Social network Graph Link Prediction - Facebook Challenge

In [0]:

```
#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 aritmetic operations on arrays
# matplotlib: used to plot graphs
import matplotlib
import matplotlib.pylab 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
#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 aritmetic operations on arrays
# matplotlib: used to plot graphs
import matplotlib
import matplotlib.pylab 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
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import f1_score
```

1. Reading Data

In [0]:

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

```
Name:
Type: DiGraph
Number of nodes: 1780722
Number of edges: 7550015
Average in degree:   4.2399
Average out degree:   4.2399
```

2. Similarity measures

2.1 Jaccard Distance:

http://www.statisticshowto.com/jaccard-index/

\begin{equation} j = \frac{|X\cap Y|}{|X \cup Y|} \end{equation}

In [0]:

```
#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 [0]:

```
#one test case
print(jaccard_for_followees(273084,1505602))
```

```
0.0
```

In [0]:

```
#node 1635354 not in graph
print(jaccard_for_followees(273084,1505602))
```

```
0.0
```

In [0]:

```
#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.predec
```

```
ssors(b)))))
        return sim
    except:
        return 0
```

In [0]:

```
print(jaccard_for_followers(273084,470294))
```

0

In [0]:

```
#node 1635354 not in graph
print(jaccard_for_followees(669354,1635354))
```

0

2.2 Cosine distance

\begin{equation} CosineDistance = \frac{|X\cap Y|}{|X|\cdot|Y|} \end{equation}

In [0]:

```
#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)))*len((set(train_graph.successors(b))))))
        return sim
    except:
        return 0
```

In [0]:

```
print(cosine_for_followees(273084,1505602))
```

0.0

In [0]:

```
print(cosine_for_followees(273084,1635354))
```

0

In [0]:

```
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))))*(len(set(tra
n_graph.predecessors(b)))))
        return sim
    except:
        return 0
```

In [0]:

```
print(cosine_for_followers(2,470294))
```

0.02886751345948129

In [0]:

```
print(cosine_for_followers(669354,1635354))
```

0

Preferential Attachment

In [0]:

```
#for followees
def pref_attach_for_followees(a,b):
    try:
        if len(set(train_graph.successors(a))) == 0  | len(set(train_graph.successors(b))) == 0:
            return 0
        pf = len(set(train_graph.successors(a)))*len(set(train_graph.successors(b)))
        return pf
    except:
        return 0
```

In [0]:

```
pref_attach_for_followees(273084,1505602)
```

Out[0]:

120

In [0]:

```
pref_attach_for_followees(273084,1635354)
```

Out[0]:

0

In [0]:

```
#for followers
def pref_attach_for_followers(a,b):
    try:
        if len(set(train_graph.predecessors(a))) == 0  | len(set(train_graph.predecessors(b))) == 0
:
            return 0
        pf = len(set(train_graph.predecessors(a)))*len(set(train_graph.predecessors(b)))
        return pf
    except:
        return 0
```

In [0]:

```
pref_attach_for_followers(848424 ,784690)
```

Out[0]:

84

In [0]:

```
pref_attach_for_followers(669354,1635354)
```

Out[0]:

```
0
```

3. Ranking Measures

https://networkx.github.io/documentation/networkx-1.10/reference/generated/networkx.algorithms.link_analysis.pagerank_alg.pagerank.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

In [0]:

```
if not os.path.isfile('fea_sample/page_rank.p'):
    pr = nx.pagerank(train_graph, alpha=0.85)
    pickle.dump(pr,open('fea_sample/page_rank.p','wb'))
else:
    pr = pickle.load(open('fea_sample/page_rank.p','rb'))
```

In [0]:

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

```
min 1.6556497245737814e-07
max 2.7098251341935827e-05
mean 5.615699699389075e-07
```

In [0]:

```
#for imputing to nodes which are not there in Train data
mean_pr = float(sum(pr.values())) / len(pr)
print(mean_pr)
```

```
5.615699699389075e-07
```

4. Other Graph Features

4.1 Shortest path:

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

In [0]:

```
#if has direct edge then deleting that edge and calculating shortest path
def compute_shortest_path_length(a,b):
    p=-1
    try:
```

```
    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 [0]:

```
#testing
compute_shortest_path_length(77697, 826021)
```

Out[0]:

```
10
```

In [0]:

```
#testing
compute_shortest_path_length(669354,1635354)
```

Out[0]:

```
-1
```

4.2 Checking for same community

In [0]:

```
#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 [0]:

```
belongs_to_same_wcc(861, 1659750)
```

Out[0]:

```
0
```

In [0]:

```
belongs_to_same_wcc(669354,1635354)
```

Out[0]:

```
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 [0]:

```
#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 [0]:

```
calc_adar_in(1,189226)
```

Out[0]:

```
0
```

In [0]:

```
calc_adar_in(669354,1635354)
```

Out[0]:

```
0
```

4.4 Is persion was following back:

In [0]:

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

In [0]:

```
follows_back(1,189226)
```

Out[0]:

```
1
```

In [0]:

```
follows_back(669354,1635354)
```

Out[0]:

```
0
```

4.5 Katz Centrality:

https://en.wikipedia.org/wiki/Katz_centrality

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 [0]:

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

In [0]:

```
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 [0]:

```
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

In [0]:

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

In [0]:

```
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 0.0
```

```
min 0.0
max 0.004868653378780953
mean 5.615699699344123e-07
```

5. Featurization

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

In [0]:

```
import random
if os.path.isfile('train_after_eda.csv'):
    filename = "train_after_eda.csv"
    # you uncomment this line, if you dont know the lentgh 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 (excludes 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 [0]:

```
if os.path.isfile('train_after_eda.csv'):
    filename = "test_after_eda.csv"
    # you uncomment this line, if you dont know the lentgh 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 (excludes 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 [0]:

```
print("Number of rows in the train data file:", n_train)
print("Number of rows we are going to elimiate 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 elimiate in test data are",len(skip_test))
```

```
Number of rows in the train data file: 15100030
Number of rows we are going to elimiate in train data are 15000030
Number of rows in the test data file: 3775008
Number of rows we are going to elimiate in test data are 3725008
```

In [0]:

