(EAGER): Creating a Cybersecurity Big Data and Analytics Sharing Platform

DGE-SaTC-1719477

PI: Dr. Hsinchun Chen: Regents' Professor; Thomas R. Brown Chair of Management and Technology; and Director, Artificial Intelligence Lab, University of Arizona.

Cybersecurity Big Data and Analytics Sharing Workshop
Women in Cybersecurity Conference,
Tucson, AZ
March 31, 2017

I. Introduction

Cybersecurity has become a significant issue that presents continual challenges to individuals, industry, and government. Incidents of large-scale hackings and data theft are now of regular occurrence, with many cyberattacks resulting in theft of sensitive personal information or intellectual property. It is estimated that cybercrime costs the global economy about $445 billion annually (Sandler & Char, 2014). To help deal with complex cybersecurity challenges, the international Intelligence and Security Informatics (ISI) community has published high-impact, Big Data driven cybersecurity research since 2003. The ISI Community consists of more than 1,500 scholars, with about 70% in Computer and Information Sciences and Engineering (CISE). For the past 13 years, ISI researchers have made significant advances analyzing terabytes of data from high-impact cybersecurity areas including hacker community, and phishing research. Despite the many novel advances in data-driven cybersecurity research, we are unaware of any current data sharing platform that aggregates and provides data and tools used and developed in cybersecurity research for the larger cybersecurity community. Such a platform would help enable researchers to publish high-impact, cutting-edge, and reproducible research. This project aims to develop a Cybersecurity Big Data and Analytics Sharing Platform designed to enable and encourage cybersecurity researchers to share their data, tools, and analytical approaches.

Prior research on cyberinfrastructure projects indicate that cyberinfrastructure must be designed to meet user needs in order to be successfully adopted by the user community and sustained after an initial grant funded development period. With this in mind, the original proposal included two workshops at relevant conferences to solicit community input. This report is on the first of these workshops.

II. Workshop Overview

The first workshop for soliciting community input was held during the Women in Cybersecurity (WiCys) Conference on March 31, 2017 in Tucson, Arizona.

Workshop title: Cybersecurity Big Data and Analytics Sharing
Target Audience: Cybersecurity professionals, faculty, and students.
Workshop Organizers: Hsinchun Chen, University of Arizona; Resha Shenandoah, University of Arizona; Victor Benjamin, Arizona State University; Bhavani Thuraisingham, UT Dallas; Latifur Kahn, UT Dallas.
Time: 3:30 – 5:30 PM
Abstract: Cybersecurity has become a significant issue that presents new challenges to individuals, industry, and government. Despite its importance and the many novel advances in data-driven cybersecurity research, there as yet exists no platform to support sharing of the data, tools, and analytics

needed for cybersecurity research. This workshop is being held to gain community input into the development of a platform to facilitate the sharing of critical cybersecurity-related data and tools for both research and education. The workshop agenda includes presentations by researchers at the cutting edge of cybersecurity R&D. In addition, audience participation, particularly from cybersecurity researchers, educators, and students, are welcomed as input to the development of a new cybersecurity analytics platform.

Workshop attendance: 110+ Workshop organizers were informed by other conference attendees that this workshop was the best attended for its time slot.

Workshop schedule: (Each speaker given 10-12 minutes to speak)

Part I:

Dr. Hsinchun Chen (University of Arizona) – opening remarks and brief overview of career

Dr. Bhavani Thuraisingham (UT Dallas) – brief overview of career

Ramkumar Paranthaman (UT Dallas) – Malware Data Collection and Analysis Using Big Data Tools

Dr. Latifur Khan (UT Dallas) - Trends and Perspectives in Big Data Research and Application

Dr. Victor Benjamin (ASU) – Blockchains for Cybersecurity Research

Q&A

Part II:

Dr. Hsinchun Chen (University of Arizona) – transitional remarks

Resha Shenandoah (University of Arizona) – DIBBs-ISI Data Repository and Research Data Management

Sagar Samtani (University of Arizona) – DIBBs-ISI Tool Inventory for ISI Research

