Behavioral Analysis of Iranian Users in a Mobile Social Network

Abouzar AbbaspourGhomi and Masoud Asadpour

***Abstract*—It's been almost more than fifteen years that social networks have become significant part of our everyday life. Social networks being so mainstream and available, owes its availability to developments in some other fields. Mobile network developments have been very significant in recent years. Introducing technologies such as 3G or LTE have been a major reason for almost every mobile subscriber to use social networks services. That's the world that on demand access is more important than ownership. There are many different mobile social network applications, which competing each other with new services and feature every other day. In order to survive this competition there can be some analysis that can be done. One of them is about the behavioral aspect of users' activities, that we can use to extract a pattern of users' activity in social network applications. These patterns can be used to analyze user acquisition or churn. In this paper we will be talking about the behaviors of Iranian users in a mobile social network. We gathered the data by a third party application for Instagram and we are going to use different methods of visualization, statistical methods and analysis to show different patterns in users behaviors.**

***Index Terms*—Mobile social networks, data visualization, social network analysis, behavioral analysis, churn, users correlation.**

I. Introduction

Mobile technologies since 2006 have had major renovating phases. Technologies such as 3G and LTE have become very reachable. In result of these developments, subscribers can access the internet with high-speed connections and that brings the ability to share contents easily and at will [1]. Mobile social networks at the other end of this spectrum are enabling users to share texts, pics and videos from everywhere and at every moment. Also mobile devices have become more powerful along the way [2]. Now they have very interactive interfaces, powerful cameras and outstanding processing power. All of these developments bundled with high-speed access to the internet. Mobile social network apps compete with each other intensely. One of the issues that they need to tackle is the users' churn or acquisition. Analyzing users' activity and creating a pattern of activities can help determining reasons for users' acquisition and churn [3]. In Iran mobile technology has started in 1994 and had a slow start. In 2008 the only mobile operator in Iran had about 7 million subscribers despite enormous demands of the population for mobile services. In 2005 there was a rival for the only mobile operator in Iran. The market became more free and competition between operators resulted in better and cheaper services. In 2014 there were three operators in Iran that started to offer 3G services to the subscribers and in result of these services subscribers data usage have been skyrocketed since [4].

In this paper we have gathered data from a third party app for Instagram. This app has about 180000 users. This data has characteristics of a big data and needs preprocessing and post processing. We used different methods toward analysis of the data such as initial data analysis and data validation, multi-dimensional graphs, correlation analysis, hypothesis testing, histograms and exploratory data analysis. Some of these steps will be discussed in this paper we will be discussing variables that we extracted from the users' data regarding acquisition and churn analysis. We draw the graph of the social connection of users in order to specify ways to gather further clues for our study. In the next step we extracted variables in order to draw histograms and calculate correlation between users. Also we are going to show the retention rate of users.

II. Data Gathering

Our third-party app uses some built-in APIs and Instagram APIs to gather users' data. This data has been gathered for the duration of 10 months. This data source consists of users' social relationship and geographical data along with analytical data such as count of activity, date of activity, etc.

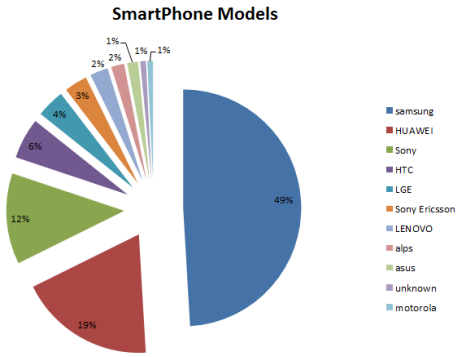


Fig. 1. Users and device models.

There are some statistics about the app that can help us know our data better. This also could be called initial data analysis (IDA), where we a get a feel for our data before we start to do any statistical analysis. In Fig. 1 you can see the distribution of users by devices model. Majority of users have high quality devices. In Fig. 2 there is percent of users using different mobile operators. This figure can be proof

Manuscript received January 7, 2017; revised May 1, 2017.

A. Abbaspour is with Mobile Telecommunication company of Iran (MCCI), Iran (e-mail: a.abbaspour@mci.ir).

M.Asadpour is with School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran (e-mail: asadpour@ut.ac.ir)

DOI: 10.7763/IJET.2018.V10.1042

that which operator has better quality of service especially at services such as mobile data, because we are considering a social network such as Instagram that is using mobile data very intensively and needs a good internet connection [5] [6].

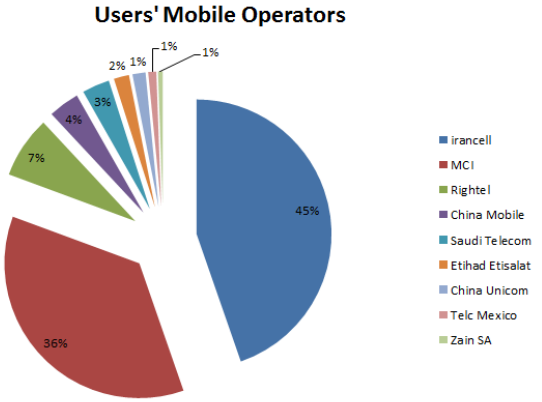


Fig. 2 users and mobile operators.