```
df_final_train = pd.read_csv('train_after_eda.csv', skiprows=skip_train, names=['source_node', 'des
tination_node'])
df_final_train['indicator_link'] = pd.read_csv('train_y.csv', skiprows=skip_train, names=['indicato
r_link'])
print("Our train matrix size ",df_final_train.shape)
df_final_train.head(2)
```

```
Our train matrix size  (100001, 3)
```

Out[0]:

	source_node	destination_node	indicator_link
0	273084	1505602	1
1	1072684	458008	1

In [0]:

```
df_final_test = pd.read_csv('test_after_eda.csv', skiprows=skip_test, names=['source_node', 'destin
ation_node'])
```

```
ation_node'])
df_final_test['indicator_link'] = pd.read_csv('test_y.csv', skiprows=skip_test, names=['indicator_l
ink'])
print("Our test matrix size ",df_final_test.shape)
df_final_test.head(2)
```

```
Our test matrix size  (50001, 3)
```

Out[0]:

	source_node	destination_node	indicator_link
0	848424	784690	1
1	121108	204025	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 [0]:

```
if not os.path.isfile('storage_sample_stage1.h5'):
    #mapping jaccrd followers to train and test data
    df_final_train['jaccard_followers'] = df_final_train.apply(lambda row:

jaccard_for_followers(row['source_node'],row['destination_node']),axis=1)
    df_final_test['jaccard_followers'] = df_final_test.apply(lambda row:

jaccard_for_followers(row['source_node'],row['destination_node']),axis=1)

    #mapping jaccrd followees to train and test data
    df_final_train['jaccard_followees'] = df_final_train.apply(lambda row:

jaccard_for_followees(row['source_node'],row['destination_node']),axis=1)
    df_final_test['jaccard_followees'] = df_final_test.apply(lambda row:

jaccard_for_followees(row['source_node'],row['destination_node']),axis=1)


        #mapping jaccrd followers to train and test data
    df_final_train['cosine_followers'] = df_final_train.apply(lambda row:

cosine_for_followers(row['source_node'],row['destination_node']),axis=1)
    df_final_test['cosine_followers'] = df_final_test.apply(lambda row:

cosine_for_followers(row['source_node'],row['destination_node']),axis=1)

    #mapping jaccrd followees to train and test data
    df_final_train['cosine_followees'] = df_final_train.apply(lambda row:

cosine_for_followees(row['source_node'],row['destination_node']),axis=1)
    df_final_test['cosine_followees'] = df_final_test.apply(lambda row:

cosine_for_followees(row['source_node'],row['destination_node']),axis=1)

    #mapping preferential attachment followers to train and test data
    df_final_train['pref_attach_followers'] = df_final_train.apply(lambda row:

pref_attach_for_followers(row['source_node'],row['destination_node']),axis=1)
    df_final_test['pref_attach_followers'] = df_final_test.apply(lambda row:
```

```
    df_final_test['pref_attach_followers'] = df_final_test.apply(lambda row:
pref_attach_for_followers(row['source_node'],row['destination_node']),axis=1)

    #mapping jaccrd followees to train and test data
    df_final_train['pref_attach_followees'] = df_final_train.apply(lambda row:
pref_attach_for_followees(row['source_node'],row['destination_node']),axis=1)
    df_final_test['pref_attach_followees'] = df_final_test.apply(lambda row:
pref_attach_for_followees(row['source_node'],row['destination_node']),axis=1)
```

In [0]:

```
df_final_test.head()
```

Out[0]:

	source_node	destination_node	indicator_link	jaccard_followers	jaccard_followees	cosine_followers	cosine_followees	pref_attach_
0	848424	784690	1	0	0.000000	0.029161	0.000000	
1	121108	204025	1	0	0.000000	0.000000	0.000000	
2	593083	1131838	1	0	0.129032	0.071429	0.229416	
3	1018582	1463758	1	0	0.000000	0.025102	0.000000	
4	557644	941303	1	0	0.250000	0.080992	0.403890	

In [0]:

```
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_followers_d, num_followees_s, num_followees_d, inter_followers, int
er_followees
```

In [0]:

```
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(d
f_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['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('fea_sample/storage_sample_stage1.h5', 'train_df',mode='r')
    df_final_test = read_hdf('fea_sample/storage_sample_stage1.h5', 'test_df',mode='r')
```

In [0]:

```
df_final_train.head()
```

Out[0]:

	source_node	destination_node	indicator_link	jaccard_followers	jaccard_followees	cosine_followers	cosine_followees	pref_attach_
0	273084	1505602	1	0	0.000000	0.00000	0.000000	
1	1072684	458008	1	0	0.000000	0.00000	0.000000	
2	122637	7211	1	0	0.017857	0.00849	0.036370	
3	521886	292052	1	0	0.000000	0.00000	0.000000	
4	1306826	1463813	1	0	0.125000	0.00000	0.223607	

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 [0]:

```
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_in(row['source_node']
,row['destination_node']),axis=1)
    #mapping adar index on test
    df_final_test['adar_index'] = df_final_test.apply(lambda row: calc_adar_in(row['source_node'],r
ow['destination_node']),axis=1)

    #---------------------------------------------------------------------------------------------
------
    #mapping followback or not on train
    df_final_train['follows_back'] = df_final_train.apply(lambda row:
follows_back(row['source_node'],row['destination_node']),axis=1)

    #mapping followback or not on test
    df_final_test['follows_back'] = df_final_test.apply(lambda row: follows_back(row['source_node']
,row['destination_node']),axis=1)

    #---------------------------------------------------------------------------------------------
------
    #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['source_
node'],row['destination_node']),axis=1)

    ##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['source_no
de'],row['destination_node']),axis=1)

    #---------------------------------------------------------------------------------------------
------
    #mapping shortest path on train
    df_final_train['shortest_path'] = df_final_train.apply(lambda row: compute_shortest_path_length
(row['source_node'],row['destination_node']),axis=1)
```

```
(row['source_node'],row['destination_node']),axis=1)
    #mapping shortest path on test
    df_final_test['shortest_path'] = df_final_test.apply(lambda row: compute_shortest_path_length(r
ow['source_node'],row['destination_node']),axis=1)

    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('fea_sample/storage_sample_stage2.h5', 'train_df',mode='r')
#     df_final_test = read_hdf('fea_sample/storage_sample_stage2.h5', 'test_df',mode='r')
```

In [0]:

