Application to the University of Michigan Institutional Review Board, Health Sciences
Application Number: REP00000030

PI Information:
Michael Kallitsis - Merit Network, Inc., University of Michigan
Joe Adams - Merit Network, Inc,. University of Michigan

Sponsor Information:
Department of Homeland Security (DHS).

Performance Sites:
Merit Network, Inc.
University of Michigan.

Project Abstract:
The Next Generation Repository for Sharing Network and Security Data is a unique effort to organize, structure, and combine the efforts of the network security research community with the efforts of the Internet data measurement and collection community. Under the umbrella of the Information Marketplace for Policy and Analysis of Cyber-risk and Trust (IMPACT)[1] initiative of the Department of Homeland Security Science and Technology directorate, the Next Generation Data Repository continues the efforts of the Virtual Center for Network and Security Data to provide a common framework for managing datasets collected from various Internet data providers. It also formalizes a process for qualified researchers to gain access to these datasets, in order to prototype, test, and improve their Internet threat mitigation techniques, while ensuring that the privacy and confidentiality of Internet users are not compromised. This common framework benefits both the data providers, as they no longer have to review, approve and monitor individual researchers that approach them for access to various datasets, as well as the security researchers, as they no longer need to rely on ad hoc and often arbitrary policies of each data provider. The repository collects both network management and security data in the form of network packets and application logs. However, this data does not contain any live payload data, or any other personally sensitive information. In addition, any information that could link the data with individuals is removed or anonymized to mitigate any potential risks. We believe that the data collected as part of this repository represents minimal risk to the privacy of individuals.

[1] IMPACT is the evolution of the Protected Repository for the Defense of Infrastructure against Cyber Threats (PREDICT) program.

Research Design:

The goal of our virtual repository is to collect, organize, and share network security data. In order to make the virtual repository feasible, it is necessary to clearly define the roles of the data providers, the repository maintainers (i.e., the data hosts), and the researchers who wish to access the datasets. We logically separate the entire virtual repository architecture into two separate processes. The first describes the process that the data consumers (researchers) must use in order to gain access to the datasets. The second describes the role of the data providers and defines how they may structure the task of publishing their collected data. In order to facilitate the virtual repository, each data provider is required to set up and maintain separate dataset servers to host the actual data. The virtual repository is responsible for maintaining an accurate catalog of the various datasets that are available. As the data providers make datasets available, they publish the metadata describing the dataset to the virtual repository catalog. Researchers then are able to search this catalog and locate datasets that are of interest to them. To help maintain uniformity across the multiple data providers, our research design describes a directory naming as well as dataset naming conventions. One of the key enabling features of the virtual repository architecture is that for the first time researchers are able to correlate datasets from different data providers, creating a robust and useful environment for the development and testing of new cybersecurity techniques for Internet threat mitigation. To remove any potential risk to individuals represented by datasets in the repository, all potentially personally identifiable information from sensitive datasets will be removed or anonymized before being stored or distributed from the virtual repository.

Significance and Benefits of Research:

As national utility infrastructures become intertwined with emerging global data networks, the stability and integrity of the two have become synonymous. This connection, while necessary, leaves network assets vulnerable to the rapidly moving threats of today's Internet. These new threats have impact beyond the scope of the individual enterprise, not only infecting vulnerable hosts (i.e., the individual computers on the network) with malicious code but also denying service to legitimate network users. Fast spreading worms have disrupted financial institutions and 911 services. Inadvertent BGP configuration changes have crashed national ISP networks. The enterprise, its upstream ISP(s), and the global Internet community address these threats differently because each has a separate view of the network. Unfortunately, research into Internet wide or infrastructure level attacks are hampered by a lack of Internet-wide datasets. While researchers can often study individual packet traces and compromised machine forensics, these datasets rarely reflect Internet-wide behaviors. Datasets that are available are often fragmented or difficult to correlate because of missing meta-data or wildly disparate time frames. The proposed virtual center will address this need by providing a virtual repository of rich, correlated datasets representing Internet scale behaviors, which will enable qualified cybersecurity researchers to test and prototype novel attack mitigation techniques. Data available from this virtual repository will include both infrastructure data and data from distributed forensic tools.

I. Intake Procedures for Acceptance of Data into the Repository

a. Identify source of data and describe data elements:

Unlike clinical research data, which is typically collected as part of a directed human subjects medical research program, the datasets that are collected and made available to researchers via the virtual repository represent activities of ordinary individuals using the Internet or activities of networking equipment that help relay Internet traffic to its destination (e.g., network routers). Different types of data can be collected, each with its own degree of risk (or non-risk). In order for the Institutional Review Board to adequately assess this risk, we provide here a general description of how the normal functioning of the Internet generates datasets that can be aggregated and published. A detailed discussion of the specific data types that will be collected in the virtual repository is contained in Appendix A. In general, we expect that all information that could potentially link an individual with a certain portion of the data will be modified to remove any risk to individuals.

