# TWOS Dataset

## 1. Description

The TWOS dataset has been constructed from real user interaction with the host machines that contains both legitimate user data and malicious insider instances (masqueraders and traitors). The dataset comprises of data collected from 6 sources (keystrokes, mouse, host monitor, network traffic, emails and logon activity). The dataset contains activity of 24 users, which was collected over a period of 5 days. It contains activities pertaining to two types of malicious insiders, namely masqueraders and traitors. There are 12 masquerader instances of 90 minutes duration and 5 traitor instances of 120 minutes duration.

NOTE: The file named "*Important\_Info.xlsx*" contains information about the composition of each team, timings and other information pertaining to each masquerader and traitor instance. In the case of wild-card periods (i.e., masqueraders) we do not indicate attacking user, as there may be more of them.

## 2. Description of Data Sources

All the logs have been anonymized in accordance to the procedure stated in our paper.

### I. Keystrokes:

Keystroke activity of each user can be found within one or more files named according to user ID followed by an optional timestamp. Depending on the amount of interaction of the user, multiple files have been created as a result of log rotation. It contains information pertaining to timestamp, key press / release event, key value and username.

NOTE: Within each log file, there could be one or more entries called "REFRESH". This has been added to indicate inactivity of the user due to absence of keystroke activity.

### **II**. Mouse:

Similar to keystroke logs, mouse activity of each user can be found within one or more files named according user ID followed by an optional timestamp. Depending on the amount of interaction of the user, multiple files have been created as a result of log rotation. It contains information pertaining to timestamp, mouse move / click / release event, coordinates of mouse pointer and username.

NOTE: Within each log file, there could be one or more entries called "REFRESH". This has been added to indicate inactivity of the user due to absence of mouse activity.

### **III**. Emails:

Email activity of all users are within several DSV files, where each file represents a particular user - rows with the label NORMAL are emails sent by that user while rows with label ANOMALY were built as rnadom mixtures of other users' emails with emphasis to include samples from all remaining users - due to this fact, the ratio of anomaly rows varies among particular users. Each file contains information such anonymized body of message, and features extracted by Linguistic Inquiry and Word Count (LIWC) tool. Note that you may use either body of the emails or LIWC features for some classification task related to authorship verification/user identification. In some cases, you might ignore the ANOMALY rows and work only with NORMAL rows according to your requirements.

### **IV**. Host Monitor:

The system calls of each user can be found within multiple files that contain the name of the user. It contains file system, registry, process and network related information. Specifically it contains information such as timestamp, process name, PID, Parent Process name, Parent PID, system call operation.

### **V**. Personality Tests:

All users were asked to fill in personality test containing 50 questions, as described in our introducing paper.

### **VI**. Network Traffic:

Network activity of all users are logged into several pcap files that are consecutive in time (they were captured one by one). The capture of the pcap files started 1 day before the competition and finished 30 minutes after official end of the competion. They contains information such as HTTP Request (e.g, GET, POST) / Response, status code, content length, content type.