```
a=read_hdf('fea_sample/storage_sample_stage2.h5', 'train_df',mode='r')
df_final_train.head(5)
```

Out[0]:

	source_node	destination_node	indicator_link	jaccard_followers	jaccard_followees	cosine_followers	cosine_followees	pref_attach_
0	273084	1505602	1	0	0.000000	0.00000	0.000000	
1	1072684	458008	1	0	0.000000	0.00000	0.000000	
2	122637	7211	1	0	0.017857	0.00849	0.036370	
3	521886	292052	1	0	0.000000	0.00000	0.000000	
4	1306826	1463813	1	0	0.125000	0.00000	0.223607	

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

\begin{equation} W = \frac{1}{\sqrt{1+|X|}} \end{equation}

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

In [0]:

```
#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%|██████████| 1780722/1780722 [00:15<00:00, 112646.16it/s]
```

In [0]:

```
if not os.path.isfile('data/fea_sample/storage_sample_stage3.h5'):
    #mapping to pandas train
    df_final_train['weight_in'] = df_final_train.destination_node.apply(lambda x: Weight_in.get(x,m
ean_weight_in))
    df_final_train['weight_out'] = df_final_train.source_node.apply(lambda x: Weight_out.get(x,mean
_weight_out))

    #mapping to pandas test
    df_final_test['weight_in'] = df_final_test.destination_node.apply(lambda x: Weight_in.get(x,mea
n_weight_in))
    df_final_test['weight_out'] = df_final_test.source_node.apply(lambda x: Weight_out.get(x,mean_w
eight_out))


    #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 [0]:

```
if not os.path.isfile('data/fea_sample/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.get(x,mean_pr))
    df_final_train['page_rank_d'] = df_final_train.destination_node.apply(lambda x:pr.get(x,mean_pr
))

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

    #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(x,mean_katz))
    df_final_train['katz_d'] = df_final_train.destination_node.apply(lambda x: katz.get(x,mean_katz
))

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

    #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(x,0))
    df_final_train['hubs_d'] = df_final_train.destination_node.apply(lambda x: hits[0].get(x,0))

    df_final_test['hubs_s'] = df_final_test.source_node.apply(lambda x: hits[0].get(x,0))
```

```
    df_final_test['hubs_d'] = df_final_test.destination_node.apply(lambda x: hits[0].get(x,0))
    #================================================================================

    #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(x,0))
    df_final_train['authorities_d'] = df_final_train.destination_node.apply(lambda x: hits[1].get(x
,0))

    df_final_test['authorities_s'] = df_final_test.source_node.apply(lambda x: hits[1].get(x,0))
    df_final_test['authorities_d'] = df_final_test.destination_node.apply(lambda x: hits[1].get(x,0
))
    #================================================================================

    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('data/fea_sample/storage_sample_stage3.h5', 'train_df',mode='r')
    df_final_test = read_hdf('data/fea_sample/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 [0]:

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

In [0]:

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

In [0]:

```
Adj = nx.adjacency_matrix(train_graph,nodelist=sorted(train_graph.nodes())).asfptype()
```

In [0]:

```
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 [0]:

```
a=np.array([1,2,3,4,5])
b=np.array([6,7,8,9,10])
a=a.reshape(1,5)
b=b.reshape(1,5)
a.dot(b.T)
```

Out[0]:

```
array([[ 6, 12, 18, 24, 30],
       [ 7, 14, 21, 28, 35],
       [ 8, 16, 24, 32, 40],
       [ 9, 18, 27, 36, 45],
       [10, 20, 30, 40, 50]])
```

In [0]:

```
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', 'svd_u_s_6']] =
\
    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','svd_u_d_6']] =
\
    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', 'svd_v_s_6',]]
= \
    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','svd_v_d_6']] =
\
    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', 'svd_u_s_6']] =
\
    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','svd_u_d_6']] =
\
    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', 'svd_v_s_6']] =
\
    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','svd_v_d_6']] =
\
    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()
```

In [0]:

```
temp1=df_final_train[['svd_u_d_1', 'svd_u_d_2', 'svd_u_d_3', 'svd_u_d_4', 'svd_u_d_5','svd_u_d_6']]
.values
print(temp1.shape)
temp2=df_final_train[['svd_v_s_1','svd_v_s_2', 'svd_v_s_3', 'svd_v_s_4', 'svd_v_s_5', 'svd_v_s_6']]
.values
print(temp2.shape)
temp3=df_final_train[['svd_v_d_1', 'svd_v_d_2', 'svd_v_d_3', 'svd_v_d_4', 'svd_v_d_5','svd_v_d_6']]
.values
print(temp3.shape)
```

```
temp4=df_final_train[['svd_u_s_1', 'svd_u_s_2','svd_u_s_3', 'svd_u_s_4', 'svd_u_s_5', 'svd_u_s_6']]
.values
print(temp4.shape)
```

```
(100001, 6)
(100001, 6)
(100001, 6)
(100001, 6)
```

In [0]:

```
temp=temp4
li1=[]
li2=[]
i=0
while(i<len(temp1)):
    li1.append(temp1[i].dot(temp4[i].T))
    li2.append(temp2[i].dot(temp3[i].T))
    i=i+1
print(len(li1))
print(len(li2))
```

```
100001
100001
```

In [0]:

```
temp1=df_final_test[['svd_u_d_1', 'svd_u_d_2', 'svd_u_d_3', 'svd_u_d_4', 'svd_u_d_5','svd_u_d_6']].
values
print(temp1.shape)
temp2=df_final_test[['svd_v_s_1','svd_v_s_2', 'svd_v_s_3', 'svd_v_s_4', 'svd_v_s_5', 'svd_v_s_6']].
values
print(temp2.shape)
temp3=df_final_test[['svd_v_d_1', 'svd_v_d_2', 'svd_v_d_3', 'svd_v_d_4', 'svd_v_d_5','svd_v_d_6']].
values
print(temp3.shape)
temp4=df_final_test[['svd_u_s_1', 'svd_u_s_2','svd_u_s_3', 'svd_u_s_4', 'svd_u_s_5', 'svd_u_s_6']].
values
print(temp4.shape)
```

```
(50001, 6)
(50001, 6)
(50001, 6)
(50001, 6)
```

