Data for AI in Network Systems Workshop Report

October 20-21, 2021

Hosted by:
Ron Hutchins, University of Virginia
Anita Nikolich, University of Illinois at Urbana Champaign
Kuang-Ching Wang, Clemson University

Introduction **3**

Executive Summary **3**

Session Takeaway Highlights **4**

NSF Data RFI Survey Results (Nicholas Goldsmith, NSF AAAS Policy Fellow) 4

Malware Detection at Scale (Josh Saxe, Sophos) 6

KC Claffy (UC San Diego/CAIDA) 7

Legal and Compliance Issues (Erin Keneally, Elchemy) 8

Breakout Session Summary **12**

Edge/IoT: 12

AI Model Development: 12

Security: 13

Governance/Legal: 13

Summary across all breakout rooms: 14

Major Takeaways **14**

Conclusions **16**

Appendix A: Session Notes **17**

Appendix B: Agenda **21**

Introduction

Machine learning techniques and AI models are proving useful across many application domains. However, the application of AI in computer networking remains challenging. Just as networking is an integral part of modern computing systems, the usefulness of machine learning techniques and AI models in this domain depends upon the ability to access, share and use good data. Data is needed to train and refine models for the ultimate implementation of good models in networks. However, data remains siloed and in many formats across institutions and in industry.

Advances in AI techniques can enhance performance and security in computer and networking systems. Significant and impactful efforts are emerging across public and private sectors to advance AI research and development. The use of AI techniques in cybersecurity - in malware and phishing detection for example - have been incorporated into mainstream tools such as endpoint detection and spam filters. But there remains a lack of focus on topics at the intersection of networking and AI. Networks and their associated data are notoriously inaccessible to researchers. This is in part due to a lack of available resources and infrastructure to collect and support such research, including data, testbeds, and benchmarks as well as the proprietary nature of networks especially in the commercial space.

This two day NSF-funded workshop sought to explore the fundamental needs that underpin the uses of AI for networking, including topics such as: what data can be made available for AI-enabled network systems? How will network data be collected, curated and used? What properties of data, and in what ways, would impact the networks and applications from a technical, legal, and ethical context? What new testbeds, labeling techniques, benchmarks, and benchmarking techniques are needed for network applications? And, most importantly, how will both the mindsets and skillsets of network researchers, and our next generation of students evolve and transform as we head into the future?

Executive Summary

We focused on three main types of network data that are amenable to ML techniques: Data about networks (observational and measurement data from the outside); Data from the network processes (telemetry); and Data payloads that traverse the networks. Network observations are easiest to collect, and are sharable but don’t scale well and remain expensive to store. Network telemetry is often considered proprietary and not sharable except by one-off agreements between researchers and providers. And Data payloads continue to be difficult to share due to potentially sensitive data or the ability to extract private information through traffic analysis.

The goal of the workshop was to explore the nature of network data to identify additional types of data in addition to the three mentioned above and understand the current obstacles to collecting, sharing and using such data sets in different contexts but focusing on AI and ML techniques and training data needs.

The main takeaways from the workshop are the following:

1. Legal and policy aspects are as daunting as the technical solutions. Governance issues are a huge hindrance. Governance issues must be addressed!
2. A community around networking and ML must be built.
3. The curation and governance of data is as important (if not more so) than its collection.
4. Ethical baselines for network data must be established. Ethical considerations should be integrated into the entire data pipeline. The community, IRBs, Funding agencies may need to be more prescriptive about best practices.
5. Data and AI Ethics should be integrated into CS and data science courses at the undergrad and graduate levels.

Session Takeaway Highlights

NSF Data RFI Survey Results (Nicholas Goldsmith, NSF AAAS Policy Fellow)

Nick presented responses to an RFI on dataset needs for data and network researchers. The RFI asked five questions of surveyed researchers:

1) What data is needed to accomplish your research?
2) To what extent can researchers contribute data to other research?
3) Are there privacy issues that impact network research data collection and access?
4) Are there issues in data collection with lack of standards and formatting for data?
5) A catch all category for other issues.

The audience for this RFI was primarily academic plus national labs and agencies. NSF received many responses on what kinds of data researchers need along with identification of what data is already available. A range of places data sets are available was described: specifically, CAIDA was identified as a major source of Internet data, plus others offering data piecemeal. Responding researchers called out a need to know where to go to find data and standards. Concerns were identified with the data: specifically, infrastructure and tools to collect data, the volume of data produced, and scaling data collection to be adequately representative.

Data quality, e.g. sample size and granularity of data sets, should be verified, specifically amid concerns of data not being representative: geographically and temporally. Also, some data can be simulated helping to relieve privacy issues. Data sharing has many barriers - license costs, proprietary nature of the data, lack of agreed upon repositories and standards. If you have your own data there may be other barriers. The most common response on barriers was: time, personnel, equipment which equates to money. Lack of agreed upon standards are a significant barrier also. If you take the time to share data, you want to know others can use it. There are incentives for sharing data: more and more sharing the data used in a paper is being required when publishing. Sharing data can be a factor in graduating students, not just “helping others

with data". Institutional issues can also produce barriers for collecting and sharing data. Privacy concerns are some of the biggest.

