

Social network Graph Link Prediction - Facebook Challenge

```
In [1]: #Importing Libraries
# please do go through this python notebook:
import warnings
warnings.filterwarnings("ignore")

import csv
import pandas as pd#pandas to create small dataframes
import datetime #Convert to unix time
import time #Convert to unix time
# if numpy is not installed already : pip3 install numpy
import numpy as np#Do arithmetic operations on arrays
# matplotlib: used to plot graphs
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns#Plots
from matplotlib import rcParams#Size of plots
from sklearn.cluster import MiniBatchKMeans, KMeans#Clustering
import math
import pickle
import os
# to install xgboost: pip3 install xgboost
import xgboost as xgb

import warnings
import networkx as nx
\
import pdb
import pickle
from pandas import HDFStore,DataFrame
from pandas import read_hdf
from scipy.sparse.linalg import svds, eigs
import gc
from tqdm import tqdm
```

1. Reading Data

```
In [2]: if os.path.isfile('train_pos_after_eda.csv'):
    train_graph=nx.read_edgelist('train_pos_after_eda.csv',delimiter=',',create_using=nx.DiGraph())
    print(nx.info(train_graph))
else:
    print("please run the FB_EDA.ipynb or download the files from drive")
```

DiGraph with 1780722 nodes and 7550015 edges

2. Similarity measures

2.1 Jaccard Distance:

<http://www.statisticshowto.com/jaccard-index/> (<http://www.statisticshowto.com/jaccard-index/>)

$$j = \frac{|X \cap Y|}{|X \cup Y|}$$

```
In [3]: #for followees
def jaccard_for_followees(a,b):
    try:
        if len(set(train_graph.successors(a))) == 0 | len(set(train_graph.successors(b))) == 0:
            return 0
        sim = (len(set(train_graph.successors(a))).intersection(set(train_graph.successors(b)))) / (len(set(train_graph.successors(a))).union(set(train_graph.successors(b))))
    except:
        return 0
    return sim
```

```
In [4]: #one test case
print(jaccard_for_followees(273084,1505602))
```

0.0

```
In [5]: #node 1635354 not in graph
print(jaccard_for_followees(273084,1505602))
```

0.0

```
In [6]: #for followers
def jaccard_for_followers(a,b):
    try:
        if len(set(train_graph.predecessors(a))) == 0 | len(set(g.predecessors(b))) == 0:
            return 0
        sim = (len(set(train_graph.predecessors(a))).intersection(set(train_graph.predecessors(b)))) / (len(set(train_graph.predecessors(a))).union(set(train_graph.predecessors(b))))
    except:
        return 0
    return sim
```

```
In [7]: print(jaccard_for_followers(273084,470294))
```

0

```
In [8]: #node 1635354 not in graph
print(jaccard_for_followees(669354,1635354))
```

0

2.2 Cosine distance

$$\text{CosineDistance} = \frac{|X \cap Y|}{|X| \cdot |Y|}$$

```
In [9]: #for followees
def cosine_for_followees(a,b):
    try:
        if len(set(train_graph.successors(a))) == 0 | len(set(train_graph.successors(b))) == 0:
            return 0
        sim = (len(set(train_graph.successors(a)).intersection(set(train_graph.successors(b)))) / (math.sqrt(len(set(train_graph.successors(a)))) * math.sqrt(len(set(train_graph.successors(b))))))
        return sim
    except:
        return 0
```

```
In [10]: print(cosine_for_followees(273084,1505602))
```

0.0

```
In [11]: print(cosine_for_followees(273084,1635354))
```

0

```
In [12]: def cosine_for_followers(a,b):
    try:
```

```
        if len(set(train_graph.predecessors(a))) == 0 | len(set(train_graph.predecessors(b))) == 0:
            return 0
        sim = (len(set(train_graph.predecessors(a)).intersection(set(train_graph.predecessors(b)))) / (math.sqrt(len(set(train_graph.predecessors(a)))) * math.sqrt(len(set(train_graph.predecessors(b)))))))
        return sim
    except:
        return 0
```

```
In [13]: print(cosine_for_followers(2,470294))
```

0.02886751345948129

```
In [14]: print(cosine_for_followers(669354,1635354))
```

0

3. Ranking Measures

https://networkx.github.io/documentation/networkx-1.10/reference/generated/networkx.algorithms.link_analysis.page_rank_alg.page_rank.html
[\(https://networkx.github.io/documentation/networkx-1.10/reference/generated/networkx.algorithms.link_analysis.page_rank_alg.page_rank.html\)](https://networkx.github.io/documentation/networkx-1.10/reference/generated/networkx.algorithms.link_analysis.page_rank_alg.page_rank.html)

PageRank computes a ranking of the nodes in the graph G based on the structure of the incoming links.



