3-19-2021

SaTC: CORE: Small: Deep Learning for Insider Threat Detection

Shuhan Yuan
Utah State University, Shuhan.Yuan@usu.edu

Data Management Plan

The outcomes and results of the proposed research will be disseminated broadly, at all stages of the project.

- **The types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project.** The research input data includes publicly available datasets, such as the CMU CERT dataset, for analysis and simulation study. No sensitive or confidential data will be used in this research. Due to the novel approach being proposed, it is anticipated that this project will result in a number of journal and conference publications. Research data generated in this work will include software tools and algorithms, simulation results, results on data sets, and related paper publications.

- **The standards to be used for data and metadata format and content.** The analysis results will be generated based on the proposed methods and will be augmented with essential metadata and narrative for interpretation. Data and results will be labeled and digitally stored in formats of plain text (.txt), comma-separated values (.csv), and javaScript object notation (.json) which have widespread usage within the research community for data storage and interchange. The developed software toolkit will be provided with necessary documentation including manual, tutorial and demonstration.

- **The policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements.** Since no sensitive or confidential data will be used in this research, we will create Github repositories to publish our experimental raw data, developed softwares, research results, and papers. We will archive the data with necessary metadata to facilitate public access to the data. For data that are expected to contribute generally to the research community, we will provide detailed descriptions, so other researchers can either reuse the data in their research projects or repeat our designs, algorithms, or experiments. On Github, other researchers can provide feedback and comments, which are expected to provide valuable information to this project. The data shared on the Github is basically open to public use, except for the conference and journal papers. For the latter, the copyright is retained by the corresponding copyright holders.

- **The policies and provisions for re-use, re-distribution, and the production of derivatives.** In this project, we will comply with NSF's data sharing directives under the Data Protection Act of 1998. We would not release any controlled-access information about any sensitive data from our website or Github repositories. However, the results learning from these data will be made publicly available.

- **The plans for archiving data, samples, and other research products, and for preservation of access to them.** Software packages include the implementation of components developed in this project. All software developed in this project will be freely available to researchers and educators, and published within a short time frame after finishing the development of software in each phase. In particular, the PI plans to publish the open-source software under GNU General Public License and permit the dissemination and commercialization of the enhanced versions of the software or incorporation of the software as components in other software packages. Our software can be transferred to other individuals or teams who can continue their own development. We plan to host a website to manage and disseminate the improvements or customization of our developed software by others.