

Project Bloom: Empowering the Security Research Community Through Data Products and Computing

Minaxi Gupta
School of Informatics and Computing
Indiana University at Bloomington
minaxi@cs.indiana.edu

Gregory R. Travis, David A.J. Ripley
Advanced Network Management Lab (ANML)
Indiana University at Bloomington
greg,daripley@indiana.edu

Douglos D. Pearson
REN-ISAC
Indiana University at Bloomington
dodpears@ren-isac.net

I. INTRODUCTION

Cybercrime is a thriving multi-billion dollar underground economy. Web sites connected to phishing, malware, and scam spring up by the millions every day, and users are lured to them through creative spam, social engineering, and search-engine manipulation. Botnets, or armies of compromised user machines, are a major enabler in much of cybercrime today. They send spam, host malicious Web sites, and scan for vulnerable Web servers and other user machines to add to the army of botnets. Each of these aspects of cybercrime need to be studied in order to devise effective solutions against the problem.

A crucial impediment encountered by the security community interested in studying cybercrime is the availability of appropriate data sets that offer insights into attackers' modus operandi. Good data sets are either not available or available only within a closed group. Even when they are available, they often do not span long enough a duration and the data collection setup is not well documented. The processes required to procure data are also lengthy or riddled with legal hassles. Nearly all of the existing or past projects [13], [9], [15], [3] that have had the goal of making cybersecurity data available to researchers suffer from one or more of these limitations. Even when researchers get access to data, storing, computing on it, and indexing it for future use often involves resources not easily available at a single institution. Combined, these reasons hamper the progress cybersecurity researchers could collectively be making.

Project *Bloom* is motivated by all the above concerns. With a focus on three key data sets widely used by researchers, our first goal is to provide well-curated, long-term, raw data sets to the security research community in an efficient manner using an established federation that can make the data acquisition process almost instantaneous. Our second goal is provide researchers compute power close to the data sources so they can avoid moving large data sets across the Internet and can focus only on taking back the key derived results. Our third goal is to offer computationally rich data products derived from raw data. The primary motivation behind this goal is to overcome the common data-processing hurdles and enhance the overall research productivity of the community.

To accomplish our goals, we have assembled the current best-possible data sources for `NetFlow`, *darknet*, and *passive*

DNS data. We have surveyed the research literature to come up with an initial offering of derived data products. To store raw data sets and derived data products, we will make extensive use of Indiana University's in-line and near-line storage infrastructures with petabytes of storage. To compute derived data products and to make computational resources available to researchers, we will again leverage Indiana University's 1120-core supercomputer cluster. We will adapt the data product offering based on feedback from the users.

Our project team consists of members from Indiana University, its Advanced Network Management Lab (ANML) [1], Research and Education Networking Information Sharing and Analysis Center (REN-ISAC) [4], Internet2, and Security Information Exchange (SIE) at Internet Systems Consortium (ISC) [11]. The team members include security researchers, operational personnel, staff with extensive systems development and integration experience, and others with a long history of policy determination and implementation.

II. DATA TYPES AND SOURCES

Project Bloom will provide computationally rich data products and Cloud Computing primarily for three types of data widely used by the research and operational community. These are *NetFlow*, *darknet*, and *passive DNS* data sets. Here, we describe each type of data, its importance in research and operations, issues in its availability, and the sources we have lined up for making it available to the research and operational communities.

A. *NetFlow* Data

Modern routers use the `NetFlow` protocol [7], originally developed by Cisco Systems, to collect IP traffic information. A typical `NetFlow` record contains information about a *flow*, where a flow is defined by the 5-tuple: (source IP address, destination IP address, source port, destination port, and IP protocol). In addition to the fields defining a flow, `NetFlow` records also typically contain the number of bytes and packets observed in the flow, the union of TCP flags if the protocol was TCP, autonomous system (AS) numbers of neighboring source and destination ASes and other routing-related information, including CIDR masks of source and destination networks and router ingress and egress interfaces.

