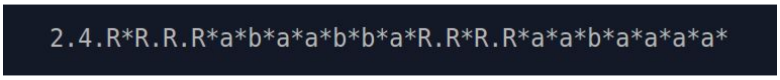
Visualizing Discriminative Power of Symbol-based Network Traffic Models

Carlos A. Catania, Jorge Guerra
Labsin, Facultad de Ingeniería, Universidad Nacional de Cuyo.

Motivation and Goal

In the context of network security, a **behavioral model** aims at capturing the long term characteristics of the network traffic following an anonymous and efficient process. Such behavioral models conform the basis of several machine learning algorithms for detecting malicious traffic. In particular, the Stratosphere IPS (http://www.stratosphereips.org) behavioral model represents the history of network connections by using a set of 50 symbols. Each time a new network flow is observed, the corresponding symbol is concatenated according the values of three features (i.e. size, duration and periodicity). An example of a behavioral model is shown in Fig. 1.



```
2.4.R*R.R.R*a*b*a*a*b*b*a*R.R*R.R*a*a*b*a*a*a*a*
```

Fig 1 . A behavioral model representing all the connection based on UDP to port 53.

Before developing a detection method for a particular malicious behavior, it is useful to have at least an intuition of the potential discriminative power of the symbol-based behavioral model. With that goal in mind, we developed STF-PATTERN-VIZ, an open source application on top of the Shiny R package. The application consists of a set of interactive visualization components for analyzing the discriminative power of symbol-based patterns present in Stratosphere IPS behavioral models (see Fig. 2).

Overview

The idea behind STF-PATTERN-VIZ consists of applying a basic n-gram analysis on the symbol-based behavioral model. Given a set of labeled behavioral models, the application provides a visualization considering the most frequent n-grams of a given size. The visual components offer information about different aspects of the selected n-grams. The aspects considered are basically:

1) The total number of times a given n-gram is observed in all the behavioral models.
2) The number of times a given n-gram is observed in different behavioral models and
3) The labels of the models where the n-gram was observed.

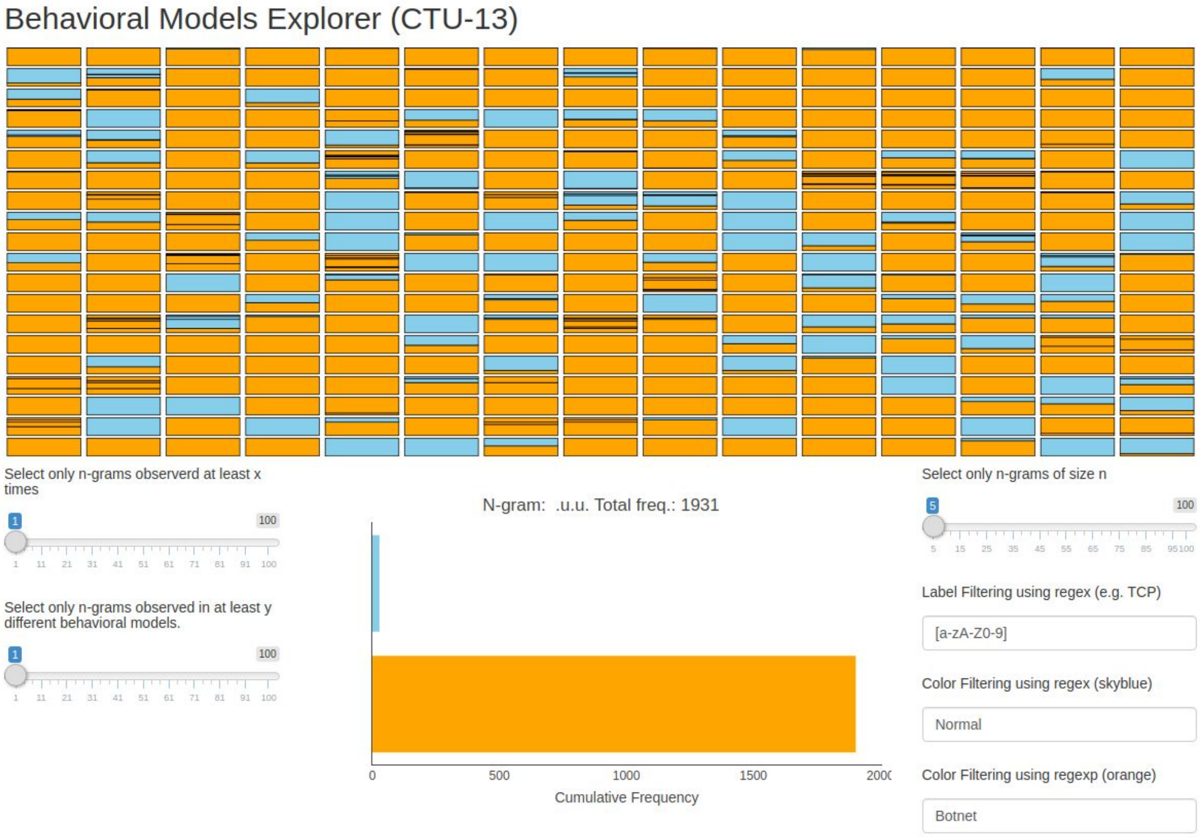
Such information is mainly provided by a simple visualization tool in the following way:

