# NSF Workshop Report: Multi-Campus Cybersecurity Data Curation for Research and Education

Jack W. Davidson, Professor of Computer Science, University of Virginia (jwd@virginia.edu)
H. Howie Wang, Professor of Computer Science, George Washington University (howie@gwu.edu)
Von S. Welch, Associate Vice President for Information Security, Indiana University (vwelch@iu.edu)

## INTRODUCTION

As part of a National Science Foundation Planning Grant entitled, "Towards Building a Community Data Infrastructure for Cybersecurity Research" (NSF Proposal #2016431), a workshop was held on July 27-29, 2021. The workshop was organized by the principal investigators, Jack Davidson (University of Virginia), Howie Wang (George Washington University, and Von Welch (Indiana University). Because of health and safety considerations, the workshop was held virtually. This report describes the workshop and the outcomes.

## WORKSHOP GOALS

The major goal of the workshop was to engage the community to formulate a vision and roadmap for the creation of a multi-campus data collection and sharing infrastructure for use by machine-learning cybersecurity and privacy researchers. Such a federated infrastructure will be invaluable for detecting zero-day (new, previously unseen) attacks and large-scale attacks with complex kill-chains, e.g., the Wannacry ransomware attack, Mirai Distributed Denial of Service (DDoS) attacks, and Advanced Persistent Threat (APT) attacks. Discussion will encompass legal, ethical, privacy, organizational and sustainability considerations**.** Another workshop goal was to identify other potential data providers and data users that would support the submission of a Grand Ensemble CCCRI proposal to build and operate the infrastructure for the benefit of the research community.

### WORKSHOP ORGANIZERS

The workshop was organized by Jack Davidson, Howie Hwang, and Von Welch with assistance from University of Virginia staff.

## INVITED SPEAKERS AND PANELISTS

K. C. Claffy, Computer Science and Engineer, University of California, San Diego
Frederick Cate, Vice President for Research, Indiana University
William Hewlett, Director, AI Research, Palo Alto Networks
Ronald Hutchins, Vice Provost of Information Technology, University of Virginia
Anita Nikolich, Director of Research and Innovation Technology and Research Scientist, University of Illinois, Urbana-Champaign
Tejas Patel, Defense Advanced Research Projects Agency
Melur K. "Ram" Ramasubramanian, Vice President for Research, University of Virginia

## WORKSHOP PARTICIPANTS

The organizers issue over 100 invitations via our networks of colleagues and collaborators. We also advertised the workshop via various mailing lists including Commonwealth of Virginia Cybersecurity Initiative (CCI), the Cyber Innovation & Society, and XXXX. We issued two types of invitations: full workshop participation and keynotes and panels only invitations. Thirty-nine people registered for the full workshop and 38 registered for the keynotes and panels only portion of the workshop.

We did not ask for demographic information, but informally, the workshop did include a diverse set of participants. Most participants were from academic institutions with a handful from industry, research laboratories and government. The full participant list is included as an Appendix.

## WORKSHOP SCHEDULE

Because the workshop would be virtual, we split the workshop over three days to avoid "Zoom fatigue." This structured worked well and the virtual aspect allowed participants

### Tuesday 7/27

| **Time Slot** | Agenda |
|---|---|
| 11:00 – 12:00 PM | Welcome and Introduction: Organizers. Opening Remarks: Marilyn McClure, National Science Foundation, CNS Program Director |
| 12:00 – 13:00 PM | Setting the Context: Von Welch, Jack Davidson, Howie Hwang. Slides are located here. |
| 13:00 – 13:30 PM | Break/Lunch |
| 13:30 – 14:30 PM | Reflections on WOMBIR: Workshop on Overcoming Measurement Barriers to Internet Research: K. C. Claffy |
| 14:30 – 14:40 PM | Break |
| 14:40 – 15:40 PM | Machine Learning and Data Privacy in Security, an Industry Perspective: William Hewlett |
| 15:40 – 16:00 PM | Wrap-up |

### Wednesday 7/28

| **Time Slot** | Agenda |
|---|---|
| 11:00 – 11:30 AM | Welcome, logistics, introduction and summary of Day 1 |
| 11:30 – 12:30 PM | Panel: Explore the benefits to using multi-campus IT data for cybersecurity research and what the barriers are to allowing that research: Fred Cate, Ronald Hutchins, Anita Nikolich, Tejas Patel, Melur K "Ram" Ramasubramanian |
| 12:30 – 13:00 PM | Break/Lunch |
| 13:00 – 14:30 PM | Concurrent Breakout Sessions |

| | |
|---|---|
| 14:30 – 14:45 PM | Break |
| 14:45 – 16:00 PM | Breakout reporting and wrap-up |

### Thursdays 7/29

| **Time Slot** | Agenda |
|---|---|
| 11:00 – 11:30 AM | Welcome, logistics, introduction and summary of Day 2 |
| 11:30 – 13:00 PM | Concurrent Breakout Sessions |
| 13:00 – 13:30 PM | Break/Lunch |
| 13:30 – 15:00 PM | Breakout reporting and wrap-up |
| 15:00 – 15:10 PM | Break |
| 15:10 – 16:00 PM | Summary of key issues, recommendations capture, next steps, and final report |

## OUTCOMES

There were many lively discussions throughout the workshop. The following sections attempt some organization of the comments. There was discussion about “What Data to Collect,” “Access and Privacy,” “Reproducibility,” and “Infrastructure.”