Mathematical PageRanks for a simple network, expressed as percentages. (Google uses a logarithmic scale.) Page C has a higher PageRank than Page E, even though there are fewer links to C; the one link to C comes from an important page and hence is of high value. If web surfers who start on a random page have an 85% likelihood of choosing a random link from the page they are currently visiting, and a 15% likelihood of jumping to a page chosen at random from the entire web, they will reach Page E 8.1% of the time. (**The 15% likelihood of jumping to an arbitrary page corresponds to a damping factor of 85%.**) Without damping, all web surfers would eventually end up on Pages A, B, or C, and all other pages would have PageRank zero. In the presence of damping, Page A effectively links to all pages in the web, even though it has no outgoing links of its own.

3.1 Page Ranking

<https://en.wikipedia.org/wiki/PageRank> (<https://en.wikipedia.org/wiki/PageRank>)

4. Other Graph Features

4.1 Shortest path:

Getting Shortest path between two nodes, if nodes have direct path i.e directly connected then we are removing that edge and calculating path.

```
In [15]: #if has direct edge then deleting that edge and calculating shortest path
def compute_shortest_path_length(a,b):
    p=-1
    try:
        if train_graph.has_edge(a,b):
            train_graph.remove_edge(a,b)
            p= nx.shortest_path_length(train_graph,source=a,target=b)
            train_graph.add_edge(a,b)
        else:
            p= nx.shortest_path_length(train_graph,source=a,target=b)
    return p
except:
    return -1
```

```
In [16]: #testing
compute_shortest_path_length(77697, 826021)
```

Out[16]: 10

```
In [17]: #testing
compute_shortest_path_length(669354,1635354)
```

Out[17]: -1

4.2 Checking for same community

```
In [18]: #getting weekly connected edges from graph
wcc=list(nx.weakly_connected_components(train_graph))
def belongs_to_same_wcc(a,b):
    index = []
    if train_graph.has_edge(b,a):
        return 1
    if train_graph.has_edge(a,b):
        for i in wcc:
            if a in i:
                index= i
                break
        if (b in index):
            train_graph.remove_edge(a,b)
            if compute_shortest_path_length(a,b)===1:
                train_graph.add_edge(a,b)
                return 0
            else:
                train_graph.add_edge(a,b)
                return 1
        else:
            return 0
    else:
        for i in wcc:
            if a in i:
                index= i
                break
        if(b in index):
            return 1
        else:
            return 0
```

```
In [19]: belongs_to_same_wcc(861, 1659750)
```

```
Out[19]: 0
```

```
In [20]: belongs_to_same_wcc(669354,1635354)
```

```
Out[20]: 0
```

4.3 Adamic/Adar Index:

Adamic/Adar measures is defined as inverted sum of degrees of common neighbours for given two vertices.

$$A(x, y) = \sum_{u \in N(x) \cap N(y)} \frac{1}{\log(|N(u)|)}$$

```
In [21]: #adar index
def calc_adar_in(a,b):
    sum=0
    try:
        n=list(set(train_graph.successors(a)).intersection(set(train_graph.successors(b))))
        if len(n)!=0:
            for i in n:
                sum=sum+(1/np.log10(len(list(train_graph.predecessors(i))))))
    return sum
    else:
        return 0
    except:
        return 0
```

In [22]: calc_adar_in(1,189226)

Out[22]: 0

In [23]: calc_adar_in(669354,1635354)

Out[23]: 0

4.4 Is person was following back:

```
In [24]: def follows_back(a,b):
    if train_graph.has_edge(b,a):
        return 1
    else:
        return 0
```

In [25]: follows_back(1,189226)

Out[25]: 1

In [26]: follows_back(669354,1635354)

Out[26]: 0

4.5 Katz Centrality:

[https://en.wikipedia.org/wiki/Katz_centrality_\(https://en.wikipedia.org/wiki/Katz_centrality\).](https://en.wikipedia.org/wiki/Katz_centrality_(https://en.wikipedia.org/wiki/Katz_centrality).)

<https://www.geeksforgeeks.org/katz-centrality-centrality-measure/>

(<https://www.geeksforgeeks.org/katz-centrality-centrality-measure/>) Katz centrality computes the centrality for a node based on the centrality of its neighbors. It is a generalization of the eigenvector centrality. The Katz centrality for node i is

$$x_i = \alpha \sum_j A_{ij} x_j + \beta,$$

where A is the adjacency matrix of the graph G with eigenvalues

$$\lambda$$

The parameter

$$\beta$$

controls the initial centrality and

$$\alpha < \frac{1}{\lambda_{max}}.$$

```
In [27]: if not os.path.isfile('katz.p'):
    katz = nx.katz.katz_centrality(train_graph,alpha=0.005,beta=1)
    pickle.dump(katz,open('katz.p','wb'))
else:
    katz = pickle.load(open('katz.p','rb'))
```

```
In [28]: print('min',katz[min(katz, key=katz.get)])
print('max',katz[max(katz, key=katz.get)])
print('mean',float(sum(katz.values())) / len(katz))
```

min 0.0007313532484065916
max 0.003394554981699122
mean 0.0007483800935562018

```
In [29]: mean_katz = float(sum(katz.values())) / len(katz)
print(mean_katz)
```

