AICS 2019 Workshop Challenge Problem:
Malware Classification under Adversarial Conditions

Task and Dataset

The proposed task for this challenge is to construct AI-based malware classifiers that are robust to adversarial evasion attacks. We define the **malware classification** task as determining the class of a sample that has previously been identified as malware. For this challenge problem, the data consists of 15,670 instances split into five classes according to a subset of malware types: Virus, Worm, Trojan, Packed Malware, and AdWare. For each sample, a dynamic analysis on a Windows virtual machine was performed and the sequence of Windows API calls was extracted. Additional processing was performed to collect unigrams, bigrams, and trigrams of API calls. We make the following assumptions about the dataset:

- All instances are malware
- The classes cover all instances under consideration
- Each instance has a single unambiguous label

In the released datasets, the class labels and API calls are obfuscated as integers, but the correspondence between any particular API call as a unigram and as part of a bigram or trigram is preserved. See below for an example.

The test dataset consists of 3133 instances with the same set of features as the previously released training data. Some, but not all, of the test data are adversarial examples, whose features have been perturbed by a variety of adversarial evasion attacks. The schema for the test data is the same as the training data (detailed below) with the label field missing.

Data Schema

The feature matrix for the 2019 AICS challenge problem is represented as an NxM sparse matrix and serialized in JSON format with the following schema:

```
{
        “data”: JSON_ARRAY(K),
        “row_index”: JSON_ARRAY(K),
        “col_index”: JSON_ARRAY(K),
        “col_schema”: JSON_ARRAY(M),
        “shape”: [N, M],
        “labels”: JSON_ARRAY(N)
}
```

The "data" field gives the values of the nonzero entries of the feature matrix, while "row_index" and "column_index" give the row and column indices corresponding to these values. So that the nonzero entries of the feature matrix are specified by

feature_matrix[row_index[i], col_index[i]] = data[i]

for i=1,…,K. The "col_schema" array provides names for the columns of the feature matrix. For unigrams, a column name will be a distinct integer corresponding to that specific API call, for bigrams (and trigrams), the column name will be a pair (triple) of integers separated by a semi-colon. For example, the trigram "27;9;150" represents the sequence of API calls 27, 9, and 150. The "shape" array indicates the dimensions of the feature matrix (note that it would be possible to infer these dimensions from the other data, but we include this field for convenience). Finally, the "labels" array indicates the class of each instance (**this field will be available in the training data, but will be absent in the test data**).

Submission

Submissions to the challenge problem will consist of two parts: 1) a CSV file containing results from applying the developed classifier to the test (sent to aics@ll.mit.edu) and 2) a paper describing the developed classifier (submitted via the workshop website: http://www-personal.umich.edu/~arunesh/AICS2019/). **The CSV file with classification results should have two columns: (1) an index field with values corresponding to the 'row_index' field of the test data and (2) a classification among the possible classes: 0, 1, 2, 3, 4.**

The results file will be evaluated based on performance on the test set while the Challenge paper submission will be selected for inclusion in the AICS workshop as full published paper. Please note that challenge paper submissions have the same submission deadlines as regular papers.

AICS 2019 Workshop Adversarial Learning Challenge Problem important dates:

- Training data released: 24 August 2018
- Test data released: 29 October 2018
- Papers due: 5 November 2018
- Challenge problem results due: 5 November 2018
- Accepted papers announced 26 November 2018

Challenge problem winners will be announced at the workshop. **Participants' solutions will be evaluated based on a harmonic mean of macro-averaged F1 scores on un-attacked pristine test data and attacked test data.**

If you have any questions about the challenge, please send an email to aics@ll.mit.edu