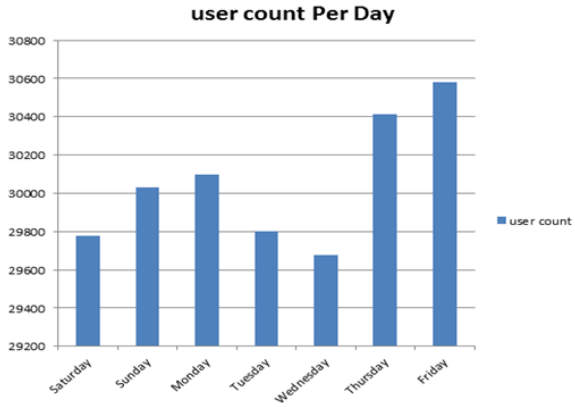


Fig. 3. Active users count per day.

III. Usage and Activity

How and when users use an app can be very useful for technical teams that are working on the app. Users activity can show the backend team that when is the peak of the traffic and they can use this information to manage and prevent service failures accordingly. We can use this data to specify events in the past and predict if it could happen in the future and what exactly is the consequence of that event will be.

As it can be seen in Fig. 3 number of active users tends to be much more at the weekend of Iranian calendar. Also very similar to Fig. 4, in Fig. 4 number of activities per day has been shown. At the weekends usually the number of shared contents increases. However this doesn't happen for every holiday, in period of time that we studied our data we had some holidays that users were much less active comparing to a working day. They were maybe on a trip or in a party, and didn't have access to internet or didn't feel like sharing. This means that every absence of activity cannot be interpreted as churn, and also obviously there is a pattern in using this app and sharing or downloading the contents.

In an hourly context we can see which hours our app has the most traffic and users check their profiles. Sharing contents may happen throughout the day but checking other people's profile and commenting or liking their photos will happen in someone's free time that best can be shown in Fig. 5.

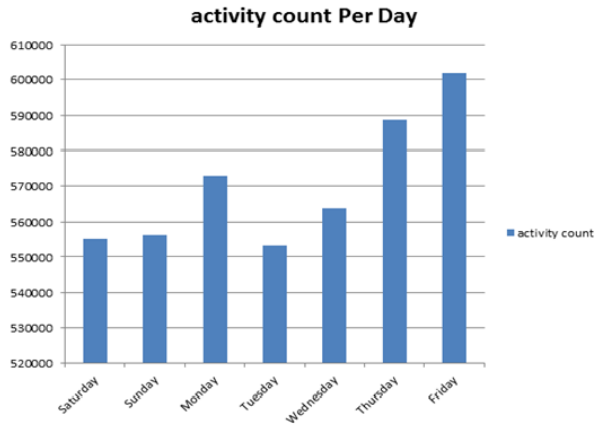


Fig. 4. Activity count per day.

Fig. 5. Users' session and screen views.

IV. Churn Variables

One of the more important issues with social network had been the process of acquiring or churning of users. By understanding the causes of these actions we can put a plan in action to increase users acquiring and decrease users churn [7]. We recognized that the first step toward this goal is finding out the variables that effect this process. We basically carved out these variables from the data that we had processed. The raw data contained of many different properties we had at least five levels of data enhancement to achieve the richness of data we had in mind. We ran different programs that we prepared mostly with JAVA language. Our raw data was stored in a relational database so we used SQL to join and gather some normal information. But we calculated out and in degree, location, Percent of installed followers, percent of installed followees, …. by using custom applications in different stages of data enhancement.

TABLE I: Churn Variables

Variables	Description
username	Username in instagram

Out Degree	Number of users' that current user follows.
In Degree	Number of users following current user.
Number of Installed Followed-by	number of followees that have getagram app.
Percent installed followed-by	Percent of followees that have getagram app
Number of installed follows	Number of followers that have getagram app
Percent installed follows	Number of followers that have getagram app
Location	The latitude and longitude of users
Number of activity	Number of usage by user
Age(Day)	User's age(number of days that user installed the app)
Last Usage in Days	Number of days since user used the app
Last Usage Date	The date that since users haven't used the app

These data have been extracted from tens of thousands of users that connected our app to Instagram. These fields would be our basic info in creating different graph and visualization for users' data.

V. Drawing Graph Based on Users' Age

In this section we used exploratory data analysis (EDA) to find and describe the main characteristics of the data. Using this method we found that in our network mutual relationship could affect behavior of a user in a social network. We used follower and followee information for users to draw a network of users in GEPHI. We used different layouts such as OpenOrd2 and ForceAtlas to achieve our desired image. As it's clear in Fig. 6 we have thousands of users separated in relatively small communities. Further in this article we are going to examine if there is a correlation between these users' action.