0.0007483800935562018

4.6 Hits Score

The HITS algorithm computes two numbers for a node. Authorities estimates the node value based on the incoming links. Hubs estimates the node value based on outgoing links.

https://en.wikipedia.org/wiki/HITS_algorithm (https://en.wikipedia.org/wiki/HITS_algorithm)

```
In [30]: if not os.path.isfile('hits.p'):
    hits = nx.hits(train_graph, max_iter=100, tol=1e-08, nstart=None, normalized=True)
    pickle.dump(hits,open('hits.p','wb'))
else:
    hits = pickle.load(open('hits.p','rb'))
```

```
In [31]: print('min',hits[0][min(hits[0], key=hits[0].get)])
print('max',hits[0][max(hits[0], key=hits[0].get)])
print('mean',float(sum(hits[0].values())) / len(hits[0]))
```

min -5.107494782050822e-20
max 0.004868653379538982
mean 5.615699699308675e-07

5. Featurization

5. 1 Reading a sample of Data from both train and test

```
In [32]: ! gdown --id 1lczxVZ0-MkPmoH3lS35Q8rRfrecKSXb1
! gdown --id 1_KN7S8zfHdkRjRYOEtBxBVq8JrGxPXD
```

'gdown' is not recognized as an internal or external command,
operable program or batch file.

'gdown' is not recognized as an internal or external command,
operable program or batch file.

```
In [33]: import random
if os.path.isfile('train_after_eda.csv'):
    filename = "train_after_eda.csv"
    # you uncomment this line, if you dont know the length of the file name
    # here we have hardcoded the number of lines as 15100030
    # n_train = sum(1 for line in open(filename)) #number of records in file (excluding header)
    n_train = 15100028
    s = 100000 #desired sample size
    skip_train = sorted(random.sample(range(1,n_train+1),n_train-s))
    #https://stackoverflow.com/a/22259008/4084039
```

```
In [34]: if os.path.isfile('train_after_eda.csv'):
    filename = "test_after_eda.csv"
    # you uncomment this line, if you dont know the length of the file name
    # here we have hardcoded the number of lines as 3775008
    # n_test = sum(1 for line in open(filename)) #number of records in file (excluding header)
    n_test = 3775006
    s = 50000 #desired sample size
    skip_test = sorted(random.sample(range(1,n_test+1),n_test-s))
    #https://stackoverflow.com/a/22259008/4084039
```

```
In [35]: print("Number of rows in the train data file:", n_train)
print("Number of rows we are going to eliminate in train data are",len(skip_train))
print("Number of rows in the test data file:", n_test)
print("Number of rows we are going to eliminate in test data are",len(skip_test))
```

Number of rows in the train data file: 15100028

Number of rows we are going to eliminate in train data are 15000028

Number of rows in the test data file: 3775006

Number of rows we are going to eliminate in test data are 3725006

```
In [36]: #https://drive.google.com/file/d/19mviN_yeJIfakb4kU5NfKdQl0QtaQ-kH/view?usp=sharing
!gdown --id 19mviN_yeJIfakb4kU5NfKdQl0QtaQ-kH
```

'gdown' is not recognized as an internal or external command,
operable program or batch file.

In [37]: `#https://drive.google.com/file/d/1H6qybuXr8i_USWu3k3uLXEOurc-SELUh/view?usp=sharing
!gdown --id 1H6qybuXr8i_USWu3k3uLXEOurc-SELUh`

'gdown' is not recognized as an internal or external command,
operable program or batch file.

In [38]: `df_final_train = pd.read_csv('train_after_eda.csv', skiprows=skip_train, names=['source_node', 'destination_node', 'indicator_link'])
df_final_train['indicator_link'] = pd.read_csv('train_y.csv', skiprows=skip_train, names=['label'])
print("Our train matrix size ", df_final_train.shape)
df_final_train.head(2)`

Our train matrix size (100002, 3)

Out[38]:

	source_node	destination_node	indicator_link
0	273084	1505602	1
1	625125	643419	1

In [39]: `df_final_test = pd.read_csv('test_after_eda.csv', skiprows=skip_train, names=['source_node', 'destination_node', 'indicator_link'])
df_final_test['indicator_link'] = pd.read_csv('test_y.csv', skiprows=skip_train, names=['label'])
print("Our train matrix size ", df_final_test.shape)
df_final_test.head(2)`

Our train matrix size (24916, 3)