All information transferred over the Internet is broken up into small chunks, called packets. Each individual packet contains a header portion as well as a payload portion. The header of a packet contains information needed to correctly route the packet from source to destination. Similar to a postal letter, it contains the address of the destination as well as a return address for sending replies. Another example is telephony; each party is associated with a telephone "address", i.e., the caller and the callee telephone number. On the Internet, these addresses are referred to as IP addresses. In addition, the header portion of the packet contains additional information used by the source and destination computers to direct packets to the correct application, for example, directing packets to a web server. The payload portion of the packet contains application-specific data, i.e. the contents of an email or the content of a web page.

The data that is generated in the normal functioning of the Internet can include three potentially sensitive data types:

1. Addresses of the source and destination computers. This information identifies a specific host on the Internet, not a specific user. While this information might possibly be mapped to a specific computer, or even a specific individual, by itself it does not reveal the identity of individuals. To remove any risk, we will anonymize these addresses from potentially sensitive datasets to remove any possibility of mapping an address to an individual user. Appendix A illustrates the types of datasets that will be anonymized.
2. Application type information. While the information contained in Internet packet headers can imply a specific application sent the data, it is not a direct one-to-one mapping. However, this information could imply potentially sensitive information. For potentially sensitive datasets, we will anonymize any address information associated with application types making it impossible to associate application information with specific individuals. Appendix A illustrates the types of datasets that will be anonymized.
3. Application payload information. This could include everything from the content of email messages to the contents of web pages. No payload data from live traffic will be stored in the repository. We will collect security related payload information from self-propagating worms and viruses as part of our "Darknet" dataset (see Appendix A);

however we will not share any of this payload information with any researchers to prevent any possibility of identifying individuals associated with this malicious traffic. The only exception for sharing Darknet payload data is when the payload only consists of computer code from Internet worms and viruses.

The Next Generation Repository for Network and Security Data will collect and publish datasets in all three categories with careful attention paid to prevent any risk to subjects. First, addresses within potentially sensitive datasets (e.g., Internet traffic flows, packet traces; see Appendix A) will be modified to prevent researchers from identifying any individuals associated with the data. In general, information about applications will not be modified, as this information is only potentially sensitive if associated with a particular individual. Finally, no payload data from live traffic, which is potentially sensitive for obvious reasons, will be included. We will publish payload information that is part of Internet attacks. Rather than containing sensitive personal information, this data will consist of computer code from Internet worms and viruses. In addition, in terms of individual privacy, the virtual repository will not include or publish any personally identifiable information or datasets that allow a particular Internet address to be associated with an individual person.

Information about individual networked computers is represented in our data by an IP address (a phone number in our analogy above). The IP (Internet Protocol) address of a networked computer represents an individual machine, not a person. As computing and networking become increasingly mobile, IP addresses are often assigned on a random basis using DHCP (Dynamic Host Configuration Protocol). DHCP makes the correlation between a machine host and an individual person even more tenuous than in a static networking environment. In addition, a common approach is to allocate IP addresses using NAT (Network Address Translation) routing. In this case, the IP addresses of the machines served by the NAT router are collated to a single IP address. With NAT addressing, an IP address does not identify a specific host on the Internet but rather there is a many-to-one mapping.

While an IP address does not easily map to a single user, the possible combination of secondary datasets with datasets that are published as part of this project could make it possible to establish this mapping. We mitigate this risk of correlating IP addresses with individual users by using standard anonymization techniques to modify the IP addresses in published data sets. A secondary dataset, such as an administrative database from an Internet Service Provider (ISP), may associate a static IP address with an individual user (i.e., the account owner) for accounting or administrative purposes. It is important to note that these secondary datasets will not be part of the virtual repository and are not typically available to researchers or published elsewhere[2]. Moreover, making the association between an IP and the registered account owner does not necessarily reveal the identity of the user that is accessing the Internet at any given moment since multiple users can share the same host. However, the possibility that secondary datasets might be combined with datasets in the virtual repository is

[2] In order to compel an ISP to reveal the mapping between an IP address and the account owner associated with that IP, a court order or subpoena is required [3].

the potential risk that the project seeks to mitigate. To remove any chance of correlation of an IP address with a single individual, we will use several different existing techniques to anonymize sensitive data before storing it for future use. One example technique involves removing some of the identifying information from the IP address, similar to blacking out the last four digits of a phone number. This technique prevents identifying an individual user on the network while still allowing researchers to reach conclusions about related groups of users. More details about the anonymization techniques are presented in Appendix B.

As noted in Appendix A, certain types of data that are generated by regular Internet traffic needs to be anonymized. However, some classes of security data, which is destined for no one in particular, do not contain sensitive information. Security data may be captured by firewalls and black hole monitors (aka network telescopes or “Darknet” monitors) and is an important part of the cybersecurity data repository. In particular, our Darknet network monitor collects Internet traffic that is directed to an unused network space that does not contain normal daily communications or any legitimate traffic. It is space that does not serve any users, but rather includes impermissible activities such as malicious network scanning, misconfigurations, backscatter from denial-of-service attacks based on source-address spoofing, and malware that self-propagates aiming to infect more machines [4,5]. This data is usually not generated by individuals, and carries no legitimate communication. It is destined to a network space where there are no computers to respond to these unsolicited messages. In the normal functioning of the Internet, security data, such as Darknet data, is collected and used by networking professionals to mitigate real-time threats to the network. This category of data, because it is inherently malicious, probing and intended to harm, does not need to be anonymized. It presents minimal risks to individual privacy (see also Appendix A, section “Internet Blackhole Data”). Furher, our research team will never interact or intervene with any machine whose address is identified in the collected datasets, and therefore there is no direct or indirect communication with any individuals. To further ameliorate any privacy risks we will not share Darknet data with payload information to any researchers to avoid the possibility of revealing sensitive private information that could be carried within the payload. In cases where we are certain that the payload information contains only computer code from worms or viruses, the payload can be disclosed.