- Each n-gram is represented by a rectangle. Right now the app shows only the first 300 n-grams
- A subdivision inside the rectangle indicates the same n-gram was seen in different behavioral models. The more subdivisions, the more observations of the n-gram pattern in different behavioral models.

By default, **Normal** behaviors are represented in Blue while **Malicious** in Orange. By using this simple strategy the user can have an idea about the discriminative power of n-grams in a given dataset. Clearly, a rectangle painted with only one color indicates that a particular pattern (of size n) was only observed in behavioral models with the same label. On the other side, a rectangle painted with different colors indicates class overlapping for that particular pattern. Notice that color differentiation is not restricted to only Normal or Malicious. If the user wants to use the color scheme for discriminating between, for instance, DNS traffic DGA and Normal, she can do it easily using regular expressions.

Detailed View

It is possible to access to detailed information about a particular n-gram by clicking the corresponding rectangle. In the bottom of the screen, a Frequency histogram for the selected n-gram is shown. The histogram you will find information about the different behavioral models where that n-gram pattern was observed.


Behavioral Models Explorer (CTU-13)

Select only n-grams observerd at least x times: 1

Select only n-grams observed in at least y different behavioral models.: 1

N-gram: .u.u. Total freq.: 1931

Select only n-grams of size n: 5

Label Filtering using regex (e.g. TCP): [a-zA-Z0-9]

Color Filtering using regex (skyblue): Normal

Color Filtering using regexp (orange): Botnet

Fig 2. A detailed view of the STF-PATTERN-VIZ interface. In this case, the first 300 n-grams with n = 5 are shown.

Filtering n-grams

The application support some basic pattern filtering.

- Filter n-grams of size n
- Filter n-grams observed at least x times
- Filter n-grams observed at least in y different behavioral models.

By default the application will consider all the behavioral models present in the dataset. However, it is possible to filter by using the Label information. Additionally, it is possible to filter by Label. Since the standard STF Label contains information regarding the different protocol layers, it then possible to filter by protocol layer 4 or 5. Such filtering is done by using standard regular expressions.

Live Demo

A live demo is available at (https://harpomaxx.shinyapps.io/stf-pattern-viz/). The demo contains a subset of the CTU-13 Dataset. Some minor modifications were made in the label description for facilitating the aggregation. A detail about the labeled modification and how the CTU-13 subset was generated is found at http://rpubs.com/harpomaxx/ctu13bis

Conclusions and Future Work

Even thought Stf-pattern-viz was developed as a tool for the Stratosphere model behaviors, it seems clear it can easily be adapted to a pattern analysis of any other n-gram oriented problem, such as text-analysis or DNA sequences, among others. Given the aforementioned extensible capacity to sequence analysis and the visual power of this solution, in the near future we plan to create an R package that will facilitate the application of the tool to new n-gram oriented problems.

Source code available at (https://github.com/harpomaxx/stf-pattern-viz)