`NetFlow` data is widely used by researchers and operators for traffic engineering, network provisioning, and for iden-

tifying network security and performance problems. Almost all organizations in the world that are connected to the Internet and run modern routers are capable of generating and collecting NetFlow. The sharing of this data, however, is almost non-existent. To our knowledge, the only source of a high-quality NetFlow dataset of any size is the Internet2 Observatory [12], which makes NetFlow records from the Internet2 Network available to researchers. The Observatory web page lists 45 projects from around the world that have used the data in the last several years. To expand on the Observatory’s offering, our project will serve NetFlow data from the following sources:

- **Internet2:** This data provides NetFlow data from Internet2 backbone, which spans over 300 member institutions, including leading U.S. universities, corporations, government research agencies, and not-for-profit networking organizations.
- **Indiana University:** The Internet2 data’s view of commercial, commodity Internet traffic is limited. We will overcome this limitation by providing telecommunications data from Indiana University.

While the Observatory is a useful source of NetFlow data, it has some limitations. The first is the proposal process required of researchers to gain access. While justified, the process is time consuming and may frustrate researchers. A second limitation is that privacy concerns prohibit distribution of *unmasked* data. Currently available data has the low 11 bits of all IP addresses set to zero. Any analysis can only be at the granularity of /21 prefixes. Yet another limitation is data *sampling*; generating NetFlow records can be computationally expensive for routers and may compromise functionality. Finally, the Observatory, in common with other available data sources, currently does not offer rich data products and high performance computing resources that can significantly enhance the utility of the data to the community. With help from Internet2 team members, we will address these concerns in the following manner:

- **Accessibility:** We will provide a comprehensive technical, policy, and federation framework for providing research access to NetFlow data, which will cut down access times and improve the quality of access.
- **Unmasked data:** While it would be ideal from a researcher’s perspective to gain unrestricted access to unmasked NetFlow data, doing so could have serious privacy implications. We propose the following trade-offs to increase the utility of NetFlow data: We assign a time-limited unique ID to the source and destination IP addresses before the NetFlow records hit persistent media, allowing researchers to draw conclusions about the behavior of individual IP addresses, and to allow the study of security and provisioning issues in depth while adhering to all applicable data management policies. In addition, only the most trusted group of researchers will be allowed access to the datasets containing these identifiers, to provide accountability and mitigate the risk of reverse mapping the identifiers.

- **Unsampled data:** Technical limitations prevent Internet2 from supplying unsampled NetFlow records. To overcome this limitation, we will make unsampled telecommunications data available from Indiana University.

- **Data products and high performance computing:** On NetFlow data, as well as the darknet and passive DNS data sets described subsequently, we will offer rich data products and high performance, parallel computing facilities, which will span across data sets.

B. Darknet Data

A darknet is an allocated, advertised prefix, but without real network hosts attached, but capable of receiving packets, which may then be collected and monitored. Packets coming to a darknet IP are by definition unsolicited, and may be the result of either misconfigured hosts elsewhere on the network, or a direct or indirect effect of malicious activity. Darknets are a valuable source of information for various kinds of nefarious activities. An example is backscatter packets generated by valid network hosts in response to packets with source addresses spoofed to an address belonging to the darknet. Backscatter packets can be an important source of information about various kinds of denial-of-service (DoS) attacks, including those targeting DNS infrastructure. Another important category of packets often seen at darknets include scanning activities, such as that from network worms which are trying to probe for vulnerable targets.

The darknet data is considered an important resource for security researchers and operators alike. Unfortunately, there are no publicly-available sources of darknet data today. The Internet Motion Sensor (IMS) project [13] monitors many IP prefixes but the data is not available to all researchers. Similarly, Team Cymru [15] offers software that allows an organization to set up the monitoring of its own darknet but we are not aware of any organizations that make their data available publicly and in real time. The Cooperative Association for Internet Data Analysis (CAIDA) intends to make the darknet data from the University of California, San Diego available to public through the DHS-sponsored PREDICT repository. We will have access to the data through a public interface. In addition to making all our darknet data available to public in a real-time manner, we would derive computationally-rich data products available publicly as discussed in detail in Section III.

