Data analysis of NFV Report

F1303027 5130309799 Yaoan Jin

1、Background

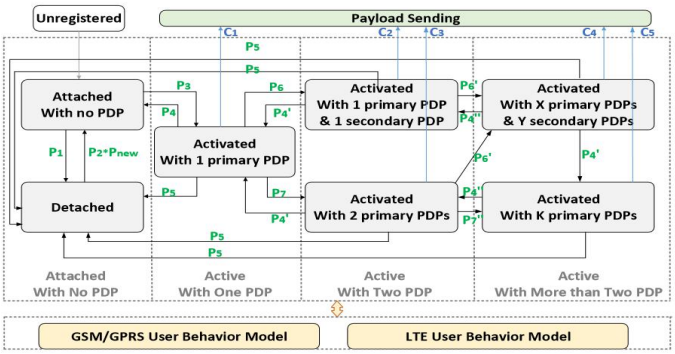
Network functions virtualization(NFV) is a network architecture concept that uses the technologies of IT virtualization to virtualize entire classes of network node functions into building blocks that may connect, or chain together, to create communication services.

NFV relies upon, but differs from, traditional server-virtualization technique, such as those used in enterprise IT. A virtualized network function, or VNF, may consist of one or more virtual machines running different software and processes, on top of standard high-volume servers, switches and storage, or even cloud computing infrastructure, instead of having custom hardware appliances of each network function.

For example, a virtual session border controller could be deployed to protect a network without the typical cost and complexity of obtaining and installing physical unis. Other examples of NFV include virtualized load balancers, firewalls, intrusion detection devices and WAN accelerators.

2、Motivation

Our group collected 20T data from a operator and tried to find some stability behind a mount of data so that we can do source allocation according to the steady data model. Finally, we built a steady user activities model based on the assumption that the number of people in any user activities is always stable. Thus, we can distribute source more efficiently according to this model. So what can I do for this topic? We already analyzed the stability of user activities distribution. Does the number of people have stability in different time or different location? Based on this thought, I came up with my ideas: 1)we can analyze the stability of the number of users in different location to help the source allocation. 2)we can analyze the stability of the number of users in different time interval to help the source allocation.



Moreover, the probabilities between different user states in picture above were calculated according to the assumption that the number of people which are out of a state is equal to the number of people which are going into a state. How about using machine learning method, MLE, to recalculate the probabilities. Maybe we can get a better steady user activities model.

3、Data analysis of NFV

I mainly did the analysis of stability of the number of users in different location to help the source allocation. I used the original data with the format(TIME, Cell_number:Users_number, Cell_number:User_number...) to do data analysis.

```
2016-01-09 20:15        0:1
2016-01-09 20:33        20402:1
2016-01-09 20:51        21621:1
2016-01-09 21:05        46261:1,23522:1
2016-01-09 21:23        16182:1
2016-01-09 21:38        10141:1,0:1,30693:1
2016-01-09 21:41        52117:1,51973:1
2016-01-09 21:56        48453:1,23691:1,11351:1,50323:1,0:1,33592:1
2016-01-09 22:13        57432:1,50952:1,53713:1
2016-01-09 22:28        21411:1,381:1,21383:1,30293:1,0:1,32361:1,20352:1
2016-01-09 22:46        34072:1,30641:1,21401:1,28993:1,51143:1,51611:1
```