Out[39]:

	source_node	destination_node	indicator_link
0	848424	784690	1
1	106849	1056866	1

5.2 Adding a set of features

we will create these each of these features for both train and test data points

1. jaccard_followers
2. jaccard_followees
3. cosine_followers
4. cosine_followees
5. num_followers_s
6. num_followees_s
7. num_followers_d
8. num_followees_d
9. inter_followers
10. inter_followees

```
In [40]: def compute_features_stage1(df_final):
    #calculating no of followers followees for source and destination
    #calculating intersection of followers and followees for source and destination
    num_followers_s=[]
    num_followees_s=[]
    num_followers_d=[]
    num_followees_d=[]
    inter_followers=[]
    inter_followees=[]
    for i, row in df_final.iterrows():
        try:
            s1=set(train_graph.predecessors(row['source_node']))
            s2=set(train_graph.successors(row['source_node']))
        except:
            s1 = set()
            s2 = set()
        try:
            d1=set(train_graph.predecessors(row['destination_node']))
            d2=set(train_graph.successors(row['destination_node']))
        except:
            d1 = set()
            d2 = set()
        num_followers_s.append(len(s1))
        num_followees_s.append(len(s2))

        num_followers_d.append(len(d1))
        num_followees_d.append(len(d2))

        inter_followers.append(len(s1.intersection(d1)))
        inter_followees.append(len(s2.intersection(d2)))

    return num_followers_s, num_followees_s, num_followers_d, num_followees_d, inter
```

```
In [41]: if not os.path.isfile('storage_sample_stage1.h5'):
    df_final_train['num_followers_s'], df_final_train['num_followers_d'], \
    df_final_train['num_followees_s'], df_final_train['num_followees_d'], \
    df_final_train['inter_followers'], df_final_train['inter_followees']= compute_features_stage1(df_final_train)

    df_final_test['num_followers_s'], df_final_test['num_followers_d'], \
    df_final_test['num_followees_s'], df_final_test['num_followees_d'], \
    df_final_test['inter_followers'], df_final_test['inter_followees']= compute_features_stage1(df_final_test)

    hdf = HDFStore('storage_sample_stage1.h5')
    hdf.put('train_df',df_final_train, format='table', data_columns=True)
    hdf.put('test_df',df_final_test, format='table', data_columns=True)
    hdf.close()
else:
    df_final_train = read_hdf('storage_sample_stage1.h5', 'train_df', mode='r')
    df_final_test = read_hdf('storage_sample_stage1.h5', 'test_df', mode='r')
```

```
In [42]: df_final_train.columns.values
```

```
Out[42]: array(['source_node', 'destination_node', 'indicator_link',
       'num_followers_s', 'num_followers_d', 'num_followees_s',
       'num_followees_d', 'inter_followers', 'inter_followees'],
      dtype=object)
```

```
In [43]: df_final_train.head()
```

```
Out[43]:
```

	source_node	destination_node	indicator_link	num_followers_s	num_followers_d	num_followees_s	num_followees_d
0	273084	1505602	1	11	15	1	1
1	132660	449350	1	4	4	0	0
2	733044	1115294	1	2	3	0	0
3	1368835	1397351	1	0	1	0	0
4	644613	1689027	1	16	14	0	0

Adding Preferential Attachment feature. It is one of the well known concept in social networks. It says that the user with many followers / friends tends to create more followers/friends later in future. It is same as saying rich gets richer.

$$\text{Score}(x,y) = |\Gamma(x)| * |\Gamma(y)|$$

```
In [44]: if not os.path.isfile('storage_sample_stage5.h5'):
    ...
    the file storage_sample_stage5.h5 is not present; so we are creating it in he
    We will store the values in train and test files.
    ...

#for train
df_final_train['preferential_attachement_followers_train']=df_final_train['nu

#for test
df_final_test['preferential_attachement_followers_test']=df_final_test['num_+'

hdf = HDFStore('storage_sample_stage5.h5') #to store data
hdf.put('train_df',df_final_train, format='table', data_columns=True) #in tra
hdf.put('test_df',df_final_test, format='table', data_columns=True) #in test
hdf.close()
else:
    ...
    reading if the file storage_sample_stage5.h5 is present
    ...
df_final_train = read_hdf('storage_sample_stage5.h5', 'train_df', mode='r')
df_final_test = read_hdf('storage_sample_stage5.h5', 'test_df', mode='r')
```

In [45]: df_final_train.head(3)

Out[45]:

	source_node	destination_node	indicator_link	num_followers_s	num_followers_d	num_followees
0	273084	1505602	1	11	15	
1	132660	449350	1	4	4	
2	733044	1115294	1	2	3	

In [49]: df_final_test.head(3)

Out[49]:

	source_node	destination_node	indicator_link	num_followers_s	num_followers_d	num_followees
0	848424	784690	1	6	6	
1	1227220	827479	1	5	0	
2	788249	1709436	1	2	57	

```
In [50]: df_final_train.columns.values
```

```
Out[50]: array(['source_node', 'destination_node', 'indicator_link',
       'num_followers_s', 'num_followers_d', 'num_followees_s',
       'num_followees_d', 'inter_followers', 'inter_followees',
       'preferential attachment followers train'], dtype=object)
```

```
In [51]: df_final_test.columns.values
```

```
Out[51]: array(['source_node', 'destination_node', 'indicator_link',
       'num_followers_s', 'num_followers_d', 'num_followees_s',
       'num_followees_d', 'inter_followers', 'inter_followees',
       'preferential_attachment_followers_test'], dtype=object)
```

```
In [52]: a=df_final_train['num_followers_s'].values  
b=df_final_train['num_followers_d'].values  
for x,y in (zip(a,b)):  
    if x==0:  
        if y!=0:  
            print('i')
```

```
In [53]: np.count_nonzero(a)
```

Out[53]: 89660

```
In [54]: np.count_nonzero(b)
```

Out[54]: 89616

```
In [55]: # https://drive.google.com/file/d/10qJ04GRcaDxc16gmJXb8rpGPmLyys7E2/view?usp=sharing
! gdown --id 10qJ04GRcaDxc16gmJXb8rpGPmLyys7E2
```

'gdown' is not recognized as an internal or external command,
operable program or batch file.