Sagar Samtani (University of Arizona) – AZSecure Hacker Assets Portal

Weifeng Li (University of Arizona) – AZSecure Hacker Underground Economy Collection and Analytics

Q&A

III. Workshop Presentations: Part I

Dr. Chen began by addressing the relatively recent development of the domain of Cybersecurity research including his own research interests that lead to participation in that domain's development and the need for a platform by which researchers can share their data, tools, and analytic approaches. Dr. Chen asked the audience to listen to the following presentations on current research in cybersecurity using different tools and analytic approaches with a focus on sharing data and tools. The following questions were presented as general guides for approaching the presentations and discussion:

- What data or tools do you consider to be most useful for you and why?
- What additional data or tools do you wish to have and why?

Dr. Chen then introduced Drs. Thuraisingham, Khan, and Benjamin.

Dr. Thuraisingham spoke briefly about her career and how she came to conduct research in cybersecurity and then introduced Ph.D. student Ramkumar Paranthaman who presented on “Malware Data Collection and Analysis Using Big Data Tools.” A team of researchers at UT Dallas is looking for datasets containing malware to make available to researchers. The datasets they identify are made available on the Intelligence and Security Informatics (ISI) domain specific repository http://www.azsecure-data.org/, a research data repository prototype currently being developed with funding from NSF and discussed in Part II by Ms. Shenandoah. For their own research, the collected malware is being used to develop a malware detection framework using a static analysis approach by employing Big Data tools and machine learning techniques. The source code for the malware detection framework they have developed is available on github.com for other researchers to access and use.

Dr. Khan spoke briefly about his career and how he came to conduct research in cybersecurity then presented on “Trends and Perspectives in Big Data Research and Application” in which he laid out some issues and solutions he sees with Big Data in cybersecurity research. Real time data processing can be addressed by the tools Apache Spark, Storm, S4, and Fink. Real time analytics can be addressed by SAMOA – Scalable Advanced Massive Online Analysis. Scalable analytics can be addressed by Spark’s Machine Learning Library and Mahout, but these only cover basic analytics algorithms. Within this tool set advanced algorithms that support relational learning are missing. The future of Big Data research is in stream mining where the learner (algorithm) is updated continuously, an issue with this currently is determining when to update the model because delayed updates may cause valuable insights to be missed, the proposed solution to this is an adaptive system that can vary the time between updates. Future research involving analytics will be based on supervised learning, but this requires labeled data.

Dr. Benjamin spoke briefly about his career and how he came to conduct research in cybersecurity then presented on “Blockchains for Cybersecurity Research,” a novel idea for secure sharing of supply chain or cybersecurity research data based on the concept of blockchains, the technological foundation for bitcoin trading. In general, industry supports sharing cybersecurity data between businesses or between businesses and government. However, there is a reluctance to share data based on perceptions of liability, accessibility, transparency, and data ownership. There is also an absence of a common platform through which such data could be shared. A common platform could encourage community building while catering to the needs of special interest groups. Blockchains are crypto-secured, used on peer-to-peer networks, managed autonomously, and are highly configurable. A computing infrastructure based on blockchains would be decentralized, resilient, immutable, offer security and privacy, and thus address many of the current reluctances in industry to sharing their cybersecurity related data with peers or the government. In this respect, blockchains posess many qualities for a cybersecurity data sharing platform.

IV. Q&A: Part I

Audience question: For people who have limited knowledge of machine learning, how do you train the data?

Panel answer: Consider the features, and extract the features that best represent the data. Feature engineering (what featurs are we considering), filter good features from noisy, choosing the right features is extremely important for results. Machine learning is becoming a huge area, and uses similar underlying techniques but feature selection is very important.

Suggestion for platform: Tutorial on how to use machine learning.

Audience question: What are the limitations of this technique for protecting against malware?

Panel answer: You may detect some of the malware, but you may classify a benign object as malware, creating a false positive. You can miss malware, but that's where updating your learning algorithm comes in. Perform dynamic analysis after conducting static analysis.