Privacy broadly includes:

1) copyright and intellectual property issues,
2) personally identifiable information,
3) cybersecurity issues (exposing vulnerabilities and locations),
4) data collection complexities that can limit collections,
5) getting permission.

Solutions include:

1) simulated data use
2) anonymization
3) aggregation of data
4) removing parts.

Policy approaches were discussed as possible partial solutions for these problems. Examples included utilizing Non Disclosure agreements (NDAs) or bringing code to data.

Challenges abound. Privacy preserving practices including anonymization can affect data quality. Training and education in these issues is not consistently offered to researchers and students. Disparate guidelines exist from agencies. This creates specific problems with cross disciplinary research like biomechanical, in which different disciplines have different funding sources and data use standards. Proprietary data was discussed as a major issue in utilizing data that had already been collected. Funding for collecting or hosting data is scarce and not consistent. Data collection difficulties (volume, tech change time, decisions on what variables we collect) are negatively impacted by this lack of collection and curation with inadequate incentives for sharing (P&T, etc).

Responses to the survey included pointers to privacy preserving algorithms, though discussions on these pointed out the limited utility of these in certain areas. New data collections and validation have been identified. The use of new tools to provide automatic annotation of data as collected is seen as an exciting tool if the annotation and labeling is accurate. Policy development can also impact data collection, curation, and use in a positive way. Having sets of standardized policy on sharing would help relieve some of the burden of risk identification. There was strong support for community based work to develop standards around metadata and formats, and to identify where to put and find data. A strong need was seen to coordinate data to work on the "not representative" issues. Publishing embargoes and data citations should be appropriately valued for promotion and tenure (P&T).

Link to the DCL on this topic: https://www.nsf.gov/pubs/2021/nsf21056/nsf21056.jsp

Malware Detection at Scale (Josh Saxe, Sophos)

Josh's presentation was divided into two parts: a discussion of a machine learning program at Sophos and his personal ideas on where research should be directed in the Network ML security space. Josh identified overlapping discussions around the fact that models run on endpoints, but threats are delivered over the network. Providing system models to the products in Sophos is a major goal.

"Not every problem is a ML problem". For a problem to be appropriate for ML, it is required that the input artifacts to the model be drawn from stable distribution, which is a tall order in security. In the area of malware identification, provider endpoint binaries in the malware software detection systems have an acceptable level of concept drift. However, going further and predicting that a particular network flow is a command and control (C2) infrastructure is not possible because the data is not stable enough. Josh mainly focused on mapping risk over the IP space. Servers in general are stable and may be a more appropriate data collection site than endpoints.

Josh's team considers a model deployment's complexity for implementation at an endpoint. The compressed size of a neural model for firewalls may be doable, but the added operational complexity causes problems. So the team didn't deploy this model. They also documented practical considerations around deployment, including staffing, pay for accuracy. A real model can't rely on static data because the model then gets stale. It needs a continuous feed of live data in order to remain responsive to changing malware. A separate team was identified that worked on cleaning the data: "garbage in garbage out". This required managing up to 30TB/day of input data. A large part of running an organization is political. Data can come from organizations that don't understand ML very well, and don't understand the dependencies, fragility, etc. Training parameters must be continually updated.

Challenges: ML tooling is changing very fast: RedShift was used as a data warehouse, but this is changing now. Hundreds of TB needed to be moved; new tooling was created around Amazon Sagemaker. This is now becoming more mature. One must keep wide peripheral vision on problems such as these. Operating in a large company brings a tremendous scope - Sophos protects more than 100M devices, creating huge data sets!. Figuring out how to squeeze out accuracy, parameter tuning, architecture search, bakeoffs is a continual challenge. Heavy hitters affect accuracy with data sent to the model. There are always people problems, cultural problems around those who write signatures. Ethical stakes are also high. Training models are challenging. They are finding a ceiling in accuracy due to label noise and data quality.

One area for academic research includes collecting and sharing benchmark data sets, which are the biggest leverage for ML/Cybersec research. Android experiments on cybersecurity for example use different data sets! One can't compare the corpus of Android research because different data is used on different experiments. This makes it hard as a practitioner to make use of literature since no benchmarks exist.

The EMBER data set is seen as good, but the community needs much more data. Differential privacy is also important. We need benchmarks.
Researchers are adapting new ideas from other areas in ML to CyberSec problems. The Transformer ML primitive is becoming popular. This tool is seen as a beautiful way to give words a contextual representation. It works well on phishing data. When one combines this with header data on email it significantly outperforms other methods. A model must train on decisions of hundreds of analysts - SOC analysts - as input. Artifact detection is a major focus. Much more territory needs to be investigated. In context we are learning to predict malicious domains. A new approach is to use "No code ML."