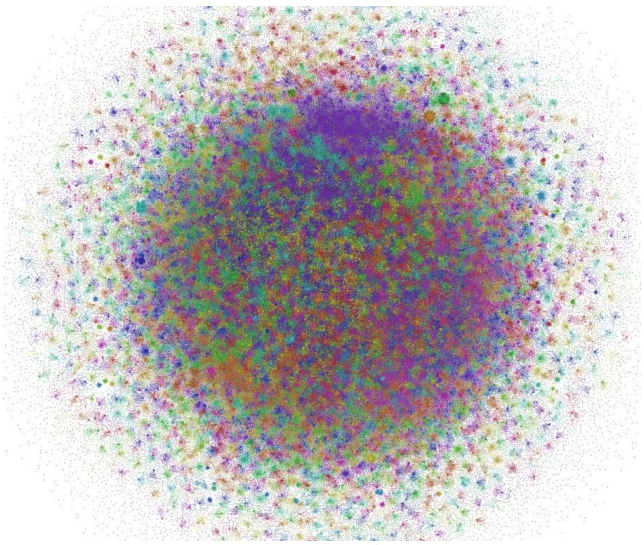


Fig. 6. Networks of users.

Connection between each user has been shown with different color. We used the data to create this graph and other two more graphs that each considers different features of users' relations.

With regard to Table I, one of the variables is user's age meaning number of days that user has been using the app. We used this variable as a property of users to draw another graph that visualizes the age of users with their connections. Minimum age with white and maximum age is shown with black color.

Another important variable shown in Table I is the number of activities that a user has done. We used this variable to draw a graph based on users' activity. So we used the same approach as the graph for age shown in Fig. 7.

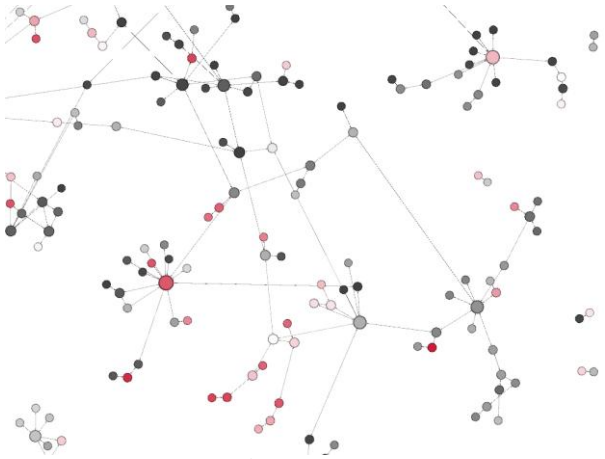


Fig. 7. Network of users colored by age property.

VI. Drawing Graph Based on Users' Activity

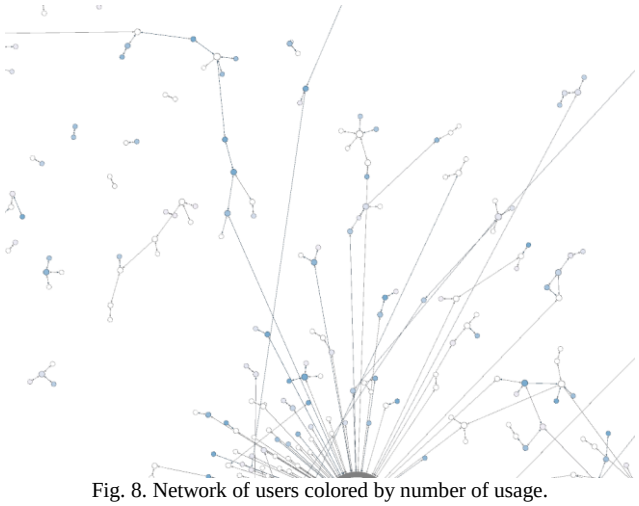


Fig. 8. Network of users colored by number of usage.

We used GEPHI's timeline feature to animate users' activity in a 90 days period. We wanted to see if there was a relation between users that are in the same network and if an active user can have effect on other users that are not active. At the end for both Fig. 7 and 8 we wanted to calculated different correlations to see if users' activity and social status are affected by each other [8], [9].

VII. Histogram

Extracting variables has helped us to see different dimensions of the data. Considering the number of activities and purposing that it has direct relation with number of followers, we drew histogram of users. So we could see the

distribution of users based on the variables defined.

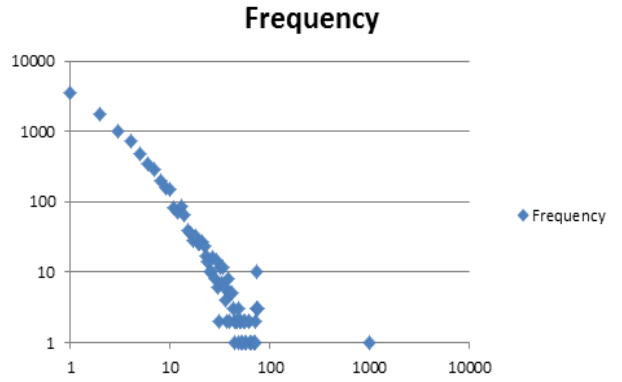


Fig. 9. Histogram on number of followers.