There is some variation in the behavior of different darknets, and some debate about what constitutes a darknet. The simplest darknets are packet sinks, and do not solicit or respond to any network traffic. At the other end of the spectrum are darknets where the sensor machines either siphon off traffic to other machines, either real or virtualized - *high-interaction honeypots*. Our current sources of darknet data are either entirely dark, or are low-interaction, carrying out limited dialog (e.g. the completion of the TCP handshake process) in order to gather additional information from source hosts. In the future, we plan to add data from high-interaction honeynets.

Darknets are most useful when their address range is not known. Because of this, our list of current darknet data sources

only contains their geographical locations and diversity aspects and not their institutional identity.

- **Midwest:** We have 6 separate /24 IP prefixes in two separate geographical locations in the Midwest.
- **East coast:** We have two /24 prefixes at two locations on the east coast.
- **South:** Four of our /24 darknets reside in the south of the U.S.
- **West coast:** We have one /8 prefix in the west coast.
- **Australian subcontinent:** Our darknets data includes two /24s in New Zealand.

C. Passive DNS Data

The DNS is a prerequisite to virtually all Internet communication, both good and bad. The term *passive DNS* is often used to refer to DNS responses obtained by the local DNS servers of an organization from (authoritative) DNS servers around the world. Though the passive DNS data is a result of DNS queries made by the clients, it does not contain any information about the clients themselves. When collected at a global scale, the passive DNS data is an invaluable resource that can help us learn about the popularity and behavior of hosts and domains on the Internet. The research community also values it for its importance in learning DNS access patterns related to malicious activity. For example, spamming bots typically require DNS to determine MX records corresponding to the addresses they are going to spam. Similarly, even though miscreants hop from domain to domain in order stay ahead of the discovery and take-down of their websites, bots such as Conficker [10] and Torpig [5] have to resolve domains in order to fetch instructions from their command and control servers. Even the users around the world who fall prey to malicious websites have to resolve the domains before they can access the scam, phishing trap, or malware. In each of these cases, intelligent correlations on the passive DNS data opens up exciting research avenues in containing malicious activities on the Internet.

Fortunately, there are a few independent sources of passive DNS data. While they allow researchers’ access to data, the process is often lengthy, undocumented, and sometimes expensive. Further, none of the passive DNS projects make derived data products available, which is a major thrust of project Bloom. By combining them, we believe we can offer an unparalleled view of DNS behavior around the world. Our data sources are the following:

- **SIE:** The Internet Systems Consortium (ISC) has set up the Security Information Exchange (SIE) [11], which collects passive DNS data from 15 large ISPs and commercial DNS service providers around the world, including a US-based Tier1 ISP, two US Cable/DSL access providers, and four US-based Universities. The rest are commercial DNS services in the US or ISPs based in Europe. The SIE is the largest resource for passive DNS data today. Our project team includes SIE personnel so we can offer rich data products derived from SIE and other data sources.

- **DHDB:** The DNS History Database (DHDB) Project [6] collects passive DNS data from 8 data sources, including University of Auckland (New Zealand), France, Norway and other locations.

III. DATA PRODUCTS

We will offer computationally rich, derived data products falling in two major categories. The first category will include products derived from each individual data type and the second will include products that are derived by combining data types. In this section, we outline the products we plan to offer initially. To derive these data products, we surveyed how researchers have used NetFlow data available through Internet2 [12] and CAIDA’s offerings of various data sets [8]. We also consulted the passive DNS data providers to learn ways in which researchers and operators were using their data [11]. To understand how darknet data sets are used, we examined research papers that document the use of that data [13], [14]. Finally, some data products are a result of the team members attending DHS-, I3P-, and Internet2-sponsored workshops where researchers from around the country gathered to discuss the data needs of the security research community (in some cases, among other things). We emphasize that these preliminary offerings will be adapted based on feedback from the user community.

A. Ready-to-use NetFlow-based Matrices

Our first set of data products are matrices based on NetFlow data. The goal behind these matrices is to relieve researchers and operators of the common data-intensive pre-processing so they can focus on actual analysis. Our initial offering will consist of five matrix-based products derived from NetFlow data. Each of these matrices will be maintained for 30 days on a rolling basis but the NetFlow data used to derive them will be maintained for 10 years, allowing researchers to request data or computation for any time period less than 10 years. *Excluding the matrices, the basic NetFlow data would cost 90TB of storage.*

IP-based matrices: For each of the UDP and TCP protocols, these matrices will capture aggregate traffic between a pair of IPs at a 15-minute granularity. The IPs will be anonymized as described in Section II. An example of such a matrix is shown in Figure 1.