KC Claffy (UC San Diego/CAIDA)

"You can't secure what you can't measure." With these opening words, KC Claffy presented her experience at CAIDA collecting and curating network telescope data at the core of the Internet. She pointed out that no actor or entity is responsible for providing internet data, and there is no oversight of others who are providing data.

Why is it so hard to collect and curate data? The two part WOMBIR workshops, held in late 2020 and early 2021, presented some of the issues and problems. Actors have to probe from the edge. It's extremely hard to get operators to collaborate with researchers to get access to this data. So, what's the role of government and policy in this space considering that packets don't know national boundaries? Governments should fund measurement as well as setting policy. With regard to policy, there is an "Activation threshold", a level of concern that it takes to move a regulator to do something. Some things are being addressed due to consumer complaints - but the FCC, for example, is constrained in what they can ask for from commercial carriers. We must focus on this challenge, identifying questions that need answering, identifying barriers, identifying how researchers can have an impact beyond the university.

The WOMBIR workshop outcomes are available. A quick summary is included here:

1) IRBs are trying but over their head - decisions are variable across institutions.
2) **IRBs are structured to eval use of dataset for a specific purpose only.**
3) Is some research not going to get done? Yes with the IRB!!! And it can be a good thing since IRBs help to protect personal information, etc.
4) www.solarium.org/report - US Bureau of Cyber Security… recommends independent analysis of data from proprietary sources.

A clear recommendation from CAIDA is to fund and continue global measurement of networks, especially for data acquisition, data management, and lowering the barriers for researchers to find and use data. Data is available at CAIDA but the research community needs help to actually use the data! For example, how does a researcher map to prefixes? Sustainability of data collection and curation is key for improving internet security - specifically the part that the market is not taking care of which is seen by commercial interests as "out of scope" for their business purposes. The internet itself isn't being protected by companies' products. Data collection efforts for the larger internet should be ratified by community workshops.

Also, there are lots of different modes of sharing data: it presents a vast challenge to manage all these. Management of data for sharing has strong dependencies on the data and use case. For a telescope, a large chunk of the address space is lightly used (IPv4). If you listen you will hear lots of noise, and you can extract interesting data for security. Stardust.caida.org/docs. This is an example of a larger data set that offers capabilities not available elsewhere. These data sets must be managed to be useful. See: github/CAIDA/bgpstream.

Legal and Compliance Issues (Erin Keneally, Elchemy)

Erin emphasized that it is important that we not put legal risk at the end and it is imperative that we broaden the lens of our AI scope to enable risk control for trusted AI. But to understand responsible trusted AI, we must first understand the origins of AI risks and the challenges inherent in it. The bottom line lays in our law and social norms: what we're allowed to do/privacy and standards.

Innovation lies in the zone of risk between capabilities and expectations. This is where the battle of rights and expectations exist. The right to privacy, for example, conflicts with the government's interest in national security, innovation and free speech. These are some areas where AI is making decisions:

1) Credit reputation and scoring - which includes credit determinations, social identity, search results, classification and prediction.
2) Crime assessment - Industry is using data we have previously protected. This assignment of risk is predicated on observation and prediction.
3) Mass collect data to fuel AI - The significance of "zone of risk" is important. AI is mediating decisions and actions, not just affordances.

All of these decisions and use cases depend on judgements and sensibilities, and many of our assumptions have been uprooted based on AI. Information and control asymmetries exist and must be visible for analysis. However today there is unilateral subjective gatekeeping by organizations who control AI systems

Another aspect is that the physical harm that can be done by these AI systems is very real. Even though the harms are low and slow and therefore hard to see, they are happening at scale. There is a widening gap in capabilities and expectations that should concern users of AI systems for real-world purposes. It is at least inefficient to ignore and avoid management of these risks. Good actors aren't sure if they're ok with the law.

There is increased tension between legitimate uses of AI and those that are less solid, undermining the trust in ordering forces: law, markets, and technology. There is a lower understanding of risk distribution in the industry today. Risk is generated when capabilities are applied in the real world. They can be manifested in scenarios that contain friction between

sides: those using AI tools, and those on the receiving end of the power of the tools. This causes a disparate impact on people and groups, more often with those who don't have the power to protest or prevent the use of these tools.

Deep fakes are now occurring on an international scale. They can be good, but they also can be bad. Facial recognition technology can be the security/identity feature of choice for phones, etc. Passports and payment apps both use facial recognition. However, deep fakes can neutralize the power of this use of images for security purposes.

AI is also revolutionizing targeted advertising. It is increasing the speed of targeting at the expense of privacy. We can ignore this expense or accept it, we can transfer the risk, or mitigate it. However, it can be very costly to ignore or absorb these risks. The transfer of risk is unlikely (contractual indemnifications) since business is loathe to accept it. Even cyber insurance is still immature in this area. So we must mitigate the risks. There are three pillars for this area:

1) Research devel and data;
2) Governance;
3) Economics.

