FaceBook Data Analysis

# Query to extract FaceBook Data

SELECT f.[ResponseXML].value('(//first\_name)[1]','varchar(40)') 'firstname',

f.[ResponseXML].value('(//last\_name)[1]','varchar(40)') 'surname',

f.[ResponseXML].value('(//birthday)[1]','date') 'birthday',

f.[ResponseXml].value('(//id)[1]','varchar(50)') 'id',

f.[ResponseXml].value('(//updated\_time)[1]','varchar(60)') 'lastupdateday',

cast(GA.app\_date as Date) appDate,

f.ApplicationId ApplicationID,

GA.cont\_first\_name appfirstname, GA.cont\_last\_name appsurname,cast(GA.cont\_date\_of\_birth as Date) appDOB

FROM fb.CollectedData f inner join WongaWholeStagingv2.greyface.applicationView GA

on GA.id=f.ApplicationId

where UserConnection ='/me'

select x.value('./id[1]','varchar(80)') friendid, f.ApplicationId into #tmp from fb.CollectedData f

cross apply f.[ResponseXml].nodes('//data') as T(x) inner join WongaWholeStagingv2.greyface.applicationView GA

on GA.id=f.ApplicationId where UserConnection = '/me/friends' order by ApplicationId

select COUNT(\*) linknumber,ApplicationID into #t from #tmp group by ApplicationID

select distinct(ApplicationID), f.[ResponseXml].value('(//id)[1]','varchar(50)') 'id' into #link from fb.CollectedData f where f.[ResponseXml].value('(//id)[1]','varchar(50)') is not null order by ApplicationId

To set a cut off for name similarity and another cut off for friends

First, I conduct the following experiments:

1. Compute the similarity of the name registered on the face book and applied in our system. I use jaro-winkler distance.
2. Look at the number of friends one applicant link and plot the cumulative distribution figure.
3. Based on decline rate 25%, we can adjust the similarity level of name and numbers of links. I plotted one curve to show these relation between the 2 value if we want to accept 75% apps.
4. For the first 100000 pairs of links, I plot the network graph.

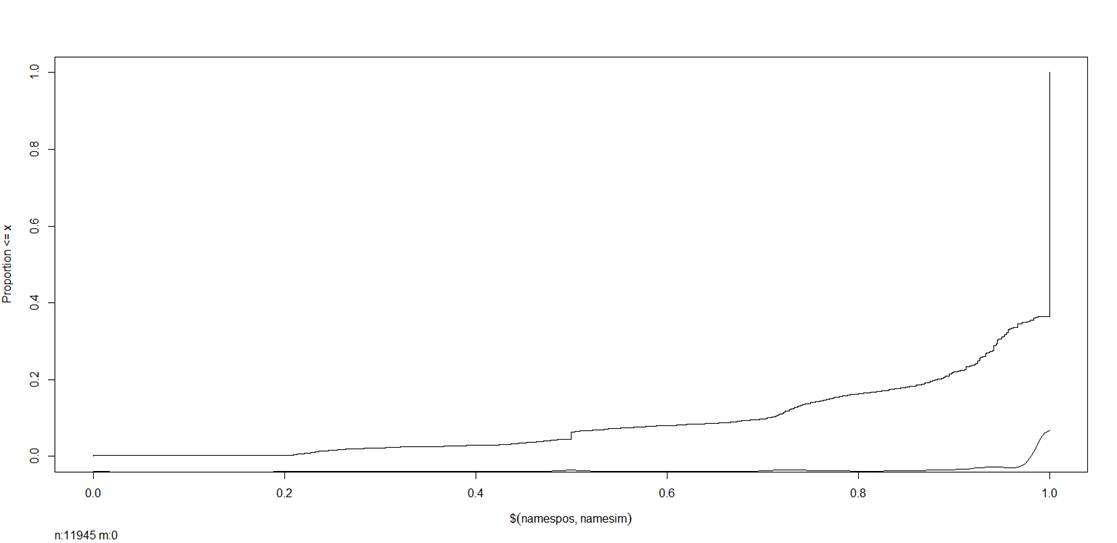


Figure 1 Name Similarity

In figure 1, x-axis is name. If one pair of names is identical, the similarity is 1, if there is nothing in common (based on definition of jara winkler distance), the similarity is 0. The top curve is the cumulative distribution.

