A Project Report on

Facebook Data Exploratory Analysis

By: Team Member Name: **Sneha Godbole**

Email id: **snegodbo@umail.iu.edu**

Table of Contents

[Abstract 2](#_Toc469434256)

[1. Introduction 3](#_Toc469434257)

[Motivation 3](#_Toc469434258)

[Existing Work 3](#_Toc469434259)

[Contribution 4](#_Toc469434260)

[2. Data & Methods 5](#_Toc469434261)

[Data Introduction: Facebook New Orleans Network 5](#_Toc469434262)

[Dataset 5](#_Toc469434263)

[Data statistics 6](#_Toc469434264)

[Data Limitations 6](#_Toc469434265)

[3. Implementation Approaches & Techniques 7](#_Toc469434266)

[Data Preprocessing 7](#_Toc469434267)

[Methods used for Exploratory Analytics / Insights 7](#_Toc469434268)

[4. Results 9](#_Toc469434269)

[Novelty 9](#_Toc469434270)

[Insights & Discussions 9](#_Toc469434271)

[Code Base Details 14](#_Toc469434272)

[5. Conclusion & Future Work 14](#_Toc469434273)

[6. REFERENCES 15](#_Toc469434274)

# **Abstract**

Online social networks have become extremely popular; many well known as well as emerging web sites provides integration with social networks like Facebook, LinkedIn, Twitter and so on, in order for users to share the content. Users of these networks often establish hundreds to even thousands of social links with other users. Many researchers are working on areas to explore use of actual interaction between users, rather than mere friendship – to distinguish between strong and weak links. There are many studies on how an activity network is structurally different from the social network itself. I found a very good research paper in regards with this, studying the evolution of activity between users in the Facebook social network to capture this notion. The paper finds that links in the activity network tend to come and go rapidly over time, and the strength of ties exhibits a general decreasing trend of activity as the social network link ages. For example, only 30% of Facebook user pairs interact consistently from one month to the next. Interestingly, the paper also finds that even though the links of the activity network change rapidly over time, many graph-theoretic properties of the activity network remain unchanged. The overall research and analysis being done on this data looks very exciting to me, and I am going to perform exploratory analysis of the dataset being used for this paper, to validate their results as well as to find few more interesting patterns / insights from the dataset based on concepts learned on Network Science course so far.

# **Introduction**

## **Motivation**

Online social networks have become a popular way for users to connect, express themselves, and share content. Popular social networks like Facebook, LinkedIn have hundreds of millions of registered users and are growing at a rapid pace. As these networks grow and mature, users have been observed to form hundreds to even thousands of friendship links.

For example, I have around 800+ active friends on Facebook. However, I do not actively interact with all of them every month! Some of them are my best friends, some average friends and some just acquaintances. Our ties too vary, from very strong to very weak. The interactions / activities we do (sharing, liking post, writing on wall, etc.) with them is dependent on the type of tie often times. So, apart from a connectivity graph – friends’ connection graph – the activity linkage / graph is very important in order to understand the actual “actual” / “active” friends among my circle. Some research work mentions this network as “activity network” which has interesting properties / patterns as compared to “connectivity networks” (example: being a friend on Facebook or LinkedIn).

I found an interesting research paper, from [Research Papers Link](https://icon.colorado.edu/#!/networks), under social network domain, studying the evolution of activity between users in the Facebook social network to capture this notion. The overall research and analysis being done on this data looks very exciting to me, and I am going to perform exploratory analysis of the dataset being used for this paper, to validate their results as well as to find few more interesting patterns / insights from the dataset based on concepts learned on Network Science course so far.

## **Existing Work**

An interesting research paper and literature review which gave me an aspiration is available @ [https://icon.colorado.edu/#!/networks](https://icon.colorado.edu/) , under Social network domain category, named Facebook Wall Posts 2009.

A research paper named  **"On the evolution of user interaction in Facebook”,** published in WOSN’09, August 17, 2009, Barcelona, Spain, is available @ <http://www.mpi-sws.org/~gummadi/papers/wosn23-viswanath.pdf> .

