Experiment: Networking Data Analytics

Earlier version prepared by: Dr. Babak Alipour (main), Dr. Roozbeh Ketabi
For: CIS4930/6930 Mobile Networking (Special Topics)
Instructor: Ahmed Helmy
(Due Oct 21, one report per group)

In this experiment, students get hands-on experience in dealing with network measurements wireless LAN, WLAN, and DHCP traces in this case. The traces can be used for analysis of network activity throughout the network (campus in this case), for activity and mobility modeling purposes, or to deploy mobile services to serve the student and faculty population.

Students are provided with data samples of WLAN DHCP logs, in addition to external data (prefixes of buildings around campus, categories of buildings, etc.).

The goal is to use simple data analytics tools and methods to read the data, analyze it, and produce representations of various statistics and distributions of activity and load over time, space, and building categories.

Following is a description of the data files (Input.a and Input.b) then four parts to the experiment to analyze the data:

- Input.a

  • **2 days** of derived-DHCP logs from campus WLAN in CSV format
  - ***outputwireless-logs-20120407.DHCP_ANON.csv***
  - ***outputwireless-logs-20120409.DHCP_ANON.csv***

  • 6 columns are:
  - •userIP: User's assigned IP address
  - •userMAC: MAC address of user device (anonymized to integers, each number uniquely identifies a user **in a single day**)
  - •APNAME: Name of AP the session belongs to
  - •APMAC: MAC address of AP
  - •startTime: Beginning timestamp of session
  - •endTime: Finish timestamp of session

User IP	User MAC	AP Name	AP MAC	Start Time	End Time
10.131.24.201,1,tur2336-win-lap3502-2,58:35:d9:9a:09:e0,1333757183,1333757190					

Fig 1. Example entry for the first input type, the WLAN/DHCP trace

```
userIP,userMAC,APNAME,APMAC,startTime,endTime
10.131.24.201,1,tur2336-win-lap3502-2,58:35:d9:9a:09:e0,1333757183,1333757190
10.128.173.51,2,unknown,00:00:00:00:00:00,1333757191,1333757191
10.128.173.51,2,brt202a-win-lap3502-1,10:8c:cf:44:80:c0,1333757191,1333757191
```

Fig 2. Example 3 lines of the first input file

```
:Experiments4ForClass> more outputwireless-logs-20120407.DHCP_ANON.csv
userIP,userMAC,APNAME,APMAC,startTime,endTime
10.131.24.201,1,tur2336-win-lap3502-2,58:35:d9:9a:09:e0,1333757183,1333757190
10.128.173.51,2,unknown,00:00:00:00:00:00,1333757191,1333757191
10.128.173.51,2,brt202a-win-lap3502-1,10:8c:cf:44:80:c0,1333757191,1333757191
10.128.173.51,2,brt202a-win-lap3502-1,10:8c:cf:44:80:c0,1333757191,1333757191
10.128.173.51,2,brt202a-win-lap3502-1,10:8c:cf:44:80:c0,1333757191,1333757191
10.130.173.243,3,rei105-win-lap1242-1,00:1d:e5:8f:38:f0,1333757186,1333757195
10.131.3.236,4,tur2342-win-lap3502-2,58:35:d9:76:54:80,1333757191,1333757195
10.131.24.201,1,turl0005-win-lap3502-2,44:e4:d9:00:f2:80,1333757190,1333757197
10.131.3.236,4,tur2342-win-lap3502-1,58:35:d9:9a:09:a0,1333757195,1333757198
10.130.123.231,5,rei105-win-lap1242-1,00:1d:e5:8f:38:f0,1333757188,1333757199
10.130.172.253,6,b1164r1walkway-win-lap3502-1,88:f0:77:ad:53:00,1333757188,1333757200
10.128.173.51,2,unknown,00:00:00:00:00:00,1333757191,1333757200
10.130.5.89,7,rei362-win-lap1231-1,00:11:5c:e5:4d:20,1333757189,1333757201
10.128.173.51,2,dau215a-win-lap1252-1,00:17:df:ab:90:b0,1333757200,1333757201
10.128.173.51,2,tur1408-win-lap1142-1,18:ef:63:fd:49:c0,1333757201,1333757203
10.131.192.46,8,fli117-win-lap3502-1,44:e4:d9:01:52:b0,1333757199,1333757205
10.132.182.27,9,socgate3-win-lap1231-1,00:17:59:5a:0f:80,1333757186,1333757205
10.130.164.188,10,b414c199a-win-lap1231-1,00:1a:a2:a1:87:90,1333757198,1333757206
10.132.182.27,9,socgate3-win-lap1231-1,00:17:59:5a:0f:80,1333757205,1333757207
10.130.85.119,11,rei122-win-lap1242-1,00:1d:e5:8f:15:f0,1333757204,1333757207
10.132.182.27,9,socgate3-win-lap1231-1,00:17:59:5a:0f:80,1333757207,1333757208
10.131.3.236,4,tur2346-win-lap3502-1,58:35:d9:9a:17:c0,1333757198,1333757208
10.130.123.231,5,rei300hall-win-lap1242-1,00:1d:e5:8f:30:e0,1333757199,1333757209
10.130.160.17,12,arch120-win-lap3502-1,c0:62:6b:d5:bb:60,1333757208,1333757209
10.131.24.201,1,tur2346-win-lap3502-1,58:35:d9:9a:17:c0,1333757197,1333757209
10.131.24.201,1,turl0005-win-lap3502-5,44:e4:d9:01:62:80,1333757209,1333757210
10.128.173.51,2,rol315-win-lap3502-1,10:8c:cf:45:87:e0,1333757203,1333757210
10.128.109.231,13,swc170-win-lap1252-1,00:22:55:e0:7b:40,1333757207,1333757211
```

