TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

institute of telecommunications

Exercise 4: Statistical Analysis of Raw Flow Records

Network Security 1 - Summer Semester 2019 (VU 389.159)

Communication Networks Group at the Institute of Telecommunications
T. Zseby, F. Iglesias, F. Meghdouri, M. Bachl, E. Zeraliu

In this section we are going to have a closer look on *Flow Record* vectors and see how some attack sources and destinations can be identified by the imprint that they leave in *Flow Record* fields. This time we will use RapidMiner for the analysis. However, you should remember that for the same goa¡l you can use any tool that suits you (Orange, Matlab, Python, R etc). RapidMiner provides some interesting functionalities for quickly importing data and visualize statistics. In the following exercises we will use only a small set of basic functions. In the Netsec laboratory, the installed version of RapidMiner is the open RapidMiner Community Edition v5.3.

4.1 Importing Data in RapidMiner

Open the RapidMiner Community Edition software. For linux users, RapidMiner requires the *Oracle* Java Runtime environment (JRE) version 7[1]. Listing 1 shows how to call the RapidMiner GUI allocating 10GB of RAM. Execute the command in the terminal from the `/rapidminder/rapidminer/lib` folder.

Listing 1: Launching RapidMiner GUI

```
1 java -jar -Xmx10000m rapidminer.jar
```

After clicking on "New Process" in the presentation panel, RapidMiner workspace is displayed. The workspace is split into 6 easily identifiable areas (Figure 1)[2].

In order to start analyzing network traffic data, we first need to import it. To do that, assuming that you have collected data in CSV format, in the "Repositories" area, unfold the options for the second menu icon ("Import data into an existing repository") and select "Import CSV File...". A wizard panel will appear.

step 1: Look for the desired CSV file, select it and click on "Next".

step 2: Select 'Comma ","' as "Column Separator". Click on "Next".

step 3: Click on "Next".

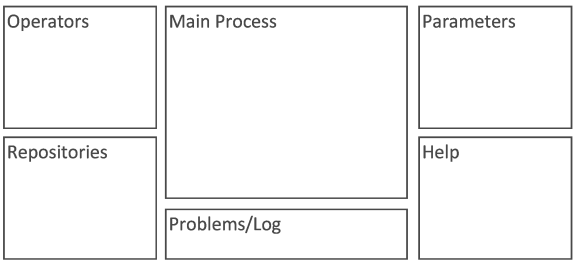

Figure 1: RapidMiner 5 workspace scheme

step 4: Write the correct *types* for the feature, e.g. "nominal" for "sIP", "dIP", "sPort", or "dPort" would be "nominal", whereas "ttl" or "duration" would be "numerical". All features are, in principle, "attributes", unless you have "labels", "ID"-features, or features that you want to discard for the analysis. Click on "Next".

step 5: Save your data into the `data` folder with a proper name. Click on "Finish".

Now a subset of your *Flow Record* file is ready to be analyzed by RapidMiner tools.

[rep-46] Retrieve the data stored in exercise 2 in the RapidMiner repository. The `Ex2flows_team_XX.csv` must be imported as `Ex2flows_team_XX`. Write in the report the "types" used for each AGM feature.

4.2 Metadata and Univariate Analysis

By double-clicking on the new "flowrecord" repository (stored into the "Repositories" area) you can move to the "Results" view and look into the loaded data. Please note that the second and third buttons of the main icon menu bar allow you to switch between the "Edition" and "Results" views (Figure 2).

In the "ExampleSet" tab you can see a scrollable table with all your data vectors. If you click on the "Meta Data View" radius selector you will get access to a new view where a basic statistical analysis of your dataset can be performed. The statistical analysis is activated by clicking on the little *calculator* icon that you will find just over the "Missings" column

[1] The *OpenJDK* Java Runtime Environment does not work properly for RapidMiner.

[2] In order to familiarize with the RapidMiner environment we recommend to have a look on `http://rapidminer.com/learning/getting-started/`

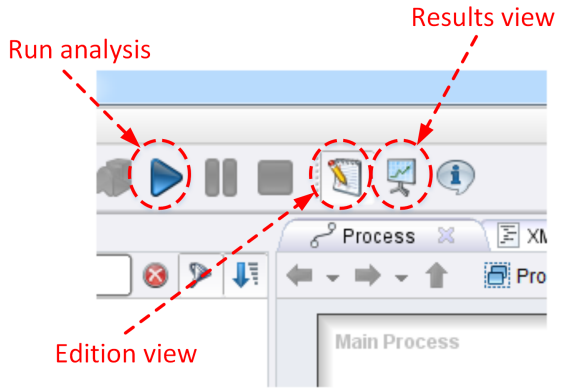

Figure 2: RapidMiner 5 main icon buttons.

of the samples table (on the right side of the screen). RapidMiner performs statistics and range calculations for every independent feature. Calculated values depend on the feature type, i.e. mean and standard deviations for *numerical* features, mode and value frequencies for *nominal* features.