In [56]: df_final_train.columns.values

Out[56]: array(['source_node', 'destination_node', 'indicator_link', 'num_followers_s', 'num_followers_d', 'num_followees_s', 'num_followees_d', 'inter_followers', 'inter_followees', 'preferential_attachement_followers_train'], dtype=object)

5.3 Adding new set of features

we will create these each of these features for both train and test data points

1. adar index
2. is following back
3. belongs to same weakly connect components
4. shortest path between source and destination

```
In [57]: if not os.path.isfile('storage_sample_stage2.h5'):
    #mapping adar index on train
    df_final_train['adar_index'] = df_final_train.apply(lambda row: calc_adar_index(row))
    #mapping adar index on test
    df_final_test['adar_index'] = df_final_test.apply(lambda row: calc_adar_index(row))

    #-----
    #mapping followback or not on train
    df_final_train['follows_back'] = df_final_train.apply(lambda row: follows_back(row))
    #mapping followback or not on test
    df_final_test['follows_back'] = df_final_test.apply(lambda row: follows_back(row))

    #-----
    #mapping same component of wcc or not on train
    df_final_train['same_comp'] = df_final_train.apply(lambda row: belongs_to_same_wcc(row))
    ##mapping same component of wcc or not on train
    df_final_test['same_comp'] = df_final_test.apply(lambda row: belongs_to_same_wcc(row))

    #-----
    #mapping shortest path on train
    df_final_train['shortest_path'] = df_final_train.apply(lambda row: compute_shortest_path(row))
    #mapping shortest path on test
    df_final_test['shortest_path'] = df_final_test.apply(lambda row: compute_shortest_path(row))

    hdf = HDFStore('storage_sample_stage2.h5')
    hdf.put('train_df',df_final_train, format='table', data_columns=True)
    hdf.put('test_df',df_final_test, format='table', data_columns=True)
    hdf.close()
else:
    df_final_train = read_hdf('storage_sample_stage2.h5', 'train_df', mode='r')
    df_final_test = read_hdf('storage_sample_stage2.h5', 'test_df', mode='r')
```

In [58]: `df_final_train.head()`

Out[58]:

	source_node	destination_node	indicator_link	num_followers_s	num_followers_d	num_followees
0	273084	1505602	1	11	15	
1	132660	449350	1	4	4	
2	733044	1115294	1	2	3	
3	1368835	1397351	1	0	1	
4	644613	1689027	1	16	14	

5.4 Adding new set of features

we will create these each of these features for both train and test data points

1. Weight Features

- weight of incoming edges
- weight of outgoing edges
- weight of incoming edges + weight of outgoing edges
- weight of incoming edges * weight of outgoing edges
- 2*weight of incoming edges + weight of outgoing edges
- weight of incoming edges + 2*weight of outgoing edges

2. Page Ranking of source

3. Page Ranking of dest

4. katz of source

5. katz of dest

6. hubs of source

7. hubs of dest

8. authorities_s of source

9. authorities_s of dest

Weight Features

In order to determine the similarity of nodes, an edge weight value was calculated between nodes. Edge weight decreases as the neighbor count goes up. Intuitively, consider one million people following a celebrity on a social network then chances are most of them never met each other or the celebrity. On the other hand, if a user has 30 contacts in his/her social network, the chances are higher that many of them know each other. credit - Graph-based Features for Supervised Link Prediction William Cukierski, Benjamin Hamner, Bo Yang

$$W = \frac{1}{\sqrt{1 + |X|}}$$

it is directed graph so calculated Weighted in and Weighted out differently

In [59]: *#weight for source and destination of each Link*

```
Weight_in = {}
Weight_out = {}
for i in tqdm(train_graph.nodes()):
    s1=set(train_graph.predecessors(i))
    w_in = 1.0/(np.sqrt(1+len(s1)))
    Weight_in[i]=w_in

    s2=set(train_graph.successors(i))
    w_out = 1.0/(np.sqrt(1+len(s2)))
    Weight_out[i]=w_out

#for imputing with mean
mean_weight_in = np.mean(list(Weight_in.values()))
mean_weight_out = np.mean(list(Weight_out.values()))
```

100% | 1
0722/1780722 [00:28<00:00, 62425.02it/s]