Fig 3. Sample from the input WLAN/DHCP trace on April 7, 2012

- Input.b
  - Mapping from AP prefixes to location
    - **prefix_lat_lon_name_category.csv**
  - 5 columns are:
    - Prefix: Prefix of APNAME (see input.a)
    - Latitude
    - Longitude
    - Building name
    - Building Category

Prefix	Latitude	Longitude	Building Name	Building Category
aaf	29.650275	-82.345720	Academic Advising Center	admin

Fig 4. Example entry for the second input type (for building codes and locations)

```
prefix,lat,lon,name,category
aaf,29.650275,-82.345720,Academic Advising Center,admin
adv,29.648792,-82.359303,Martin H. Levin Advocacy Center,academic
aer,29.643272,-82.348300,Mec. Aerospace Engineering,academic
aerv,29.643272,-82.348300,Mec. Aerospace Engineering,academic
alf,29.642948,-82.348636,Alfred A. Ring Tennis Pavilion,sports
```

Fig 5. Example 5 lines of the building codes and locations file

```
:Experiments4ForClass> more prefix_lat_lon_name_category.csv
prefix,lat,lon,name,category
aaf,29.650275,-82.345720,Academic Advising Center,admin
adv,29.648792,-82.359303,Martin H. Levin Advocacy Center,academic
aer,29.643272,-82.348300,Mec. Aerospace Engineering,academic
aerv,29.643272,-82.348300,Mec. Aerospace Engineering,academic
alf,29.642948,-82.348636,Alfred A. Ring Tennis Pavilion,sports
and,29.651573,-82.341939,Anderson Hall,academic
ans,29.631292,-82.351768,Animal Science,academic
ansc,29.631292,-82.351768,Animal Science Building C,unknown
apl,29.646905,-82.343942,Aquatic Products Lab,academic
apopka,28.638361,-81.548856,UF Apopka,academic
arch,29.647656,-82.341748,School of Architecture,academic
aud,29.648887,-82.342909,University Auditorium,academic
bar,29.643915,-82.344403,Bartram Hall,academic
barB,29.643915,-82.344403,Bartram Hall,academic
bcf,29.639509,-82.359952,BioContainment Facility,unknown
bcftr,29.639518,-82.359949,Bio-Containment Facility,academic
ben,29.647961,-82.351393,Tolbert Hall,housing
bgh,29.649216,-82.359055,Bruton-Geer Hall,admin
blk,29.641892,-82.347876,Black Hall,academic
bro,29.646552,-82.342104,Broward Hall,housing
broEastOIR,29.646552,-82.342104,Broward Hall,housing
broOIR,29.646552,-82.342104,Broward Hall,housing
brt,29.648906,-82.345784,Bryant Space Science Center,academic
brtb,29.648906,-82.345784,Bryant Space Science Center,academic
brttr,29.648906,-82.345784,Bryant Space Science Center,academic
bry,29.648906,-82.345784,Bryan Hall,admin
bwc,29.645354,-82.347903,UF bookstore,social
bwcbookstore,29.645256,-82.347710,UF Bookstore,social
```

Fig 6. Sample from the input trace for building classification and locations

- Part I. (25%)
  1. Produce the time series plot of events in the file
     "*outputwireless-logs-20120409.DHCP_ANON"*
     - x-axis: Time (15-minute bins)
     - y-axis: number of events in the bin

     A. What is the number of records in the trace, unique devices, unknowns?
     B. What time of the day is the most active (in terms of number of events)?
     C. What time of the day is the least active*?
     D. Considering the class periods (start/end times), please explain your observations in the time series around beginning or end of classes.
  2. Try different bins (1m, 5m, 10m, 30m, 60m) and discuss how it affects your analyses.