We also wish to note that the data collected in the repository represents information collected from the Internet community at large. While we do not target any specific groups, it is likely that this community includes vulnerable subjects such as minor children and pregnant women. However, there is no straightforward way to correlate any subset of the data with these individuals. In addition, the anonymization techniques we will use should make it impossible to correlate data with any individual vulnerable subjects. For this reason we believe that this project represents no particular risk to vulnerable subjects.

Data collected for the virtual repository comes from globally distributed Internet measurement points. While in some cases, this data might show a bias towards information about local network traffic, there is no effort made to include or exclude any specific groups of individuals for this measurement. The only identifiable group of individuals are those using computers with

similar IP addresses, representing a common service provider. However, the data included in the virtual repository is not targeted even at these groups, and is instead focused on understanding global Internet behavior.

b. Describe the application process for submitting materials to the repository:
There is no application process for submitting materials to the repository. Data will be collected by the project team on a longitudinal basis. The virtual repository will use existing Internet measurement devices to collect data. Data will be stored on dedicated servers accessible through a private web portal.

c. Waiver of Informed Consent
The research involves no more than minimal risks to subjects: In our assessment, the risks described above are minimal, and represent an impact on privacy that is no greater than the risk that individuals face in their regular use of the Internet. The datasets included in the repository represent large aggregations and pose no additional risk to individual privacy; we do not track or target any specific individuals. Any information within sensitive datasets (see Appendix A) that could potentially be linked to an individual will be anonymized prior to distribution to researchers to remove any risk to individuals. We utilize standard anonymization techniques described in Appendix B. In addition, we do not interact, intervene with any persons, we do not manipulate the environment of any individual nor we collect or distribute any private information that is individually identifiable.

We further diminish any potential risk by employing the policy framework that IMPACT provides. Any researcher requesting repository data is required to abide by certain data use terms (see Section II). When a dataset is requested, the researcher is required to justify the reason that the data is needed, and specify any other researchers from their organization that will be accessing the data. Researchers are required to safeguard the data using reasonable security practices. Further use or disclosure of the data is prohibited by these policy and binding agreements. In addition, probing, interaction or any other communication with a machine or a machine operator identified in the disclosed data is also prohibited. This policy framework enforces responsible data use, and holds researchers accountable for their actions.

Waiver of Informed Consent: The waiver of informed consent will not adversely affect the rights and welfare of subjects.

The research could not be practically carried out without the waiver: As we are studying the behavior of large populations, we do not expect to request or receive individual waivers or consents. This research could not be practically carried out without the waiver of informed consent. Subjects will not be recruited individually in this project.

Pertinent information after participation: This does not apply to our project.

II. Distribution and Dissemination

a. Application process for releasing data

The sponsoring organization, Department of Homeland Security, has put the following process in place that would apply to us as well as other data providers in the virtual repository. Application requests for data are only accepted from vetted researchers with valid IMPACT accounts at the IMPACT portal [11]. Vetting is performed by IMPACT's Coordination Center (ICC). RTI International, an organization based in North Carolina, serves currently as the ICC. This ICC committee will only accept account applications from researchers sponsored by valid research organizations, and will require letters of support from those organizations before an IMPACT account is granted. Further, before any data request is processed (1) the request needs to be approved by the ICC committee (and the data provider when necessary–see below) to ensure that the requested data is reasonable for the stated research goals, and (2) the user needs to agree on terms and conditions that highlight the acceptable use of data and the user obligations. More details are given next.

In order to streamline the process of data dissemination and congruent to the fact that some provided datasets have lower privacy risks than others, the sponsoring organization has asked data providers to categorize their data into the following classes: (1) Unrestricted; (2) Quasi-Restricted and (3) Restricted. Before this process change, all IMPACT data were treated as Restricted data and required the execution of a bilateral memorandum of agreement (MOA) between the ICC and the researcher *and* a separate memorandum of agreement between the data provider and the researcher. With the new procedure, a request for restricted data requires the execution of a *single* memorandum of agreement between the ICC and the researcher; all the terms and conditions and all the policy safeguards that were enforced with the foregoing process are now present in the ICC-Researcher MOA. In addition, all requests for Restricted data need to be approved by the data provider and the ICC committee. Data providers still have the ultimate veto when deciding if users are allowed to access their restricted data. This will ensure that this process does not violate any conflicts of interests or violate any agreements made when originally collecting the data. For datasets that are classified into the Quasi-Restricted category, the application process for releasing data is similar to the Restricted case with the exception that users are not required to execute a bilateral MOA. Instead, users must agree to an online "click-through" agreement that states the terms and restrictions for data use. As with the Restricted case, prior to any data release the data provider and the ICC committee need to approve the data request, and data providers have the ultimate veto when deciding if users are allowed to access this data. Finally, requests for Unrestricted data also enforce the researcher to accept the terms of the "click-through" agreement before any data is released. The difference with the Quasi-Restricted category is that only the ICC needs to approve the data request. We are providing (1) the template of the MOA signed by the researcher and the ICC and (2) the "click-through" terms of use as supplementary materials to this document. Categorization of our datasets is depicted in Appendix A.