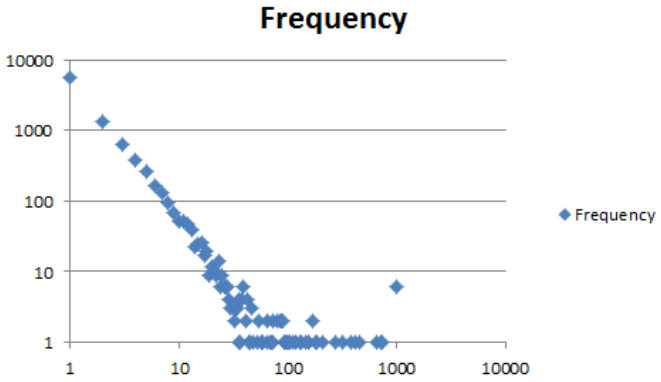


Fig. 10. Histogram on number of followees.

As it's clear in the Fig. 9 and 10 both figures are showing power law distribution and also 20-80 law. Basically these figures show that users that follow lots of people or followed by lots of people are very few. Also active users usually have higher than normal amount of followers.

VIII. Correlation

As we showed in previous sections by modelling different graphs, correlation may exist between users that are in the same network and a user being active in a network could affect the usage of other users, in a nutshell we suppose that our variables are not independent. So we wanted to do the correlation analysis. That's the basis of hypothesis that we had in mind for this section. We returned to our raw data and extracted some fields that we thought most effective for this subject. Extracted variables could be seen in Table II.

TABLE II: Correlation Variables

Variable	Description
UserId	Userid
MaxMedia	Max Number of user's followers uploaded media
MinMedia	Max Number of user's followers uploaded media
SumMedia	Sum of user's followers uploaded media
Followers	Number of Followers
Average(sum/followers)	Sum of media uploaded by followers / number of followers
Usage count	Count of activities each user

As its part of big data classification, it's important to always process the data and make the needed data. In this section we cleaned the data and aggregated records with a python script. Our data contains records from 3580 users that in average follow 190 other users, average count of usage is 48 times, maximum usage count is 1061 and minimum count is 0.

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2}\sqrt{\sum_i (y_i - \overline{y})^2}}$$

Equation 1 Pearson Correlation

We used Pearson correlation equation to calculate the correlation between some of the variables.

The results have been gathered in Table III. The results show that correlations are positive but there is no strong correlation between variables.

TABLE III: Correlation Calculation Result

Usage count and follows	Usage count and Summedia	Usage count and maxmedia	Usage count and avg	Usage count and avg*log
0.06123	0.06479	0.06170	0.01041	0.04195

IX. Activities Visualization

In this part we used MapBOX studio to create a map of Iran as a shape file and used it in PROCESSING. In order to visualize users' activities we gathered users' activities for duration of 2 weeks. We animated the activities during this period and also drew two diagrams showing the users activities hourly as could be seen in Fig. 11. Yellow circles show the usage of the app and red circles show a new user that just started using the app. Also we uploaded this visualization in youtube[1].

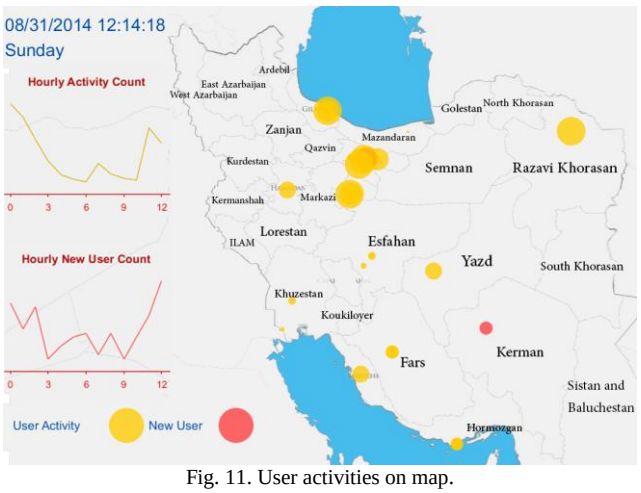


Fig. 11. User activities on map.

[1] https://www.youtube.com/watch?v=j5DuOCifX9s

We also showed the users connections using their geographical data. In Fig. 12, blue dots are followers and red dots are followees and the line between them shows the relation between them. We can see that how users are spread in different parts of the country. Fig. 12 shows that effective users are mostly in more modern cities such as Tehran and Shiraz. Good communication infrastructure can help people to connect easily and they will be more connected [10]. Generally visualization is storytelling, and one of the jobs of data analyst is to find out whether if there is a hole in the story.