The bottom curve near x-axis is the density of similarity, we can see majority of customers have exact match.

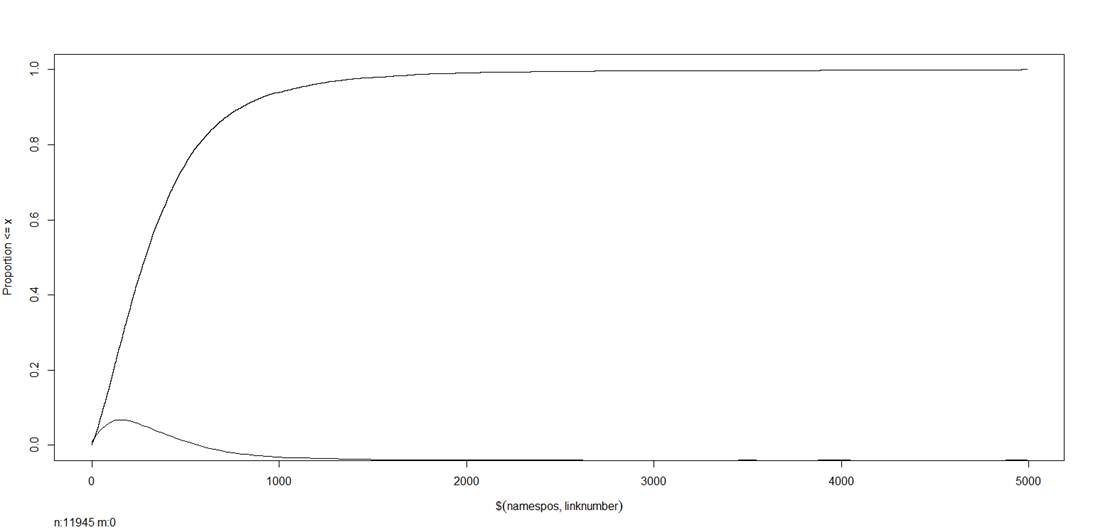


Figure 2 Number of Friends

Figure 2 shows the cumulative distribution and density  of number of friends. The x-axis is the number of links one person have. In this dataset, the mean is 390 and median is 287. The maximum is 4990 and the minimum is 1 friend.

The density is right skewed with some outlier far away on the tail.