The paper finds that links in the activity network tend to come and go rapidly over time, and the strength of ties exhibits a general decreasing trend of activity as the social network link ages. For example, only 30% of Facebook user pairs interact consistently from one month to the next.

Interestingly, the paper also finds that even though the links of the activity network change rapidly over time, many graph-theoretic properties of the activity network remain unchanged.

## **Contribution**

I did study the paper I used for implementation, along with the other papers referred in the paper to understand what research and insights has been proposed by the authors. I studied the connection and activity datasets and came up with the scope of my project work.

I could successfully implement and verify the research work insights mentioned in the referred paper, the details about this work are present in [data & methods section](#_Data_&_Methods).

I also proposed, designed and implemented **3 new additional insights** on the same datasets as follows –

1. Most popular user in the given activity network
2. Most active user in the given activity network
3. Number of wall post per year by all the users.

I found interesting insights and patterns on all these **3 new insights** and thedetails are present in [insights & discussions](#_Insights_&_Discussions) section.

# **Data & Methods**

## **Data Introduction: Facebook New Orleans Network**

The research paper focuses on the New Orleans regional network in Facebook for the study. The data is collected over two separate periods. First, during December 29th, 2008 and January 3rd, 2009. In this period, the information about friendship links (i.e., who is friends with whom) was collected from a single user in the New Orleans network and visited all friends of the user and their friends in a breadth-first-search (BFS) fashion, consistent with previous social network crawls. Second, the wall feature in Facebook as a form of user interaction in the social network were crawled. In Facebook, a user’s friends can post comments to the user’s wall; these comments appear on the user’s wall and can be seen by others who visit the user’s profile. Therefore, wall posts represent a broadcast-style messaging service within the site. This second crawl between January 20th, 2009 and January 22nd, 2009, and had downloaded the entire wall history for all users whom were previously discovered in earlier crawl. Each wall post entry in the data set contains information about the wall owner, the user who made the post, the time at which the post was made, and the post content.

## **Dataset**

The dataset which was used for this research work is available at [Dataset Link](http://konect.uni-koblenz.de/networks/facebook-wosn-wall) , [Dataset1](http://socialnetworks.mpi-sws.org/data-wosn2009.html), and I will use same dataset for my validations as well as experiments.

There are two datasets which I have used for this project Connection(Friendship) Dataset and Activity Dataset

1. **Connection dataset** is formed from undirected network which contains friendship data of Facebook users. A node represents a user and an edge represents a friendship between two users.

This dataset contains three columns **node1, node2 and Timestamp** so here node1 and node2 are users and timestamp shows their connection time.

I have used this connection dataset to calculate node degrees and found out pattern of degree distribution happened in this network. Also, I have used this dataset to get sampled user pairs to calculate interaction rate.

1. **Activity dataset** is formed from the directed network of a small subset of posts to another user's wall on Facebook. The nodes of the network are Facebook users, and each directed edge represents one post, linking the users writing a post to the users whose wall the post is written on**.**

Since users may write multiple posts on a wall, the network allows multiple edges connecting a single node pair. Since users may write on their own wall, the network contains loops.

This dataset contains three columns **Source, Target and Timestamp** so here when source node post something on Target nodes wall then the interaction happens and that time is recorded.

Using this dataset, I was able to calculate Popular Users (Who got most incoming posts) and Active Users (Who initiated most number of wall interactions).I have also calculated number of wall interactions happened every year and interaction rate per year.

## **Data statistics**

The total data gathered information about 90,269 users and 3,646,662 friendship links between those users. This accounts for 52% of the users in the New Orleans network based on the statistics provided by Facebook. However, not all users make their profiles public; this data could view the wall postings of 60,290 (66.7%) users. These 60,290 users were connected together by 1,545,686 links in the social network with an average node degree of 25.3. In the paper I studied and the implementation I am targeting to do, I consider only this set of users. The wall post data spans from September 26th, 2006 to January 22nd, 2009. In total, 838,092 wall posts were observed, for an average of 13.9 wall posts per user. This covers communication between 188,892 distinct pairs of users, representing 12.2% of the links in the social network. The remaining 87.8% of the link in the social network did not exhibit any wall activity.

