Network Traffic Group Final Proposal

Emily Andrulis, Jeffrey Klow, Paul German, and Ben Oakley

As Ben showed us in his presentation, the data was easily available to us through contacting Tim Weber and Tim Messick and having them pull the log files for us. Interpreting the log files is relatively simple: they start with a header line, and from there each line gives the timestamp, average incoming transfer rate (in bytes per second), average outgoing transfer rate, maximum incoming transfer rate, and maximum outgoing transfer rate for the current interval. Reading this into R is easy because we can separate by spaces, and the only adjustment we need to make is to delete the first header row that is unique from the rest of the data. Once the data set is in R and the variables are labelled, we need to convert the time from the UNIX timestamp to a usable date and time that we can better use to plot and view the data. Additionally, we decided to create another variable that measures the percentage of total bandwidth that is being used on average for each interval. We made this using the average incoming transfer rate divided by the total possible bandwidth, which we found through questioning Tim Weber.

In this past week, we followed these steps to get the data into R and cleaned up so that we have something to work with for making visualizations and examining what the data can tell us about network traffic at Cornell. However, in discussing what we wanted to accomplish with looking at the data, we realized that due to the decreasing resolution of the data we would want to collect new data sets each day and merge them so that we would have the best resolution possible for each day. With this in mind, we have asked Tim Weber to send us new log files at approximately the same time each day for the next week. As the new log files come in, we will work to merge them and update our visualizations. We anticipate this might give us some issues we can work out the first few days, but once we figure out a system it should be easy to incorporate new data daily. Also, we believe that the maximum bandwidth is in megabytes, but the data seems to show that it might be measured in

megabits. To clear this up, we will be contacting the Tims again to make sure our information is correct, and to see if we need to change the percentage of bandwidth variables in any way.

Our main focus with this data will be to investigate what the internet usage looks like for the average week at Cornell. With this, we will want to delve into figuring out when Cornell has the most internet traffic, outgoing and incoming, as well as when we are least busy. In this, we would want to highlight any spots where we use our maximum bandwidth, and any times when the network is particularly quiet. Basically, we are looking at how much of our maximum data cap we are using at any time, and with this we can extrapolate our results to talk about if we are paying too much for our current bandwidth, or if we might possibly need more. To work with this data, examining the percentage of total bandwidth used at different times throughout different days will be critical because then we can compare from there at what times we are using an above average percentage of our bandwidth or below average percentage.

In presenting our data, we envision a presentation of multiple graphs that tell about network traffic over different time spans. Our main graph would be a simple line graph that would show the percentage of bandwidth usage for every time interval over a single day. We also want to present a graph that adds onto the previous graph lines showing the percentage of usage for different days over the same time periods to better highlight what the trends are throughout the week. At the same time, we also want to simultaneously show the average incoming and outgoing data for both the primary and secondary internet providers. Along with that, we can also present different boxplots revealing the average amounts of usage throughout the day for the different days of the weekday, and compare them side by side. We also want to do similar graphs for the months of the year instead of just days to see what the average year looks like. Animating a graph that plots the average usage of incoming and outgoing data for the different providers is another way we hope to make our data more palatable. Our

premier visualization that we hope to accomplish would be an interactive graph that allows the user to input a particular time of day, and possibly day of the week, and the output would be a pipe-like figure that fills up the amount of incoming/outgoing data usage that is typical of that time. To create this interactive visualization we would utilize Shiny and try to make it as intuitive and informative to the viewer as possible.

After Ben's presentation, we were all very excited to work on the project, but we were a bit dismayed by the unfortunate timing of Ben's absence from class. However, we still came together and came up with a game plan to tackle how we would go about getting this data, getting it into R and cleaning it. Emily was sent as our representative to talk to the Tims throughout the week, and was tasked with getting the data from them, inquiring about the total bandwidth, and requesting additional data logs from them each day. Paul and Jeff worked together from there creating a script to put the data into R and clean it up a bit from there, so we could actually use it and work with it a bit more. Afterwards, they also drafted up a first attempt at plotting some graphs showing the average data usage against time over a one day period and over a longer time period, so that we could use them in our presentation. While we all collaborated on how to deliver our fifteen minute presentation, Emily was assigned the role of scribe, who would take note of our project plans and write out our proposal and put together our presentation materials.

In the coming week, most of our work will be revolving around inspecting the data we have and creating our visualizations to present what we find. As mentioned previously, we plan to have three main types of graphs: static graphs, animated graphs, and interactive graphs. Most of our visualizations will be static graphs depicting the network traffic over different days or months, but we feel that adding other types of graphs will make our story more compelling and easier for readers to get more out of our information. We found that the easiest way to split up the workload would be to allow each person to

have one specific type of graph that they focused on making, i. e. Emily does static graphs, Jeff has animated graphs, and Paul is responsible for the interactive graph. For some, like the static graphs, there might be multiple graphs that are needed, but they should be easier to generate in R. Other graph types might have only one final product, but the work is expected to be much more challenging since the graph type is more advanced. Overall, each workload should be about equal, but if one does finish sooner they can help whoever is struggling most. In addition, we will still need someone to make sure we are easily merging the newly acquired data each day as the new data sets come in, and Jeff has volunteered to take on that responsibility.