***Suggestion for platform**:* Understanding the limitations of techniques opens new avenues for research. Community challenges to identify and address current limitations could be used.

Audience question: Do you manually create labels or is it dynamic?

Panel answer: Depends on knowledge of the data and what type of analytics you're performing.

Suggestion for platform: Clearly indicate collections that are labeled already.

Audience question: How can we build a model with closed-source code?

Panel answer: There are tools capable of reverse-engineering binary code into source code. With these tools, we can conduct dynamic analysis on malwares to study their behaviors.

Suggestion for platform: See Part II on data and tools sharing.

V. Workshop Presentations: Part II

Dr. Chen briefly introduced the presenters for Part II.

Resha Shenandoah, a master's student at the University of Arizona, spoke briefly about how she came to work on digital archives and cybersecurity then presented on digital archiving using the example of the Digital Infrastructure Building Blocks for Intelligence and Security Informatics (DIBBs-ISI) research data repository under development. The prototype repository is available at http://www.azsecure-data.org/ and currently contains 14 collections, over 200 GB of data, for a variety of ISI related topics, in nine different languages, and containing a wide variety of file types. This heterogeneous set of open access collections is available to and used by researchers worldwide. The first prototype of the repository was quickly developed to establish user need. Since need and use have been established (over 1,404 GB of data have been downloaded in the past eight months) the project is now working on plans to migrate the data into the archival content management system DSpace. DSpace will allow for greater discoverability, persistent identifiers for citation and access purposes, and more detailed use analytics. Ms. Shenandoah then addressed the importance of data preservation, Data Management Plans, and the absence of standardized data management skills among cybersecurity researchers. A platform could be a place to discuss domain standards for data management skills and disseminate tools and tutorials to support best practices in cybersecurity for data management as well as assist researchers and data curators by automating some metadata creation. Any platform must meet user needs.

Sagar Samtani, a Ph.D. student at the University of Arizona, spoke briefly about how he came to conduct research in cybersecurity then presented on "DIBBs-ISI Tool Inventory for ISI Research," a project the University of Arizona team is conducting to identify tools used in ISI research and make them discoverable in the http://www.azsecure-data.org/ repository. They identified three categories of tools: collection and storage tools, pre-processing and analytics tools, visualization tools. They present these tools along with IEEE-ISI, FOSINT-SI, and ISI-ICDM papers to show how the tools are used in research. This information is now available (http://www.azsecure-data.org/tools---tutorials.html) and the UA team has plans to seek permission as appropriate to host and index the tools in the planned DSpace based repository.

Mr. Samtani then presented on "AZSecure Hacker Assets Portal," a browser based tool that allows researchers to search the contents of hundreds of international hacker forums and gain insights into emerging threats. On the collected forums, hackers share assets such as malicious tutorials, code, and exploits. The portal provides some built in visualization tools and allows users to download materials for further research. Identified forums are automatically collected and the portal is updated monthly. Students and researchers can use the assets on the portal to understand how tools are created, implemented, and operated. Researchers who wish to access the portal can request access from Mr. Samtani and must provide their name, organization, position, and intended use to gain portal access.

Weifeng Li, a Ph.D. student at the University of Arizona, spoke briefly about how he came to conduct research in cybersecurity then presented "AZSecure Hacker Underground Economy Collection and Analytics," a research project collecting information about underground economies that includes malware (encrypter/ransomware, Trojan, exploit), zero-day vulnerabilities, POS/ATM skimmer, stolen credit/ debit card, fake documents (driver's license, SSN), and prices for these malicious services and wares. Examples of research that can be conducted with this data include identification of key actors and comparison of ware availability to data breach events.

VI. Q&A: Part II

Audience question: Do you think you're a target for the hackers, and have there been any attempts to attack the AZSecure Hacker Assets Portal

Panel answer: We have seen attacks against the AI Lab, but as we are still in the early stages of releasing the Hacker Assets Portal we have not seen anything directed at it, though we are concerned with people using the hacking tools found on the website to turn around and attack the website.

Audience question: Is the data available to everyone?