In addition, we intend to distribute data we collect to a broader community of researchers. This data sharing will happen as part of our present and future collaborations with researchers at other institutions. In these cases, we will follow a similar procedure as above. We will

personally check the relevance of the research to the data requested. In addition, we will expect recipients to sign an agreement with the University of Michigan and/or Merit Network, Inc. stating their intended use of the data.

b. Ensure that recipient has obtained IRB approval for use of identifiable data

This does not apply to our virtual repository. Our repository does not include any data with personally identifiable information, and we will not share any datasets with identifiable information.

c. Procedures for verifying that the proposed use is consistent with the informed consent

This does not apply to our project as no informed consent is needed. By anonymizing all potentially sensitive data before storing it for research purposes, we expect that all data covered under this document will represent no privacy or security risk.

d. Data transfers under appropriate institutional agreements

As described above, the sponsoring organization has put in place a procedure in which researchers (data requestors) need to conform with in order to obtain access to the requested data. This policy framework adds extra risk control mechanisms to help diminish further any privacy risks that may be associated with network traffic data. Prior to data disclosure researchers must state their research objectives and why they are requesting the corresponding dataset. They must also enlist all other employees of their organization that they will have access to the shared data, and they must ensure that every team member safeguards the data using commonly accepted security practices. The binding legal agreements enforce authorized researchers to treat data with confidentiality, and avoid any attempts to reverse engineer, decrypt any anonymized data or link data back to any individual or group. Further, any potentially sensitive information must be de-identified before appearing to any publication or disseminated otherwise, and all restricted datasets must be securely purged upon completion of the project. The terms and conditions also prohibit interaction or probing with any machine that can be identified from the requested dataset. Finally, researchers agree that conformance with these terms might be audited by the ICC committee.

e. Methods for ensuring security and confidentiality of data

The only potentially identifiable information included in the virtual repository are IP addresses. To minimize the risk to subjects, all IP addresses from sensitive datasets included in our repository will be anonymized (see Appendix A). In particular, the last 11 bits of the IP address will be zeroed out. This aggregates groups of 2048 IP addresses to one anonymized address, preventing any correlation of the data with individual host computers and hence individual users.

Data collected as part of the virtual repository will be collected from existing Internet measurement devices. The data will then be transferred to and from the repository using secure

communications tools such as ssh and ssl-encrypted web transfers. All data will be stored on dedicated, secured servers that will only be accessible to authorized personnel. In addition, any data stored in the repository will have all personally sensitive information removed (e.g., all payload data from live traffic will be removed), and any potentially personally identifiable information anonymized (i.e., IP addresses from traffic flow data will be anonymized). The repository does not maintain any private information that is individually identifiable. Any data that is transferred from the repository to researchers will also be transferred over the Internet using secure techniques.

Data collected as part of the virtual repository represents normal data collected as part of the day to day operation of the Internet. While we believe that representatives of this community have some expectations of privacy, we believe that our methods of removing private information and anonymizing any potentially identifiable information would meet with general expectations of privacy on the Internet.

III. Data security Plan

a. Access plan and controls

Who accesses data: A subset of the researchers at Merit Network Inc. and at the University of Michigan that are participating in this project will have access to the local, dedicated server that is responsible for distributing information and datasets to interested researchers. This access will be used to administer and maintain the server and the associated applications. Researchers that are interested in accessing data collected as part of the virtual repository will be granted access to a specific subset of the data on a case by case basis.

Requirements for access: Vetted researchers are granted access to the requested dataset after following the procedure outlined in section “Application process for releasing datasets”. In addition, all data use agreements should be fully executed by all involved parties before any data is released. Finally, the researchers are able to access the subset of data requested only through secure communication channels (such as ssh).

b. Data security procedures

As described above, data will be transferred to and from the repository using secure communications tools. All data will be stored on dedicated, secured servers that will only be accessible to authorized personnel. In addition, any data stored in the repository will have all personally sensitive information removed, and any potentially identifiable information anonymized. Any data that is transferred from the repository to researchers will also be transferred over the Internet using secure techniques.

c. Subject identifiers

Plan to protect identifiers from improper use or disclosure: The only potentially identifiable information included in the virtual repository are IP addresses. To minimize the risk to subjects, all IP addresses from sensitive datasets included in our repository will be anonymized (see Appendix A). In particular, the last 11 bits of the IP address will be zeroed out. This aggregates

groups of 2048 IP addresses to one anonymized address, preventing any correlation of the data with individual host computers and hence individual users.

Subject identifiers: coded materials
No encoded materials will exist in this repository.