## **Data Limitations**

First, in Facebook, users can interact in many ways (e.g., messaging, applications, photo uploads, and chat). While wall posting is one of the most popular methods of user inter- actions, I do not know if it is representative of other forms of interaction. Second, our Facebook crawl was limited to the subset of people who made their profiles visible to the people in the same regional network, which is the default privacy setting in Facebook. Third, the data collected is only for users in the giant connected component of the New Orleans networks.

# **Implementation Approaches & Techniques**

## **Data Preprocessing**

1. As a requirement of doing exploratory analytics by year / month / date level, I standardized all the timestamp values in the given dataset into “yyyy-mm-dd” format. This was achieved by implementing a custom function and using it on entire dataset.
2. I used step #1 to populate two extra columns namely “year” and “month”, for each row in a given dataset.
3. Result of steps #1 and #2 was stored in the form of DataFrame and was used for all further analytics / insights exploration

## **Methods used for Exploratory Analytics / Insights**

I have validated the research and literature survey published in [Paper](http://www.mpi-sws.org/~gummadi/papers/wosn23-viswanath.pdf). Implementation led to interesting exploratory analysis and I have got some meaningful insights from the friendship(connection) as well as activity network.

I have visualized following meaningful insights from the data after implementation –

1. **Most popular user in activity network.**

I analyzed activity network to get meaningful insights on who is a more popular user in this friendship(Connection) network, based on user activities (wall posts). I built a network graph for activity network and found out user/s with most incoming walls for this insight.

**Approach**:

Popular user is the user with most number of incoming wall posts means we have to select Target node which has most wall interactions as this is directed network. So, I grouped all Target nodes based on month because dataset was too huge and I found out that in 2008 rate of interactions increased so I had selected 2008 as a year.

I plotted this insight using a bar chart for each month of year 2008 and the details are present in following section.

1. **Most active user in activity network.**

I analyzed the activity network to get meaningful insights on who is a more active user in this friendship(Connection) network, based on user activities (wall posts). I built a network graph for activity network and found out users with most number of Facebook wall interactions initiated by a user.

**Approach:**

To get number of active users I grouped all the nodes who initiated wall interactions i.e. Source. As the dataset was very huge and I was interested to see active users in 2008 because 2008 year is special because Facebook launched new design that year so I had selected year 2008 and plotted active users on bar chart for each month and the details are present in following section.

1. **Number of wall post per month/year by all the users.**

From Activity network, I have calculated number of wall posts for each month of corresponding year. I choose data from 2005 because number of wall posts increased from 2005(got this result in #6 below).

I plotted this pattern in line chart, each line indicate year and shows how wall interactions increased or decreased month by month. I figured out that in particular months’ number of posts spiked it may be because of birthdays or anniversaries.

1. **Calculated following network properties for Friendship Network**
   1. **Degree Distribution:**

I wanted to see friendship links between users and observed that each user has average 22 friends. When I plotted the degree distribution graph, I got power law distribution of node degrees’ results are available in following section.

* 1. **Clustering Coefficient**

I have calculated clustering coefficient of the graph and it is 0. 1975.This is very low than I expected, which shows nodes are very sparsely connected to each other.

1. **Rate of interaction over time based on fraction of wall posts sent since link formation between / among users.**

I analyzed the activity and connection network to figure out how does the rate of interactions varied over time, since the users got connected to each other.

* **Data Sampling** 
  + 1. In order to meet processing time / computing requirements (using my single node laptop for execution), I sampled 1000 connections from given connection dataset.
    2. I then populated appropriate activity dataset for these 1000 connections.
* Then I analyzed the connection time & wall interactions information, to find a pattern on how does the rate of interactions varied over time.
* The pattern was categorized based on year.
* I plotted this pattern using bar chart, and the results are available in following section.

