Github Collaborator Recommendation

Albert Jian[a], **Bill Wen**[a], **and Shan Zhong**[a]

[a]University of California, Los Angeles

This manuscript was compiled on June 11, 2020

Recommender systems in the context of social networks recommends a likely acquaintance to a user. In this paper, we use concepts from network theory to develop features for machine learning to produce a recommender system to suggest GitHub followees to users. Using training and testing sets, we are able to demonstrate highly accurate recommender system using machine learning models. From the result, we found that the most important features are personalized PageRank scores, and how similar is the suggested followee to the current user's followees.

Machine Learning | Recommender System | Social Networks

During the recent rise of web services, many companies utilize different version of recommender system to recommend personalized items to their users, including Google's infamous YouTube algorithm. In social networks, recommender system recommend the more probable connections to the user.

In this paper we produce a recommendation system for a GitHub social network using machine learning approaches. We will first analyze our results and discuss their relevance, then we will describe the dataset and concepts from network theory and machine learning we used to develop the recommender system.

Results

Table 1. Model Hyperparameters and Training and Testing F_1 Score

Model	Hyperparameters(max_depth, n_estimators)	Train F_1	Test F_1
Random Forest	(14, 121)	0.96	0.92
Gradient Boosting Decision Tree	(10, 200)	0.99	0.99

Discussion

Materials

Dataset. The dataset GitHub Social Network is taken from SNAP (1). This dataset contains a large social network of GitHub developers collected from the public API in June 2019. Nodes are developers who have starred at least 10 repositories and edges are mutual follower relationships between them. The dataset consists of 37700 nodes and 289003 edges.

Exploratory Data Analysis. The degree distribution of the network is shown in figure 1. The largest number of mutual follows is 9458 and the minimum number is 1. 5045 users in the network have 1 mutual follow and around 25000 people have less than 5 mutual follows. As seen from both of the graphs, the degree distribution is quite left-skewed.

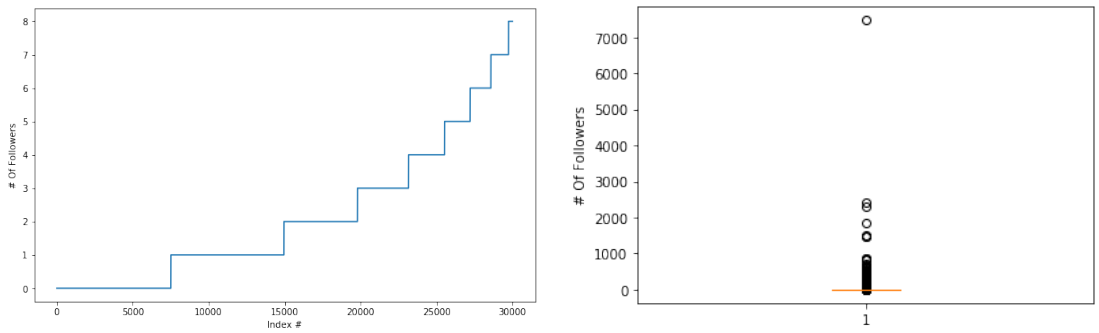
In order to visualize the network, we picked a central node from the training set and we performed a breadth-first walk of the group to a maximum distance of 3 as shown in figure 2. Nodes are sized according to their distance from the center and colored by their personalized PageRank with the central node.

Methods

The methods in this paper is mostly inspired by this research in link prediction in social networks (2). Binary classification method (Random Forest, Support Vector Machine and Gradient Boosting) was adopted to classify good links and bad links.

Significance Statement

Recommender system in social network aims to predict the most probable hidden friendship within the network and recommend it to the users. In this paper, we use network theory feature to develop machine learning model for a recommender system and analyzes the importance of different features within the model.



(a) Mutual follows for the least 30000 users

(b) Mutual follows box plot

Fig. 1. Mutual follows per user

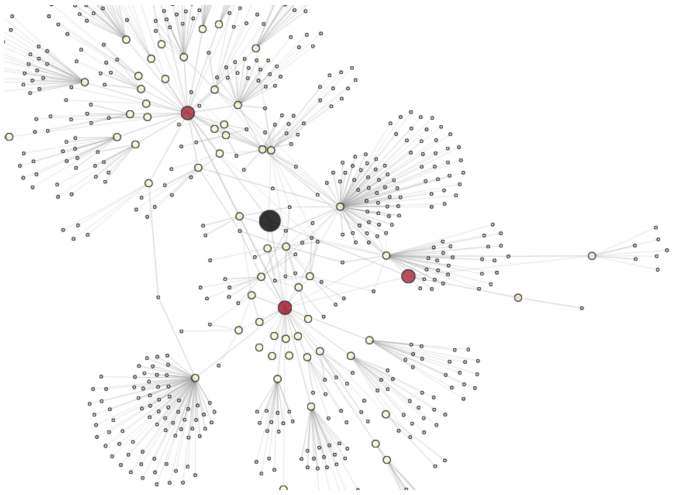


Fig. 2. Example Local Network with Distance of 3

SVD was applied to reduce the dimension of the dataset. Similarity measures such as jaccard distance, cosine distance and preferential attachment were calculated as features. Ranking measures (page rank), centrality measures (Katz centrality) and other metrics such as shortest path and common neighbors were also added for model training. Confusion matrix was plotted at the end to evaluate model performance. We trained for two separate models, one with Random Forest, RF, and one with Gradient Boosting Decision Tree, GBDT.

Features. (3)

Jaccard's Coefficient.

Preferential Attachment.

Machine Learning Models.

ACKNOWLEDGMENTS. We would like to acknowledge our professor Dr. Heather Zinn-Brooks and our teaching assistant Abigail Hickok for their accomodations during the remote instructions this quarter.

1. B Rozemberczki, C Allen, R Sarkar, Multi-scale attributed node embedding (2019).
2. W Cukierski, B Hamner, B Yang, Graph-based features for supervised link prediction in *The 2011 International Joint Conference on Neural Networks*. pp. 1237–1244 (2011).
3. A Sadraei, Link prediction algorithms (2014).