In [0]:

```
li3=[]
li4=[]
i=0
while(i<len(temp1)):
    li3.append(temp1[i].dot(temp4[i].T))
    li4.append(temp2[i].dot(temp3[i].T))
    i=i+1
print(len(li3))
print(len(li4))
```

```
50001
50001
```

In [0]:

```
df_final_train["Dot_product_U"]=li1
df_final_train["Dot_product_V"]=li1
df_final_test["Dot_product_U"]=li3
df_final_test["Dot_product_V"]=li4
```

In [0]:

```
print(df_final_train.shape)
print(df_final_test.shape)
```

```
(100001, 59)
(50001, 59)
```

In [0]:

```
df_final_train.head(5)
```

Out[0]:

	source_node	destination_node	indicator_link	jaccard_followers	jaccard_followees	cosine_followers	cosine_followees	pref_attach_
0	273084	1505602	1	0	0.000000	0.00000	0.000000	
1	1072684	458008	1	0	0.000000	0.00000	0.000000	
2	122637	7211	1	0	0.017857	0.00849	0.036370	
3	521886	292052	1	0	0.000000	0.00000	0.000000	
4	1306826	1463813	1	0	0.125000	0.00000	0.223607	

In [0]:

```
df_final_train.columns
```

Out[0]:

```
Index(['source_node', 'destination_node', 'indicator_link',
       'jaccard_followers', 'jaccard_followees', 'cosine_followers',
       'cosine_followees', 'pref_attach_followers', 'pref_attach_followees',
       'num_followers_s', 'num_followers_d', 'num_followees_s',
       'num_followees_d', 'inter_followers', 'inter_followees', 'adar_index',
       'follows_back', 'same_comp', 'shortest_path', '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', 'Dot_product_U', 'Dot_product_V', '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'],
      dtype='object')
```

In [0]:

```
y_train = df_final_train.indicator_link
y_test = df_final_test.indicator_link
```

In [0]:

```
df_final_train.drop(['source_node', 'destination_node','indicator_link'],axis=1,inplace=True)
df_final_test.drop(['source_node', 'destination_node','indicator_link'],axis=1,inplace=True)
```