1. **Number of wall post per year by all the users.**

To implement this, I used the timestamp of wall posts for a given set of users, to examine the growth of user interactions over time by analyzing wall interactions for all users and then categorized the interactions every year.

The insight here is a pattern of activity rate over the period of time.

I plotted this pattern using line chart and the result is available in following section.

# **Results**

## **Novelty**

I proposed, designed and implemented **3 new additional insights** on the same datasets as follows –

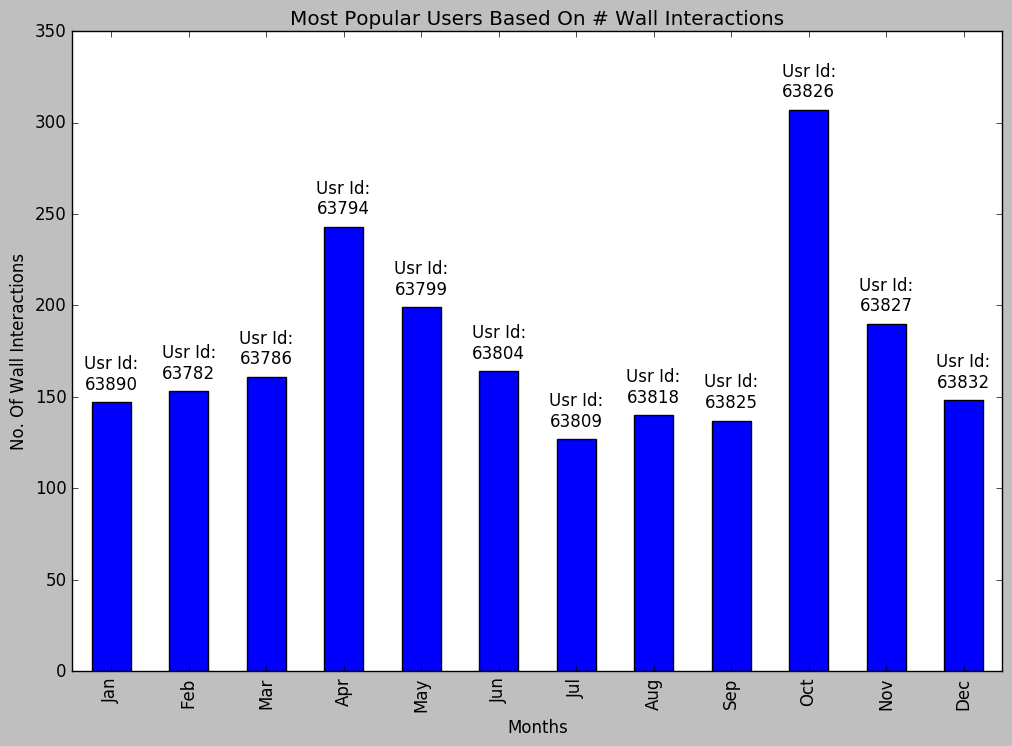
1. Most popular user in the given activity network
2. Most active user in the given activity network
3. Number of wall post per year by all the users.

I found interesting insights and patterns on all these **3 new insights** and thedetails are present in section below.