The field of AI needs sustained investment in research and development. There is a strong advantage for pre-competitively addressing risk and control so market pressures can be shared across corporations, not borne by one. And there are ways we can optimize choices around humans through asking questions like, "what is augmented control?" Building a human into the loop is a necessary component of AI systems since we still don't have "provable AI." We need to be able to map collective problems that affect everyone into forms that are addressable by AI. Additionally, cybersecurity is not a well defined area for AI. Today it is too dynamic. It's hard to define problems for machines to tackle, which provides a great space in which to do R&D. Consider addressing the valley of death problem. When one takes AI research and throws it over the wall to industry, a piece is missing - that of relationship, cooperation, and community.

The open secret is real world data (labeled training data) needs much more private data sharing. If we don't get this, we chase toy problems. Researchers then get marginalized by not being able to work on the larger issues.

The Impact programs [Impactcybertrust.org] formed the basis on this. There is more than just data at play to enhance responsible AI. DHS has a need for data to evaluate the space. R&D is impossible without quality data. We live in the Big Data era, but it's hard to find good "Big Data." Fruit must be picked and washed and packed. Data is similar in that it must be gathered, curated and stored.

NSF is also working to help solve this problem. The NSF program has 5 components:

1) Metadata discovery,
2) how do you find the data,
3) matchmaking capability,

4) tools to extract value,
5) social feedback loop.

Success elements for this program include:
1) cost savings from the research community,
2) findability and diversity of data.
3) Enablement of tools,
4) responsible framework.
5) Adding high value data sets to the mix.

AI regulation in the US is basically non-existent. There is no federal law (algorithmic accountability act was not passed), and no other regulation since then. Existing laws are fragmented and sectoral (health, credit, education, children). Recent activity (2021) includes the US Innovation and Competition Act, focused on competition with China ($200B investment, $80B carveout for AI) where development aligns with US values. The National AI Research Resource Taskforce (OSTP and NIST) Defense Authorization Act is another new initiative in this area. This group is providing a coordinated roadmap for national AI research.

NIST has undertaken trustworthy AI documentation that is open for comments via an RFI process. In order to measure bias in AI systems, the AICT act pushes to increase transparency in government AI systems and recommends that NSF propose AI intelligence institutes. The AIA in Europe (like GDPR) provides a much more comprehensive model. This regulation grants government full access to AI providers training datasets. It is currently a draft legislation, with the goal to provide legal certainty to protect innovation and private rights. It regulates the use of AI systems, going beyond GDPR. The regulation includes definitions for three types of systems:
1) prohibited (subliminal, social scoring, biometrics),
2) high risk (a safety component with impact on rights and safety, such as medical and transport),
3) low risk, outside of the other two.

AI Ethics should be seen as a three legged stool: principles, applications, enforcement. We have seen the most advancement in the principles side. What is needed now and for the near future is more convergence instead of more principles. We will never get harmony but we must at least get past the “not invented here” syndrome. We must pick something and do something. From the application standpoint, it’s about anticipating the harm from technology (products, services, features). The application component side comes in here. The Impact program (Creds tool) is attempting to address this to a degree. There exists a shortage of ethics and risk management focus and personnel for companies. RAIL (Responsible AI Licenses) is taking an end user license and source code license approach. IBM is delivering fact sheets. Enforcement is still lacking but we are making some progress with regulatory approaches. Enforcement appears to be the least stable today. One can take a bottom up approach (ethically defensible), a top town approach (IRB, ERB, Regulation, tie funding to rewarding behavior, “carrot or stick”). A sideways approach can be used to engage the reputation lever: name and shame. We need a Cyber Risk decision support tool. We could consider this to be a bottom up tool, for

example something like the ”Menlo report” operationalized into the Creds tool. This may be produced and utilized through a wizard approach for simplicity and bring to bear respect for persons, beneficence, law, public policy, etc. We’ve got to have a framework for applying these principles.

Economics is a strong forcing function. If the incentives are not aligned with the outcomes we want, then the data won’t matter. What is driving AI risk accountability right now? Government, venture capital funding, and market forces are. Are these strong enough incentives? Will market forces prevent bad things from happening, “Let ‘em out and deal with harms after,” is the most common attitude today. What data and tools are needed to help make better decisions for Risk and Benefits?

Question from audience: Should we require a process like IRB for AI research? Hopefully this exists!! Yes, it should exist! NSF can help to effectuate this.

Answer from Erin: A “Menlo report” for AI should include credentials and a framework. Thoughts on how a framework might evolve are being generated - we’ve gotten Creds to an alpha version but this fell short because it didn’t tackle the benefits. We need to present benefits along with risks to our institutions! If they only see the risks, they will say no! Then there is still a gaping hole around analytics and model risks.

The IEEE held a focus groups on AI Ethics. These groups reported that it’s not about pressing a button and getting a right/wrong answer. The value of the technology is in helping reason and articulate how to think about risk, how to justify decisions from an implementation standpoint. From a regulatory standpoint, fear, uncertainty, and doubt, still rule. It should not be “did you get it wrong” but did you think about all the issues and document the outcomes.