Subject identifiers: disposal of identifiers or the code
For datasets that contain identifiable information the anonymization techniques described in Appendix B will be employed. This project will not make use of encryption keys.

IV. Governance and Oversight

a. Role of repository director
The repository director will be responsible for data collection, data curation (e.g., anonymization of the data), data dissemination and data access. The director is also part of the authorizing committee that reviews data requests for Restricted and Quasi-Restricted datasets and ensures that the requested data is reasonable for the stated research goals. The committee then decides whether the Researcher should be granted access to the dataset requested. The repository directory is also responsible for IRB compliance and execution.

b. Role of governance committee (if any)
There is no external governance committee for this project. The repository is governed by the project team.

c. Plans for sustaining the repository over time
Continuing funding: The current phase of this project is expected to be completed by the end of 2017. Should the sponsoring organization (DHS) is interested in extending the lifecycle of the project, the project team will submit a proposal for a renewal.

Plans for data destruction: Data will remain in the repository as long as their scientific merit continues to exist and the project team has the resources to sustain them. Otherwise, the data will be securely purged. Non-functioning hard disks will be physically destroyed.

Appendix A: Internet Data Types and Their Risk Assessment

Summary Table:

	Data Type	Description	Risk Potential[1]	Security Data[2]	Mitigation[3]	Data Class[4]
1.	BGP Updates	Internet wide routing protocol messages	N/A	N	N/A	U
2.	BGP Dumps	Internet wide routing table information	N/A	N	N/A	U
3.	Traffic Flows	Sampled information representing traffic flows at routers	2	N	Anonymizing lower 11-bits	QR
4.	Packet Traces	Header information sampled from traffic flowing across the Internet	2	N	Packet contents not collected. Anonymizing lower 11 bits	N/A
5.	Blackhole Data ("Darknet" data)	Information gathered by observing unoccupied portions of the Internet	2	Y	Remove payload	R
6.	Network Mgmt	Summary information collected from Internet infrastructure equipment	N/A	N	Anonymizing lower 11 bits	N/A
7.	HTTP Logs	Log files from web servers	2	N	Anonymization using work by Vern Paxon	N/A
8.	HTTPS certificates	SSL/TLS certificates from web servers	1	N	N/A	U
9.	RADb data	Routing Arbiter Database archival records	N/A	N	N/A	U

[1] **Risk Potential –** This field is based upon the type and amount of information that is present in the data traces; (1) implies only information about the origin end point is present, (2) implies that information regarding the origin and destination end point is present as well as the category of applications that generate that traffic, (3) implies that the complete contents of packets are present.

[2] **Security Data –** This field is used to indicate whether the information present in this data trace is essential and useful for network security administrators to identify potentially malicious activity. It is generally pre-filtered to only include malicious events.

[3] **Mitigation –** The mitigation techniques mentioned here are described in greater detail in Appendix B, "Anonymization Techniques for Mitigating Risk Potential of Internet Data"

[4] **Data Class –** The data categories that IMPACT data are classified into: Unrestricted (U); Quasi-Restricted (QR) and Restricted (R). Data that are currently not offered by our repository our marked with N/A.

1. BGP Update Messages

The BGP routing protocol is used to exchange routing information between different autonomous systems in the Internet. Each BGP router sends information describing its own part of the Internet in the form of BGP Update messages, which are sent to every BGP device. These messages are the way that network reachability knowledge is propagated around the Internet. When various networks are added or if they lose connectivity with the rest of the Internet, the BGP protocol is responsible for propagating this information throughout the Internet. Monitoring these messages can reveal important properties of the Internet, specially related with stability. Due to various inherent inadequacies in the BGP protocol, it is has been shown to be vulnerable to various security attacks which can have a significant impact on the performance of the Internet. Monitoring these update messages can also reveal critical information regarding various ongoing attacks.

By its very distributed nature, the BGP protocol messages need to be visible to the entire Internet. These messages only describe aggregate network topology and therefore do not pose any privacy risk to any individuals or organizations. This information is already being made publicly available by some institutions via sites such as www.routeviews.org.

2. BGP Routing Table Dumps

BGP routing table dumps are simply attempts to capture the state of the BGP routing protocol at a particular instance of time for future study. The BGP routing tables at any instance in time represent the state of various parts of the Internet, and by studying them at various instances in time, and form various locations on the Internet, one can attempt to reconstruct the behavior and evolution of the Internet.

Similar to BGP update messages, these routing table dumps only represent the reachability of large aggregates of networks. In no way do they represent any privacy risk to any individuals or to any organizations.

3. Traffic Flows via Netflow

As traffic flows across the Internet, it is important to be able to study some of its key characteristics which can then be used to enhance and improve the performance, reliability, and security of the Internet. However, due to the sheer volume of the traffic, it is only possible to collect small samples of this traffic. Netflow is a commonly used technology to aid in this collection. Netflow provides coarse statistics on the types and amount of traffic that traverses a particular point in the Internet. It is important to note that using Netflow it is not possible to capture the detailed contents of traffic as it flows across the Internet. One can only capture very basic information, the addresses of the end points exchanging information and a very coarse estimate of what type and how much traffic traversed that link in the network.

