Spatial Data Analysis For Cell-based Mobile Network

SUN Yiping 5130309764

June 25, 2016

Nowadays, more and more people spend more and more time on cell phones, which may raise a problem that what exactly their behaviors are and relatively what Cellular bases act. We fortunately have a data set from China Unicom, whose raw data is 60GB. As a part of NFV(Network Function Visualization) Group, Data Analysis is the first work we have done, but our previous work doesn't cover spatial data analysis. So in this project, we use heatmap, distribution and CDF(Cumulative Distribution Function) to find out what's spatial performance behind the data set. The result tells us that users have a fixed pattern for activities, High delay happens on small and medium cells, Top 10% of cells controls 70% of user activities and so on. More results will be reveled in this report.

1. Data Set Introducion

The data set is from China Unicom, which covers user activities from 2016-1-1 to 2016-1-20. There are many labels in this data set, such as User Activities, Response Time, Traffic, Application Count,PDP and so on. In all labels, only two labels have the information about spatial data - User Activities and Response Time. Other labels all have been analyzed in previous work. Here we give the detailed data structure of this two data in this data set. The first one is User Activities.

User-Activities labels	detailed data
Record-structure	MINUTE NUM
Record-example	2016-01-11 04:00 9
Record-structure	MINUTE CELL:NUM,CELL:NUM...
Record-example	2016-01-11 04:02 40611:1,21862:1,20951:1,0:3

The next one is Response Time.

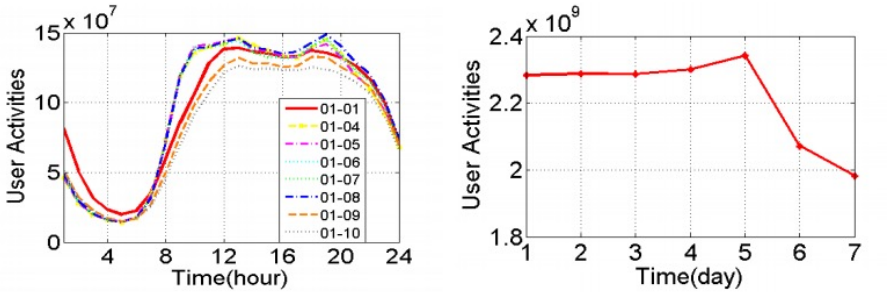


Figure 1: User Activities in day and hour

Response-Time labels	detailed data
Record-structure	MINUTE MIN-MAX-LENTH-SUM
Record-example	2016-01-11 04:00 20-68-9-367
Record-structure	MINUTE CELL:MIN-MAX-LENTH-SUM
Record-example	2016-01-11 04:02 40611:6-6-1-6,21862:47-47-1-47

We can see that both User Activities and Response Time have two kinds of record-structure, one of them has cell number. It can tell us where this user activity happens and what's the response time of this cell right now. The number is from 1 to more than 60,000. What should be noted that we don't know what relations between the cell number. It just a number for us and we can't get other information because it's some kind of secret information.

2. Previous work

As we have told before, we have done a lot of data analysis work previously except for spatial data. So in this section we want to spend some space talking about our work briefly. We have analyzed Fig.[1] about what user activities change in day and hour range, Fig.[2] about what Response change in day and hour range, Fig.[3] about what Throughput and DataSize change in hour range and Fig.[4] about what Duration and Appusers change in hour range, there are also PDPDuration in day, PDP in day, Delay in hour and PDP Users in hour. Due to space, we list 8 of them rather than all of them.