Breakout Session Summary

Breakout sessions were organized to facilitate participant discussions about the needs, gaps, and research opportunities. Given the diverse domains and expertise of our participants, four breakout rooms were convened centered around the general topics of: Edge/IoT, AI Model Development, Security, and Governance/Legal. All four rooms were seeded with the same three questions:

- What are data needed for using AI in networked systems and the gaps in data available today?
- Should a new NSF research program be created around the topic of ML Data for Networking?
- What specific disciplinary data sets are useful in this area?

The discussions in the four breakout rooms turned out to have significant intersections. The following summarizes key topics discussed from all breakout rooms:

Edge/IoT:

1. With Edge/IoT research, collected data increasingly involves actions or information of human subjects. Better understanding of their privacy protection requirements, Institutional Review Boards (IRB) best practices, legal and cost constraints are necessary.
2. NSF guidance on data sharing requirements and solutions is very important.
3. Privacy preserving techniques, e.g., homomorphic encryption, can be used for data requiring stronger privacy protection. However, anonymizing data may be an acceptable approach for protecting privacy; however, certain research, especially AI/ML research, can be negatively impacted by anonymization due to loss of identifier information. Some research has begun to explore trade-off between collecting accurate data while preserving privacy.
4. Data collection across distributed testbeds is no small task. Well designed instrumentation tools are important.
5. Streaming data is increasingly important for research of ML, especially when it concerns dynamic decisions in operating systems and/or on real-time systems. Making streaming data available and accessible (e.g., through FABRIC) is useful.
6. Traffic capture, SFLOW data, Perfonar, traceroute, metadata on routers and end hosts are examples of data collected.
7. Edge data has tremendous volumes. A program focused on tools to collect, transport, and make data accessible to researchers is useful.

AI Model Development:

1. As it stands, academics see a shortage in data, while industry does not. Instead, industry has too much data and is more focused on seeking faster, more memory-efficient and compute-efficient data.
2. As it stands, industry has plentiful data, but sharing the data is not easy due to legal, technical, and public perception challenges. On perception, there is public concern of their data being shared by big corporations (e.g., telcos). Having a public, neutral data sharing intermediary (e.g., NSF) can help with public concerns.
3. New NSF-industry partnerships start to offer ways for industry to share data with NSF sponsored research. e.g., NSF RINGS.
4. Labeling data is a big challenge. Frameworks are needed to explore labeling methods. Lag between data collection and labeling needs to be shortened.
5. Network systems are complex and there is far too much data.

Security:

1. A community-agreed labeling framework is the first step for data analysis across data sets. Furthermore, researchers can identify useful features to collect (a “recipe”) and share them with the community.

2. Data sharing needs to navigate around vendor non-disclosure-agreement (NDA) protected information.
3. A data sharing consortium is also possible. Data from Internet2 or NSF-sponsored facilities (e.g., FABRIC) may have less stringent sharing restrictions.
4. Research has been done on inferring information from network data (I.e., de-anonymizing).
5. Involving lawyers to discern boundaries for fair use is required. Being able to clearly show and affirm that critical information is not revealed is important. Past research has analyzed corporate data to ensure no sensitive data is leaked.
6. Defining a taxonomy of data vs. threats may be useful for vetting data sets for sharing.
7. Work is needed to explore data sharing mechanisms, repositories, and governance.

Governance/Legal:

1. Data management is essential for all universities, but not all universities have established data management protocols.
2. Out of schools that have data management plans, there has not been best practices in consensus.
3. Can there be a consistent data management plan and risk assessment approach across universities?
4. How to release private, noised, and anonymized data for usage? The Department of Transportation has a tiered model and physical facility for accessing data securely with some data anonymized.
5. NSF as a centralized authority can provide guidelines for data management.
6. Currently, there is not a common standard for training university researchers how to handle data. Industries and government, on the other hand, have long established processes in place for data man agreement.

Summary across all breakout rooms:

Network systems are complex with large numbers of types and volumes of data about them, both static and streaming, that can potentially be used to derive AI models for network control, security, and other applications. Data labeling is one of the most important open challenges for them to be usable for AI.

There is a major disparity of accessible data across academics, industry, and government. For academics, data is scarce because both commercial and R&E operators are hesitant to share. Business confidentiality, personal privacy and regulatory issues prevent sharing. One of the practical barriers to data sharing is that there is a *lack of clarity* on what can be shared, both on the industry side and academic side. The risk/reward of sharing is perceived to be too high.

For industry and government, there is an overwhelming amount of data yet not enough innovation in the creation of the underlying algorithmic models that function at high speed and with accuracy. This presents an area of opportunity to partner with academia. To overcome the

bottlenecks, reducing concerns and overheads of academic, industry, and government collaboration is key, and having clear guidance for data governance and data management requirements, standards, and tools are a top priority.

