## Research – Cognitive Bias Data Mining

**Background**

Cyber defenders and attackers are humans; like all people, they are subject to cognitive biases. For the defender, cognitive bias can negatively impact the discernment of threats and the ability to evaluate courses of action. For the attacker, it can lead to time consuming mistakes or non-optimal sequences of decisions, which creates opportunities to frustrate, curtail, or halt the attack completely. Research is needed to better understand these cognitive biases, provide better safeguards against error and misjudgment, and provide a way to reduce the costs and damages associated with cybersecurity incidents. The need for data is a profound limiting factor in the development of cognitive models, simulations, experiments, and other approaches to research these phenomena.

Curating data to satisfy multiple use cases is a non-trivial endeavor. Development of a data repository to aid researchers trying to build models and simulations, develop experiments, develop metrics, and test solutions would greatly aid in these efforts. There are three primary obstacles to creating such a repository: locating and identifying relevant data; proprietary concerns that prevent data sharing; and the need for a set of standards, formats, and other specifications that would enable usability of the data for a variety of relevant use cases.

Research Proposal:  
Identify existing data sets suitable for use in modeling cognitive bias in cybersecurity, simulating cyber-attack behaviors, train artificial intelligence models, and develop human-computer teaming for cyber defense.

The objective of this project is to locate and describe relevant data resources in open source and proprietary domains for the purpose of building models, developing solutions for high priority cases, and related efforts in the area of cybersecurity.

The offeror should be prepared to gather information on relevant data resources, including those that might require NDAs, anonymization and other projections that would be available for these purposes. What data exists? What are the barriers to accessing or making these data stores available for multiple purposes (including cog security projects) and other challenges?

Research Questions

What are the existing data sets, what questions can they be used to answer, and what are the constraints on using them? Could or should such data sources be placed in a data repository? What would be barriers to collecting a given data source into a common repository? Would researchers need an actual repository, or would a simple library or index of resources be sufficient?

One example of existing open-source data is Protective Capacity (PCAP) data (see <https://www.netresec.com/?page=PcapFiles>*,* <https://sugerent.tistory.com/459>) which includes sources originating from:

* Computer defense exercises
* CTFs, red/blue team competitions
* Malware traffic from real-world intrusions
* Network forensics training
* SCADA/ICS captures

Many different data sources exist relevant to the problems of cybersecurity (insider threat; Advance Persistent Threat (APTs); etc.). We will explore PCAPs, which are files used to determine network status, study data communications, and allow forensic investigators to attend to potential network issues. Network traffic data, PCAP data and other open-source data exist in volume -- if one knows where to find it. Some of this data includes the handiwork of amateurs (“script kiddies” and amateur hacking groups) who may generate traffic containing common and well-known tactics, techniques, and procedures, such as network scans, host software version lookups, and common automated attacks like Metasploit or Mimikatz. APTs and nation-states may instead attack in more subtle ways and could even utilize exploits that have not been disclosed publicly (i.e. zero days).

Much of the most useful data in this area is held by private corporations and is proprietary. And yet, companies, as well as universities and research labs, would benefit from data sharing. Some modifications might enable proprietary data sharing, such as anonymization of data, which might make data sharing more possible while preserving customer privacy, trade secrecy concerns, and similar barriers. Where data is proprietary, what would businesses be willing to do in order to make data shareable?

Whether from open or proprietary sources, data would need to be mapped to taxonomies and frameworks in order to organize, catalog and index data to use-cases and use case types. Such taxonomies could include:

* Reconnaissance (Scan/Ping) e.g. Nmap, Netstat
* Exploitation (page access, authentication, buffer overflow, remote access, etc.)
* Communication (exfiltration, command and control, etc.)

Data could also be generated, for example, the C2 Windtunnel, an AFOSR supported project developed by George Mason University, University of California (Berkeley), and Vanderbilt University, might be capable of producing suitable data.

Additional Research Questions:

* What would or should that data look like?
* Are there other capabilities available with similar or better capacity to generate simulated data for use-cases?
* What are the features needed for a dataset to be of use in understanding the human cyber attacker, the human responder, and the attack itself?
* How long a time period does the data need to cover for given use-cases? Malware, e.g., can evolve in the space of 2 years - does the data set need to cover multiple years?
* What types of augmentations are needed so that the data can be used with multiple modeling paradigms?
* How can qualitative data be provided along with quantitative data to support improved analysis?
* Which type of data is needed for which cybersecurity problem; e.g., when can netflow data, incident response data, vulnerability data, and so forth be best employed?
* How can the biases inherent in the data be expressed and conveyed to the users?
* Can data sharing standards be established at different levels, so that data providers can feel comfortable in opening up data to the scientific community for research purposes?
* Universities, companies, and research labs often have relevant data sets.
* Lack of common format and taxonomy. Consequently, when data is shared, it is always in a different format, or labeled differently than prior data, thus inhibiting comparison and accumulation.
* Data that is shared is often done with data so obfuscated it provides little insight. Rather, what is needed is data labeled with type of company, country, time, type of machine (e.g. printer, server, etc).
* Data is often collected on the dependent variable; however, science needs balanced data.
* How can we collect the volume and diversity of traffic seen in production/enterprise networks?

**References**

<https://pdfs.semanticscholar.org/6bbc/b7ac5076f85c3004c2085c8fac1aca62a1d1.pdf>