Lecture 7: You take the high road, and I’ll...

Last time

We looked at one final data structure in Perl, the hash or associative array; we found it was particularly well suited for text applications.

This structure proved to be a cleaner way to handle the Roberts transcripts; we can avoid the creation of a large number of intermediate files and handle things entirely in Perl.

Finally, we looked at a new data source; logs describing wireless accesses on the Dartmouth campus.

Today

We will go over the answers to your homework and talk about how you computed the various summary measures; we will then get a new assignment that relates to the flow of people on campus as measured through their traces.

15 Minutes of Unix: Emacs!

The Dartmouth wireless data

Recall that these data are generated automatically by wireless access points (APs) serving the Dartmouth campus community.

The data are collected in time order and represent activities of thousands of people; we will apply some of our basic Perl and Unix skills to see if we can uncover something about usage patterns.

The project is still very active and you can find information about their aims for the data at the URL below:

http://crawdad.cs.dartmouth.edu/
Your assignment

1. Initially, consider just the data from the first week of March in 2003. How many different APs were used in Baker Hall during this period? (Recall that these would all have a specification of baker#-ap.)

2. Create a time series of the number of unique MAC addresses associating with baker15-ap per day.

3. Determine which AP is the busiest (in terms of the number of unique MACs accessing it) and create a time series of daily unique MAC counts for it.

4. Find the 25 users who access the most unique APs during the time we have in March.

Logging specifics

Recall that to start using an AP, the wireless device has to either Associate or Reassociate with it

That means to focus on usage, we might restrict our attention to just these lines in the log

Before we do that, let’s compute some simple statistics for our file

Usage

But we’re really after usage; does this big spike mean that there were more users on the system that day?

Recall that we are interested in devices that are “Associated” or “Reassociated” with the AP

How do we extract those lines?
Does this give us what we want?

Looking at these lines, a better regular expression might be given by

```
\(Info\): Station \w+ (Rea|A)ssociated$
```

Let's see what our original crude pattern captured in addition to the lines we really want

We can add another pipe and leave out things that we expect to see; this is done with a `-v` flag for `egrep`
Leftovers

Armed with this information, we can try to make sense of some of the other leftovers; the pattern is similar, keep scraping out the things you understand

At some point the numbers of lines involved drop to a level that you start to feel comfortable that you haven’t left much out

% gunzip -c *.gz | egrep '(Rea|A)ssociated' | egrep -v '\(Info\): Station \w+ (Rea|A)ssociated$|Disassociating|Associated to Parent|Disassociation|Deauthenticating' | more

1046645829 Mar  2 17:57:19 ns1.dartmouth.edu berry13-ap (Warning): Station 0006256de9c7 Associated with Encryption, then attempted to send an Unencrypted packet t
1046645833 Mar  2 17:57:13 ns1.dartmouth.edu berry13-ap (Warning): Station 0006256de9c7 Associated with Encryption, then attempted to send an Unencrypted p
1046645840 Mar  2 17:57:20 ns1.dartmouth.edu berry13-ap (Warning): Station 0006256de9c7 Associated with Encryption, then attempted to send an Unencrypted p
1046645848 Mar  2 17:57:28 ns1.dartmouth.edu berry13-ap (Warning): Station 0006256de9c7 Associated with Encryption, then attempted to send an Unencrypted p
1046645856 Mar  2 17:57:36 ns1.dartmouth.edu berry13-ap (Warning): Station 0006256de9c7 Associated with Encryption, then attempted to send an Unencrypted p
1046645889 Mar  2 17:58:09 ns1.dartmouth.edu berry13-ap (Warning): Station 0006256de9c7 Associated with Encryption, then attempted to send an Unencrypted p
1046719045 Mar  3 14:17:25 ns1.dartmouth.edu sachem-center1-wb (Warning): Station 0040968c18da Associated with Encryption, then attempted to send an Unencrypted p
1046719053 Mar  3 14:17:33 ns1.dartmouth.edu sachem-center1-wb (Warning): Station 0040968c18da Associated with Encryption, then attempted to send an Unencrypted p
1046721222 Mar  3 15:08:57 ns1.dartmouth.edu sachem-center1-wb (Warning): Station 0040968c18da Associated with Encryption, then attempted to send an Unencrypted p
1046722132 Mar  3 15:09:15 ns1.dartmouth.edu sachem-center1-wb (Warning): Station 0040968c18da Associated with Encryption, then attempted to send an Unencrypted p
1046722154 Mar  3 15:09:24 ns1.dartmouth.edu sachem-center1-wb (Warning): Station 0040968c18da Associated with Encryption, then attempted to send an Unencrypted p
1046750765 Mar  3 23:06:05 ns1.dartmouth.edu chase2-ap (Info): Deauthentication from 000625497969, reason “Not Associated"

% gunzip -c *.gz | egrep '(Rea|A)ssociated' | egrep -v '\(Info\): Station \w+ (Rea|A)ssociated$|Disassociating|Associated to Parent|Disassociation|Deauthenticating' | wc

3429   78804  674537

Just pulling out “usage” lines

% gunzip -c 200303*_log.gz | egrep '\(Info\): Station \w+ (Rea|A)ssociated$' | cut -d" " -f4 | uniq -c

3429 78804 674537
Some new questions

How many users were there during this period?

What was going on differently on March 1 than the other days in March?

The Dartmouth wireless project has been running for many years now and you have access to it via the computer lab-compute

More data

Rather than pull all the data to your machine, you can grab pieces of it

As with your selection from march, the files are named according to the date they were collected

```
scp lab-compute.stat.ucla.edu:/Data/wireless/data/20021016.log.gz .
```

copy it to a file of the same name (in this case '20021016.log.gz') on your local machine

Your assignment

1. Initially, consider just the data from the first week of March in 2003. How many different APs were used in Baker Hall during this period? (Recall that these would all have a specification of baker#-ap.)

2. Create a time series of the number of unique MAC addresses associating with baker15-ap per day.

3. Determine which AP is the busiest (in terms of the number of unique MACs accessing it) and create a time series of daily unique MAC counts for it.

4. Find the 25 users who access the most unique APs during the time we have in March.
Your assignment

You are going to create matrices that describe the flow of users between the different buildings on campus. Given two buildings, A and B, for each day i, let $n(A,B,i)$ represent the number of times people using an access point in building A move to an access point in building B. In terms of our log file, this means someone has associated (or reassociated) with an access point in A and then associates (or reassociates) with an access point in B. Consider only those people who move from A directly to B; that is, we will not count in $n(A,B,i)$ someone who moved from A to C and then to B. Notice also that $n(B,A,i)$, the number of people using B who move to A, will be a different number. To start, you might consider creating a Perl program that will do this for a single day in March and for two specific buildings. Then, repeat the process for all the buildings carrying traffic in March. Right now, we will count the number of transitions, no matter who made them; a single student might be responsible for all the changes between the buildings. If you are feeling ambitious, you might want to instead tabulate $n'(A,B,i)$, the number of unique people (MAC addresses) that move from A to B.

Baker Library

What commands did we use to identify the number of APs in this building?

... here’s what the oracle has to say