Major Takeaways

Data Accessibility - Data sets are not easily accessible was the pervasive theme of the workshop. This encompassed problems with *identifying* data sets that were available, finding robust labeled data, and easily accessing such data sets. A topic that came up repeatedly is that network or systems providers provide data only on a case by case basis, leaving out a lot of researchers who don't have the connections or institutional structures (legal, contracting, risk management personnel/practices) to assist with obtaining the data.

Data Sharing - There are existing technical solutions for sharing what might be considered sensitive or private but these solutions come with a lot of system and administrative overhead. For example, an often cited solution for sharing data anonymously, Multi Party Computation (MPC), doesn't scale well in practice.

Data Cyberinfrastructure - Funding is perpetually needed for collecting and hosting data. The funding cycles often don't match the research needs. Funding for data storage tends to be on a case by case or project by project basis, when in reality it's a long term effort to achieve sustainability. The tension remains between funding agencies, institutions, PIs and other researchers in terms of who has ownership of the data storage funding and process. Infrastructure for data collection is complex and doesn't scale well. And often this infrastructure is constructed and maintained as a result of a short term grant which doesn't offer scalability or longevity. Data collection difficulties include (volume, technical rate of change, decisions on what variables we collect).

Data Ethics - Ethics is a three legged stool (triad): principles, applications, enforcement. The most advancement to date is in the principles side, as people are becoming familiar with ethical concepts around data use. What's needed is more convergence between the three instead of just more principles. We will likely never get completely harmonized across the three, but still must get past the "not invented here" syndrome and adhere to some basic accepted standards around data.
From an application standpoint, researchers must anticipate the harm from technology (products, services, features). Network application developers must participate in the principles and enforcement part of the triad. There is a shortage of prescriptive ethics and risk management guidance for companies with regard to data. Most risk management guidance focuses on the resulting systems or data storage devices. Some examples are: RAIL - Responsible AI Licenses - which is an end user license approach, and source code license approach. IBM has fact sheets.

However, enforcement of data ethics is still lacking despite some progress with regulatory approaches. A framework for applying these principles is needed for any realistic enforcement to occur. Enforcement is the least stable - several examples exist for modeling enforcement. A top down approach (IRB, ERB, Federal Regulation, or tying funding to ethically principled behavior in data collection and use), i.e. a “carrot or stick” approach. A sideways approach to enforcement (engaging the “reputation lever” - name and shame) is also used on occasion. At the foundation, a bottom up tool - such as the Menlo report (which was operationalized into the Creds tool) should be seen as basic including respect for persons, beneficence, law, public policy, etc.

Data Governance - We often assign the same level of risk to disparate types of data. The nature of medical data is much more codified. However, what comprises PII is still in flux with regard to data collected from/about our networks, leaving IRBs to often follow the much stricter protocols assigned to medical, human subjects work. Several participants mentioned that the Belmont Principles underlying IRBs don’t account for the nuances of network data. The Menlo Report was a good start to account for such nuances in network and CS data but is already out of date and should be revisited. The network community itself should draft ethical concerns and guidelines for network data for IRBs, a “Network Menlo Report.”

Data Sharing with Network Providers - ISPs and Content Providers hold a lot of power - their data is generally not shared with researchers except on a case by case basis. This is viewed as a huge obstacle to better network and security research. Several participants recommended more robust Industry-University cooperation in this area. This will require publishing of best practices and a formalization of personnel needed to accomplish the risk management framework needed.

Privacy Issues. There is currently no technically sound, yet easy way to share data that’s considered private. However, even the definition of private data varies highly from PI to PI and institution to institution. There is a lack of agreement on what data fields need to be anonymized. There isn’t a canonical definition of privacy with respect to network and network-adjacent data and therefore a resulting strong justification for anonymity. It was mentioned by several participants that current technical anonymization techniques negatively affect data quality and therefore impacting training of AI models because so much data has to be taken out to enable its use.

Data Sharing and Labeling Standards. There is a **l**ack of standards, techniques and trained people to do the work of data labeling. This is often left to the discretion of PIs or researchers and thus varies. Because of the large variation across data labeling, data sharing is made more difficult.

Reproducibility. Without standards around data collection, labeling, and long term storage, reproducibility of machine learning research continues to be problematic.

Funding for Data Collection and Sharing - The current method of funding data collection and sharing on a per project basis is ineffective. The workshop participants urged federal agencies to create programs and opportunities at a national or larger scale for long term data collection and curation.

Static Data - Researchers can't only rely on static data because network data patterns evolve quickly and today's data may be too quickly of date to be useful. A continuous feed of live data is more effective, especially in areas such as network security in which adversaries change techniques very quickly.

Synergies between Network, Security and Systems Data - Rather than collect network data alone, we should be able to correlate network data with other research areas and systems data.