While it is conceivable that simply revealing the presence of Internet traffic between certain end points may constitute some privacy risk, it should be noted that the identity of an end point generating traffic destined for the Internet does not directly translate to a unique individual without the use of secondary data sets. We will not publish any of these data sets as a part of this project. Moreover, we will use various well-known anonymization techniques to further remove any potential risks. Anonymization techniques essentially serve to mask even the identity of the end points that are generating Internet traffic, thereby completely mitigating any potential privacy risks to individuals or organizations.

4. Packet Traces

Using specialized hardware, it is possible to capture entire traffic flows at a specific point in the Internet. While it is possible, with these devices to capture the contents of the entire traffic flow, we do not operate these devices in that capability. We limit ourselves to only capturing information contained in the packet headers, which characterizes the source and destination of the traffic, as well as information regarding what category of traffic the packets represent. The information contained in packet traces is a little bit more detailed than the information obtained from the Netflow capability described above. The information obtained from the study of packet flows can be extremely useful in the study of the nature of traffic flows in the Internet, which can lead to insight into the evolving nature of the Internet and play an important role in the design of future Internet systems.

Similar to Netflow traffic logs, the information contained in packet traces, can at best be used to identify a traffic end point. It is not possible to further identify an individual as the source of the traffic represented in a packet trace on the basis of information represented in the packet trace without the use of secondary data sets. These data sets will not be published as a part of this project. We further mitigate any potential privacy risks by using anonymization techniques to mask the identities of the end points generating the traffic.

5. Internet Blackhole Data (Darknet data)

The Internet consists of a collection of nodes each of which is assigned a unique address. Various organizations can request and obtain permission to use certain address ranges for their computers. However, it is often the case that an organization is assigned a certain range of addresses and it does not use all of it. Such assigned but unused address space is known as a blackhole or a Darknet. An Internet blackhole represents an address range that has been assigned but is not being used by any legitimate Internet devices. One would expect that as there are no active nodes in these dark address ranges, there would be no traffic destined for them. However, this is not the case; computer viruses, worm and other self propagating harmful software, cannot distinguish between used and unused address spaces. Therefore, if we were to carefully watch any activity directed towards a blackhole, we would most likely observe such malicious software. It is important to study traffic data obtained by monitoring a blackhole, as it can provide useful insight into the propagation of viruses and worms that would otherwise be much more difficult to do. It is further a tool for network operators and researchers to help them identify malicious scanning, victims of distributed-denial-of-service attacks and other types of misconfigurations [4,5].

As a blackhole by definition does not contain any active legitimate nodes, the privacy risks to individuals are minimal. Any traffic that is being directed towards a blackhole, represents malicious activity that exceeds the limits of normal, daily communications. It is unauthrorized traffic, destined to an unused/ungoverned network space that should not receive any traffic. We further mitigate the privacy risks of disclosing any data that can be linked to an identifiable individual by filtering out the payload. Any packet payload data collected from a blackhole will be removed unless it is part of a well-known worm, virus, or other malware infection attempt. This data is generally available from other security sources and represents no risk to individuals using the Internet.

Because any traffic arriving to our Darknet monitor is unsolicited, inherently malicious and does not represent regular Internet communications, we will not anonymize the IP addresses of the machines that this traffic originates from. Anonymizing the source IP information significantly diminishes the research utility of this dataset and hampers its applicability to scientific experiments. This, however, enables some potential risks that we believe are tolerable due to the data access mechanisms we enforce and the policy controls that IMPACT provides (described in the next paragraph). First, there is some potential legal liability for the source of traffic reaching a blackhole. Second, since Darknet data involves traffic generated from compromised, vulnerable hosts running malicious software, there is the risk of exposing the IP addresses of those vulnerable machines.

We diminish any potential risks by employing the policy framework that IMPACT provides. Any researchers requesting repository data are vetted by the ICC committee and are required to abide by certain data use terms. When a dataset is requested, the researcher is required to justify the reason that the data is needed, and specify any other researchers from their organization that will be accessing the data. Any actions beyond the intended data use are

forbidden. Further, disclosure of the data to any persons other than the authorized ones is prohibited by the policy and binding agreements that IMPACT enforces, and data must be sanitized, summarized or anonymized before being published. In addition, probing, interaction or any other communication with an IP identified in the disclosed Darknet data is also prohibited. Interactions with any individuals are also not allowed. In addition to these policy controls, we enforce privacy-preserving access controls; secure access to data is granted only to authorized researchers, and researchers are not permitted to copy or disseminate the data out of their hosting location (i.e., data are prohibited from leaving Merit Network). With these controls in place, we aim to hold researchers accountable for their actions and, thus, mitigate any plausible legal liability risks and the risk of interaction with exposed machines encountered in the Darknet data. At the same time, we do not deteriorate the scientific quality of the dataset by employing heavy anonymization/aggregation.

Notably, a similar paradigm of sharing Darknet data with the networking community (raw data that includes IP addresses and, unlike our case, packet payload as well) has been established since 2009 by CAIDA at University of California at San Diego [7, 8]. CAIDA implements the PS2 (Privacy-Sensitive Sharing) framework that integrates privacy-enhancing techniques with a policy framework (similar to the legal framework that IMPACT offers) that enforces researchers to abide by various data use terms and conditions before data access is granted. This enforces responsible research, holds researchers accountable regarding the data use, and has helped CAIDA to mitigate any probable privacy risks to individuals. At the same time, it significantly enables the search for scientific knowledge.