## **Insights & Discussions**

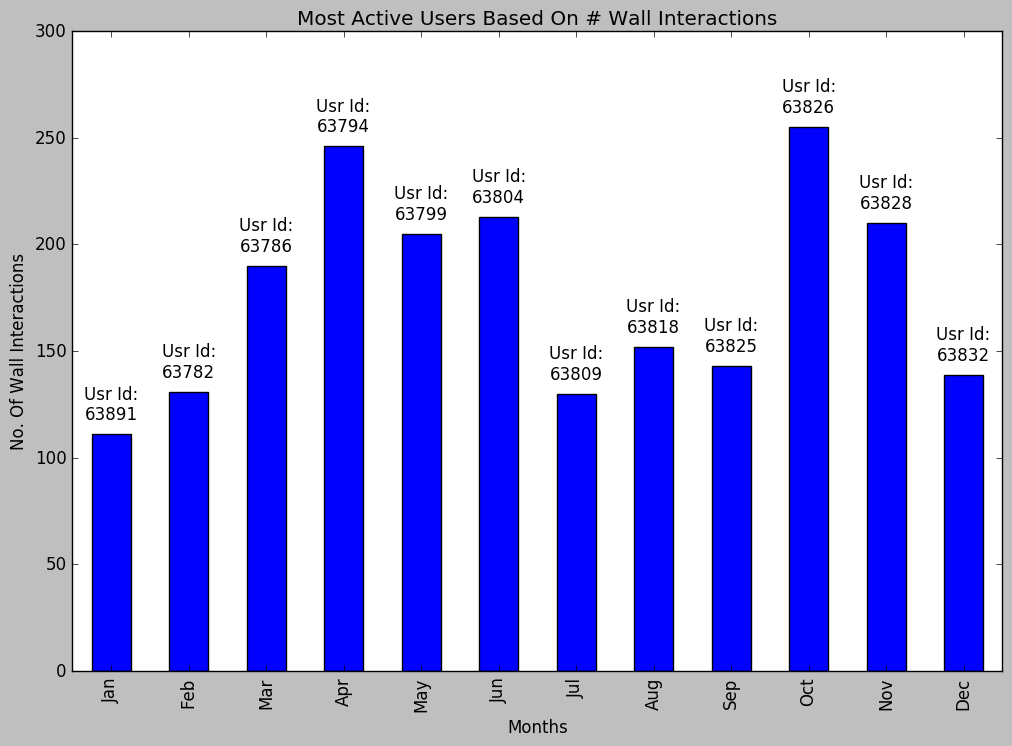
In this section, we will go over all the exploratory analytics / insights I have found after implementation –

1. In 2008, based on activity network, User Id 63826 is most popular user and total maximum incoming wall interactions per user is ~300.

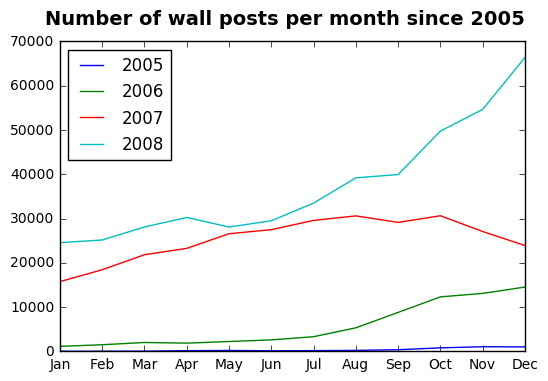


In 2008, based on activity network, User ID 63826 is most active user and total interactions happened are ~250.

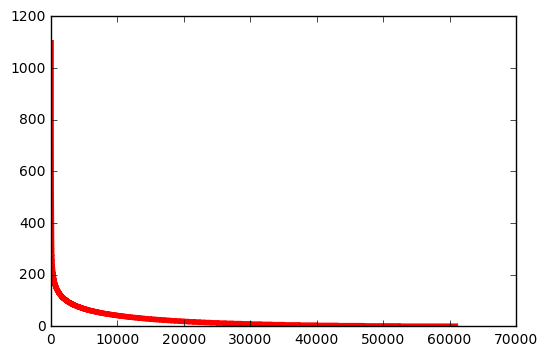
If we closely see the most active and most popular users graph, we see that many “**most active**” users are “**most popular**” and vice versa. This is correct as active users are popular, and popular users are active



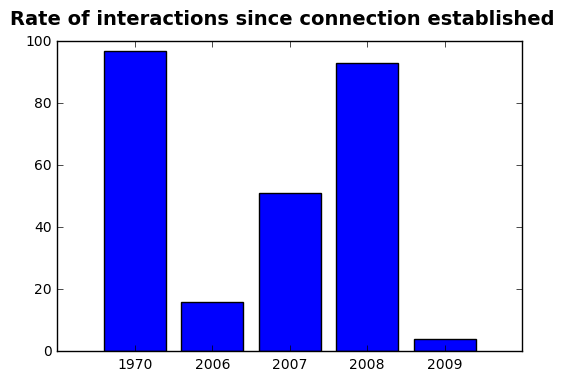
1. This graph shows in 2005 number of wall interactions were very low and interactions increased from 2007 and 2008.In 2008 wall interactions spiked to highest point because of Facebook’s new design launch.