In [60]: *if not os.path.isfile('storage_sample_stage3.h5'):*

```
#mapping to pandas train
df_final_train['weight_in'] = df_final_train.destination_node.apply(lambda x: Weight_in[x])
df_final_train['weight_out'] = df_final_train.source_node.apply(lambda x: Weight_out[x])

#mapping to pandas test
df_final_test['weight_in'] = df_final_test.destination_node.apply(lambda x: Weight_in[x])
df_final_test['weight_out'] = df_final_test.source_node.apply(lambda x: Weight_out[x])

#some features engineerings on the in and out weights
df_final_train['weight_f1'] = df_final_train.weight_in + df_final_train.weight_out
df_final_train['weight_f2'] = df_final_train.weight_in * df_final_train.weight_out
df_final_train['weight_f3'] = (2*df_final_train.weight_in + 1*df_final_train.weight_out)
df_final_train['weight_f4'] = (1*df_final_train.weight_in + 2*df_final_train.weight_out)

#some features engineerings on the in and out weights
df_final_test['weight_f1'] = df_final_test.weight_in + df_final_test.weight_out
df_final_test['weight_f2'] = df_final_test.weight_in * df_final_test.weight_out
df_final_test['weight_f3'] = (2*df_final_test.weight_in + 1*df_final_test.weight_out)
df_final_test['weight_f4'] = (1*df_final_test.weight_in + 2*df_final_test.weight_out)
```

```
In [61]: if not os.path.isfile('storage_sample_stage3.h5'):
```

```
#page rank for source and destination in Train and Test
#if anything not there in train graph then adding mean page rank
df_final_train['page_rank_s'] = df_final_train.source_node.apply(lambda x:pr.
df_final_train['page_rank_d'] = df_final_train.destination_node.apply(lambda x:pr.

df_final_test['page_rank_s'] = df_final_test.source_node.apply(lambda x:pr.get
df_final_test['page_rank_d'] = df_final_test.destination_node.apply(lambda x:pr.get
#=====

#Katz centrality score for source and destination in Train and test
#if anything not there in train graph then adding mean katz score
df_final_train['katz_s'] = df_final_train.source_node.apply(lambda x: katz.get
df_final_train['katz_d'] = df_final_train.destination_node.apply(lambda x: katz.get
#=====

df_final_test['katz_s'] = df_final_test.source_node.apply(lambda x: katz.get
df_final_test['katz_d'] = df_final_test.destination_node.apply(lambda x: katz.get
#=====

#Hits algorithm score for source and destination in Train and test
#if anything not there in train graph then adding 0
df_final_train['hubs_s'] = df_final_train.source_node.apply(lambda x: hits[0].get
df_final_train['hubs_d'] = df_final_train.destination_node.apply(lambda x: hits[0].get
#=====

df_final_test['hubs_s'] = df_final_test.source_node.apply(lambda x: hits[0].get
df_final_test['hubs_d'] = df_final_test.destination_node.apply(lambda x: hits[0].get
#=====

#Hits algorithm score for source and destination in Train and Test
#if anything not there in train graph then adding 0
df_final_train['authorities_s'] = df_final_train.source_node.apply(lambda x: hits[1].get
df_final_train['authorities_d'] = df_final_train.destination_node.apply(lambda x: hits[1].get
#=====

df_final_test['authorities_s'] = df_final_test.source_node.apply(lambda x: hits[1].get
df_final_test['authorities_d'] = df_final_test.destination_node.apply(lambda x: hits[1].get
#=====

hdf = HDFStore('storage_sample_stage3.h5')
hdf.put('train_df',df_final_train, format='table', data_columns=True)
hdf.put('test_df',df_final_test, format='table', data_columns=True)
hdf.close()
else:
    df_final_train = read_hdf('storage_sample_stage3.h5', 'train_df', mode='r')
    df_final_test = read_hdf('storage_sample_stage3.h5', 'test_df', mode='r')
```