Original data

Firstly, I processed the original data and calculated the number of people in every cells among ten days.

```
41      0       221     191     1319    731     4       77      14      185     1
129     21      248     3       47      722     1137    792     19      390     2
337     617     1750    0       350     199     122     0       158     348     3
1052    54      0       0       0       678     2       5       130     210     4
0       0       0       0       3699    3       7       468     1       17      5
101     1076    2       15      0       6       1333    657     1       472     6
0       0       0       0       0       0       1       0       0       0       7
0       0       5       0       0       3       0       0       0       0       8
0       3       61      0       74      10      0       0       0       0       9
71      399     129     200     15      83      1091    76      246     1698    10
125255  109828  118716  101852  107745  86910   91254   86564   97010   72179   11
81999   99918   83373   75625   58158   43297   49727   58696   81372   67887   12
97435   125977  102682  100850  89329   86117   71945   80274   100130  105546  13
7819    4445    2704    3247    5913    5078    5717    3600    6211    1524    14
7837    6786    3036    3606    3842    6319    1904    2856    1678    5300    15
2064    3769    1339    7202    6735    1643    7857    2781    2754    4466    16
9616    2668    3596    3158    1882    3602    3269    2295    3209    7326    17
202     956     353     197     389     110     389     2005    582     241     18
```

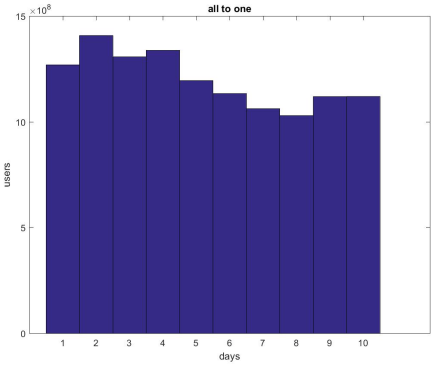
Processed data

I get a ten dimensions vector as the information of a cell. Each dimension indicates the number of people in this cell one day.

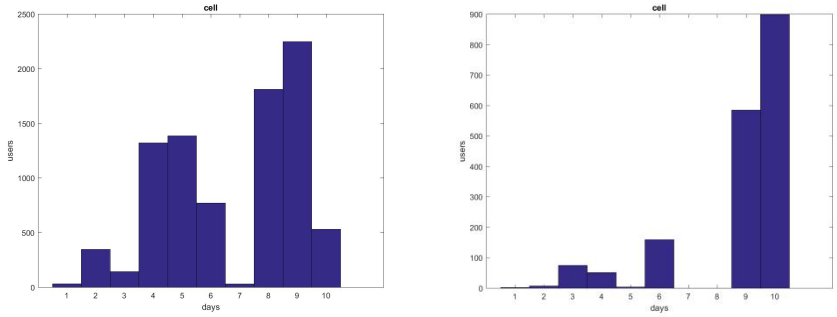
I proposed two method to cluster these cells to prove the stability of location distribution. The First one is KMEANS FOR CLUSTERING CELLS. KMEANS is a classical method to cluster data points in machine learning field. In my scheme, each vector of cell is a point and the variance of the number of people in different days after making any two cells a group is the distance between two cells. We cluster these cell points so that the changes of the number of people of any group in different days are steady by KMEANS. The second one is TOP_TO_DOWN. We make all cells as a group firstly and then remove each cell from this group so that the stability of the group can be better. After the remain cells in group satisfying our need, we stop removing and get two new groups. For each group, we repeat the steps until we getting a number of groups of cells. I think the second method is better than the first one. Because we only test making two cells as a group in the first method. It is hard to present stability when the number of cells in group is small.

4、Experiment

In experiment, we first made all cells as a group to observe the stability.