In [0]:

```
#i made use of code from previous assignments
import xgboost as xgb
from sklearn.model_selection import GridSearchCV
parameters =     {
            'max_depth': [1,3, 4, 5],
            'n_estimators': [1,16,64,100,200,300,400]
```

```
            }
xgb_model = xgb.XGBClassifier()
clf = GridSearchCV(xgb_model, parameters, scoring = 'f1', verbose=5,return_train_score=True)
clf.fit(df_final_train,y_train)
```

```
Fitting 3 folds for each of 28 candidates, totalling 84 fits
[CV] max_depth=1, n_estimators=1 ....................................

[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.

[CV]  max_depth=1, n_estimators=1, score=(train=0.712, test=0.714), total=   0.6s
[CV] max_depth=1, n_estimators=1 ....................................

[Parallel(n_jobs=1)]: Done   1 out of   1 | elapsed:    0.7s remaining:    0.0s

[CV]  max_depth=1, n_estimators=1, score=(train=0.712, test=0.714), total=   0.6s
[CV] max_depth=1, n_estimators=1 ....................................

[Parallel(n_jobs=1)]: Done   2 out of   2 | elapsed:    1.4s remaining:    0.0s

[CV]  max_depth=1, n_estimators=1, score=(train=0.714, test=0.710), total=   0.6s
[CV] max_depth=1, n_estimators=16 ...................................

[Parallel(n_jobs=1)]: Done   3 out of   3 | elapsed:    2.1s remaining:    0.0s

[CV]  max_depth=1, n_estimators=16, score=(train=0.876, test=0.875), total=   1.7s
[CV] max_depth=1, n_estimators=16 ...................................

[Parallel(n_jobs=1)]: Done   4 out of   4 | elapsed:    3.9s remaining:    0.0s

[CV]  max_depth=1, n_estimators=16, score=(train=0.876, test=0.875), total=   1.7s
[CV] max_depth=1, n_estimators=16 ...................................
[CV]  max_depth=1, n_estimators=16, score=(train=0.874, test=0.876), total=   1.8s
[CV] max_depth=1, n_estimators=64 ...................................
[CV]  max_depth=1, n_estimators=64, score=(train=0.926, test=0.925), total=   5.0s
[CV] max_depth=1, n_estimators=64 ...................................
[CV]  max_depth=1, n_estimators=64, score=(train=0.926, test=0.925), total=   4.9s
[CV] max_depth=1, n_estimators=64 ...................................
[CV]  max_depth=1, n_estimators=64, score=(train=0.925, test=0.927), total=   5.0s
[CV] max_depth=1, n_estimators=100 ..................................
[CV]  max_depth=1, n_estimators=100, score=(train=0.929, test=0.927), total=   7.9s
[CV] max_depth=1, n_estimators=100 ..................................
[CV]  max_depth=1, n_estimators=100, score=(train=0.928, test=0.926), total=   8.5s
[CV] max_depth=1, n_estimators=100 ..................................
[CV]  max_depth=1, n_estimators=100, score=(train=0.927, test=0.928), total=   8.5s
[CV] max_depth=1, n_estimators=200 ..................................
[CV]  max_depth=1, n_estimators=200, score=(train=0.957, test=0.957), total=  15.9s
[CV] max_depth=1, n_estimators=200 ..................................
[CV]  max_depth=1, n_estimators=200, score=(train=0.957, test=0.955), total=  15.9s
[CV] max_depth=1, n_estimators=200 ..................................
[CV]  max_depth=1, n_estimators=200, score=(train=0.956, test=0.957), total=  15.9s
[CV] max_depth=1, n_estimators=300 ..................................
[CV]  max_depth=1, n_estimators=300, score=(train=0.964, test=0.963), total=  21.8s
[CV] max_depth=1, n_estimators=300 ..................................
[CV]  max_depth=1, n_estimators=300, score=(train=0.964, test=0.963), total=  23.9s
[CV] max_depth=1, n_estimators=300 ..................................
[CV]  max_depth=1, n_estimators=300, score=(train=0.964, test=0.965), total=  23.6s
[CV] max_depth=1, n_estimators=400 ..................................
[CV]  max_depth=1, n_estimators=400, score=(train=0.968, test=0.968), total=  32.5s
[CV] max_depth=1, n_estimators=400 ..................................
[CV]  max_depth=1, n_estimators=400, score=(train=0.969, test=0.967), total=  31.5s
[CV] max_depth=1, n_estimators=400 ..................................
[CV]  max_depth=1, n_estimators=400, score=(train=0.968, test=0.968), total=  28.6s
[CV] max_depth=3, n_estimators=1 ....................................
[CV]  max_depth=3, n_estimators=1, score=(train=0.896, test=0.896), total=   0.8s
[CV] max_depth=3, n_estimators=1 ....................................
[CV]  max_depth=3, n_estimators=1, score=(train=0.896, test=0.895), total=   0.8s
[CV] max_depth=3, n_estimators=1 ....................................
[CV]  max_depth=3, n_estimators=1, score=(train=0.896, test=0.897), total=   0.7s
[CV] max_depth=3, n_estimators=16 ...................................
```

```
[CV]  max_depth=3, n_estimators=16, score=(train=0.932, test=0.930), total=   3.8s
[CV] max_depth=3, n_estimators=16 ....................................
[CV]  max_depth=3, n_estimators=16, score=(train=0.931, test=0.930), total=   3.7s
[CV] max_depth=3, n_estimators=16 ....................................
[CV]  max_depth=3, n_estimators=16, score=(train=0.929, test=0.930), total=   3.5s
[CV] max_depth=3, n_estimators=64 ....................................
[CV]  max_depth=3, n_estimators=64, score=(train=0.970, test=0.969), total=  12.8s
[CV] max_depth=3, n_estimators=64 ....................................
[CV]  max_depth=3, n_estimators=64, score=(train=0.971, test=0.969), total=  12.4s
[CV] max_depth=3, n_estimators=64 ....................................
[CV]  max_depth=3, n_estimators=64, score=(train=0.971, test=0.972), total=  12.9s
[CV] max_depth=3, n_estimators=100 ...................................
[CV]  max_depth=3, n_estimators=100, score=(train=0.975, test=0.974), total=  19.9s
[CV] max_depth=3, n_estimators=100 ...................................
[CV]  max_depth=3, n_estimators=100, score=(train=0.975, test=0.975), total=  19.4s
[CV] max_depth=3, n_estimators=100 ...................................
[CV]  max_depth=3, n_estimators=100, score=(train=0.975, test=0.975), total=  19.2s
[CV] max_depth=3, n_estimators=200 ...................................
[CV]  max_depth=3, n_estimators=200, score=(train=0.979, test=0.977), total=  38.8s
[CV] max_depth=3, n_estimators=200 ...................................
[CV]  max_depth=3, n_estimators=200, score=(train=0.979, test=0.977), total=  41.2s
[CV] max_depth=3, n_estimators=200 ...................................
[CV]  max_depth=3, n_estimators=200, score=(train=0.978, test=0.978), total=  38.0s
[CV] max_depth=3, n_estimators=300 ...................................
[CV]  max_depth=3, n_estimators=300, score=(train=0.982, test=0.979), total=  57.5s
[CV] max_depth=3, n_estimators=300 ...................................
[CV]  max_depth=3, n_estimators=300, score=(train=0.982, test=0.978), total= 1.0min
[CV] max_depth=3, n_estimators=300 ...................................
[CV]  max_depth=3, n_estimators=300, score=(train=0.981, test=0.980), total=  57.1s
[CV] max_depth=3, n_estimators=400 ...................................
[CV]  max_depth=3, n_estimators=400, score=(train=0.985, test=0.981), total= 1.3min
[CV] max_depth=3, n_estimators=400 ...................................
[CV]  max_depth=3, n_estimators=400, score=(train=0.984, test=0.979), total= 1.3min
[CV] max_depth=3, n_estimators=400 ...................................
[CV]  max_depth=3, n_estimators=400, score=(train=0.984, test=0.982), total= 1.3min
[CV] max_depth=4, n_estimators=1 .....................................
[CV]  max_depth=4, n_estimators=1, score=(train=0.924, test=0.923), total=   0.7s
[CV] max_depth=4, n_estimators=1 .....................................
[CV]  max_depth=4, n_estimators=1, score=(train=0.924, test=0.924), total=   0.7s
[CV] max_depth=4, n_estimators=1 .....................................
[CV]  max_depth=4, n_estimators=1, score=(train=0.924, test=0.924), total=   0.7s
[CV] max_depth=4, n_estimators=16 ....................................
[CV]  max_depth=4, n_estimators=16, score=(train=0.927, test=0.926), total=   3.9s
[CV] max_depth=4, n_estimators=16 ....................................
[CV]  max_depth=4, n_estimators=16, score=(train=0.927, test=0.926), total=   4.1s
[CV] max_depth=4, n_estimators=16 ....................................
[CV]  max_depth=4, n_estimators=16, score=(train=0.925, test=0.925), total=   4.0s
[CV] max_depth=4, n_estimators=64 ....................................
[CV]  max_depth=4, n_estimators=64, score=(train=0.974, test=0.974), total=  16.3s
[CV] max_depth=4, n_estimators=64 ....................................
[CV]  max_depth=4, n_estimators=64, score=(train=0.973, test=0.972), total=  16.2s
[CV] max_depth=4, n_estimators=64 ....................................
[CV]  max_depth=4, n_estimators=64, score=(train=0.974, test=0.973), total=  14.6s
[CV] max_depth=4, n_estimators=100 ...................................
[CV]  max_depth=4, n_estimators=100, score=(train=0.977, test=0.976), total=  22.8s
[CV] max_depth=4, n_estimators=100 ...................................
[CV]  max_depth=4, n_estimators=100, score=(train=0.977, test=0.976), total=  24.7s
[CV] max_depth=4, n_estimators=100 ...................................
[CV]  max_depth=4, n_estimators=100, score=(train=0.977, test=0.976), total=  25.0s
[CV] max_depth=4, n_estimators=200 ...................................
[CV]  max_depth=4, n_estimators=200, score=(train=0.982, test=0.979), total=  46.6s
[CV] max_depth=4, n_estimators=200 ...................................
[CV]  max_depth=4, n_estimators=200, score=(train=0.982, test=0.978), total=  45.9s
[CV] max_depth=4, n_estimators=200 ...................................
[CV]  max_depth=4, n_estimators=200, score=(train=0.982, test=0.980), total=  44.7s
[CV] max_depth=4, n_estimators=300 ...................................
[CV]  max_depth=4, n_estimators=300, score=(train=0.987, test=0.981), total= 1.1min
[CV] max_depth=4, n_estimators=300 ...................................
[CV]  max_depth=4, n_estimators=300, score=(train=0.986, test=0.979), total= 1.1min
[CV] max_depth=4, n_estimators=300 ...................................
[CV]  max_depth=4, n_estimators=300, score=(train=0.986, test=0.982), total= 1.1min
[CV] max_depth=4, n_estimators=400 ...................................
[CV]  max_depth=4, n_estimators=400, score=(train=0.991, test=0.982), total= 1.5min
[CV] max_depth=4, n_estimators=400 ...................................
[CV]  max_depth=4, n_estimators=400, score=(train=0.990, test=0.981), total= 1.5min
[CV] max_depth=4, n_estimators=400 ...................................
[CV]  max_depth=4, n_estimators=400, score=(train=0.990, test=0.983), total= 1.5min
```

```
[CV] max_depth=5, n_estimators=1 ....................................
[CV]  max_depth=5, n_estimators=1, score=(train=0.927, test=0.925), total=   0.8s
[CV] max_depth=5, n_estimators=1 ....................................
[CV]  max_depth=5, n_estimators=1, score=(train=0.927, test=0.926), total=   0.7s
[CV] max_depth=5, n_estimators=1 ....................................
[CV]  max_depth=5, n_estimators=1, score=(train=0.926, test=0.927), total=   0.7s
[CV] max_depth=5, n_estimators=16 ...................................
[CV]  max_depth=5, n_estimators=16, score=(train=0.960, test=0.960), total=   4.7s
[CV] max_depth=5, n_estimators=16 ...................................
[CV]  max_depth=5, n_estimators=16, score=(train=0.943, test=0.942), total=   4.8s
[CV] max_depth=5, n_estimators=16 ...................................
[CV]  max_depth=5, n_estimators=16, score=(train=0.943, test=0.943), total=   4.7s
[CV] max_depth=5, n_estimators=64 ...................................
[CV]  max_depth=5, n_estimators=64, score=(train=0.975, test=0.975), total=  17.2s
[CV] max_depth=5, n_estimators=64 ...................................
[CV]  max_depth=5, n_estimators=64, score=(train=0.975, test=0.974), total=  17.3s
[CV] max_depth=5, n_estimators=64 ...................................
[CV]  max_depth=5, n_estimators=64, score=(train=0.975, test=0.975), total=  17.0s
[CV] max_depth=5, n_estimators=100 ..................................
[CV]  max_depth=5, n_estimators=100, score=(train=0.979, test=0.978), total=  26.9s
[CV] max_depth=5, n_estimators=100 ..................................
[CV]  max_depth=5, n_estimators=100, score=(train=0.979, test=0.977), total=  27.5s
[CV] max_depth=5, n_estimators=100 ..................................
[CV]  max_depth=5, n_estimators=100, score=(train=0.979, test=0.978), total=  27.7s
[CV] max_depth=5, n_estimators=200 ..................................
[CV]  max_depth=5, n_estimators=200, score=(train=0.987, test=0.980), total=  55.1s
[CV] max_depth=5, n_estimators=200 ..................................
[CV]  max_depth=5, n_estimators=200, score=(train=0.987, test=0.979), total=  54.3s
[CV] max_depth=5, n_estimators=200 ..................................
[CV]  max_depth=5, n_estimators=200, score=(train=0.986, test=0.981), total=  56.4s
[CV] max_depth=5, n_estimators=300 ..................................
[CV]  max_depth=5, n_estimators=300, score=(train=0.992, test=0.982), total= 1.4min
[CV] max_depth=5, n_estimators=300 ..................................
[CV]  max_depth=5, n_estimators=300, score=(train=0.992, test=0.980), total= 1.4min
[CV] max_depth=5, n_estimators=300 ..................................
[CV]  max_depth=5, n_estimators=300, score=(train=0.992, test=0.983), total= 1.4min
[CV] max_depth=5, n_estimators=400 ..................................
[CV]  max_depth=5, n_estimators=400, score=(train=0.996, test=0.983), total= 1.8min
[CV] max_depth=5, n_estimators=400 ..................................
[CV]  max_depth=5, n_estimators=400, score=(train=0.996, test=0.981), total= 1.8min
[CV] max_depth=5, n_estimators=400 ..................................
[CV]  max_depth=5, n_estimators=400, score=(train=0.996, test=0.984), total= 1.8min
```

```
[Parallel(n_jobs=1)]: Done  84 out of  84 | elapsed: 42.7min finished
```

