developing new techniques to protect Web servers against denial of service attacks and could use example hostile workloads to help evaluate new DDoS detection and mitigation techniques.

**Malware samples** Malware is a program that carries the malicious actions intended by attackers. By analyzing malware, we can gain information of what machines are vulnerable, and what malicious actions will be carried out. Such information can be used to help network defense (e.g., blocking malware spread and attacks), recovery (e.g., cleanup the compromised machines), and attribution or removal of the attacks (e.g., "follow" the malware's network connections to locate a botnet's command-and-control servers). Currently, every anti-virus company runs its own malware sample collection and performs limited exchanges with other companies. Industry, academia, and the government need a centralized malware clearinghouse that collects malware from all parties, runs state-of-the-art and automated analysis, and provides a unified labeling and threat assessment of the malware instances. Without such a clearing clearinghouse, we will never have a good understand the threats posed by these malware instances to help coordinate and prioritize our network defense and investigations. Just as public health has a CDC to monitor and track outbreaks and recommend appropriate actions, the security community should track malware outbreaks.

**A large production software base with a history of bug fixes** Software systems are continually subject to programming errors and bugs that can introduce threats to vulnerabilities. Information about the bugs (and bug fixes) for a large production piece of software could help researchers better understand the nature of programming bugs in large software systems today. For each bug that is repaired, data would include documentation of the bug, the source code before and after the bug was repaired, and a set of regression tests.

**Human user event data on various platforms and operating systems** This data could be used for insider threat studies, usability of security technology, anomaly detection and profiling, and a host of other applications. The data to be acquired could be as coarse as Unix commands (with and without arguments), or as detailed as the stream of Windows events per click. The data does not need to include any personally identifiable information. The length of time would ideally span an important epoch, such as a standard business cycle (a week and a weekend or far longer). The user roles should be identified and data should span across clerical, administration, technical, professional, system administrators, front office personnel, and other major categories of user within a modern organization. The sensors that acquired the data should be especially useful, for example those available at `http://www.cs.columbia.edu/ids/RUU/study.html`.

**Data on circumvention of security requirements** Case studies and interviews in the investment banking and health care industries indicates that operational requirements trump security controls in real-world, mission-critical environments. Unfortunately, the nature of the mismatch between "de jure" and "de facto" security policies makes it difficult to gather data on when and how security controls are being violated. Users and administrators have a strong incentive to keep quiet about the workarounds they have developed—understandably, as those workarounds enable them to meet the operational goals of their organization. Traditional security approaches encourage regulators and compliance staff to root out the "bad" users who violate policies—yet too often, audits consist of bureaucratic checklists filled with tacit understanding that the real world is messier than a clean-cut security policy can represent. Quantitative data on circumvention would enable us to design security controls that would make circumvention unnecessary.

**Power Grid Related Data** The power grid is vulnerable due to outdated technologies and architectures in its networking equipment. The security community has proposed several solutions tailored to the SCADA networks on legacy hardware, as well as new architectural solutions but both the target scenarios (shaping the design) as well as evaluation remain too hypothetical, because of data. These data include: Network traces, including SCADA messages and their times of transmission, which should be representative of typical traffic scenarios, including netflows, Time-stamped, fine-grained power line measurements; protocol execution traces (e.g., those for IEC 61850, DNP3, ICCP); intrusion detection logs, and the (power system specific rules) that generated them; meter data for consumers; data regarding power outages; state estimation data snapshots; bus/grid topologies corresponding to the data; and metadata associated with devices being measured.