3. User Activities

Let's come back to this project. First thing we think about is to figure out what exactly user activities between all cellular bases. So we want to visualize all the user activities in every cellular base. To reach this, we have constructed the Heatmap, range from 2016-1-1 to 2016-1-20 all 20days every 1 hour, so there are all 480 figures.

$$Color = \log(User - Activies)$$

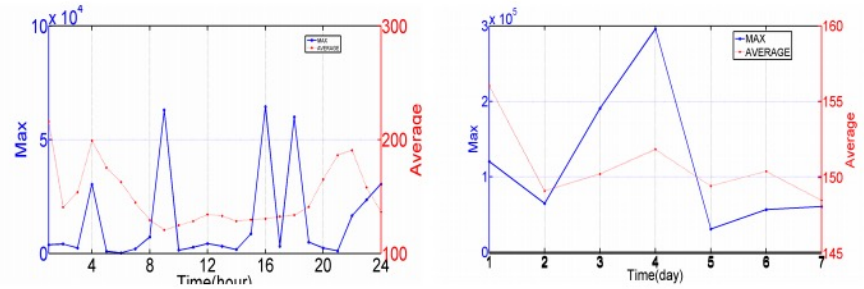


Figure 2: Response Time in day and hour

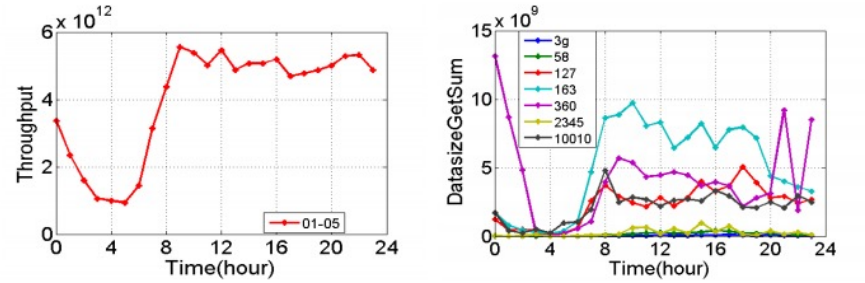


Figure 3: Throughput and DataSize in hour

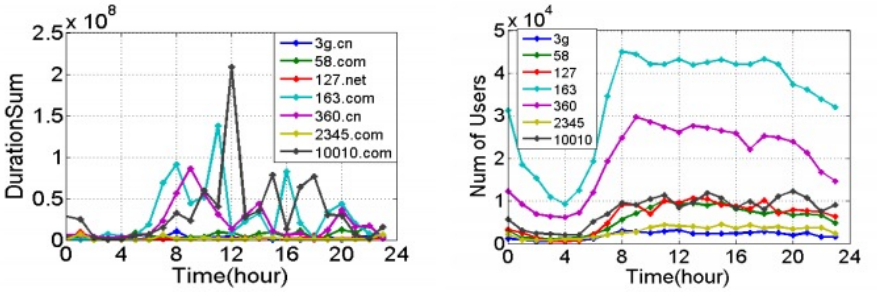


Figure 4: Duration and Appusers in hour

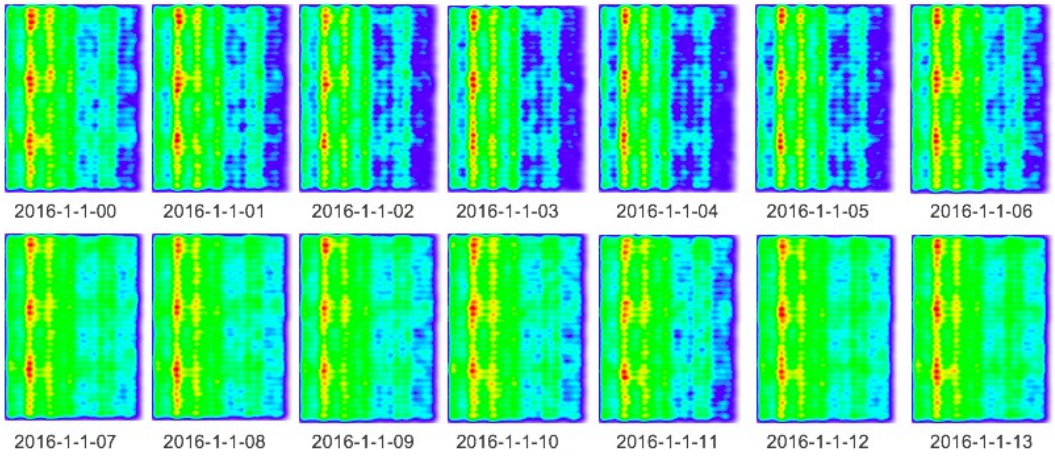
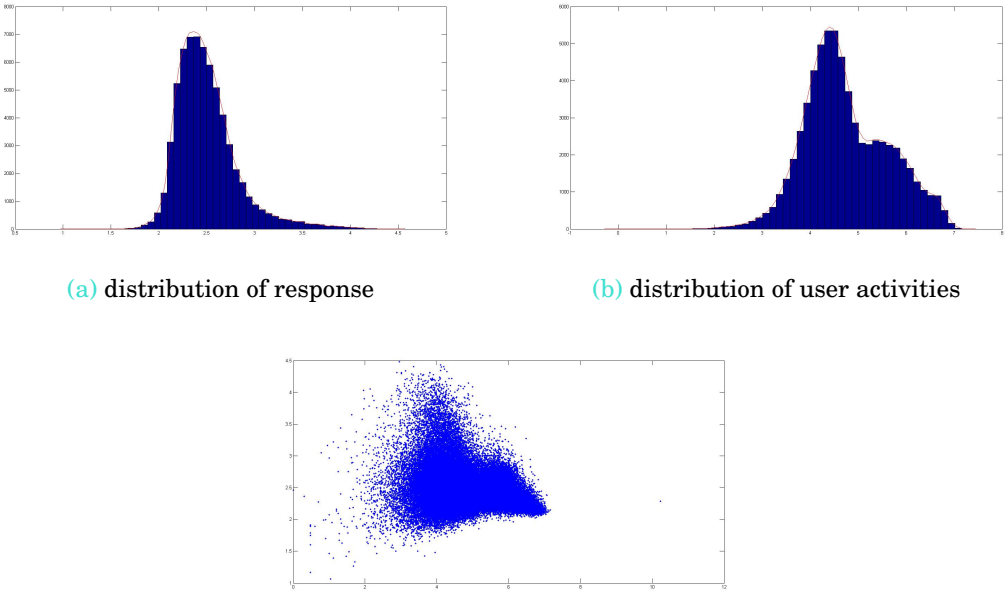


Figure 5: Heatmap of UserActivities in hour

The color distribution is high number is red and low number is blue. We treat every cell as a single pixel. Next is how we construct the heatmap in Python.

```
for ln in sdata:
    data = []
    a = ln.split("\t")
    time=a[0]
    cell=eval(a[1])
    for i in cell:
        cellrange=int(math.log10(cell[i]))
        if cellrange<=0:
            cellrange=1
        data.append([int(i/300),i%300,cellrange])
    # start painting
    data.append([0,0,1])
    data.append([300,300,1])
    hm = HeatMap(data)
    hm.heatmap(save_as='heat-images/'+time+'.png')
    return
```

Here we list 14 figures from 0:00 to 13:00, we can clearly find that from 0:00 to 7:00 user become inactive, but from 7:00, users become more active and in 11:00 maybe most of users start to have lunch so they become less active than 10:00 and 12:00. Let's compare this heatmap with Figure.[1], we can see that the result is almost the same. But what we can't find in it is that there are a column of cellular bases in Figure.[5] are always active. So this make us try to know how much distribution is this part controls and how many user activities it has controlled.



(a) distribution of response

(b) distribution of user activities

Figure 7: distribution of user activities and reponse

4. Cellular Base Distribution

To figure out what the active part distributes and controls, we draw the distribution figure of User Activities and Response Time. The figure is easy to draw, we divide them into 50 parts. And what we can find is obvious that the distribution is similar to normal distribution.

However, if we combine the two pictures together, the amazing things happen. What we haven't thought about is that we always think the most user a cell has, it will be likely to have high delay. But the interesting thing is that in our analysis, the cells with the most users have the acceptable delay, but unfortunately, the cells with the middle part of user activities have the unbelievable high delays, which may inform the operators of upgrading the cells or assigning more resources to them. But we still don't know the exact number of it. So we need a detailed mathematical analysis rather than figures.