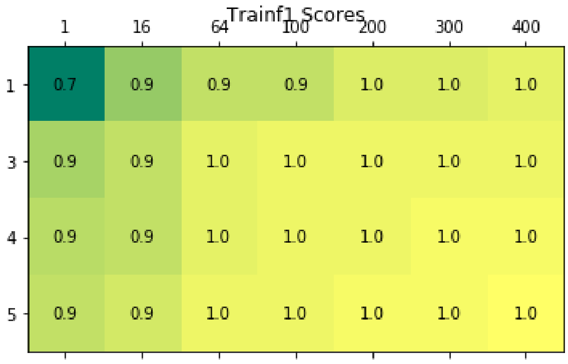
Out[0]:

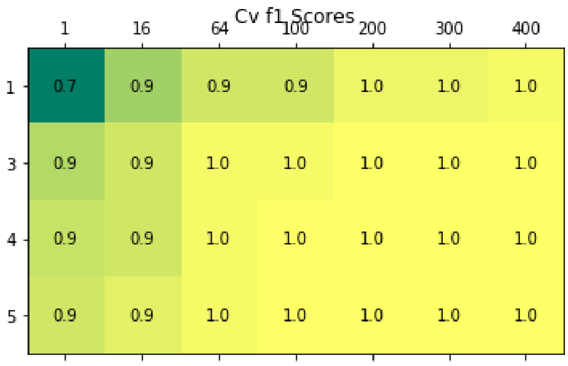
```
GridSearchCV(cv='warn', error_score='raise-deprecating',
             estimator=XGBClassifier(base_score=0.5, booster='gbtree',
                                     colsample_bylevel=1, colsample_bynode=1,
                                     colsample_bytree=1, gamma=0,
                                     learning_rate=0.1, max_delta_step=0,
                                     max_depth=3, min_child_weight=1,
                                     missing=None, n_estimators=100, n_jobs=1,
                                     nthread=None, objective='binary:logistic',
                                     random_state=0, reg_alpha=0, reg_lambda=1,
                                     scale_pos_weight=1, seed=None, silent=None,
                                     subsample=1, verbosity=1),
             iid='warn', n_jobs=None,
             param_grid={'max_depth': [1, 3, 4, 5],
                         'n_estimators': [1, 16, 64, 100, 200, 300, 400]},
             pre_dispatch='2*n_jobs', refit=True, return_train_score=True,
             scoring='f1', verbose=5)
```

