Paul German, Jeffrey Klow, and Emily Andrulis

Network Traffic Final Presentation

Every Cornell student knows what it feels like to have to wait on the slow internet speeds that they run into during the peak network traffic hours, but when exactly are those hours? At what times are students typically doing the most downloading and uploading on the Cornell network? When is the network least busy? These questions among others led us to look into getting data from Network Services about what the network traffic really looks like on Cornell's campus. Once we realized we could actually get some data to look at, we refined our focus until we were left with two main questions we wanted to explore: "What does network traffic at Cornell look like on an average day?" and "Assuming the pattern holds, at what point should we consider getting more bandwidth because we will be frequently coming close to our maximum allotted?"

When we went to Network Services the Tims (Tim Messick and Tim Weber) were gracious enough to give us the information they could about the Cornell's network usage. We were given three different types of log files: one for each the primary internet provider, the secondary internet provider, and the internal network traffic. Each log file contained data with a given UNIX timestamp, averages of the upload and download usage since the last timestamp, and the maximum upload and download usage over that same time interval. Problems arose though when we realized that not all the time intervals were the same, and as one went farther back in the data the time intervals got increasingly larger, going from 5

minute intervals for the newest data all the way to day-long intervals towards the end of the oldest data. We really wanted to be able to analyze one week's worth of data at least with the best resolution, meaning smallest time intervals, so to do this we asked the Tims to send us new log files every day for a week at approximately the same time.

Once we had this data, we started immediately putting it into R and trying to analyze it. To do this we created a script called createLogDF.R, which initially takes some log files and creates data frames for them in R. From there we could use our updateLogDF.R script to update our data frames daily as we got new data coming in from the Tims. In cleaning the data our main two tasks were to make the UNIX time field more usable and to create variables for each line showing how much of the bandwidth we were using for either upload or download. In making graphs and analyzing the data we used many different time intervals to examine what was going on with the network traffic, so from the UNIX time stamp we also created day, time, day of the week, decimal time, and Central Standard Time fields to help aid us in these efforts. Creating a percentage of bandwidth used variable was a simple task once we inquired to the Tims about how much our maximum bandwidth was for each provider, and Jeff also put these new variables into our create and update R scripts so they would be made and continuously updated from the start.

After cleaning our data, we were able to start looking at what information we could really glean from our data set. Examining this amount of data can be overwhelming though, and to keep us on track we referred back to our main questions, allowing ideas about the average day at Cornell and the future of Cornell's

bandwidth usage to lead us in making our visualizations. We wanted to create an array of different types of graphs to better tell the story that we were getting from the data. We decided to make static graphs to show main ideas about the average usage in a day at Cornell, the differences between the three types of log files in a typical day, and the future projections for how much of our bandwidth we will be using on a daily basis. We found that animated graphs worked well to show how different days compared to the average day's usage. For example, we have animations that cycle through every single Friday in our data and compare that to the average curve that we created using data from each day for the past week. Using animated graphs also allowed us to get a glimpse at what an average block's usage would look like. However, it should be noted that we could only analyze block 4's data, which most would agree is not an average block due to its weird schedule. Furthermore, we went a step further and created interactive graphs of all our network traffic data in high resolution (5 minute intervals) over the past week, and of all the network traffic data for any log file that you give it.

To analyze the data, we really had to scrutinize our graphs to see if we could decipher any weird trends that were going on that went against the typical data usage. What we found was that although certain days may have a few abnormalities, such as a spike on Friday around 9 AM that could perhaps be due to students downloading something for class around that time, these abnormalities do not seem to have any distinct patterns and as far as we can tell it seems that they would each be due to some individual reason based on that specific circumstance. This means that while we can expect that there will usually be abnormalities in the data with

peaks and dips in odd places, for the most part we can expect that the data will roughly follow our average day curve that we created with a Loess curve for the past week's data.

This average curve is characterized most notably by its slow dip after midnight until it's low point around 6 AM, at which point it gradually increases until 11 AM where it starts to sharply increase and continues to until it peaks around 4 PM. After this there is a noticeable dip during dinner time, centered on 6 PM, after which it starts to increase steadily again until it hits peak usage again around midnight. This pattern seems reasonable for a college campus because we would expect most students, even with weird sleep schedules, are sleeping around 6 AM, and then as more and more wake up the internet usage steadily increases. It is understandable that students suddenly start to use the Web again once class ends at 11, and they continue over lunch and especially once most people have finished eating after noon. The drop in usage over dinner is to be expected, since the dining hall seems to be the one place at night where it is rare to find students working on laptops and browsing the internet. Increasing usage until midnight is also within the realm of reason since this is a college campus and most students prefer to stay up later and tend to do their best procrastinating by surfing the Web later at night.

We had initially hypothesized that the weekend's usage might look dramatically different than the data from during the week, but even when comparing the two intervals side by side we see that the weekend usage still seems to follow the average usage curve just as well as the data from the middle of the week.

Looking at the future usage of our bandwidth for downloads at Cornell, we noticed that by taking an average download usage for each day in our data set we could plot a scatterplot and use linear regression to examine where we really are headed. At first, our data seemed a bit skewed due to the fact that our bandwidth cap had been changed this past September. With the increase in bandwidth, our maximum downloads had immediately risen, and our averages rose a bit, too. Taking this into account, we created another graph only using the data after the increase in bandwidth, but even from this we see that the average download usage per day is slowly but surely increasing. With this linear regression model, assuming the pattern holds, we would predict that we would regularly be reaching 85% of our bandwidth as the average download usage at worst by September of 2016, and at best by January of 2017. This is extremely helpful information because it gives the Network Services workers hard evidence that shows how our bandwidth usage is increasing and that it would definitely be wise to start considering buying more bandwidth for the campus's network.

As we noted, the applications for our work are especially important for the workers in Network Services because our data can give them better insight into how the network is actually being used and at what times, and it can clue them in to looking at when they might need to upgrade for more bandwidth. We plan on giving all our documented code to the Tims so that they can use it later for more analysis later on. On top of that, we have created an interactive graph using Shiny that allows them to simply point and click to download a log file and then they will

immediately be able to look at all the data plotted over any time interval that they care to look at.

Our big takeaways from our work are that the average network traffic at Cornell follows the same relative pattern for any day, and that the average download usage is slowly increasing as the days go on. This means that eventually we will end up hitting our cap, and to anticipate this we may want to purchase more bandwidth within the next two years. Also, if you're looking for faster network speeds and are trying to guess when the network is least busy, we recommend getting up at 6 AM because no matter what the day, no Cornell students seem to want to be up at this time, even to use the internet.