5. Cumulative Distribution Function

Here we raise the cumulative distribution function to solve the problem about finding out how much the active parts distribute and control.In probability theory and statistics, the cumulative distribution function (CDF) of a real-valued random variable X, or just distribution function of X, evaluated at x, is the probability that X will take a value less than or equal to x. In the case of a continuous distribution, it gives the area under the probability density function from minus infinity to x. Cumulative distribution functions

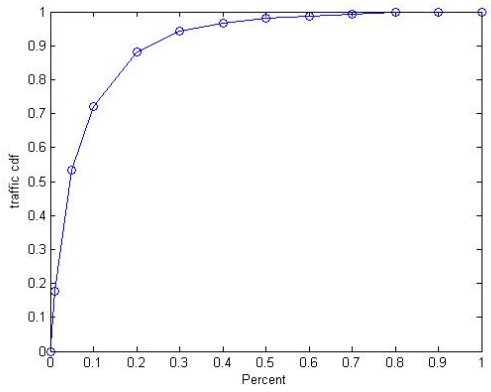
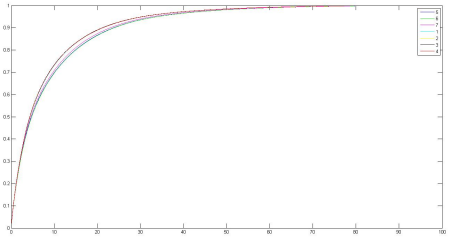
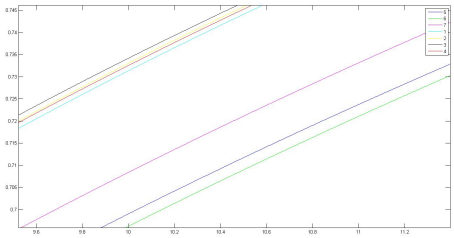


Figure 8: cdf of user activities



(a) cdf of user activities in week



(b) cdf of user activities in week

are also used to specify the distribution of multivariate random variables. Its equation is:

$$F_x(x) = P(X \leq x)$$

Now we can get the exact number, from the figure, the top 5% of cells control 50% of user activities and the top 10% of cells control 70% of user activities. However, in paper [1] the result is top 10% of cells control the 50%-60% of user activities. Such difference may reveal that our chinese are more aggregated than people in the USA. Besides, we also draw the cdf in time series.

From this figure, we can also clearly find that on weekdays, people are more aggregated than on weekends.

6. Conclusion

In our project, we try to analyze the spatial data to figure out the performance of cellular bases and behaviors of users. From the heatmap of user activities, we find a fixed pattern of user activities and a special part is that user activities are still very high at night. From the distribution part, we find that the highest delays happen on the middle part of cellular bases. From the CDF part, we find that the top 10% of cells control 70% of user activities

and on weekdays, people are more aggregated than on weekends. These results may tell operators that when allocating the resources, they should consider them and allocate more reasonably. Besides, operators should upgrade some of cells or allocate more to them. We hope our analysis can help operators to improve a little. Meanwhile, Zhu XuanYu is responsible for the spatial data prediction and Jin YaoAn is responsible for the spatial data clustering. They will also give their report.

References

[1] Paul U, Subramanian A P, Buddhikot M M, et al. Understanding traffic dynamics in cellular data networks[C] INFOCOM, 2011 Proceedings IEEE. IEEE, 2011: 882-890.

[2] Zhang Y, Årvidsson A. Understanding the characteristics of cellular data traffic[J]. ACM SIGCOMM Computer Communication Review, 2012, 42(4): 461-466.

[3] Lee D, Zhou S, Zhong X, et al. Spatial modeling of the traffic density in cellular networks[J]. IEEE Wireless Communications, 2014, 21(1): 80-88.

[4] Laner M, Svoboda P, Schwarz S, et al. Users in cells: a data traffic analysis[C] 2012 IEEE Wireless Communications and Networking Conference (WCNC). IEEE, 2012: 3063-3068.

[5] Heatmap Python website, http://jjguy.com/heatmap/, 2016