Panel answer: For the Hacker Assets Portal, yes it is, once you've gone through the vetting process. The ISI repository is currently Open Access but the migration to DSpace could allow for controlled access to sensitive collections if deemed necessary by the cybersecurity community.

Audience question: Are you worried about collecting all of this hacker information and tools and putting it all together?

Panel answer: The vetting process will hopefully stop a lot of the malicious users. Some existing cyberinfrastructures use login systems such as Shibboleth where users belong to existing institutions and use their institutional login credentials to access a linked system.

Suggestion for platform: Vetting of access to sensitive data is clearly a concern. A portal containing sensitive data needs robust vetting and access procedures.

Audience question: How can we in the industry use your portals and the results to defend our infrastructure?

Panel answer: Once you understand your infrastructure and what assets you are trying to defend, you can search for the types of tools that can attack you. This will help you defend your systems.

Audience question: How is the data stored on the new chip enabled credit cards?

Panel answer: It depends, some chip cards hold the information statically (which has already been hacked) and some cards generate the information dynamically.

Suggestion for platform: A platform needs not only granular metadata to allow users to discover the materials specific to their needs, but automated notification systems where users can subscribe to topics of interest and receive timely updates when new information on those topics are available could also be of utility. Users are interested in how cybersecurity effects their daily lives as well as research and industry. A broader service available via a platform could allow other users who are not vetted for data access but allowed to subscribe to timely updates on new technological developments and cybersecurity threats of interest to the general public.

VII. Summary of suggestions raised by presenters and audience

The invited presenters spoke about their current research, how they are making data, tools, and analytical approaches available to the cybersecurity research community. Presenters also directed attention to areas that need better tools such as advanced algorithms that support relational learning for real time analytic applications and labeled data to support supervised learning.

Some presenters made suggestions for infrastructure to support a research sharing platform. Dr. Benjamin suggested blockchains technology could form the foundation for secure data sharing and Ms. Shenandoah implied that archival repository infrastructure such as DSpace can make data and tools more discoverable. A platform could also be a location to discuss domain standards for data management skills and disseminate tools and tutorials to support best practices in cybersecurity for data management as well as assist researchers and data curators by automating some metadata creation.

The UA team, presented by Mr. Samtani, identified three categories of tools: collection and storage tools, pre-processing and analytics tools, visualization tools. This list of tools is available in the DIBBs-ISI repository located at (http://www.azsecure-data.org/tools---tutorials.html). More such tools, their associated tutorials and links to the research papers that utilize them could be identified and added to an index for easy discoverability. With permission from appropriate intellectual property owners, the repository could also preserve these tools and tutorials. The source code for the malware detection framework developed by the UT Dallas team is available on github.com for other researchers to access and use. This is an example of dispersed shared resources that would be more discoverable if indexed in a central location.

The audience was clearly concerned about the process of vetting access to sensitive data, such as malware collections and the malicious code and tutorials contained in the Hacker Assets Portal. A platform containing sensitive data needs robust vetting and access procedures.

Many audience questions focused on research methods, suggesting the need for practical tutorials to support student learning. A tutorial on how to use machine learning was specifically requested. Audience questions also highlighted a desire to understand the limitations of techniques. This understanding opens new avenues for research. Community challenges coordinated by a community platform could be used to identify and address current limitations.

A platform needs not only granular metadata to allow users to discover the materials specific to their needs, but automated notification systems where users can subscribe to topics of interest and receive timely updates when new information on those topics are available could also be of utility. Users are interested in how cybersecurity effects their daily lives as well as research and industry. A broader service available via a platform could allow other users who are not vetted for data access but allowed to subscribe to timely updates on new technological developments and cybersecurity threats of interest to the general public.

VIII. Next Step

We will assemble a team to evaluate and implement these community suggestions for a Cybersecurity Big Data and Analytics Sharing Platform. Our plan is to have the platform ready for community evaluation and discussion prior to the IEEE-ISI conference in 2018. We will host a second workshop at IEEE-ISI 2018 to receive community feedback on the platform.