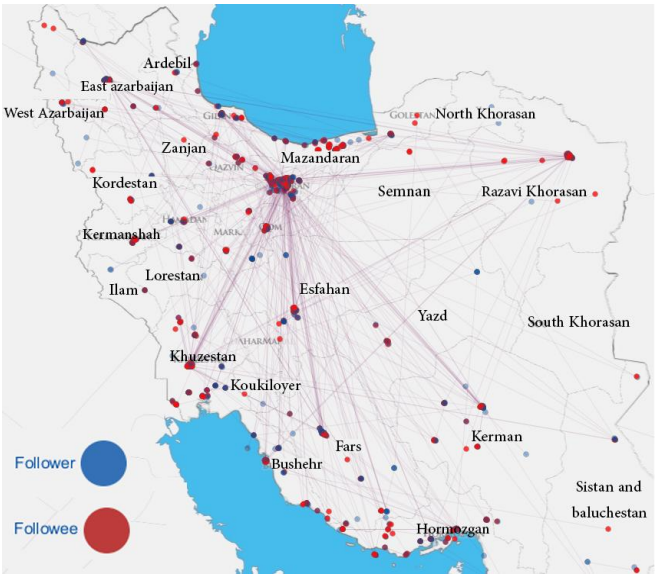


Fig. 12. Users network on map.

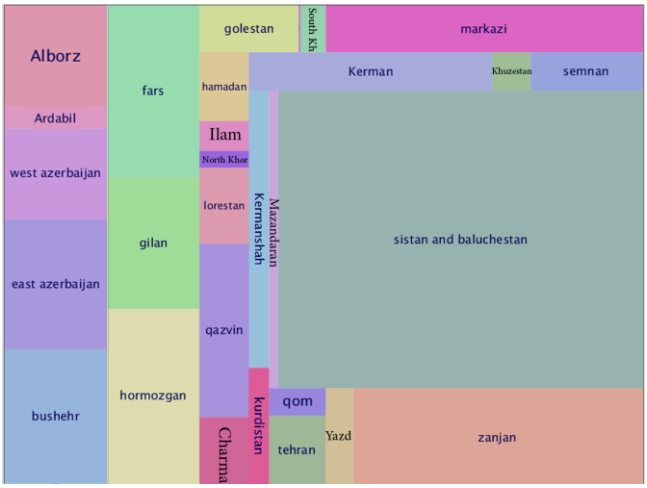


Fig. 13. Users registration based on location.

The exploratory data visualization method helps us to detect influential observations, meaning that this method help us find the provinces that are affecting the social and behavioral patterns. From Fig. 11 and 12 we can see that some provinces had very small percentage of active users. We chose to also get into this problem and see whether it's possible to gather data and study the efficiency of different locations infrastructures based on users need and also benchmarking different mobile operators that our users subscribed to.

Fig. 13 shows a tree map that shows number of users based on their location. We can see that for instance Sistan and Baluchistan has the most number of users that downloaded the app. And Tehran at the bottom of the map doesn't have many downloads for sure. But in Fig. 14 it's clear that Tehran has most active users and Sistan has one of the lowest numbers of active users.

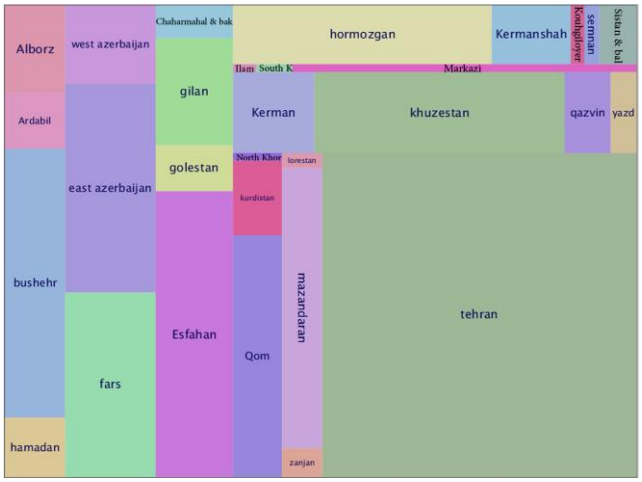


Fig. 14. Active users based on location.

Bear in mind that our users are all using social network and the basis of joining our app is using some kind of social network that need good internet connection. In Fig. 15 it's shown that how users spread across different operators. Despite that Rightel doesn't have large amount of user base but it has the same percent of users using Hamrahaval network that has over 50 million subscribers. Rightel has offered 3G services exclusively for about 3 years, and has been challenging 2G services that the other 2 operators offered.

