Jeffrey Klow, Paul German, and Emily Andrulis

Network Traffic Initial Findings

Last week we presented on how we first got our data from the Tims in the IT department. This past week we have been continuously getting new log files each day, which we have had to merge with our previous data frames to create a single data frame for each type of log file (Internal or primary/secondary network providers). When we originally put our first log files in we used our createLogDF.R script, but this week we continuously used the new updateLogDF.R script to put the new log data into our already established data frames. This new script goes through the newest log file and uses a helper function called getRelevantLines to pick out only the data that is newer than the most recent time from the old data frame, and then adds the new lines of data onto the front of the already existing data frames. Using this script has made merging the data into a fairly simple process, and for the most part our data acquisition has been one of the easier portions of our project.

In cleaning the data, we decided to create a few more variables from our four given ones so that it would be easier to work with the data. In our prior presentation we talked about how we added fields that showed the percentage of the total bandwidth that the average download or upload usage was using for any time interval. Along with that we created various fields to clean up and work with the UNIX timestamp data in different ways. This includes fields for central standard time (CST time), the date, the time, the time in decimal format (2:30 would be 2.5) and the day of the week. This is how we went from our original data frames that

contained five variables each to our current data frames with twelve each. With cleaning the data we had to consult the Tims again to make sure we had the correct amounts for total possible bandwidth, but other than that we did not run into many issues. The reason we created different time variables was because it made it easier later in trying to pick out different pieces of data and graph things more neatly.

Splitting up the work as a team, we decided last week that Emily would focus on static graphs, Jeff on animated graphs, and Paul would tackle Shiny and our interactive graphs. So far, we have stuck to this work distribution quite well, with the exception that Jeff has also taken on many extra responsibilities in making helper functions for our code and creating variables in the data set when updating the data so that it is easier for us to graph and pull out certain pieces of information.

Our main static graph this week is a purposefully busy graph that shows the upload and download usage (both maximums and averages) over a 24 hour time period for each from one week. With this we hope to create a plot showing what the "average" day looks like for upload and download usage at Cornell. We also have plans to try and make this incredibly cluttered graph into an interactive graph where the user would be able to show only certain variables or days at a time, or possibly hover over lines and learn more about what each line represents and what it looks like. Good portions of Emily's time this week were spent looking into making the data into a time series object, but after a lot of research on this topic she realized that it would be easier to just keep it as the UNIX timestamp and change the axis labels. The reason we did not go with time series is because they mainly focus on

plotting data over the course of years, and finding functionality to bring this down to hourly or minutely data was quite a hassle.

Another static graph we came up with this week was one that showed our main four network traffic variables for each of the three types of log files we had. We wanted to use this to show what type of data we were getting from each log file, and how they might be useful in different ways to see what was going on around campus as far as network traffic goes. This graph makes it very clear that the secondary data shows incredibly lower amounts of data usage than the primary internet provider data, and therefore a network analysis would not be possible if one simply looked at the secondary provider's network traffic data. Also, we found that the primary log data followed the internal data very closely, with the exception of the 9 AM and midnight spikes that we voiced concern about last week. After speaking with the Tims, they explained to us that these were indeed a consequence of a new backup system, as we had previously suspected. The slight inconsistencies in the times of the spikes are due to the fact that these are new systems and they have been making slight changes to try and test out different aspects. Due to these backups, we usually can use the primary internet provider's data to see the most accurate representation of Cornell's network traffic at anytime.

With animated graphs, we thought might be helpful to see what the usage looked like for a certain day of the week over a few months. To do this, Jeff worked on an animated graph that looped through all the Mondays, for example, that we have data for and each individual graph shows our main four variables (average and maximum upload and download usage) over the 24 hour period of that Monday. He

has done this for each day of the week, and for each of our three network data frames.

After creating animated graphs for each day of the week across our whole data set, Jeff moved on to create an animated plot that went through every day from the very beginning of our log data to our most recent data. Since this graph goes so far back, he also decided to take 30 minute time intervals towards the end of the data where he could have instead taken 5 minute time intervals. He chose to do this because otherwise the sudden increase in resolution created a jarring effect and seemed to only further confuse the viewer and hinder comprehension of what the graph was trying to convey. From looking at this graph and the graphs of the specific weekdays the main thing we can take away is that the data usage seems to stay roughly the same for each day in each week. Of course, there are some drastic changes that can easily be explained by winter break, or times when students are off campus, but for the most part our data usage seems to be pretty consistent.

We also thought it would be interesting to see what the different weeks in a certain block might look like for network traffic. The only block we had full data for was fourth block, which was unfortunate because it had the unusual schedule. However, we still thought it might be cool to see what the data usage looked like over this period, so Jeff made an animated graph that shows week by week what our data usage looked like. From analyzing this we can clearly see where Thanksgiving break is as our usage makes a definite overall decrease, and for winter break the data usage almost stops completely after Friday. It did surprise us though that the drop in network traffic over Thanksgiving break was not as sharp as winter break,

and it did still have peaks above 20 MB per second for maximum download speeds. This is surprising since that is about half the amount for usual peaks when all the students are on campus, but more than half the students leave campus for Thanksgiving break. This leads us to believe that the students who stay on campus over Thanksgiving break are using the internet a disproportionately high amount, which makes sense to us since we are not sure what else one would do with free time on a mostly empty college campus.

Paul has spent most of the week learning what Shiny has to offer, and how we could use it to create our interactive graphs. First he tackled the problem of making a graph where the user was allowed to set the time frame that was being graphed. This means that users could choose to just show network traffic between 9-11 AM or 3-8PM or whatever time frame they were interested in seeing. After receiving the main static busy graph he also worked on trying to make this interactive so users could choose which variables they saw. After playing around with the different widgets that Shiny has to offer, we settled on having him implement checkboxes for the days of the week so that a user could choose which combination of days they wanted to view at one time. He also included a drop down box to choose which of our three data sets the user wanted to look at. From here, we still want to also modify it so that the user can choose which of the four variables (upload/download maximums and averages) they want to see at one time. We feel that this graph is going to be really helpful in analyzing our data usage because it allows us to compare many different variables in many different ways with one simple interface. For example, if we suspect that download maximums might be

similar for Tuesdays and Thursdays, we can choose to plot the two on that same graph and by isolating these variables it will be easy to tell how alike they really are. For our final analysis we plan to do some of this work ourselves to see if there are any days that seem relatively different or any times that have particularly similar data usage.

Beyond making these visualizations, are main work up to Tuesday will be to further analyze the graphs we have now made. To help with analysis we may also like to add average curves or loess curves to some of our graphs to make it easier to illustrate when the data is straying drastically from the usual pattern. We hope that in our further analysis we will be able to shed light on any particularly interesting irregularities, as well as show what the average usage really looks like for network traffic at Cornell.