In [0]:

```
# i made use of code from the previous assignment
#https://seaborn.pydata.org/generated/seaborn.heatmap.html
train_f1= clf.cv_results_['mean_train_score']
cv_f1 = clf.cv_results_['mean_test_score']
train_f1=np.around(train_f1, decimals=2, out=None)
cv_f1 = np.around(cv_f1, decimals=2, out=None)
train_f1=train_f1.reshape(4,7)
cv_f1=cv_f1.reshape(4,7)
#https://matplotlib.org/tutorials/colors/colormaps.html
```

```
#https://stackoverflow.com/questions/20998083/show-the-values-in-the-grid-using-matplotlib

def showAucPlot(text,data):
  labels = [['1','16','64','100','200','300','400'],['1', '3','4','5']]
  fig = plt.figure()
  ax = fig.add_subplot(111)
  cax = ax.matshow(data,cmap="summer")
  #https://matplotlib.org/tutorials/colors/colormaps.html

  #https://stackoverflow.com/questions/20998083/show-the-values-in-the-grid-using-matplotlib

  for (i, j), z in np.ndenumerate(data):
    ax.text(j, i, '{:0.1f}'.format(z), ha='center', va='center')
  plt.title(text)
  ax.set_xticklabels([''] + labels[0])
  ax.set_yticklabels([''] + labels[1])
  plt.show()

showAucPlot("Trainf1 Scores",train_f1)
showAucPlot("Cv f1 Scores",cv_f1)
```





In [0]:

```
print(clf.best_estimator_)
```

```
XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree=1, gamma=0,
              learning_rate=0.1, max_delta_step=0, max_depth=5,
              min_child_weight=1, missing=None, n_estimators=400, n_jobs=1,
              nthread=None, objective='binary:logistic', random_state=0,
              reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
              silent=None, subsample=1, verbosity=1)
```

In [0]:

```
clf=xgb.XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree=1, gamma=0,
              learning_rate=0.1, max_delta_step=0, max_depth=5,
              min_child_weight=1, missing=None, n_estimators=400, n_jobs=1,
              nthread=None, objective='binary:logistic', random_state=0,
              reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
              silent=None, subsample=1, verbosity=1)
```

In [0]:

```
clf.fit(df_final_train,y_train)
y_train_pred = clf.predict(df_final_train)
y_test_pred = clf.predict(df_final_test)
```

In [0]:

```
from sklearn.metrics import f1_score
print('Train f1 score',f1_score(y_train,y_train_pred))
print('Test f1 score',f1_score(y_test,y_test_pred))
```

```
Train f1 score 0.9939782420448099
Test f1 score 0.9195895762675481
```

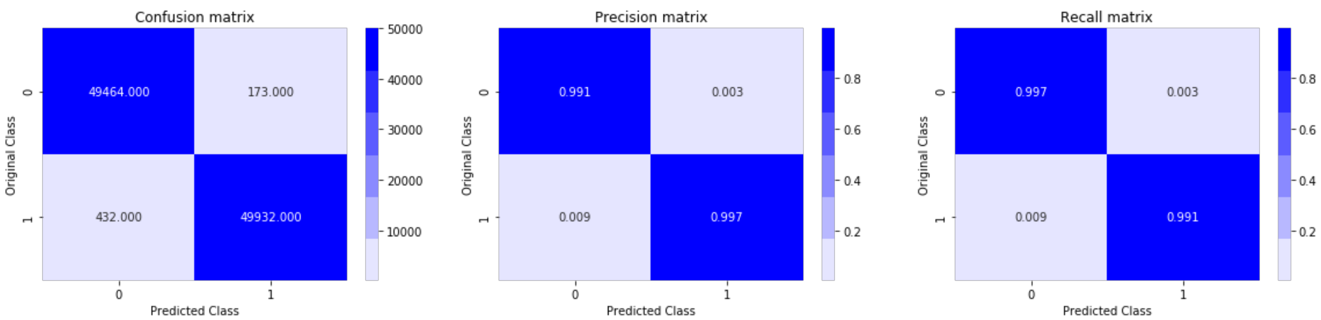
In [0]:

```
from sklearn.metrics import confusion_matrix
def plot_confusion_matrix(test_y, predict_y):
    C = confusion_matrix(test_y, predict_y)

    A =(((C.T)/(C.sum(axis=1))).T)

    B =(C/C.sum(axis=0))
    plt.figure(figsize=(20,4))

    labels = [0,1]
    # representing A in heatmap format
    cmap=sns.light_palette("blue")
    plt.subplot(1, 3, 1)
    sns.heatmap(C, annot=True, cmap=cmap, fmt=".3f", xticklabels=labels, yticklabels=labels)
    plt.xlabel('Predicted Class')
    plt.ylabel('Original Class')
    plt.title("Confusion matrix")

    plt.subplot(1, 3, 2)
    sns.heatmap(B, annot=True, cmap=cmap, fmt=".3f", xticklabels=labels, yticklabels=labels)
    plt.xlabel('Predicted Class')
    plt.ylabel('Original Class')
    plt.title("Precision matrix")

    plt.subplot(1, 3, 3)
    # representing B in heatmap format
    sns.heatmap(A, annot=True, cmap=cmap, fmt=".3f", xticklabels=labels, yticklabels=labels)
    plt.xlabel('Predicted Class')
    plt.ylabel('Original Class')
    plt.title("Recall matrix")

    plt.show()
```