5.5 Adding new set of features

we will create these each of these features for both train and test data points

1. SVD features for both source and destination

```
In [62]: def svd(x, S):
    try:
        z = sadj_dict[x]
        return S[z]
    except:
        return [0,0,0,0,0,0]
```

```
In [63]: #for svd features to get feature vector creating a dict node val and inedx in svd
sadj_col = sorted(train_graph.nodes())
sadj_dict = { val:idx for idx,val in enumerate(sadj_col)}
```

```
In [64]: Adj = nx.adjacency_matrix(train_graph,nodelist=sorted(train_graph.nodes())).asfpt
```

```
In [65]: U, s, V = svds(Adj, k = 6)
print('Adjacency matrix Shape',Adj.shape)
print('U Shape',U.shape)
print('V Shape',V.shape)
print('s Shape',s.shape)
```

```
Adjacency matrix Shape (1780722, 1780722)
U Shape (1780722, 6)
V Shape (6, 1780722)
s Shape (6,)
```

```
In [66]: if not os.path.isfile('storage_sample_stage4.h5'):
    #=====

    df_final_train[['svd_u_s_1', 'svd_u_s_2', 'svd_u_s_3', 'svd_u_s_4', 'svd_u_s_5',
    df_final_train.source_node.apply(lambda x: svd(x, U)).apply(pd.Series)

    df_final_train[['svd_u_d_1', 'svd_u_d_2', 'svd_u_d_3', 'svd_u_d_4', 'svd_u_d_5',
    df_final_train.destination_node.apply(lambda x: svd(x, U)).apply(pd.Series)
    #=====

    df_final_train[['svd_v_s_1', 'svd_v_s_2', 'svd_v_s_3', 'svd_v_s_4', 'svd_v_s_5',
    df_final_train.source_node.apply(lambda x: svd(x, V.T)).apply(pd.Series)

    df_final_train[['svd_v_d_1', 'svd_v_d_2', 'svd_v_d_3', 'svd_v_d_4', 'svd_v_d_5',
    df_final_train.destination_node.apply(lambda x: svd(x, V.T)).apply(pd.Series)
    #=====

    df_final_test[['svd_u_s_1', 'svd_u_s_2', 'svd_u_s_3', 'svd_u_s_4', 'svd_u_s_5',
    df_final_test.source_node.apply(lambda x: svd(x, U)).apply(pd.Series)

    df_final_test[['svd_u_d_1', 'svd_u_d_2', 'svd_u_d_3', 'svd_u_d_4', 'svd_u_d_5',
    df_final_test.destination_node.apply(lambda x: svd(x, U)).apply(pd.Series)
    #=====

    df_final_test[['svd_v_s_1', 'svd_v_s_2', 'svd_v_s_3', 'svd_v_s_4', 'svd_v_s_5',
    df_final_test.source_node.apply(lambda x: svd(x, V.T)).apply(pd.Series)

    df_final_test[['svd_v_d_1', 'svd_v_d_2', 'svd_v_d_3', 'svd_v_d_4', 'svd_v_d_5',
    df_final_test.destination_node.apply(lambda x: svd(x, V.T)).apply(pd.Series)
    #=====

    hdf = HDFStore('storage_sample_stage4.h5')
    hdf.put('train_df', df_final_train, format='table', data_columns=True)
    hdf.put('test_df', df_final_test, format='table', data_columns=True)
    hdf.close()
else:
    df_final_train = read_hdf('storage_sample_stage4.h5', 'train_df', mode='r')
    df_final_test = read_hdf('storage_sample_stage4.h5', 'test_df', mode='r')
```

```
In [67]: # prepared and stored the data from machine Learning models
# please check the FB_Models.ipynb
```

Add feature called svd_dot. you can calculate svd_dot as Dot product between source node svd and destination node svd features.

In [68]: `df_final_train.columns.values`

Out[68]: `array(['source_node', 'destination_node', 'indicator_link', 'jaccard_followers', 'jaccard_followees', 'cosine_followers', 'cosine_followees', 'num_followers_s', 'num_followees_s', 'num_followees_d', 'inter_followers', 'inter_followees', 'adar_index', 'follows_back', 'same_comp', 'shortest_path', 'weight_in', 'weight_out', 'weight_f1', 'weight_f2', 'weight_f3', 'weight_f4', 'page_rank_s', 'page_rank_d', 'katz_s', 'katz_d', 'hubs_s', 'hubs_d', 'authorities_s', 'authorities_d', 'svd_u_s_1', 'svd_u_s_2', 'svd_u_s_3', 'svd_u_s_4', 'svd_u_s_5', 'svd_u_s_6', 'svd_u_d_1', 'svd_u_d_2', 'svd_u_d_3', 'svd_u_d_4', 'svd_u_d_5', 'svd_u_d_6', 'svd_v_s_1', 'svd_v_s_2', 'svd_v_s_3', 'svd_v_s_4', 'svd_v_s_5', 'svd_v_s_6', 'svd_v_d_1', 'svd_v_d_2', 'svd_v_d_3', 'svd_v_d_4', 'svd_v_d_5', 'svd_v_d_6'], dtype=object)`

In [69]: `df_final_test.columns.values`

Out[69]: `array(['source_node', 'destination_node', 'indicator_link', 'jaccard_followers', 'jaccard_followees', 'cosine_followers', 'cosine_followees', 'num_followers_s', 'num_followees_s', 'num_followees_d', 'inter_followers', 'inter_followees', 'adar_index', 'follows_back', 'same_comp', 'shortest_path', 'weight_in', 'weight_out', 'weight_f1', 'weight_f2', 'weight_f3', 'weight_f4', 'page_rank_s', 'page_rank_d', 'katz_s', 'katz_d', 'hubs_s', 'hubs_d', 'authorities_s', 'authorities_d', 'svd_u_s_1', 'svd_u_s_2', 'svd_u_s_3', 'svd_u_s_4', 'svd_u_s_5', 'svd_u_s_6', 'svd_u_d_1', 'svd_u_d_2', 'svd_u_d_3', 'svd_u_d_4', 'svd_u_d_5', 'svd_u_d_6', 'svd_v_s_1', 'svd_v_s_2', 'svd_v_s_3', 'svd_v_s_4', 'svd_v_s_5', 'svd_v_s_6', 'svd_v_d_1', 'svd_v_d_2', 'svd_v_d_3', 'svd_v_d_4', 'svd_v_d_5', 'svd_v_d_6'], dtype=object)`