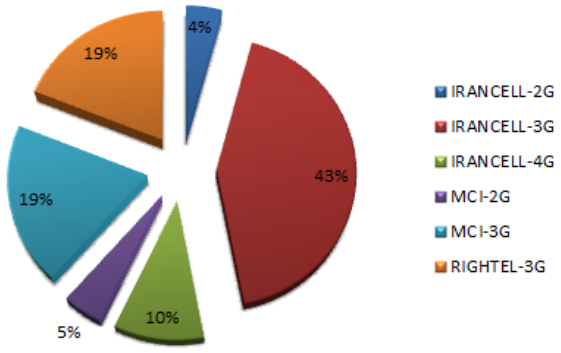


Fig. 15. User distribution by operator and network type.

Fig. 16. User network type.

It has been less than a year since Irancell and Hamrahaval rolled out their 3G network coverage. In Fig. 16 we can see

that most of our users benefit from 3G network and that solely shows how fast mostly young users adopt themselves to the new technology. There is a thirst in Iranian young society to have internet connection with high speed but obviously there are some huge obstacles in the way and Iran remains among the countries that have lowest internet speed. These problems may be solved by placing in some policies by the government or other related organizations but we are not going to discuss these subjects here and it's out of scope of our expertise.

We have also gathered data and calculated data connection speed based on provinces and grouped them by operators. We created radar chart for each operator. All of the three charts are based on the test that has been done in 3G networks. These diagrams are result of over 5000 tests done by users all over the country. Some of provinces may not be in chart that's because we didn't have enough or sufficient data regarding to that location.

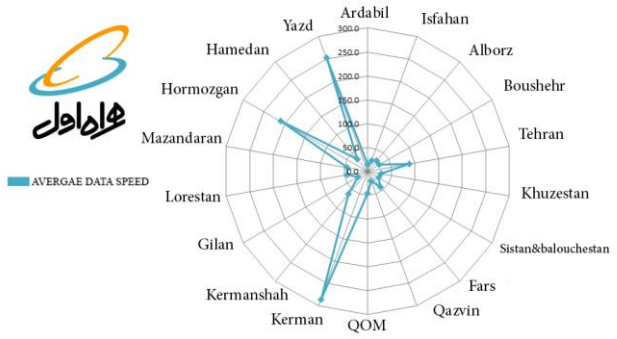


Fig. 17. Mobile data speed by province (Hamrah aval).

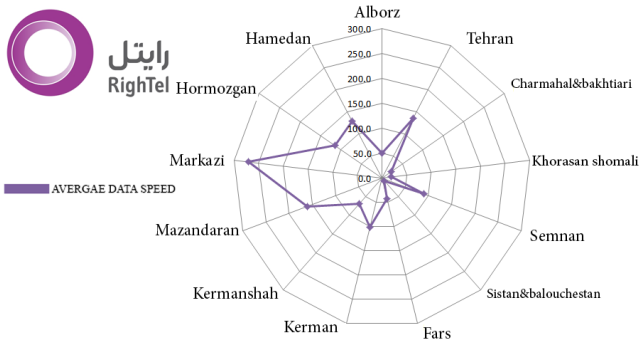


Fig. 18. Mobile Data speed by province (Rightel).

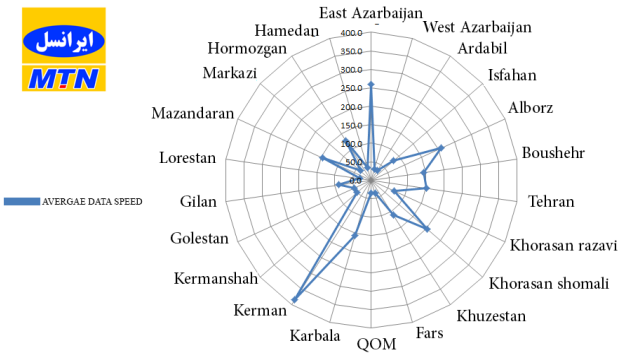


Fig. 19. Mobile data speed by province (Irancell).

As we previously said and showed in Fig. 14 sistan province despite being a province with one of highest registered users, have lowest active users, Fig. 17 and 18 shows that also sistan province has one of the lowest internet connection speed, intentionally these benchmarks that is shown in Fig. 17-19 have been done only in 3G network so that means users should have gotten higher bandwidth.

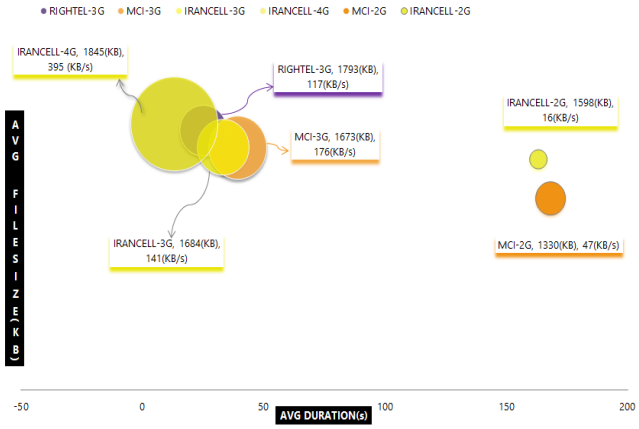


Fig. 20. Comparing operators' mobile data speed by avg speed- avg filesize- avg duration.