		Source IP (Anonymized)					
Dest. IP (Anonymized)							
				Aggregate			
				Traffic Values			

Fig. 1. Matrix containing traffic volume for each (anonymized) IP seen over a 15-minute time period.

This basic matrix type has many uses. A researcher could combine these matrices to determine aggregate traffic between a set of IPs at a daily, weekly, or monthly granularity. They can view TCP or UDP separately or together. Similarly, researchers interested in learning about traffic aggregated by IP address or network can combine appropriate rows and columns. Temporal analysis on these matrices could be used to spot cases when traffic to or from a particular IP, prefix, or protocol exhibits statistically or historically anomalous behavior.

Accounting for the number of distinct IP addresses seen on the Internet2 network during a typical 15-minute interval, retaining these matrices on a 30-day rotating basis would require approximately 3TB of persistent storage. Separating TCP and UDP traffic would multiply this storage requirement by a factor of two. *The total persistent storage for this product would be 6TB per month.*

Port-based matrices: These matrices represent aggregate traffic between a pair of ports at a 24-hour granularity. (In our experience, these matrices tend to be dense, so keeping them at a 15-minute granularity is not feasible.) These can be used to study temporal variations or to detect anomalous behavior on specific ports. *Retaining matrices for this data product for both TCP and UDP traffic would require about 2TB of persistent storage per month.*

Router interface-based matrices for each port: For each port and protocol, these matrices will capture aggregate traffic information per router interface at a 15-minute granularity. The Internet2 core routers have a total of approximately 850 hardware and virtual interfaces - five orders of magnitude smaller than the number of IP addresses seen during a typical Internet2 day. The relatively small size of the resulting matrices will allow us to calculate them separately for IP destination port numbers of particular interest or researchers (e.g. TCP port 22, UDP 137, and other commonly attacked ports.) Assuming approximately 200 ports of interest, the relatively small size of these matrices means their storage would require *approximately 108GB per day, for a total of 3.25TB per month.* This product will allow researchers to make per-port deductions about network behavior not possible with other data products; they will be able to aggregate traffic by router interface, destination port, or IP protocol. Temporal analysis will also be applicable for detecting anomalous traffic.

ASN-based matrices: For both TCP and UDP protocols, these matrices will capture aggregate traffic information per autonomous system number (ASN) at a 15-minute granularity. Researchers will be able to view aggregate traffic at an ASN (for TCP, UDP, or both) to gain insights into inter-organizational traffic behavior. While the actual number of allocated ASNs in the Internet (approximately 40,000) would result in a relatively large and densely-populated matrix, the number of ASNs seen on the Internet2 network is a magnitude smaller. *Using a 15-minute calculation and a 30-day retention policy, this data product would require approximately 1TB of persistent storage.*

IP-based, per-protocol matrices for specific ports: The IP-

based, per-protocol matrices outlined earlier do not offer a per-port view. While it is not feasible to enumerate these matrices for each of the 65K ports, we will provide per-port matrices for ports interesting from the point of view of security. For example, the list of ports of interest maintained by the REN-ISAC ¹ could be the basis of these matrices. This list includes approximately 100 (port, protocol) pairs. The availability of these matrices will be for 30 days also but due to storage issues, we will compute them at daily granularity instead of doing so every 15 minutes. *This data product will require an estimated 6TB of persistent storage for a period of 30 days.*

B. Ready-to-use Darknet-based Products

Our second set of data products are based on darknet data.

