Network-based restaurant link prediction

Aditya Shah: 202003045
Mentor : Prof. Mukesh Tiwari

Introduction

In this project, we focus on predicting restaurant check-ins within the Foursquare network in New York City. Link prediction is an important problem in numerous network applications. Applications such as customized restaurant suggestions and strategic commercial alliances can be greatly enhanced by the capacity to predict future connections between eateries and anticipate customer preferences. Our methodology, dataset specifics, visualization of data, problem description, and score schemes are all described in this analysis.

Dataset Used

Data Summary	
Users	3,112
Venues	3,298
Check-ins	27,149
Tips	10,377

Data Files	
NY.Restaurants.checkins.csv	User ID, Venue ID
NY.Restaurants.tips.csv	User ID, Venue ID, Tip
NY.Restaurants.tags.csv	Venue ID, Tag Set

Network

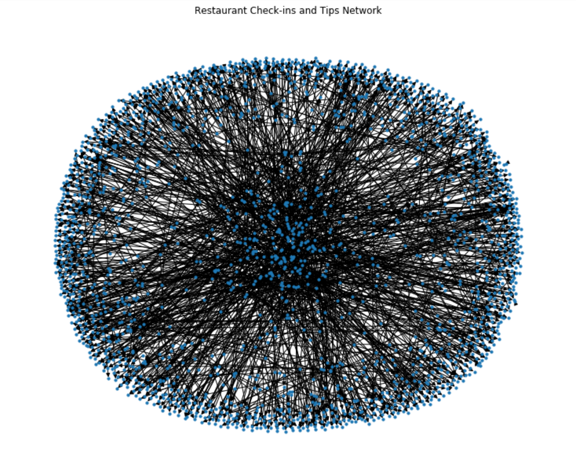


Figure 1: Restaurants Check-ins and Tips network

Degree Distributions

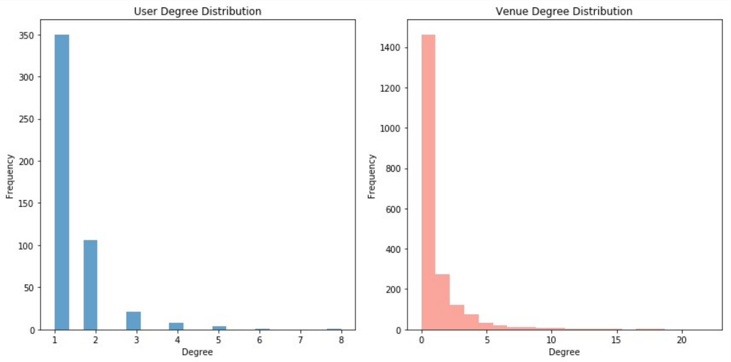


Figure 2: Degree distribution of User and Venue nodes

Problem Statement

The idea is to use the current Foursquare network to predict future linkages, or restaurant check-ins. Users and restaurants represent separate vertices in this bipartite network. Our goal is to create and evaluate scoring systems that forecast possible user-venue relationships in the future. Taking into account the changing dynamics of the social network, we must determine the most efficient scoring technique that strikes a balance between recall and precision.

Scoring Methods

- Distance Score
- Common Neighbors
- Tip-based Prediction
- Preferential Attachment-based Prediction
- Community Detection Score
- k-Nearest Neighbors (k-NN)

Distance Score

The scoring mechanism, denoted as score(x, y), for the connection between user x and venue y is defined as the negative of the shortest distance path. This emphasizes the preference for shorter distances in establishing edges.

$$\text{score}(x, y) = -\text{shortest\_distance\_path}(x, y)$$

Common Neighbors (User and Venue)

Objective: Forecast the probability of a future relationship by utilizing common neighbors.
Scoring Method(score(x,y)) for Unipartite Graph: It is the number of shared neighbors between nodes x and y.
Scoring Method(score(x,y)) for Bipartite Graph: For both users and venues in the bipartite graph:

- User Common Neighbours Score ($Score(x, y)_{user}$): It takes into account shared neighbors between users. It strengthens the link if user x and another user visit y and share a large number of visited venues.

- Venue Common Neighbours Score ($Score(x, y)_{venue}$): It takes into account shared neighbors between venues. The strength of the connection is affected by a different venue that draws a lot of the same people as the venue y

.

Tip-based Prediction

Objective: Predicting connections by utilizing suggestions from users.

Approach: To measure and forecast the textual similarity between tips, cosine similarity, and the TF-IDF representation are used. The measures for precision and recall offer useful data about the predicted accuracy of the tip-based method. Changing the cosine similarity threshold could affect how recall and precision are given off.

Preferential Attachment-based Prediction

Objective: Predicting connections based on node degrees. Attachment Preference Edges are scored in predictions by multiplying their degrees of connection with other nodes. The theory is that nodes with more connections, or higher degrees, have a higher probability of attracting new connections.

$$PAP(x, y) = N(x)N(y)$$

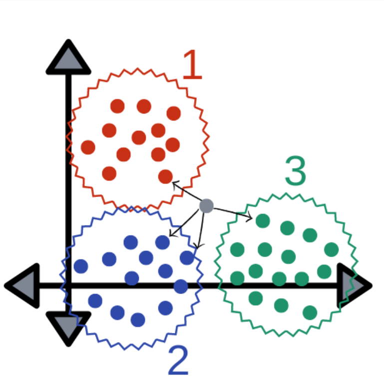
Community Detection Score

It involves identifying groups or communities of nodes within the network that have a higher degree of interconnectedness.

Scoring Mechanism (CDS(x, y)): The Community Detection Score aims to evaluate the possibility of a link between nodes x and y. If nodes belong to the same community, the score suggests a higher likelihood of a connection.

k-Nearest Neighbors (k-NN)

k-NN facilitates the identification of possible connections between users and venues by calculating the closeness of nodes based on shared attributes.



Results

Scoring Method	Precision (%)	Recall (%)
Shortest Distance	0.282	17.273
User Common Neighbors	34.029	8.854
Venue Common Neighbors	54.741	13.797
Tip-based Prediction	0.029	11.298
Preferential Attachment	0.270	4.780
Community Detection	2.849	15.970
k-NN	2.87	83.72

Table 1: Results of Link Prediction Techniques