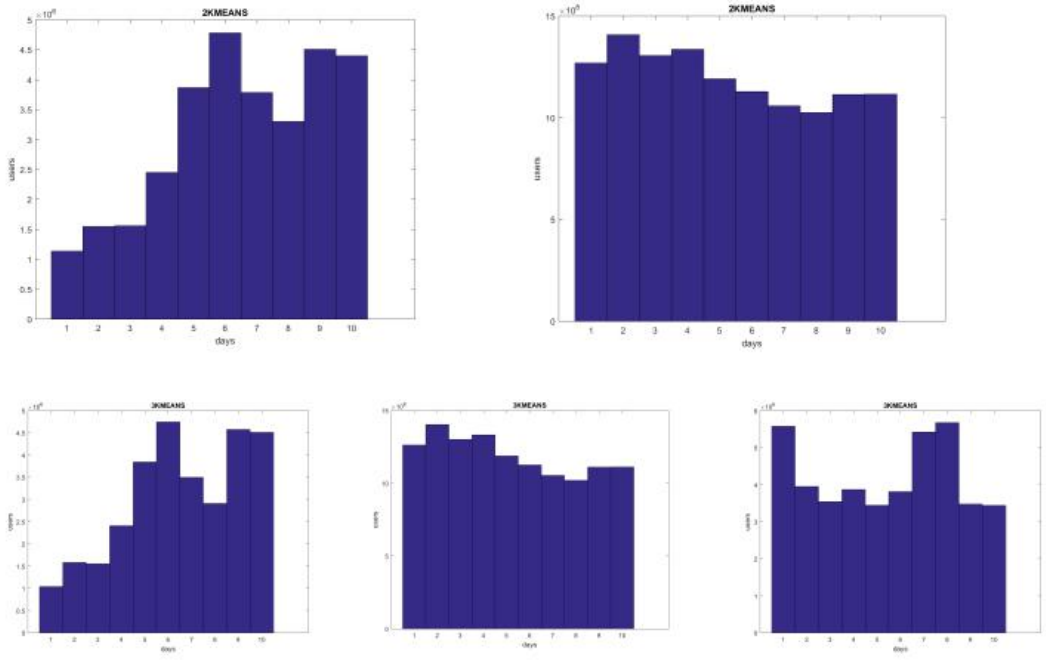


We can see the number of users distributing every days is steady. However we would like to find the stability of small group to do source allocation.



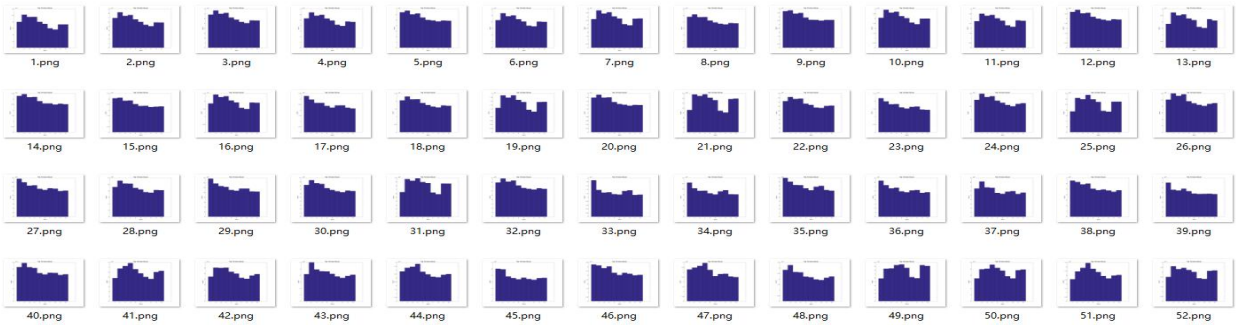
Users distribution in one cell

From pictures above, we can see the number of users distributing every days in one cell is unstable.



Users distribution in groups by KMEANS

We use KMEANS method to cluster the cells as two groups and three groups. From pictures above, we can see the stability of users distribution is much better than it in one cell.



Users distribution in groups by TOP_TO_DOWN

We use TOP_TO_DOWN to cluster cells as more groups(250groups). Every groups have great stability.

5、Conclusion

To achieve analyzing the stability of the number of users in different location, we design two methods. Both methods cluster the cells as several groups and work out great results. For future work, we can improve our clustering algorithm to get better clustering groups. Moreover, we can do dynamical clustering of cells so that we can achieve prediction of source allocation to make performance of NFV better. There are also a lot of ideas remaining to try. We can do more in data analysis of NFV!