Annotating request and response packets: Packets received by a darknet are one of two types: a direct request (e.g. an attempted attack or reconnaissance) or response from a third-party machine to source-spoofed packets from an attacker's connection attempts. It is not always possible to differentiate request packets from response packets, but the distinction can be important for research. Since this distinction is not present in darknet data packets by default, we will develop heuristics to annotate each packet as "request", "response", or "unknown". There are two basic heuristics that are likely to make the distinction in many cases: If the protocol is TCP, the presence of SYN or ACK flags can help distinguish request packets from response packets. Similarly, for UDP port 53 (DNS), the response packets contains answers to DNS queries from DNS servers and can be distinguished from request packets. Notice that for low-interaction darknets, the *only* response packets in our darknet data would be those where the original request contained a spoofed IP address belonging to a darknet IP address. These are often referred to as *backscatter* packets. We will quantify backscatter packets on a daily, weekly, and monthly basis, along with the number of servers targeted around the world and the number of potential miscreants seen in request packets. This data product requires only a small amount of annotation overhead to the 39GB per day of darknet data we receive from all our data sources combined. So, we will annotate all raw data with this information. *Overall, this product would require over 13TB of persistent storage for a period of one year and 135TB to store it for 10 years.*

Tables of per-port darknet activity: For each active port, we will maintain tables that will provide 15-minute snapshots of darknet activity across "request", "response", and "unknown" categories. A researcher or operator could aggregate these tables to view darknet activity across all ports. We will maintain this table for one year to allow analyses over longer time frames. *Keeping per (port, protocol) pair traffic levels measured in both bytes and packets at a 15-minute granularity would generate about 1TB of data per year.*

¹<http://www.ren-isac.net/cgi-bin/monitoring/Internet2TGa.cgi>

C. Darknet and NetFlow-based Products

The ability to correlate data derived from darknet sensors with other network data, such as NetFlow offers significant opportunities for new, derived, data sets of value to cybersecurity research. For example, the NetFlow data contains information regarding network transit and topology information, such as router ingress and egress ports and information about ASNs. We based our next data product on annotating the darknet data with NetFlow records.

Annotating darknet data with NetFlow information: As we discussed in Section III-B, the source IP addresses on the darknet “request” packets can be spoofed. In order to add more credibility to IPs contained in darknet packets, we will annotate the source and destination IPs with information from NetFlow packets. By matching NetFlow records with darknet packets we would add router ingress, ASN, and TCP flags (when applicable). Notice that offering this product requires access to machines with a large amount of memory due to the privacy constraints on NetFlow data which disallow copying it on persistent storage, and the fact that the darknet data is not real time. These two constraints imply that the processing machine has to be able to hold NetFlow data for several hours to ensure that matching with delayed darknet data is possible. *With NetFlow sampling, we anticipate that this annotation will not add much beyond the 13TB persistent storage we estimated per year in Section III-B.*

D. Passive DNS-based Products

The primary goal behind our third set of data products is also to eliminate often-used pre-processing steps on passive DNS data. In addition, we will offer statistics describing numbers of hosts, domains, and IP addresses in the Internet. Many such estimates today are less precise than the research community would like. *Storing the raw passive DNS data will require 240TB of storage over a period of 10 years.* All derived data products would be stored for a month on a rolling basis.

Unique host and domain names in the Internet: The passive DNS data contains responses from DNS servers. It provides an opportunity to tabulate unique hosts on the Internet and their corresponding DNS records. Using this information, we will release unique server and domain names present in the Internet on a daily basis, maintaining this information online for one month while providing near-line archives for up to 10 years. *We expect this product to cost 2TB of storage per month.*

Unique IP addresses in the Internet: Similar to host names, the passive DNS data contains IP addresses of servers in the Internet. We will provide daily tables of unique IP addresses for a month and the facility to compute them over 10 years. Additionally, we will use routable BGP prefixes from the RouteViews project [16] to associate IP addresses with BGP prefixes. This information will enable researchers to assess the fraction of hosts within each routable IP prefix. We will provide routines to assist with such analysis. *We anticipate*

that one month of this product will amount to about 2TB of persistent storage.