Fig. 20 we tried to show average file size, average duration and average internet speed in different networks. The x axis is the average duration of downloads, the y axis is the average file size downloaded by users and the bubble size is the average speed. Each operator has been shown by associated colors. In 2G and 3G network Hamrahaval has the highest speed by average. If we take a look at the violet (Rightel) bubble we can see that the average file size is bigger than other operators that may be the result of subscribers' perception that has been mirrored in different advertisements, that Rightel subscribers are using this operator usually for its data service. In 4G market Irancell is alone for the meantime and seems to be the best solution for subscribers who want fastest mobile data connection but there may be some compromises such as network coverage and pricing as well.

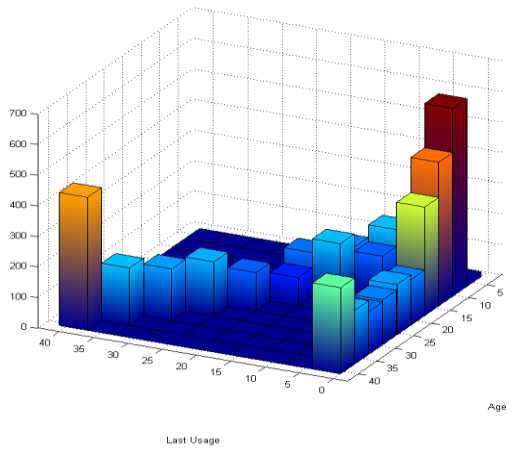


Fig. 21. Histogram of isolated users.

X. Isolated Users

After different studies that we have done on users'

network of relations in this section we are going to calculate correlation for users that are being selected as isolated users [11] [12]. We separated them into two groups Target and source. We made a matrix out of some correlation variables according to Table II such as Age and Usage Count.

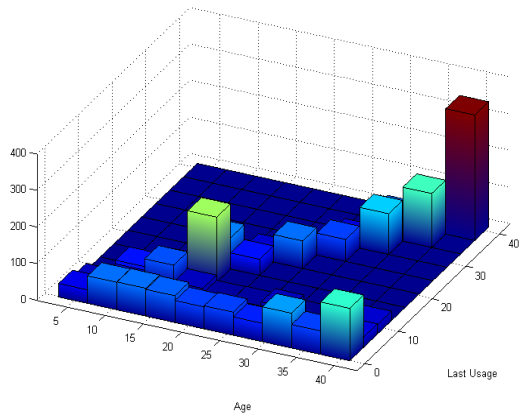


Fig. 22. Histogram of target users.

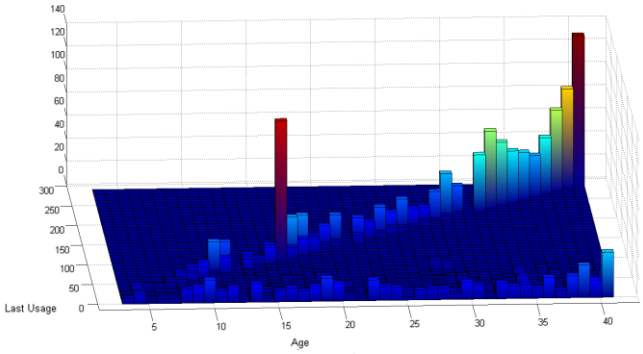


Fig. 23. Histogram of source users.

Comparing Fig. 21-23 we can extract different patterns from users behaviour. Group of users are older and also have high usage count and recently have used the app. Comparing source and target based on the variables extracted and shown Table IV we can see that their behavior is almost totally different from each other. For isolated users seems that last usage and age should have correlation with each other, we are going to check that. Also we can pinpoint some event that has happened in these figures. For instance the peak that shows the time that we sent notification for users or the time that number of user acquisition increased and how could've affected the activity churn and time of releasing a new version.

TABLE IV: CORRELATION VARIABLES

Source		Target		Isolated Users	
Age Source	Usage Count	Age target	Usage Count	Age	Usage Count

TABLE V: RESULT OF CORRELATION CALCULATION

	Source	Target	Isolated
r	-0.0068	0.00821	-0.0452
R^2	0.005	0.0007	0.002

Correlation between isolated users is negative based on the result in Table V, for the target users the correlation value is positive but very low; however these values can change by taking some actions accordingly.

XI. RETENTION RATE

Retention rate means those users that were on the verge of leaving the app or already had left but had returned. We chose two periods one of them from February to September 2014 and the other one from August 2014 to February 2015.

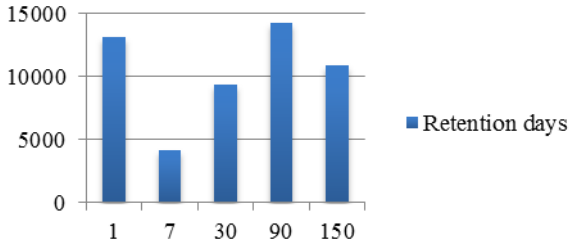


Fig. 24. Retention days between 2/2014 to 9/2014.

