*Draft*

From: Daniel Egger

To: Data Science Class

RE: Cisco Data-Analysis Project

Date: October 25, 2017

Cisco has created a test network (the “Testbed”) to represents a small section of the network architecture used by Facebook.

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 and 4 spines of machine type NCS 5011, for a total of 12 machines for which 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. 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 nodes which allows a machine to send on data. Deletion is done with a “Border Gateway Protocol Clear” or “bgpclear” command to that machine. Once BGP is cleared, 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 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.

The seven files record five 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

(v) Three spines fail in sequence

Note that many (most!) of the columns contain data that is not relevant or helpful when locating a potential failure. We recommend generating histograms of all the data in a column 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 with some confidence.

Assignment 1

For each of the seven files, you should be able to identify that it contains a failure, or represents normal network operation.

For the five files with one or more failures, you should be able to specify *which* machine or machines received bgpclear commands, and in what sequence.

You should be able to identify what evidence in the data led you to classify the file as you did.

What other machines are most influenced by the failure, and how? Describe what you “see” in the data.

Explain your reasoning in a short paragraph. An illustration or graphic to accompany each file 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. Good luck!

-Daniel