- Part II. (25%)
  1. Repeat the same tasks of Part I above, but on the other file:
     "***outputwireless-logs-20120407.DHCP_ANON"***
  2. In addition, summarize the similarities and differences in your observation?
     - Explain why you think these similarities and differences exist.
     - Include any supporting graphs and plots

- Part III. (25%)
  Using the file "*outputwireless-logs-20120409.DHCP_ANON"*
  A. Can you identify the user devices that have many sessions in early morning hours 6:30-9:30 (the early-birds)? How about during lunch time 11:30-14:30 (the munchers)? Or the evening 17:30-20:30 (the stompers)?
  B. Use *Input.b* to map these devices (and corresponding sessions) to locations. Which buildings are the most popular among 'early-birds'? Or 'munchers? Or 'stompers'?
     1. Sample output of this task: 50% of munchers munch at the hub.
     2. Please include any necessary graphs or plots to support your output.
  C. Does the most popular building of different users change as the day progresses? What are the most popular *building categories* in the morning? How about in the evenings?

- Part IV. (25%)
  From part III.C., let us define '*user density*' as the number of devices/sessions in a building.
  A. Get the distribution of user density across all the buildings in the trace (plot those in order from lower to higher density) for the morning hours
  B. Repeat A above for the lunch and evening hours, and observe the density distribution
  C. Overlay the density values on a map and visualize (or animate) it. You can use 'folium' with python, Google maps/earth, or other tools

- Part V. Encounters/Contacts (optional, bonus up to 15%)
  Encounter/contact pair-wise data: define an 'encounter' between two devices as the event when they log into the same AP at the same time. Also, define the 'encounter duration' as the time span of that encounter.
  A. From the data in Part I (April 9), how many encounters occurred on that day with any duration
  B. Repeat A above for various values of encounter durations to get the distribution of encounter durations
  C. Get the values x and y such that we can get three equi-frequency categories of encounter duration (d) as follows:
     - Short: $d < x$
     - Medium: $x<d<y$
     - Long: $d > y$

- Part VI. Flutes vs Cellos (optional, bonus up to 15%)
  - Call smartphones 'flutes' (can be used on the move), and call laptops 'cellos' (sit-to-use devices). These are the two device types in the trace
  - From encounter point-of-view we now have three types of encounters: flute-flute, cello-cello, flute-cello

  A. Repeat Part V for each encounter type (you will need extra info about device types, existing in an extra file. Ask for that file.)

B. Comment on the *x* and *y* values for each of the encounter types in 'A' above. What can we infer about encounters from this part of the experiment?

- Part VII. Feature-based Device-type Classification (optional, bonus up to 15%)
  - Define a mobility metric/feature for each device based on the number of APs visited in a time window (say per hour)

  A. Use a machine learning (ML) classifier (say support vector machine SVM) to attempt to classify devices into flutes or cellos based on their mobility
  B. How successful is 'mobility' as a lone feature in classifying device types?
  C. What do you suggest to improve the classification?

- Part VIII. (optional, bonus up to 15%)

  Include your own extra part in detail, with concepts/definitions, questions, goal, answers, code, data, etc.

- Files and datasets:

  • The experiment relies on a few files available (zipped) here (UserMAC anonymized): https://uflorida-my.sharepoint.com/:u:/g/personal/helmy_ufl_edu/ET02HLb2Z6NMiWAx2j-fvVkBj1YCq87GNP5_eUuqZNdRtA?e=VEzt1c

  • If you opt for parts VI or VII, you will need extra file for the device types

- Additional submission notes:

  - Note1: please indicate what each member of your group/team did for the experiment. Example, person A wrote the code to parse the input files into python panda data frames. Person B and C worked on the code for analyzing the time series in parts 1 & 2 … so on.

  - Note2: please submit the commented code you used along with your submission, noting what each person/sub-group worked on in the code comments, along with explanation of what you were trying to do in each piece of the code.