In [70]: `type(df_final_train['svd_u_s_1'])`

Out[70]: `pandas.core.series.Series`

In [71]: `df_final_train['svd_u_s_1'].head(2)`

Out[71]: `0 1.666310e-13
1 -7.051469e-13
Name: svd_u_s_1, dtype: float64`

SV_dot feature = source_node*destination_node

```
In [72]: if not os.path.isfile('storage_sample_stage6.h5'):
    ...
    If the file storage_sample_stage6.h5 is not present.
    Then we are creating it in here and storing the file using HDFStore.
    HDFStore is use to store large files.
    In this file we are storing the values in train and test files of Svd_dot fea
    ...
#train
svd_dot_train= (df_final_train['svd_u_s_1']*df_final_train['svd_u_d_1'])+(df_
(df_final_train['svd_v_s_1']*df_final_train['svd_v_d_1'])+(df_
#test
svd_dot_test= (df_final_test['svd_u_s_1']*df_final_test['svd_u_d_1'])+(df_fi
(df_final_test['svd_v_s_1']*df_final_test['svd_v_d_1'])+(df_fi
df_final_train['svd_dot_train']= svd_dot_train
df_final_test['svd_dot_test']=svd_dot_test
hdf = HDFStore('storage_sample_stage6.h5')
hdf.put('train_df',df_final_train, format='table', data_columns=True)
hdf.put('test_df',df_final_test, format='table', data_columns=True)
hdf.close()
else:
...
If the file is present then we are just reading it.
...
df_final_train = read_hdf('storage_sample_stage6.h5', 'train_df',mode='r')
df_final_test = read_hdf('storage_sample_stage6.h5', 'test_df',mode='r')
```

```
In [73]: df_final_train.columns.values
```

```
Out[73]: array(['source_node', 'destination_node', 'indicator_link',
   'jaccard_followers', 'jaccard_followees', 'cosine_followers',
   'cosine_followees', 'num_followers_s', 'num_followees_s',
   'num_followees_d', 'inter_followers', 'inter_followees',
   'adar_index', 'follows_back', 'same_comp', 'shortest_path',
   'weight_in', 'weight_out', 'weight_f1', 'weight_f2', 'weight_f3',
   'weight_f4', 'page_rank_s', 'page_rank_d', 'katz_s', 'katz_d',
   'hubs_s', 'hubs_d', 'authorities_s', 'authorities_d', 'svd_u_s_1',
   'svd_u_s_2', 'svd_u_s_3', 'svd_u_s_4', 'svd_u_s_5', 'svd_u_s_6',
   'svd_u_d_1', 'svd_u_d_2', 'svd_u_d_3', 'svd_u_d_4', 'svd_u_d_5',
   'svd_u_d_6', 'svd_v_s_1', 'svd_v_s_2', 'svd_v_s_3', 'svd_v_s_4',
   'svd_v_s_5', 'svd_v_s_6', 'svd_v_d_1', 'svd_v_d_2', 'svd_v_d_3',
   'svd_v_d_4', 'svd_v_d_5', 'svd_v_d_6', 'svd_dot_train'],
  dtype=object)
```

```
In [74]: df_final_test.columns.values
```

```
Out[74]: array(['source_node', 'destination_node', 'indicator_link',
       'jaccard_followers', 'jaccard_followees', 'cosine_followers',
       'cosine_followees', 'num_followers_s', 'num_followees_s',
       'num_followees_d', 'inter_followers', 'inter_followees',
       'adar_index', 'follows_back', 'same_comp', 'shortest_path',
       'weight_in', 'weight_out', 'weight_f1', 'weight_f2', 'weight_f3',
       'weight_f4', 'page_rank_s', 'page_rank_d', 'katz_s', 'katz_d',
       'hubs_s', 'hubs_d', 'authorities_s', 'authorities_d', 'svd_u_s_1',
       'svd_u_s_2', 'svd_u_s_3', 'svd_u_s_4', 'svd_u_s_5', 'svd_u_s_6',
       'svd_u_d_1', 'svd_u_d_2', 'svd_u_d_3', 'svd_u_d_4', 'svd_u_d_5',
       'svd_u_d_6', 'svd_v_s_1', 'svd_v_s_2', 'svd_v_s_3', 'svd_v_s_4',
       'svd_v_s_5', 'svd_v_s_6', 'svd_v_d_1', 'svd_v_d_2', 'svd_v_d_3',
       'svd_v_d_4', 'svd_v_d_5', 'svd_v_d_6', 'svd_dot_test'],
      dtype=object)
```

```
In [ ]:
```

```
In [ ]:
```

```
In [ ]:
```