6. Network Management Data

Network Management Data is generally represented in the form of messages that are transmitted using the Simple Network Management Protocol (SNMP). These messages essentially convey gross aggregated statistical information from various network devices. Using these messages it is possible to study aggregated Internet traffic behavior. Aside from basic statistical information that is recorded via these messages, SNMP messages are also used to transmit information conveying the physical well being of these devices. Therefore it is important to study these data logs, as they may often reveal important causal information that subsequently is observed in various other types of data logs.

Due to the coarse grained nature of these messages, they do not constitute any privacy risk to individuals. These messages do however reveal information regarding the operation of the network, which is often considered proprietary information by commercial Internet service providers. To further mitigate any risk we will employ anonymization techniques in a way that critical network operator information is not leaked to unauthorized parties.

7. HTTP Logs

Hyper Text Transfer Protocol (HTTP) is the foundation on which the World Wide Web is based. It is used to transmit information from web sites to individual users. When a user visits a web site using a browser on a computer, the web site is able to record the sequence of activities and interactions that represent the communication between the computer and the web site. It is common practice in the current Internet to log such activity as it can be used to improve the operation of the web site. By studying the interaction of various users with a web site we can hope to achieve a better understanding of how users access information. This in turn can lead to significant insight I the design of user interfaces and the design of information distribution systems.

Using HTTP logs it is possible to identify an end point that is attempting to access data from the web site. Using the HTTP logs alone it is not possible to identify a particular individual. Moreover, we will utilize anonymization techniques that mask the origination end point to further mitigate any privacy risks to individual users. This data is not critical to maintaining the security and integrity of the Internet, and can therefore easily be anonymized.

8. HTTPS Web Certificates (X-509 Certificates)

Communication between a user's browser and a Web server over the Hyper Text Transfer Protocol (HTTP) can sometimes be vulnerable to security attacks. Due to the fact that exchanged messages are not encrypted, connections over HTTP could be susceptible to eavesdropping. Sophisticated attackers might also launch man-in-the-middle attacks in which the true Web server that the user is intended to connect to is impersonated with a fake one; this fraudulent activity aims at identity stealing by tricking users to reveal their passwords or other sensitive information. HTTPS (i.e., HTTP Secure) secures the HTTP communication by employing the SSL/TLS protocol. SSL/TLS encrypts the messages exchanged and provides a reasonable guarantee to the user for the authenticity of the intended server.

A key component for the operation of HTTPS is the installation of X-509 certificates on the Web server. Every time an HTTPS connection is initiated from a user, the Web server presents its SSL/TLS certificate. The browser uses this information to validate the identity of the Web server and get prepared for an encrypted communication session. In this dataset, we provide the X-509 certificates of publicly available HTTPS-serving Web servers, as collected by the study of the HTTPS ecosystem presented at [9]. It is important to study these data since it provides information about the security posture of publicly available Web servers, and can help identify possible misconfigurations.

This dataset includes information about Internet assets (i.e., Web servers) that are publicly available to anyone with a Web browser, and therefore does not identify any individuals nor it contains any private information. Note that there is a potential reputation risk for organizations

or network administrators that run Web servers with insecure (e.g., expired) HTTPS certificates[3]. However, this poses no greater risk than the risk imposed with regular use of the Internet. Nonetheless, to further mitigate these risks, vetted researchers requesting this dataset (as well as any dataset) are agreeing to terms of use before any data get released (see Section II). Such terms restrict data use beyond reasons other than their stated research intends, allows data access only to specific individuals disclosed during the data request, and asks the researchers to agree on protecting and securing the data. It is important to note that this dataset is also directly available from its authors [9].

9. Routing Arbiter Database archival data (RADb data)

The Routing Arbiter Database (RADb) is a service offered and hosted by Merit Network, Inc. to the Internet operational community [10]. It is a public *routing registry* that provides up-to-date information about the routing policies of Autonomous Systems (i.e., an Autonomous System (AS) is an IP network run by a network operator). Autonomous systems exchange routing information using exterior gateway protocols such as the BGP protocol described above. This exchanged routing information is needed to "glue together" these networks and enables the traversal of Internet packets from one network to the other.

Routing decisions, though, can be complex and are not always based on technical parameters like topology and link speeds. Much like the aviation industry in which certain airline operators follow policies that prevent them from "routing" their fleet into certain countries or regions due to geo-political reasons, Autonomous Systems may also impose constraints on routing. Such constraints might include "routing policies" that prevent forwarding packets through certain networks or "access policies" that block packets originating from certain ASes. The RADb system provides a centralized database where network operators can register their routing polices. This enables organizations to troubleshoot routing problems, automatically configure backbone routers, generate access lists, and perform network planning [10]. Studying archival records of this dataset can help researchers understand the stability of BGP routing and the Internet.