Tables of per-host, per-domain, and per-DNS server activity: For host names, domain names, and DNS servers seen over the course of a day, we will maintain a table each that summarizes the first and last time they appeared in a response packet and the frequency of appearance. In cases when name (both for host names and DNS servers) to IP address mappings change, we will provide annotations to reflect the change. These tables will be useful in DNS-related investigations of cybercrime infrastructure. *We anticipate that storing a month of these tables will require 2TB of persistent storage.*

E. Nightly Zone Files for all TLDs

Domains on the Internet may be divided into two categories: generic top-level domains (gTLDs), such as .com and .net, and country-code TLDs (ccTLDs), such as .de (Germany) and .hk (Hong Kong). Researchers often use *zone files*, which enumerate the domains contained in a TLD, while investigating DNS aspects of cyberfraud infrastructures. Unfortunately, access to zone files today is far from perfect. While many of the prominent gTLDs, including .com and .net, can be obtained on a nightly basis through contracts with registrars like Verisign, the availability of ccTLD zone files is practically non-existent. This is a severe limitation because roughly half of the 183 million domains in the Internet today [17] belong to ccTLDs, and some ccTLDs, including .cn, .de, .uk, and .eu, are among the biggest contributors to the current growth in domain registrations. Lack of access to ccTLD zone files restricts researchers’ view of roughly half of the domains in the Internet. This limitation has been a recurrent theme across all the workshops that have discussed aspects of availability of security data to the research community.

Using our three passive DNS data sources and all the darknet data sources, we will construct all zone files, including ccTLD zone files, on a nightly basis and make them available to the community. Our passive DNS data sources see data pertaining to globally distributed clients. This allows us to enumerate virtually all domains of interest in each TLD in the world. A typical zone file contains the names and IP addresses of DNS servers corresponding to each domain. This information is already contained in DNS responses, requiring no additional queries on our part. A limitation of this methodology is the complete opaqueness of the domains that are not accessed by any client, which can be useful information in certain research scenarios. When possible, we will supplement the derived zone files with DNS responses reaching our darknets. We will annotate the zone files so the researchers would be aware of the source of data for each domain listed.

Daily zone files for all TLDs could occupy up to 100GB, based on the fact that there are 183 million domains in the Internet today [17]. However, our investigations have shown that constructing the zone file on a weekly basis, and calculating daily incremental files should require approximately 160GB/week, for a total consumption of 10 TB/year. *Since we*

would like to maintain zone files for a 10-year period based on community feedback, this data product will be our most expensive derived product at 100TB.

IV. ARCHITECTURE

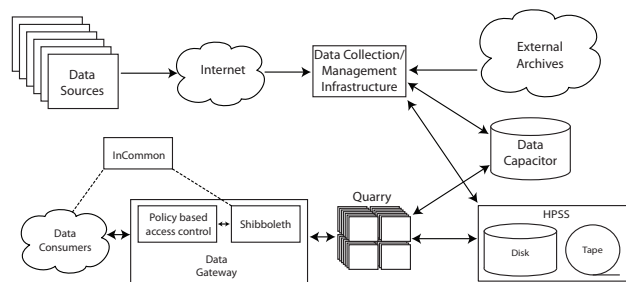


Fig. 2. Architecture overview for Bloom.

Figure 2 shows a diagram of our architecture. The processing component of Bloom will be carried by Indiana University’s IBM e1350 distributed shared-memory supercomputer cluster, called *Quarry*. Quarry currently consists of 1120 processor cores sharing a total of 1.2TB of RAM at 8GB per processor. It is directly coupled via a 14.5GB/s interconnect to Indiana University’s storage subsystem called *Data Capacitor* which provides a total of over 1PB of in-line mass storage. The Data Capacitor will be used to spool incoming data, to provide temporary processing workspace, and to store result data products for delivery to end users. Long-term and archival storage will be provided by Indiana University’s *Massive Data Storage Service (MDSS)*. The MDSS is a geographically redundant, with mirrored installations in both Bloomington and Indianapolis, storage system currently configured with over 2.8PB of magnetic tape and front-end caching disk storage. The MDSS appears to the Quarry system as a mounted file system similar to the Data Capacitor but with far greater capacity and larger latency times. Bloom will use the MDSS as both an operational backup to the Data Capacitor as well as the long-term archival facility for project data sets. With an exception for access time, end researchers will not need to differentiate between an access for data stored on the Data Capacitor versus that on the MDSS. The access API will make the difference transparent to the end user. The data we serve may be located in multiple geographical locations. A *data management system* will be in charge of channeling it to the Data Capacitor or HPSS.