Conclusions

Machine Learning techniques for cybersecurity have been steadily progressing over the last decade. Advances in object recognition and pattern matching, powered by more robust ML techniques, including deep learning, have enabled the development of security systems with improved accuracies that protect systems. Machine learning techniques as applied to networks have also matured somewhat in this timeframe, but there remain systemic issues around the collection, curation, sharing and use of data sets that are inhibiting research progress. Much of this was pointed out in the NSF DCL around data sets needed to conduct research on computer and network systems. One of the largest challenges is that data from networks, especially commercial networks, is hard to get unless you have a relationship with the company. Most participants agreed that commercial and/or large R&E networks have the most interesting and useful data. However, even when data is shared, privacy and liability concerns remain. Rather than being addressed in a one-off (short term, ad hoc) fashion, our workshop participants urged a collective set of standards around governance.

Networking researchers need stronger industry/academic cooperation, which will produce better research for everyone. This is particularly important for research into ML techniques because access to large amounts of data is vital for progress. In addition to funding for basic cyberinfrastructure around the collection, curation and storage of ML data that can be used by both the cyber and network communities, NSF might consider an AI Institute around network data. The bottom line is that federal agencies need to direct more attention, programs and funding towards long-term data collection. governance and storage in the area of networking.

Appendix A: Session Notes

Intro Session: Why now? What new science can be enabled by collecting more data?

- Fundamental insights into the nature of networks.
- Science of streaming digital networks at unprecedented scale.
- Privacy preserving analytics.
- Essential data for developing and testing new models/theories of sparse/graph AI/ML.
- Foundations and tools necessary to improve network safety, security, and surety.
- Future wireless/mobile networks design (xG) that are more self-managed, leverage AI.
- Improve accuracy of performance predictions in large scale applications, leveraging AI/ML.
- Future architectures that facilitate large scale coordination and synchronization of IOT for spatially- distributed measurements, perhaps across continents, such as the CORS network for measuring continental drift.

Topics of interest to the attendees:

- Streaming analytics on real-time data sets, not just on stored data (IoT).
- Securing data. Privacy preserving techniques.
- AI training for near-real-time digital twins for networking; the digital twin can then be run forward in time at a faster than real-time rate to provide probabilistic predictions
- Logical and probabilistic AI in addition to ML; networks often have structures on which logic works as well as or better than ML neural nets
- Network routing and rate accommodation algorithms that learn by making a small percentage of deliberate mistakes (evolutionary strategy)
    - (Some communities related to this: “Self-Driving Network” and “Autonomous Networks”)
- Intersection of AI and networks in cyber physical systems
- Edge networks and real-time AI
- Trustworthy AI for networks (e.g., NIST Trustworthy AI initiative)
- Harmonizing EU-US “AI for Networking” policies” (e.g., EU AI Act which has been proposed)
- Life-long learning, concept/data drift
- Data labeled differently by different providers and label changes over time
- AI for synthesizing realistic network data
- Coordination of distributed AI systems such as networks acting on local information in addition to delayed global information

Gaps identified:

Technological gaps

- Applying/Adapting statistical methods to computer security when they were developed/intended, and often perform better, for other applications/problems
- Building datasets serving the interest of the wider community considering their application in AI/ML methods. Perhaps guidelines on how to generate datasets?

- Data management and long term archival - policy and cost issues.

Policy gaps

- Cross administrative domain sharing
- IRB approval — What needs IRB approval? What are the risks of various types of research?
- Training of university personnel/students in standard industry/gov't cyber security necessary for proper data handling. Most companies/gov't entities require ~2 hours/year of cyber training for all personnel.
- Greater training for researchers on ethical use in use of data, training around IRB
- Lack of awareness of data approval practices for different universities
- Most IRBs are specialized for biomedical/clinical or social science. Does a third category for networks, smart cities, and human IOT need to be stood up?
- Build a conversation around privacy/ethics vs. value of the data and the need for the research.
- NSF Data Management Plan could require IRB approval for appropriate data in research?
- Social and human aspects beyond the network systems
- We tend to work in stovepipes and need to bring: Legal, compliance, risk management, security, and beyond to come together.

Data sets currently available:

- SOREL-20M
- EMBER 2.0
- ISCX-2016
- CAIDA
- Internet Topology Zoo (http://www.topology-zoo.org/) — albeit stale at this point
- Internet2 Network NOC (https://noc.net.internet2.edu/i2network/index.html)

Data sets needed?

- Data and dynamic metadata on TCP flows - correlation of network data with application performance data and events (augmented network datasets)
- Router configurations — anonymization of security-sensitive aspects (e.g., access control lists) is fine; more enhanced version of Netconan (https://github.com/Intentionet/netconan) could help with anonymization
- Network management and control policies such as ACLs, routing policies, and configurations and other related control data
- Related to the above, given the wide adoption of (virtualized) network functions and service function chains in networks today, it would be useful to have use cases and policy examples for various types of networks

Suggestions for moving forward.