The RADb dataset contains information about networks (ASes) and their routing polices, and does not include any personally identifiable information nor any private information about persons or organizations. Orthogonal to the fact that this dataset is provided via the IMPACT repository, Merit offers archival RADb records via public, anonymous FTP access (see [10]).

[3] Note that several companies offer reputation lists that provide "scores" for the security posture of a network are readily available (see references within [6]). Therefore, merely sharing this dataset does not uniquely expose mismanaged networks or hosts. In any case, exposing unsecure networks can help network administrators strengthen the security defenses of their networks by employing firewall and filtering rules that block access from unsecure, vulnerable, mismanaged and ill-behaving networks/hosts.

Appendix B: Anonymization Techniques for Mitigating Risk Potential of Internet Datasets

Summary:
Here we provide a brief description of the two well-known anonymization techniques that we intend to use to mitigate the risk potential of Internet datasets. The first is most useful for anonymization of source and destination endpoints that are generating Internet traffic. It should be noted that even though the identity of the source and destination end points does not uniquely identify individuals without secondary data sets, in some cases we may wish to mask even that information, via the use of this tool. The second tool that we intend to use is essentially a trace transformation tool, which can be used to overwrite or delete specified fields in a dataset that might be sensitive. This tool is most useful for anonymizing log files that contain potentially sensitive information. Further details regarding these tools can be found in the references at the end.

1. Low order bits anonymization

This methodology eliminates the "low order" 11 bits of an IP address, destroying information encoded in that part of the address. Recall that IP addresses consist of 32 bits, which identify a network endpoint. When the last 11 bits of an address are removed, IP addresses whose first 21 bits are the same, but vary only in the last 11 bits will appear identical. This approach essentially aggregates the possible 2048 end points reflected by the last 11 bits into a single entity. For example, 141.213.4.4, 141.213.4.5, 141.213.5.1 would all appear as 141.213.0.0 in this scheme. This aggregation value, and anonymization approach, has been used successfully by other academic and non-profit institutions, such as Internet2 (http://www.internet2.edu/observatory/archive/data-collections.html) to publish data containing IP addresses. A software suite that can be used to accomplish this is CAIDA's CoralReef [2].

2. Packet Trace Transformation

Vern Paxon in his paper describing a high-level programming environment for packet trace anonymization and transformation [1], describes a general framework which can be used to mask important information from various Internet datasets. Though the framework is general enough to be useful in anonymization of source and destination end points of data, it is most useful for anonymization of various log files such as those generated by FTP, SMTP or HTTP. The trace transformation technique can be used to search and replace specific fields in a log file with user specified values. If the user specifies that a blank transformation value be used, then the various fields in the dataset will simply be blanked out after processing via this tool.

Though we intend to use the "low order bits anonymization" technique to anonymize datasets that describe packets or flows, we will use the trace transformation techniques and tools described in [1] to anonymize any log files that we publish via the virtual data repository.

References:

[1] R. Pang and V. Paxson, A High-level Programming Environment for Packet Trace Anonymization and Transformation, Proc. ACM SIGCOMM 2003, August 2003.

[2] David Moore, Ken Keys, Ryan Koga, Edouard Lagache, and k claffy, White paper: The CoralReef Software Suite as a Tool for System and Network Administrators, 2001
http://www.caida.org/tools/measurement/coralreef/

[3] Aaron Burstein. "Amending the ECPA to Enable a Culture of Cybersecurity Research". *Harvard Journal of Law & Technology*, 22(1):167-222, December 2008

[4] Zakir Durumeric, Michael Bailey, and J. Alex Halderman. An Internet-wide View of Internet-wide Scanning. *In 23rd USENIX Security Symposium (USENIX Security '14)*, San Diego, California, August 20-22, 2014

[5] Eric Wustrow, Manish Karir, Michael Bailey, Farnam Jahanian, and Geoff Houston. Internet Background Radiation Revisited. *In Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement (IMC '10)*, Melbourne, Australia, November 2010

[6] Jing Zhang, Zakir,Durumeric, Michael Bailey, Mingyan Liu, Manish Karir. On the Mismanagement and Maliciousness of Networks, NDSS 2014

[7] K. Claffy and E. Kenneally, "Dialing Privacy and Utility: A Proposed Data-Sharing Framework to Advance Internet Research," in *IEEE Security & Privacy*, vol. 8, no. 4, pp. 31-39, July-Aug. 2010.

[8] CAIDA Research - Institutional Review Boards (IRB) Approval Process, http://www.caida.org/home/about/irb/, 2012 (Last visited: August 2nd, 2016)

[9] Zakir Durumeric, James Kasten, Michael Bailey, J. Alex Halderman, Analysis of the HTTPS Certificate Ecosystem, Proc. of the 13th Internet Measurement Conference (IMC'13), Oct. 2013.
Publicly available by its authors [3]: https://scans.io/study/umich-https (Last visited: August 2nd, 2016)

[10] Routing Arbiter Database, Merit Network, Inc. http://www.radb.net
Publicly available at: ftp://ftp.radb.net/radb/dbase (Last visited: August 2nd, 2016)

[11] Information Marketplace for Policy and Analysis of Cyber-risk and Trust (IMPACT), Web portal: https://www.impactcybertrust.org (Last visited: August 2nd, 2016)