### *What Data to Collect*

Generally, everyone was agreement that the more data available the better. Keynote speaker William Hewlett from Palo Alto Networks commented that “Having more data makes your classifier more accurate.” Another anecdotal comment was that when researchers were asked, “What data do you want?”, the response was often “What data do you have?”

Additional themes that were discussed were standardization of formats, data with ground truth, and historical data versus real-time data.

Tejas Patel, an invited panelist from DARPA, mentioned that University data is attractive as DoD network data is heavily encumbered. However, other participants raised the question as to whether University data sets are representative. William Hewlett discussed the issues with using data collected by companies such as Palo Alto Networks, CISCO and other security firms. Their customers are very protective of their data both from a privacy standpoint, but also from a standpoint that a customer does not want any information about their network operations to be public because of the risk it may prove useful to adversaries.

It was viewed as critical that datasets need to be curated. In particular, there needs to be ground truth regarding what is malicious.

There was discussion of whether access to historical data was enough, or if real-time data is needed. The consensus was that historical data was good enough for now. The cost to collect operational data in real-time is too high and not worth the cost at this time.

### *Access and Privacy*

Concerns were raised regarding privacy of data as it relates to potential identification of LGBTQIA group members. There needs to be careful consideration of what are the potential harms to members of threatened groups. A participant mentioned the Asilomar Conference on Recombinant DNA research and wondered if there needs to be a similar gathering of researchers, ethicists, and privacy experts.

Another concern was fair access to data and avoiding “haves” and “have nots”. Tier 1 institutions with greater level of resources at their disposal may have preferential access which could then be bootstrapped into accumulating more resources, thereby entrenching the privileged positions of Tier 1 institutions. Federation of universities was mentioned as a potential solution to mitigate this concern.

Panelists mentioned that data was already easy to buy, e.g., 10000 attributes on a single person, but others observed the lack of sufficiently rich data sets for cyber security research.

Haphazard IRB processes (both inter and intra universities) inhibit research and collaboration. Further data sharing policies between institutions, or between institutions and commercial entities, are typically negotiated bi-laterally. Re-negotiating term of use data is challenging and time-consuming. Thus, legal aspects may be more of a barrier to collaboration than purchasing and/or managing hardware resources to host the data. One panelist noted that sometimes competition within a university means that it is easier to share and collaborate between two different universities!

What guidelines are there for who (legally) is allowed to access data w/o violating laws? What are the applicable state and federal laws? An issue that was raised is that counsel at various institutions will have different opinions as to what should be allowed. Furthermore, data crossing university networks may not be owned by the University themselves. Leadership is needed from government otherwise we will end up with data silos. Ron noted that data silos exist even within a university!

Panelist Ron Hutchins mentioned ACCORD. In particular, negotiations for data sharing, risk management, legal compliance is done once at the state level, thus all Virginia universities benefit and have a pre-negotiated framework for accessing and sharing data.

What does it mean for data to be “open”? Ron suggests that universities should do better at characterizing data sets, as “openness” covers a wide spectrum of policies. What does it mean to have “access” to data? Wide open model is viewed as unrealistic. But what are the alternatives? Queries on internal data? Release of anonymized data sets?

Panelist Anita Nikolich raised the problem that academic studies often have unrealistic models and thus the potential for harm is understated. What is the applied reality of what can happen when collecting/analyzing data sets? The Atlas Internet project was cited as example where the benefits (understanding resilience of the Internet) might have come at the cost of a large-scale attack map.

One protection that was discussed to address privacy concerns was enforcing a code-to-data model, where data is maintained in an access-controlled environment with layered security. Research would only be conducted within that environment by known individuals who have consented to abide by security policy. This could be further augmented through the anonymization of PII where feasible as an additional layer of security. The PCORE project at the University of Virginia is an example of this approach.

### *Reproducibility*

Participants noted that reproducibility of experiments is an essential component of the scientific method. To facilitate reproducibility, some portion of stored data would need to be maintained and made available indefinitely.

One participant mentioned that any project for curating data sets should be aware of the FAIR (Findable, Accessible, *I*nteroperable, and *R*eusable) principals for scientific data management and stewardship.

