## Social network Graph Link Prediction - Facebook Challenge

**Problem statement:**
Given a directed social graph, have to predict missing links to recommend users (Link Prediction in graph)

**Data Overview:**
Data is taken from the facebook's recruiting competition hosted at Kaggle in 2012. It has 2 columns, a source_node (int64) and destination_node(int64)

**Mapping the problem into supervised learning problem:**
We first generate training samples of bad and good links from the given directed graph. Then we got some feature for each link, I.e total number of followers, total followee, page rank, Katz score Adar index hit score and svd features of adjancy matrix and some weight features etc. We the train two separate machine learning models on these features to get the best performing model.

**Performance metric for supervised learning:**
Both precision and recall are important so F1 score is good choice and plotted the Confusion matrix.

**To run the sample code,**

- Clone the repository: git
- Raw data is already present in zip form: Run the *0_utilities.ipynb* to unzip data
- The *data* folder contains all the processed files containing all the necessary features for model training. Trained models are also present in data folder.
- Make sure to have all the required libraries present in *lib.txt*
- First Time Run: Run the code files in the following sequence: 1_EDA, 2_SplitData, 3_Data_Featurization,4_PredictionModel_RandomForest, 5_Prediction Model_XGBoost

**References:**

Data: https://www.kaggle.com/c/FacebookRecruiting

Problem approach:

https://www.cs.cornell.edu/home/kleinber/link-pred.pdf

https://www3.nd.edu/~dial/publications/lichtenwalter2010new.pdf

https://www.youtube.com/watch?v=2M77Hgy17cg

https://kaggle2.blob.core.windows.net/forum-message-attachments/2594/supervised_link_prediction.pdf