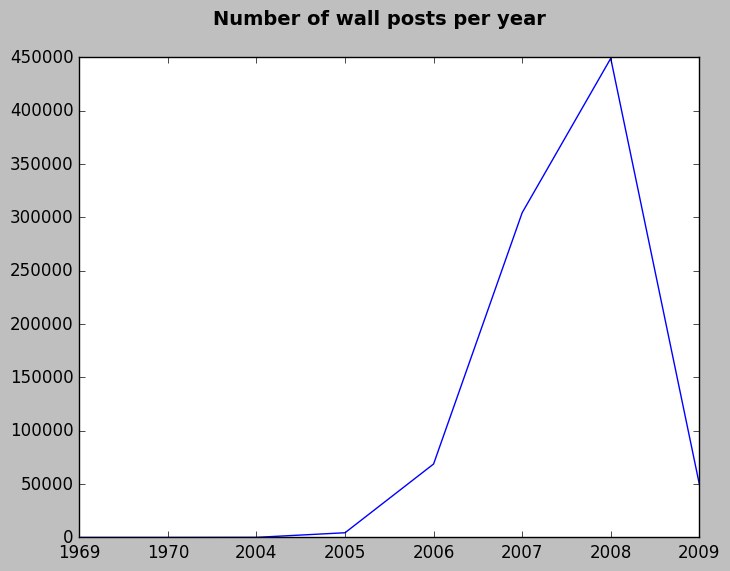
1. Degree distribution of connection network follows power law distribution. Some nodes have more than 1000 friends and some have less than 5 or no friends.



1. This bar chart shows number of interactions happened since 1970 because I have sampled users who connected each other in 1970 so sampled users who started connection in 1970 had lot of interaction in that year but after that they had no interactions in between years 1970 and 2006.Suddenly from 2006 t0 2008 interaction rate increased and in 2009 it again dropped to its lowest.



1. In 2nd result I have plotted wall post by month for each year and here I have plotted them yearly. In this diagram we can see number of wall posts per year increased drastically from 2007 to 2008 and again decreased from 2008 to 2009.



## **Code Base Details**

The code base for this implementation is available [here](https://github.com/snehagodbole/Project-FB-Data-Exploratory-Analysis)

# **Conclusion & Future Work**

* I could successfully implemented, and verified most of the research work, insights and literature survey published in <http://www.mpi-sws.org/~gummadi/papers/wosn23-viswanath.pdf> paper.
* The implementation led to interesting exploratory analysis and I got many meaningful insights from the friendship as well as activity network. I found the overall end to end experience very exciting and interesting. I could relate network science academics with the real use cases after this successful implementation.
* One of the future work item could be to implement and validate remaining insights proposed in the referred paper / research work.
* Another future work item could be to apply these explorations and techniques to my Facebook connection and activity graph and find the interesting patterns / insights!

# **REFERENCES**

[1]. [B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi. "On the evolution of user interaction in Facebook." Proc. 2nd ACM SIGCOMM Workshop on Social Networks, 37-42 (2009)](http://www.mpi-sws.org/~gummadi/papers/wosn23-viswanath.pdf)

[2]. <https://www.cs.ucsb.edu/~ravenben/publications/pdf/interaction-eurosys09.pdf>

[3]. <https://www.cs.ucsb.edu/~ravenben/publications/pdf/graphs-tweb12.pdf>

[4]. <https://docs.python.org/3/tutorial/>

[5]. <http://pandas.pydata.org/>

[6]. <http://www.google.com>