### *Infrastructure Requirements*

The participants agreed that a substantial cyber infrastructure would be warranted making the data useable and accessible. It was noted the data volumes are large, therefore substantial data capacity is needed. To support the code to data model, there would need to be support for developing and running machine-learning algorithms. There would need to be some user support for accounts and infrastructure support. In addition, high-speed network access for both users, but also to ingest new data sets would be needed (pushing data).

### *Sustainability*

There were several issues/question raised regarding sustainability. How do we incentivize data set generation/curation/maintenance? How does one reconcile the relative short-time horizon of funding (3-4 years) with sustaining longevity of long-term infrastructure? Tejas Patel suggested that perhaps the role of government is to bootstrap such an infrastructure and let market forces take over in establishing long-term viability. What is the role of the US Govt? Government funding is on the order of $15B whereas Venture Capital is one order of magnitude higher. NSFNet was mentioned as an example of a project that was handed over to the commercial sector. While price went down due to the efficiency of the marketplace, “all data disappeared” from a research perspective.

Should we consider different models for procuring resources? Instead of bringing everything in-house at various universities, might it better to use cloud providers, and/or use a co-location model? Would government agencies be receptive to such approaches?

Panelists observed that funders are OK with capital costs. However, the challenge is in funding support for personnel and sustaining that funding.

Participants noted that research over datasets may yield valuable information that could be operationalized by participating institutions once a certain level of maturity is reached. This could incentivize some subsidization by participating data providers to augment their own security portfolios. The unique multi-institutional (or global) vantage of contained data could also be used to market curated threat intelligence to non-producers as part of a subscription model. The Stingar project was mentioned as an example of this.

## SUMMARY

The participants were largely unanimous regarding the need for collecting and generating significant cyber security data sets that would be easily accessible. Some noted that the various datasets for computer vision significantly accelerated the pace of innovation in that space. Unlike image datasets where there are standardized formats, such standardization is not the norm for cybersecurity datasets. One issue is that there many different types of cybersecurity data—network captures are only one.

These datasets need to be as diverse as possible.  Good synthetic datasets would be useful, but realistic data is the real goal.  There is a real need for large, realistic datasets that include a diversity of attacks, and where ground-truth is known.

It was suggested that a valuable project would be to create a list of existing resources. The current state is that while there are some well-known data sets, there is no comprehensive catalog of available cybersecurity data sets and the characteristics.

# APPENDICES

## *Participant List*

| Name | Affiliation | Email Address |
|---|---|---|
| Salman Ahmed | Virginia Tech | ahmedms@vt.edu |
| Thomas Ambrosi | Washington State University | tambrosi@wsu.edu |
| Damon Armour | North Carolina State University | damon_armour@ncsu.edu |
| Ilya Baldin | Renaissance Computing Institute/University of North Carolina Chapel Hill | ibaldin@renci.org |
| Tom Barton | Internet2 | tbarton@internet2.edu |
| Gregory Bell | Corelight | greg@corelight.com |
| Kathy Benninger | Carnegie Mellon University/Pittsburgh Supercomputing Center | benninge@psc.edu |
| Omkar Bhat | University of Virginia | odb6pz@virginia.edu |
| Richard Biever | Duke University | richard.biever@duke.edu |
| John Board | Duke University | john.board@duke.edu |
| Judi Bowers | University of Virginia | Jrb5z@virginia.edu |
| Molly Buchanan | University of Virginia | mkb4vb@virginia.edu |
| William Burke | George Washington University | wburke@gwu.edu |
| Miles Chung | University of Toronto | mhmchung@gmail.com |
| Tijay Chung | Virginia Tech | tijay@vt.edu |
| Andrew Cormack | Jisc | Andrew.Cormack@jisc.ac.uk |
| Ian Courtney | University of British Columbia | ian.courtney@ubc.ca |
| Bala Desinghu | Rutgers University | bala.desinghu@gmail.com |
| Hongying Dong | University of Virginia | hd7gr@virginia.edu |
| Ingy ElSayed-Aly | University of Virginia | ie3ne@virginia.edu |
| Myles Frantz | Virginia Tech | frantzme@vt.edu |
| Cal Frye | Case Western Reserve University | cxf244@case.edu |
| Hao Fu | New York University | hf881@nyu.edu |
| Tracy Futhey | Duke University | futhey@duke.edu |
| Peng Gao | Virginia Tech | penggao@vt.edu |
| Mark Gardner | Virginia Tech | mkg@vt.edu |
| Brendan Gilbert | Africa Health Research Institute | brendan.gilbert@ahri.org |
| Rigel Gjomemo | University of Illinois at Chicago | rgjome1@uic.edu |
| Daniel Graham | University of Virginia | Dgg6b@virginia.edu |
| James Griffioen | University of Kentucky | griff@netlab.uky.edu |
| Zichuan Guo | University of Virginia | zst2ym@virginia.edu |
| Shuai Hao | Old Dominion University | shao@odu.edu |
| Dan Hardisty | Novetta | dhardisty@novetta.com |
| Scottt Henwood | CANARIE | scott.henwood@canarie.ca |
| Jason Hiser | University of Virginia | hiser@virginia.edu |
| Tonmoy Hossain | University of Virginia | pwg7jb@virginia.edu |
| Frank Hu | Norfolk State University | yhu@nsu.edu |
| Howie Huang | George Washington University | howie@gwu.edu |
| Liling Huang | George Mason University | lhuang20@gmu.edu |
| Andy Ingham | Duke University | andy.ingham@duke.edu |
| Indraneel Joshi | CanSSOC | indraneel.joshi@canssoc.ca |
| Elizabeth Kinney | University of British Columbia | em.kinney@ubc.ca |
| Inna Kouper | Indiana University | inkouper@indiana.edu |
| Prashanth Krishnamurthy | New York University | prashanth.krishnamurthy@nyu.edu |