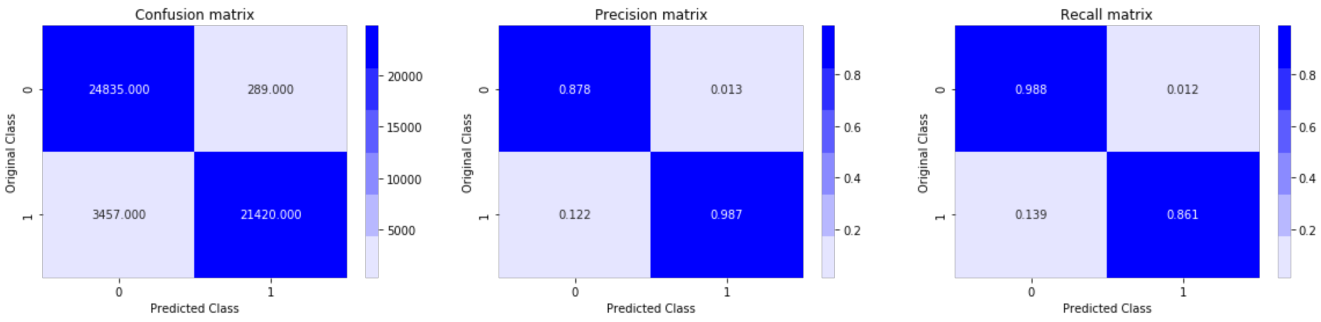
In [0]:

```
print('Train confusion_matrix')
plot_confusion_matrix(y_train,y_train_pred)
print('Test confusion_matrix')
plot_confusion_matrix(y_test,y_test_pred)
```

```
Train confusion_matrix
```

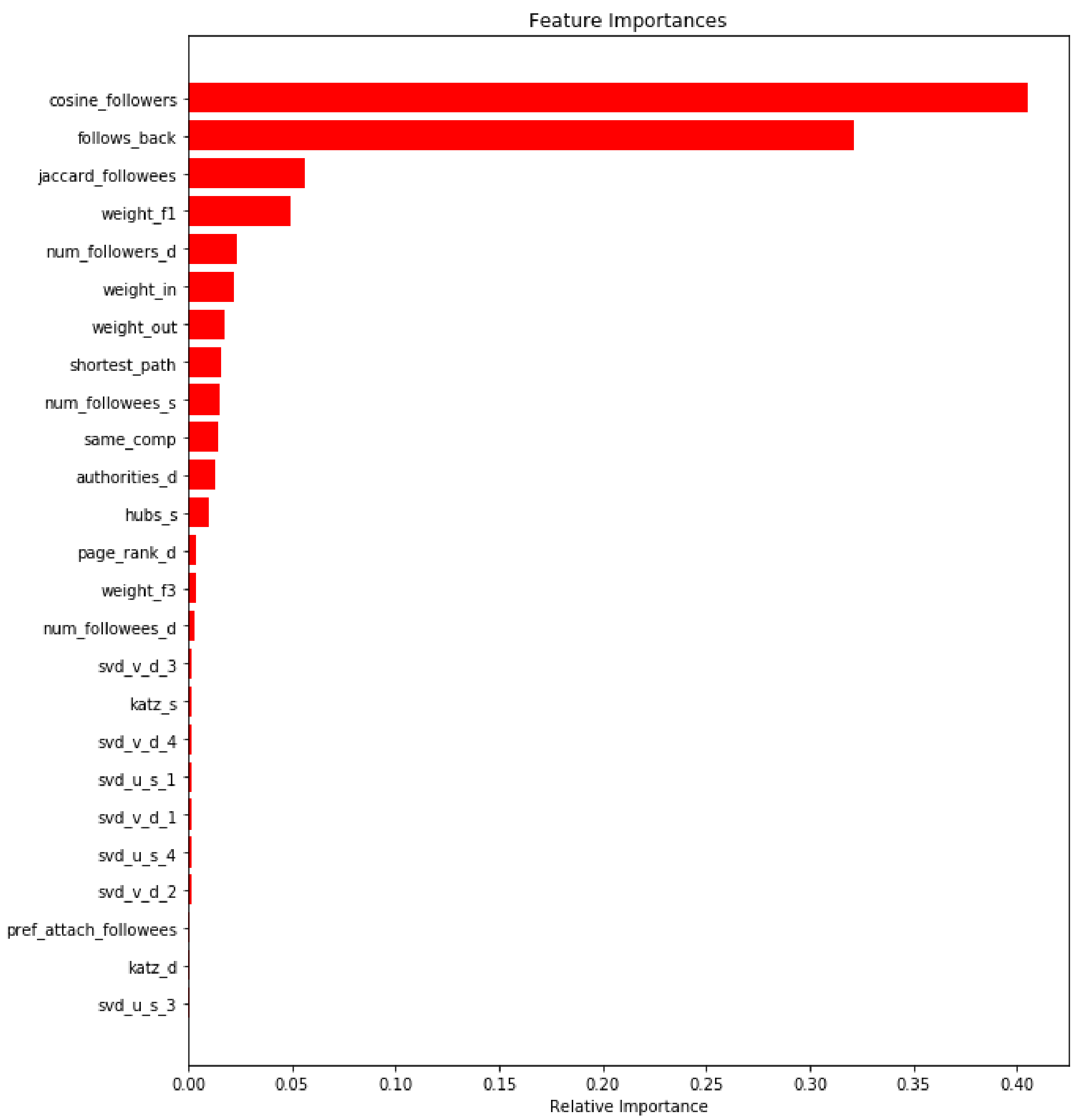


```
Test confusion_matrix
```



In [0]:

```python
features = df_final_train.columns
importances = clf.feature_importances_
indices = (np.argsort(importances))[-25:]
plt.figure(figsize=(10,12))
plt.title('Feature Importances')
plt.barh(range(len(indices)), importances[indices], color='r', align='center')
plt.yticks(range(len(indices)), [features[i] for i in indices])
plt.xlabel('Relative Importance')
plt.show()
```



Conclusion:

1. we have added two ne w feature namey preferential attachments for followers and followes and also svd_dot as Dot product between sourse node svd and destination node svd features
2. But both the features really are not contributing much to the model
3. we tuned the hyperparameters for Xgboost using Gridsearch.
4. the train and test fi scores are 0.99 and 0.92.