Assignment # 3: Rich Club

October 13, 2020

Goals:

- To be able to process real network data using its netflow dump.
- To learn to load netflow data.
- To learn to decompose netflow data into service subnetworks.
- To learn to detect (potential) rich club.

Assignment:

1. Prepare input data:
   (a) Decompress given netflow dump into a folder of your choice;
   (b) Identify what netflow protocol (its number) and what endian (order of bytes) have been used;
   (c) Transform input data into a format suitable for use with your libraries (such as `networkx`).
2. Load data into memory and compute basic characteristics of the newtwork:
   - total amount of netflow entries;
   - total amount of packets transferred;
   - total amount of source IPs;
   - total amount of destination IPs;
   - distribution of the protocol usage for the transport layer.
3. Select only netflow related to TCP and UDP protocols.

4. Create a visualization of the network between IP addresses.

5. Compute a destination-port distribution in its whole scope (0-65535) and in the range of system ports (0-1023).

6. For every system destination port with the amount of netflow entries higher than 200 and the amount of destination IPs higher than 10:

   (a) Create a network subgraph between IP adresses that were contacted on the given port; visualize it;

   (b) Calculate its node-degree distribution;

   (c) Calculate a rich-club coefficient for every degree and visualize this dependency;

   (d) For the value that generates the highest rich-club coefficient with the highest amount of IPs, generate subgraph and if its order is at least 5 visualize it;

   (e) Discuss roles of IP addresses of such a club.

Advice:

1. The file `2016-05-26.tgz` contains binary NetFlow data. You can convert them into `.csv` files on Linux using `flow-tools`. They can be installed on Ubuntu 18.04 LTS by

   ```
   sudo apt-get install flow-tools
   ```

   They are not available on Ubuntu 20.04 LTS and newer Linux distros because they require Python 2 to work. If you have them, use

   ```
   export-flow -f2 < INPUTNAME > OUTPUTNAME.csv
   ```

   Otherwise, I created a new archive `exported-flows.zip`, you can use it and just skip first step of the assignment.

2. The amount of transported packets is in the column `dpkts` (it's the 4th one - indexing from 0).

3. The source address in in the column `srcaddr` (10th).

4. The destination address is in the column `dstaddr` (11th).

5. It is a netflow with TCP protocol if its 17th column is equal to 6.

6. It is a netflow with UDP protocol if its 17th column is equal to 17.