Draft 2.0

From: Daniel Egger
To: Data Science Class
RE: Cisco Data-Analysis Project
Date: November 1, 2017

Cisco has created a test network (the "Testbed") to represents a small section of the network architecture used by large companies to facilitate communication between their users.

The network architecture includes nodes called "leaves" [using NCS 5011 machines] and nodes called "spines" [using both NCS 5011 and NCS 5508 machines].

Leaves are generally found at the edges of a network. Data from outside the network enters at a leaf, is routed through a spine, and is then directed to the leaf nearest to the data's ultimate destination.

A diagram of the network structure, "Testbed Topology.pdf" is provided to you and should be studied.

The Testbed has 8 leaves (labeled leaf1 through leaf8 in the diagram) and 4 spines (labeled spine1 through spine4 in the diagram) for a total of 12 machines for which telemetry is collected and failure can be simulated. The network also contains two machines of type NCS 5508 that are a different type of spine, labeled DR01 and DR02. Although some telemetry data from these machines is available, no failures of these two machines occur during the tests.

A machine "failure" in the Testbed is simulated by a machine deleting the list of nearby nodeswhich allows a machine to send on data. Deletion is done with a "Border Gateway Protocol Clear" or "bgpclear" command to that machine. When that command is executed, data sent to that machine cannot be sent on until that machine has re-identified its neighbors and repopulated the bgp file. In the meantime, network traffic will be re-routed through other, nearby machines. A failure in one machine is therefore marked by a sharp but temporary reduction in data traffic through it, and a comparable increase in data traffic elsewhere in the network.

Each of the seven files you are receiving contain time series data for an interval of time when all 12 machines are providing telemetry simultaneously.. The third column contains a numerical time stamp. Note that the files, as received, are not in chronological order. They can be ordered chronologically for each machine separately; this is recommended. [Note also that because of latency issues in receiving the data, the series of time stamps for the different machines are not identical].

The seven (7) files record five (5) distinct circumstances:

(i) Normal operation of the network, no fault (two examples),
(ii) One leaf fails (two examples),
(iii) Two leaves fail in sequence,
(iv) One spine fails, and
(v) Three spines fail in sequence.

Note that many (most!) of the columns contain data that is not at all relevant or helpful when locating a potential failure. We recommend generating histograms of all the data of a particular type to determine whether it has sufficient range of values, in at least some of the files, that it may be associated with a machine failure or failures.

Step one in your data-analysis task is to learn the data sufficiently well that you can *classify* each of the seven files correctly and support your classification with a reasoned argument from evidence.

Assignment 1

For each of the seven files, you should first, be able to identify whether or not it contains a failure, or represents normal network operation. In other words, your first step should be to do binary classification and try to identify the two "negatives" and five "positives" in the data set.

Second, within the group of five (5) files with one or more failures, you should be able to specify *which* of the 12 *machines* received bgpclear commands, and in what sequence.

Write a short explanation of your classification. It should include the following:

(1) what specific evidence in the data led you to classify the file as you did?

(2) What other machines are most influenced by the failure (or failures), and how? Describe what you "see" in the data.

(3) How long did it take the network to return to "normal" functioning after the initial failure?

(4) An illustration or graphic (which can be done in Excel or Tableau) to accompany each of the five files, and illustrate the pattern you used to make your classification, would be helpful and is highly recommended.

We could tell you much more about networks and their protocols, but it is not necessary to do a good job. This project is all about exploratory learning from the data itself.

Future Goals

Our initial goal is to be able to describe distinct visual or other patterns in the data that characterize each type of file, upon manual inspection.

In future, we would seek to automate the classification process using machine learning, so that a new time series of telemetry data could be classified in a completely automated manner as containing no event, or one or more of the above failure events. Full automation will of course require significantly more examples of each type of fault to use as training and test sets.

Eventually it may even be possible to include subtler faults in the system that reflect potential future failure, and forecast failures before they occur.

We will discuss with Cisco appropriate next steps for the project.