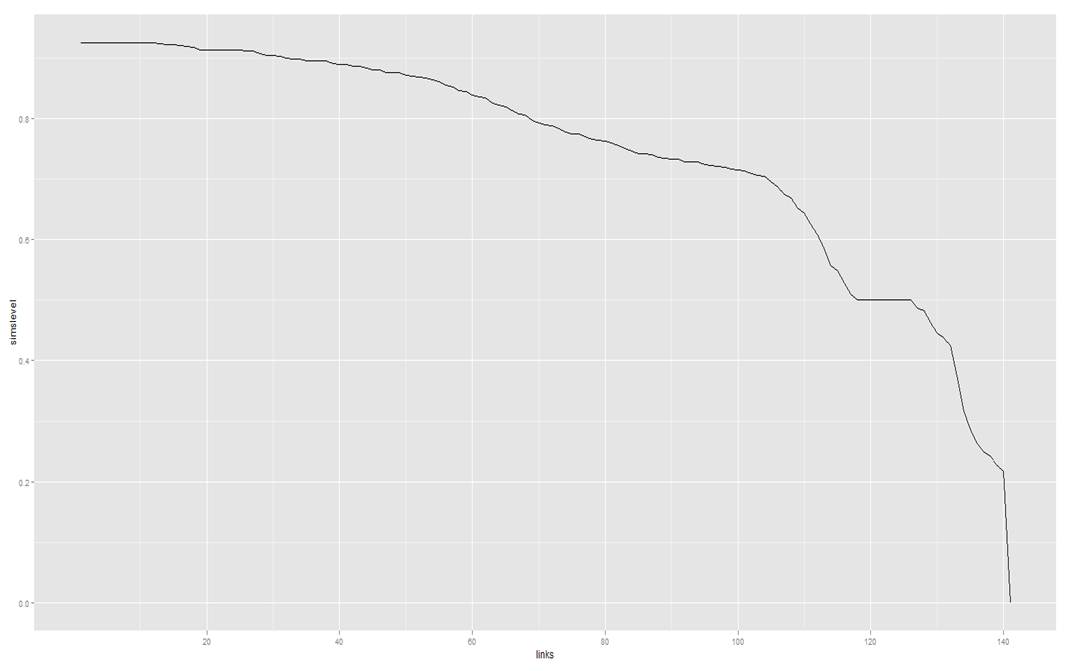


Figure 3 links and similarity level for decline 25%

In figure 3, the x-axis is the number of links and y-axis is the similarity level of 2 names. The curve shows that if we want to have 25% decline rate, what is the number of links one person should have and corresponding

Name similarity. For example, if we ignore the name similarity ( set it as 0), we need about set the cut off for number of links at 141. And if we ignore the number of links (corresponds to x=0), then we need name similarity level. From my personal view,

I think we need have slightly higher name similarity threshold and low number of links, like (0.8381,60).

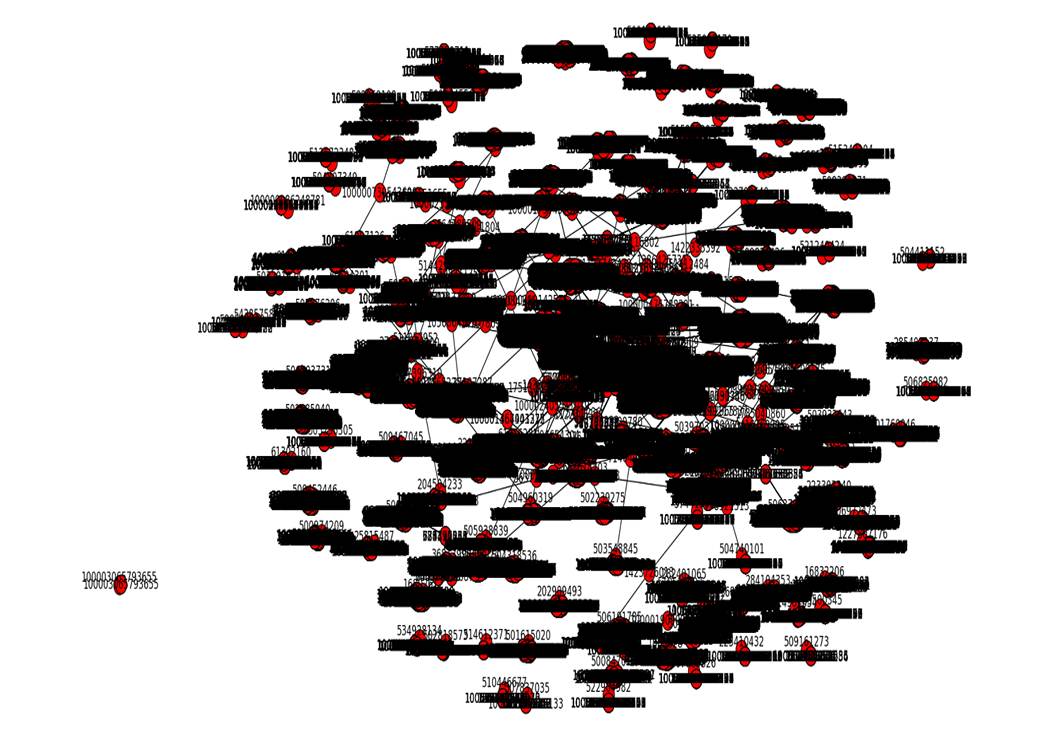


Figure 4 FB network graph

In figure 4, I plot the Facebook network graph, the label on the node is the user id. If there is an edge between 2 nodes, which implies that they are friends. I took 100000 pairs to plot the above figure. Note not every node in the graph is our customers; they could be the friends of our customers.