| | | |
|---|---|---|
| Nina Lewin | University of Witwatersrand | nina.lewin@wits.ac.za |
| Wenjing Lou | Virginia Tech | wjlou@vt.edu |
| Lucky Matjelo | Central University of Technology | elmatjelo@cut.ac.za |
| Siun-Chuon Mau | CACI / LGS Labs | siun-chuon.mau@caci.com |
| Mimi McClure | National Science Foundation | mmcclure@nsf.gov |
| Sue McGlashan | University of Toronto | sue.mcglashan@utoronto.ca |
| Inder Monga | ESnet | Imonga@es.net |
| Roderick Mooi | SANReN | roderick@sanren.ac.zax |
| Diane Murphy | Marymount University | dmurphy@marymount.edu |
| Stuart Murray-Smith | Wits University | stuart.murray.smith@gmail.com |
| Anh Nguyen | University of Virginia | nguyen@virginia.edu |
| Alastair Nottingham | University of Virginia | atn5vs@virginia.edu |
| Alina Oprea | Northeastern University | a.oprea@northeastern.edu |
| Angela Orebaugh | University of Virginia | Ado4v@virginia.edu |
| Pegah Parsi | University of California San Diego | Pparsi@ucsd.edu |
| Sean Peisert | Berkeley Lab | sppeisert@lbl.gov |
| Tanmoy Sarkar Pias | Virginia Tech | tanmoysarkar@vt.edu |
| Wirawan Purwanto | Old Dominion University | wpurwant@odu.edu |
| M. Rosen | University of Virginia | rosen@virginia.edu |
| Sagar Samtani | Indiana University | ssamtani@iu.edu |
| Alireza Sarmadi | New York University | as11986@nyu.edu |
| Scott Seaborn | University of California, Berkeley | sseaborn@berkeley.edu |
| Zain Shamsi | University of Texas at Austin | zain.shamsi@arlut.utexas.edu |
| Chris Smith | Southern Methodist University | chris@smu.edu |
| Plato Smith | University of Florida | plato.smith@ufl.edu |
| Lisa Snyder | University of California, Los Angeles | lms@ucla.edu |
| Stephen Streng | University of Minnesota | stephen@umn.edu |
| Yixin Sun | University of Virginia | ys3kz@virginia.edu |
| Anton Verlygo | Northwestern University | anton@northwestern.edu |
| Kent Wada | University of California, Los Angeles | kent@ucla.edu |
| Gang Wang | University of Illinois Urbana-Champaign | gangw@illinois.edu |
| Cheryl Washington | University of California, Davis | washington@ucdavis.edu |
| Zachary Whitley | Novetta | zwhitley@novetta.com |
| Shengjie Xu | Delaware State University | shengjie.xu@dsu.edu |
| Shanchieh (Jay) Yang | Rochester Institute of Technology | jay.yang@rit.edu |
| Daphne Yao | Virginia Tech | danfeng@vt.edu |
| Aidong Zhang | University of Virginia | aidong@virginia.edu |
| Yizhe Zhang | University of Virginia | yz6me@virginia.edu |
| Danella Zhao | Old Dominion University | dzhao@odu.edu |

## ACKNOWLEDGEMENTS

This workshop was supported by National Science Foundation award 206431, “CCRI: Planning: Towards Building a Community Data Infrastructure for Cybersecurity Research.”

We thank Marilyn McClure for her support and for her participation in the workshop. We would also like to thank Emily Stacy at the University of Virginia for all her efforts in helping organize the workshop. In particular, she made sure all the speakers, panelists, and participants had the right Zoom links. She also made sure Jack, Von, and Howie had all the information (Zoom links, e-mail addresses, etc.) so the workshop ran smoothly. We also thank Molly Buchanan, Alastair Nottingham, and Anh Nguyen-Tuong for taking detailed notes and helping write the workshop report.