- **Data sharing** - governance, technical

- Convene a **community** of networking folks to talk about data (data needs for research, data sharing)
    - Include networks science and data at scale experts.
    - Community needs to produce a **research agenda specific to AI for Networking**. What new research would be possible if we get these gaps filled?
    - What do you need to perform research in this area?
- Build foundational information to encourage proposals.
- Build **guidelines for IRBs** about how to judge network research proposals
    - Menlo report provides a set of principles that can be used to evaluate the harms from experiments in the CS context, but IRBs may be challenged to apply those principles to specific cases, especially since the sorts of experiments that are done in the CS can differ widely in their character, which makes it hard for IRBs to reason by analogy. It might be useful to organize some sort of advisory group with deep experience in the area to prepare an assessment of a specific research proposal that could be given to IRBs as guidance.
    - It is problematic to have program committees raise ethical concerns after a work has been completed and submitted
    - Allman/Paxson IMC 2007 paper *Issues and Etiquette Concerning Use of Shared Measurement Data* (https://conferences.sigcomm.org/imc/2007/papers/imc80.pdf)
    - Commission/motivate writing a meta-report about different universities'/institutions' policies on data acquisition and sharing.
- **Ethics** curricula for networking students at an earlier time, before they are collecting data including data sharing.

Suggested Priorities for NSF including research agenda :

1. Create a Data Institute to complement the funded AI Institutes.
2. Encourage data sharing and provide incentives in solicitations
3. Possible foci for new NSF funding programs:
    a. Data pipeline/lifecycle - new methods, etc on collecting, sharing, curating - methodologies, governance,
    b. Program on legal and ethical (accountability) issues with network data.
    c. Research is needed on AI techniques for optimizing local action to facilitate global coordination in the face of uncertainty and delayed information about global state. Networks are complex and dynamic distributed systems which have different information at different places in the network.
    d. The supporting CI in support of this problem
4. Cross agency solicitation: DoE, NIH, NIST, NSF
5. Fundamental insights into the nature of networks.
6. Science of streaming digital networks at unprecedented at scale.
7. Privacy preserving data sharing and analytics.
8. Essential data for developing and testing new models/theories of sparse/graph AI/ML.
9. Foundations and tools necessary to improve network safety, security, and surety.

10. Future wireless/mobile networks design (xG) that are more self-managed, leverage AI.
11. Improve accuracy of performance predictions in large scale applications, leveraging AI/ML.
12. Future architectures that facilitate large scale coordination and synchronization of IOT for spatially- distributed measurements, perhaps across continents, such as the CORS network for measuring continental drift.
13. Research is needed at the intersection of cyber physical systems (CPS) and AI in the large sense (not just ML). There will be applications in networking as well as many other areas.
14. We recommend that this workshop be re-run annually (bi-annually?) given the rapid changes in theory and application.
15. Support for long term management of collection, curation, and training around Internet data (CAIDA)
16. Invite research aimed at supporting the NIST Trustworthy AI initiative.
17. Fund a CCRI focused on acquiring and disseminating research-ready datasets. The CCRI would have expertise in de-identification, homomorphic encryption, and similar anonymization or uncertainty-increasing techniques.
18. Ensure sustainability for current AI networking measurement data collection and repositories.

Appendix B: Agenda

Day One | October 20, 2021		
Time (ET)	*Topic*	*Presenter*
11:00 AM	Welcome	Deep Medhi, NSF and Kuang-Ching Wang, Clemson University
11:10 AM	NSF Data RFI Survey Results	Nicholas Goldsmith, NSF
11:20 AM	Malware Detection at Scale	Josh Saxe, Sophos
11:50 AM	Transforming Mindsets in STEM Education	Anita Nikolich and Ron Hutchins on behalf of Wendy Newstetter, Georgia Tech
12:20 PM	Lightning Talks	Engin Arslan, University of Nevada Reno Suman Banerjee, University of Wisconsin - Madison Ram Durairajan, University of Oregon Erick Galinkin, Rapid7 Michele Polese, Northeastern University Ness Shroff, Ohio State University
1:15 PM	Break	
1:25 PM	Breakouts on Topics - Edge/IoT - AI Model Development - Security - Governance/Legal	Facilitators: - Ness Shroff & Suman Banerjee - Sven Cattell - Anita Nikolich - Dave Clark
2:15 PM	Breakout Group Discussions	Facilitator: Dave Clark, Massachusetts Institute of Technology
3:00 PM	End	
Day Two | October 21, 2021		
Time (ET)	*Topic*	*Presenter*
11:00 AM	Welcome	Kuang-Ching Wang, Clemson University
11:20 AM	Getting Data from Network Systems	KC Claffy, University of California San Diego

11:50 AM	Data Legal and/or Compliance Issues	Erin Kenneally, Elchemy, Guidewire
12:20 PM	Lightning Talks	Christophe Diot, Google John Heidemann, University of Southern California Hongxin Hu, University at Buffalo Yingjie Lao, Clemson University Georgios Papadimitriou, University of Southern California Sagar Samtani, Indiana University
1:20 PM	Break	
1:30 PM	Open Discussion Forum	Facilitator: Dave Clark, Massachusetts Institute of Technology
2:30 PM	Report Out (link to report)	Ron Hutchins
3:00 PM	End	