From here we can see many small islands isolated from others and also a few connected nodes.

In FB data, we also have DOB information, we can compare this DOB information to the DOB from our internal system.

Amongst 8947 apps, there are 7762 apps whose DOB is exact match, which is about 86.76%.

The dob difference is shown in below figure

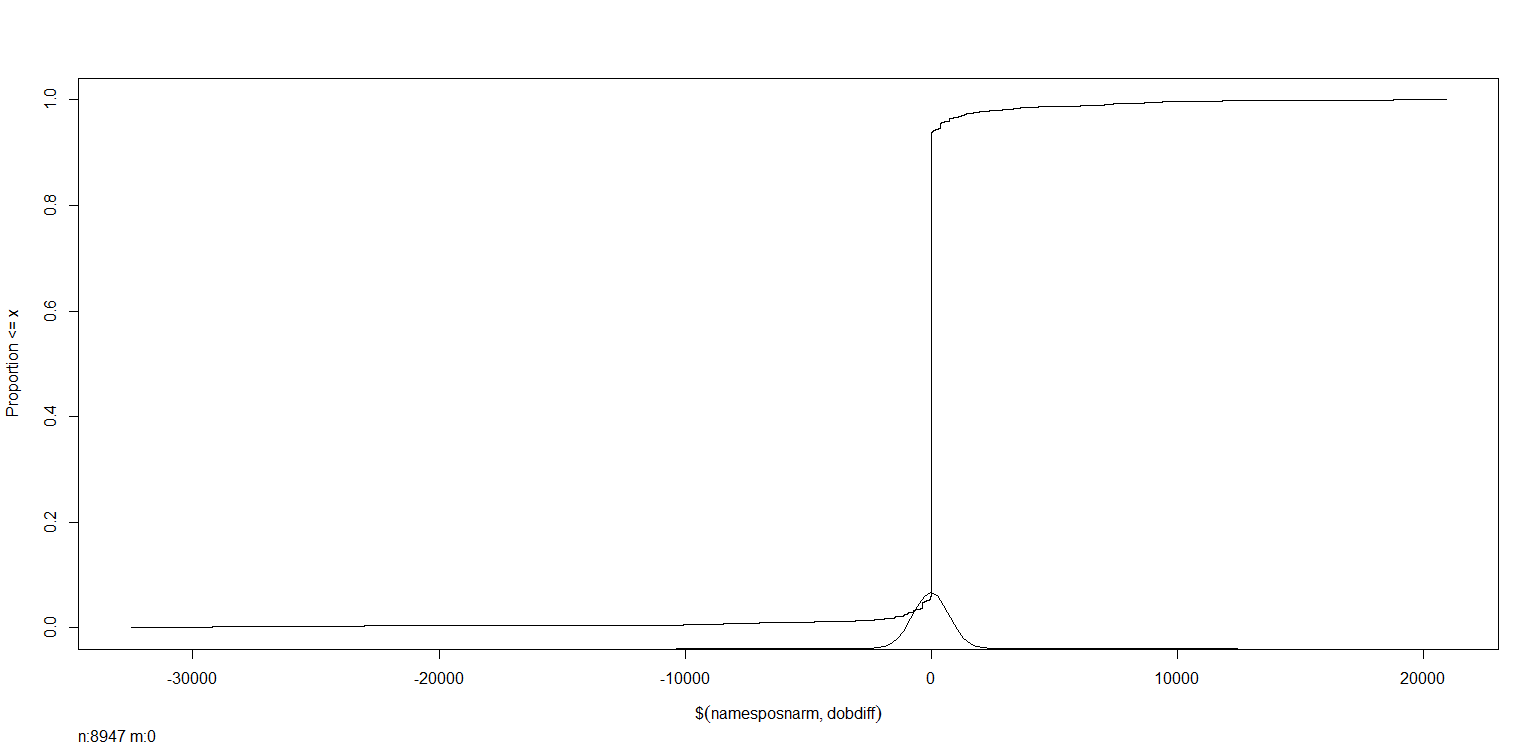
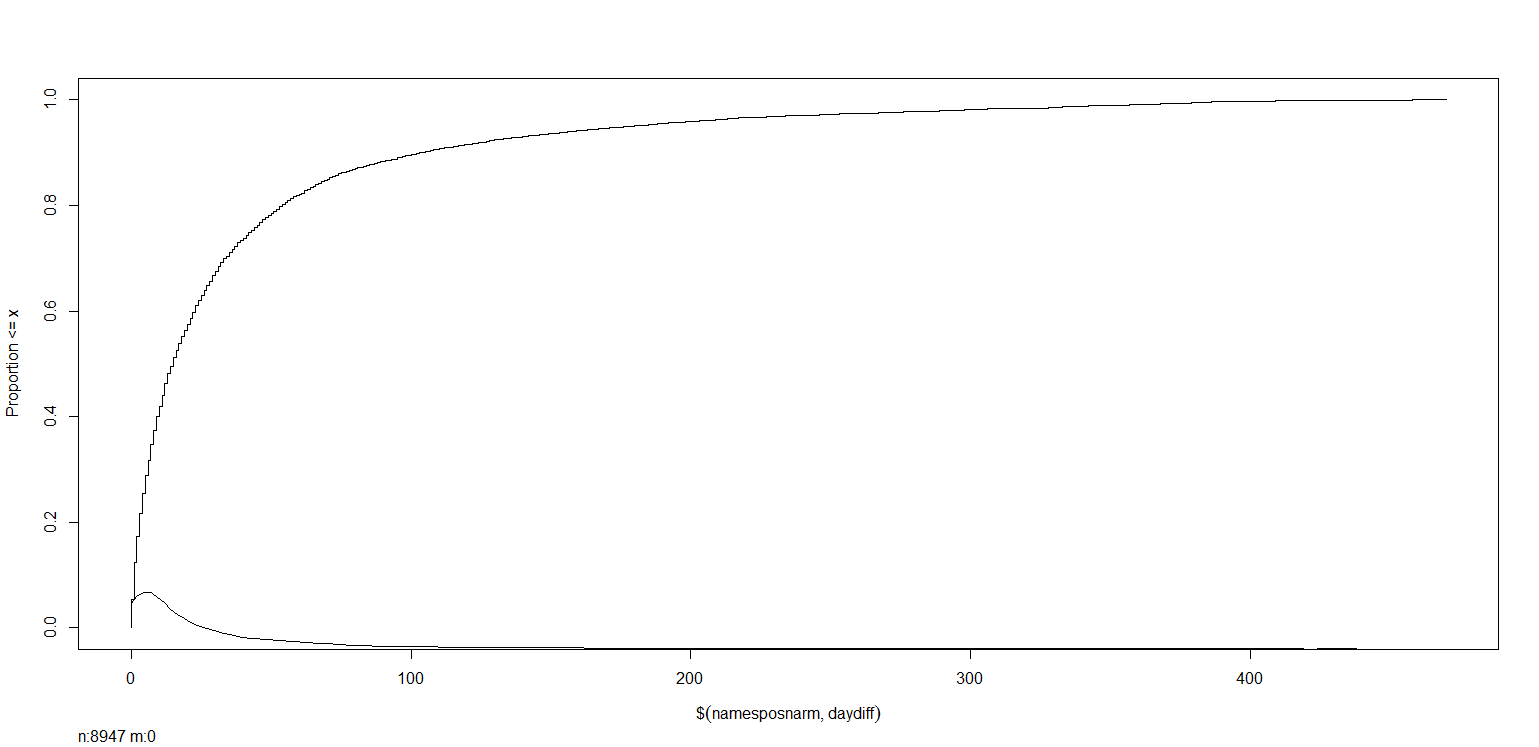