You can also get frequency distributions for the features if you select "Plot View" in the drop down menu and look for the "Histogram" plotter. There you can study the distribution of every feature in your specific dataset. If the number of bins is fewer than the amount of different nominal values (or categories), nominal values will be shared out into the existing bins (note that the maximum number of bins available is 100). This limitation can be **misleading** for some features with many different nominal values (e.g. IP and Port sources and destinations); in such cases we recommend to directly resort to the metadata analysis results or use alternative operators or plots.

[rep-47] What are the top three most used values for the "TCP" flags and in which percentage? Does it make sense for you?
[rep-48] Plot the histogram for the "mode TTL" feature and add it to the report. Explain the shapes that you see.

4.3 Bivariate Analysis

Metadata, univariate and bivariate analysis are techniques that in most cases perfectly suffice to obtain a suitable knowledge and understanding of datasets under study. In RapidMiner you can run a bivariate analysis by generating a scatter plot ("Plot View>Scatter"). In the scatter plot, features for the x-axis and the y-axis must be specified; additionally, a third feature can also be added by using colors. This allows us to get a proper interpretation of feature dependencies, correlations and interactions in the dataset. Furthermore, a very useful function provided by RapidMiner is the "Jitter" selector. By increasing jitter you can separate points which are actually on top of one another, obtaining a much better impression about the concentration of samples around specific coordinates. Finally, to visualize specific values in the scatter plot axis you will probably have to zoom in the area of interest (mouse, left-click).[3]

[3]RapidMiner plotters can be difficult to handle sometimes. We recommend to perform "Auto Range>Both Axes" (mouse, right-click) whenever you modify values or switch to another plotter, as well as to check outcomes with metadata results.

Visualize a scatter plot with the following configuration: (x-Axis: destination IP address), (y-Axis: total count of packets), (color: mode of protocol).
[rep-49] Add the plot in your report.
[rep-50] How many anomalies/outliers do you see? Do you think that outliers/anomalies are always illegitimate/-malicious traffic?
[rep-51] Filter and isolate the flow corresponding to the source that is sending packets to the highest number of IP addresses. Explain how you do it in RapidMiner. Can you figure out which type of traffic it is based on the AGM profile? Is it malicious?
[rep-52] Filter and isolate the flow corresponding to the source that is sending more packets per IP address. Explain how you do it in RapidMiner. Can you figure out which type of traffic it is based on the AGM profile? Is it malicious?

4.4 Did we leave traces?

In exercise 1 you performed three types of malicious activities in the company's network: two types of scanning and the brute force attack itself on a webserver login page. In this exercise we will take the role of a security engineer and see if your activity was detectable using the same methodologies discussed before.

You should have the Wireshark captures from the brute force part available. You will here use your `hydra_team_XX.pcap` and `nmap2_team_XX.pcap` captures in addition to `hscan_team_XX.pcap` which is available in your working directory. `nmap1_team_XX.pcap` is not used here because, depending on the nmap scan that you performed, the output may not contain IP traffic and uses host discovery which is based on the "ARP" link layer protocol.

- Extract the flow CSVs from all three pcaps using Go-Flows and the configuration file `pcap2flows.json`. Make sure that you keep the same time window in the configuration file (10 seconds).
  [rep-53] Explain in the report the values of the AGM profile for each row (i.e., instance) in the `hscan_team_XX.csv`.
  [rep-54] Explain again in the report the values of the AGM profile for each row (i.e., instance) in the `nmap2_team_XX.csv`.
- Load the `hydra_team_XX.csv` CSV file into Rapidminer and make sure to use the correct settings (be careful with the features types).
- **[rep-55]** Join the imported dataset with the `Ex2flows_team_XX` data imported before by using the "Append" block in Rapidminer. Before that, create a new feature called "capture" with the "Generate

Attributes" block. This "capture" feature must be 0 for all instances from the MAWI dataset and 1 for all instances from the hydra dataset. Plot the joined dataset in a scatter plot with the following keys, (x-Axis: distinct source transport port), (y-Axis: total count of packets), (color: capture).

[rep-56] Add in the report a capture of the Rapid-Miner final design.

[rep-57] Add the plot to your report. Are hydra instances outliers? Why? What does it mean?