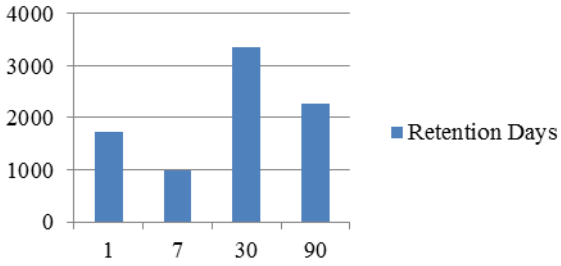


Fig. 25. Retention days between 9/2014 to 2/2015.

In the second period we had put in place some features in app like notification and different tutors so users experience could be improved. Comparing two Fig. 24 and 25 we can see a lot of differences for instance percent of people leaving the app in the first day is hugely reduced but 30 days retention increased compared to 90 days.

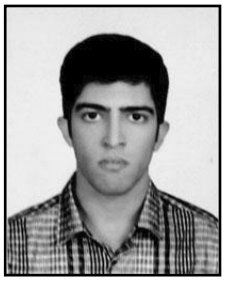
XII. CONCLUSION

We gathered a lot of data and processed most of them. Amount of data exceeded 7 million records. We tried to tell the story of users' behavior through visualization of data. A big part of this research was finding the best way to see the lost dimensions that we need to consider, if we needed to study specific user behavior. Studying users' behavior can help us to predict and maintain desirable events. For instance find the best time to send notification to users, specifically when to send contents that may have higher size than just a text. Finding out what results in high amount of user acquisition and user churn. If there is relation between users' churn and their followers or if a user is active does that mean his/her followers could be as active or not. Gathering analytic data and prepping the raw data for the backend team, that gives inside to peak traffic time and if they need to reroute the traffic or think about changing their server side architecture. We can go on and on about these examples. Matter of the fact is users' behavioral data is very important and the analytical view of them plays a major role in how to make benefit from them.

REFERENCES

[1] "The Mobile Consumer: A Global Snapshot," Nielsen, 2013.
[2] Statistics and facts about Android. (2013). [Online]. Available: http://www.statista.com/topics/876/android/.

[3] M. Richeldi and A. Perrucc, "Churn analysis case study," in *Proc. of the Workshop on Data Mining and Business (DMBiz) at the 9th European Conference on Principles and Practice in Knowledge Discovery in Databases (PKDD)*, Porto, Portugal, 2005.
[4] *Index Statistics of Telecommunications of Iran*, 2014.
[5] J. Golbeck, Introduction to Social Media Investigation, Chapter 15 – Instagram, Syngress, 2015.
[6] C. Smith. By the numbers: Interesting Instagram statistics. (2015). [Online]. Available: http://expandedramblings.com/index.php/important-instagram-stats/3/
[7] A. Backiel, B. Baesens, and G. Claeskens, "Mining telecommunication networks to enhance customer lifetime predictions," *Artificial Intelligence and Soft Computing*, vol. 8468, 2014.
[8] A. Klein, V. Sharmac, and H. Ahlfb, "Social activity and structural centrality in online social networks," *Telematics and Information*, vol. 32, no. 2, 2015.
[9] X. Sun, H. Lin, and K. Xu, "A social network model driven by events and interests," *Expert Systems with Applications*, vol. 42, no. 9.
[10] J. Rana, J. Kristiansson, J. Hallberg, and K. Synnes, "Challenges for mobile social networking applications," *Communications Infrastructure. Systems and Applications in Europe*, vol. 16, 2009.
[11] J. Heidemann and M. Klier, "Identifying key users in online social networks: A PageRank based approach," in *Proc. of the 31th International Conference on Information Systems (ICIS)*, University of Missouri, 2010.
[12] J. Blechar, I. D. Constantiou, and J. Damsgaard, "Understanding behavioural patterns of advanced mobile service users," *Electronic Government*, vol. 93, no. 104, 2006.



Abouzar Abbaspour Ghomi received his B.S. in computer engineering from Azad University of Sari, Iran in 2012 And M.S in software engineering from university of Tehran, Iran in 2015. He is a senior OSS engineer at MCCI. His major interests are working on mobile social network, big data, visualization of data and user behavioral analysis.

Masoud Asadpour received his Ph.D. degree in machine learning and collective robotics, autonomous systems lab at Ecole Polytechnique Fédéral de Lausanne (EPFL), Switzerland in 2006-2007 And postdoc in modular robotics, biorobotics Lab at Ecole Polytechnique Fédéral de Lausanne (EPFL), Switzerland in 2006-2007. He is assistant professor at School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Iran. His current research interests are social networks, complex networks, machine learning and modular robots.