Figure 5 DOB difference cumulative distributions

In additional to that, the FB data also provide last update information. We can compare the difference between last update days with app date.

The result is shown in Figure 6

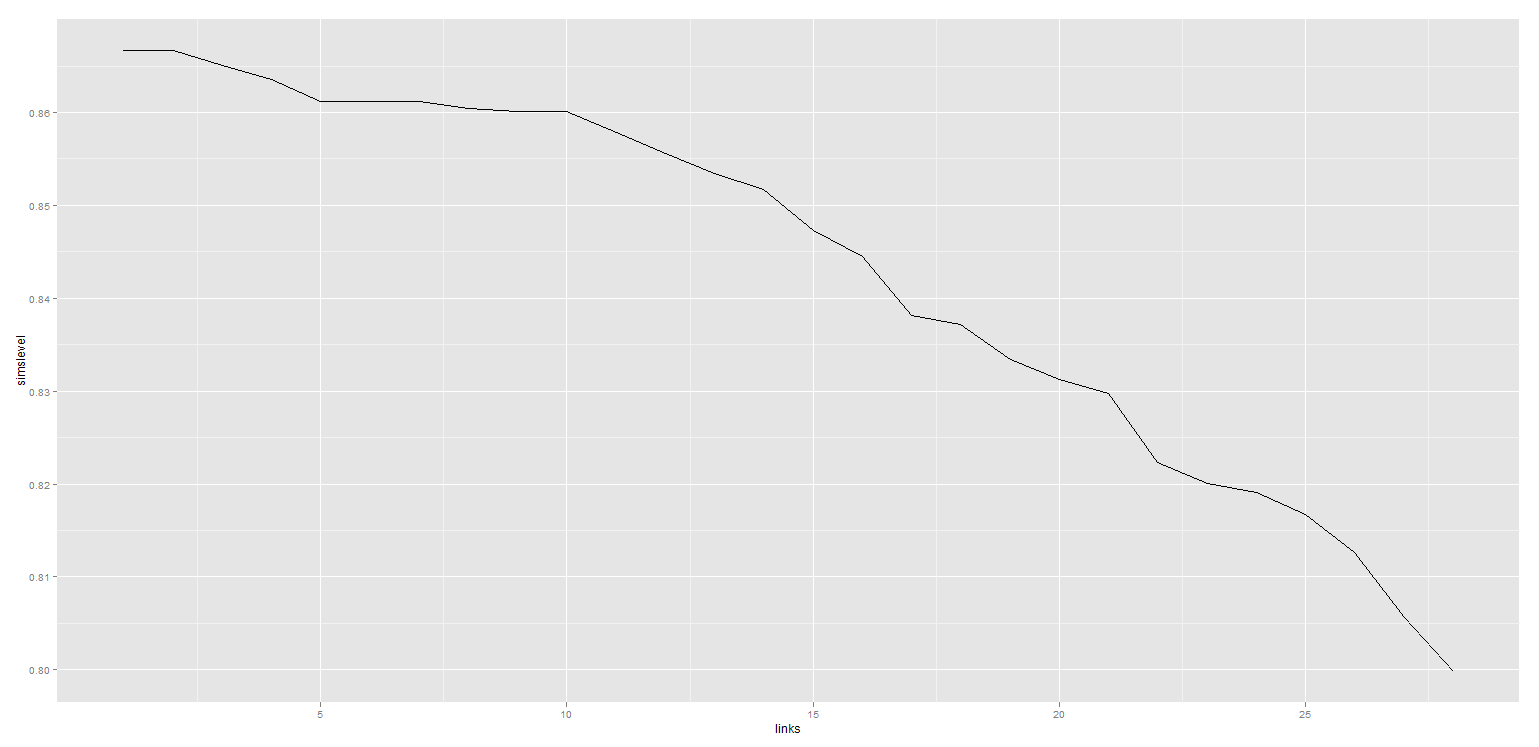


To verification the user using facebook, we can set up a few rules. Amongst 8947 apps, if we want to decline 25% (2236 apps). We can have require the following:

1. DOB should be matched.
2. Name similarity should above some threshold
3. The number of friends and the days since last update should be greater than some value.

If we fix the name similarity level to be 0.8, we can find the relation between the number of friends and days since last update.

The name similarity and number of friends curve after DOB match is shown in below



One possible suggestion based on declined 25% is

1. If DOB doesn’t match, decline it.
2. If Name similarity is lower than 0.83, decline it.
3. If numbers of friends is less than 20, decline it.

In other words, we will accept apps whose DOB match, and name similarity is above 0.83 and number of friends is greater than 20.