A *data gateway* will control access to data. It will provide a front end to Bloom, accept data requests, authenticate users, apply policy-based access rules, and most importantly construct code to submit to Quarry. All machines and storage units are connected with dual, bonded 10Gb/s Ethernet links into a dedicated 10Gb/s Ethernet switch, providing an aggregated throughput of 20Gb/s between the individual units and the Data Capacitor, MDSS, and Quarry. This will ensure that Bloom’s performance is not limited by network throughput.

Access to Bloom will be through an instance of the Shibboleth system, *InCommon* [2]. InCommon is a “single sign-on” authentication and authorization system developed by the Internet2 community. It includes over 150 educational institutions, government agencies (including the NSF), and corporate research labs, representing a total of over three million users. The authentication aspect of InCommon will allow us to establish the identity of users in a secure manner without having to maintain a pool of locally administered accounts: users can authenticate based on their existing credentials at their home institution. The authorization aspect will allow us to provide different views of data sources to different users depending on the nature of their relationship to their research.

V. CONCLUSION

In order to execute Bloom, we need financial resources. The biggest resource required is personnel time to provision and maintain Bloom, offer data products, and collect feedback to improve the service. Further, even though Indiana University has committed much of the computing and storage infrastructure needed to realize it, access to those resources will not be exclusive for Bloom users. Hence, some addition to those resources is necessary to ensure that Bloom users can have dedicated sources and hence good quality of service.

REFERENCES

- [1] Advanced Network Management Lab (ANML). <http://www.anml.iu.edu/>.
- [2] InCommon. <http://www.incommonfederation.org/>.
- [3] Protected REpository for the Defense of Infrastructure Against Cyber Threats (PREDICT). <https://www.predict.org/>.
- [4] Research and Education Networking Information Sharing and Analysis Center (REN-ISAC). <http://www.ren-isac.net/>.
- [5] Torpig botnet hijacking reveals 70gb of stolen data, May 2009. <http://www.darknet.org.uk/2009/05/torgpig-botnet-hijacking-reveals-70gb-of-stolen-data/>.
- [6] BFK. DNS History Database (DHDB) Project. http://www.bfk.de/bfk_dnslogger.html.
- [7] B. Claise. Cisco systems NetFlow services export version 9. IETF RFC 3954, October 2004.
- [8] CAIDA: The Cooperative Association for Internet Data Analysis. Non-CAIDA Publications using CAIDA Data. <http://www.caida.org/data/publications/bydate/index.xml>.
- [9] CAIDA: The Cooperative Association for Internet Data Analysis. The IPv4 Routed /24 Topology Dataset and IPv4 DNS Names Dataset. <http://www.caida.org/data/>.
- [10] Kelly Jackson Higgins. Widespread Conficker/Downadup worm hard to kill, January 2009. <http://www.darkreading.com/security/attacks/showArticle.jhtml?articleID=212901489>.
- [11] Internet Systems Consortium (ISC). SIE@ISC: Security Information Exchange. <https://sie.isc.org/>.
- [12] Internet2. Research projects using the internet2 observatory. <http://www.internet2.edu/observatory/archive/research-projects.html>.
- [13] F. Jahanian J. Nazario D. Watson M.D. Bailey, E. Cooke. The Internet Motion Sensor: A distributed blackhole monitoring system. In *Internet Society Network and Distributed System Security Symposium (NDSS)*, 2005.
- [14] F. Jahanian S. Sinha, M.D. Bailey. The Internet Motion Sensor: A distributed blackhole monitoring system. In *IEEE International Conference on Technologies for Homeland Security (HST)*, 2009.
- [15] Team Cymru. IP to ASN lookup v1.0. <http://asn.cymru.com/>.
- [16] University of Oregon Advanced Network Technology Center. Route Views project. <http://www.routeviews.org/>.
- [17] VeriSign. Domain name industry brief, June 2009. http://www.verisign.com/static/DNIB_09_0529web.pdf.