Network-based restaurant link prediction

Aditya Shah (202003045)

Dhirubhai Ambani Institute of Information & Communication Technology, Gandhinagar, Gujarat 382007, India
SC 435, Introduction to Complex Networks

I. 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 customised 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, visualisation of data, problem description, and score schemes are all described in this analysis.

II. DATASET DETAILS

The FourSquare - NYC Restaurant Check-Ins Dataset includes check-in, tip, and tag data of restaurant venues in NYC collected from Foursquare from 24 October 2011 to 20 February 2012. It contains 3,112 users, 3,298 venues with 27,149 check-ins, and 10,377 tips [1][2].

- **NY.Restauraunts.checkins.csv** has two columns. Each line represents a check-in event. The first column is the user ID, while the second is the venue ID.
- **NY.Restauraunts.tips.csv** has three columns. Each line represents a tip/comment a user left on a venue. The first and second columns are the user ID and venue ID, respectively. The third column is tip text.
- **NY.Restauraunts.tags.csv** has two columns. Each line represents the tags users added to a venue. The first column is venue ID, while the second column is a tag set of the corresponding venues. Empty tag sets may exist for a venue if no user has ever added a tag to it.

III. DATA VISUALIZATION

We use degree distributions to explore the structure of our social network dataset and uncover patterns of connection between venues and users. We first look at the overall user-venue interactions, and then we concentrate on the views of individual users and venues.

The user-venue interactions network is subsequently illustrated using a spring arrangement, which offers a graphical representation of the complex links between venues and users. The resulting pattern demonstrates the intricacy of the social network structure by graphically capturing the interaction of links. The log-log scale highlights the existence of hubs or strongly linked nodes, which may represent significant users or well-liked locations within the network, and improves the clarity of distribution tails.

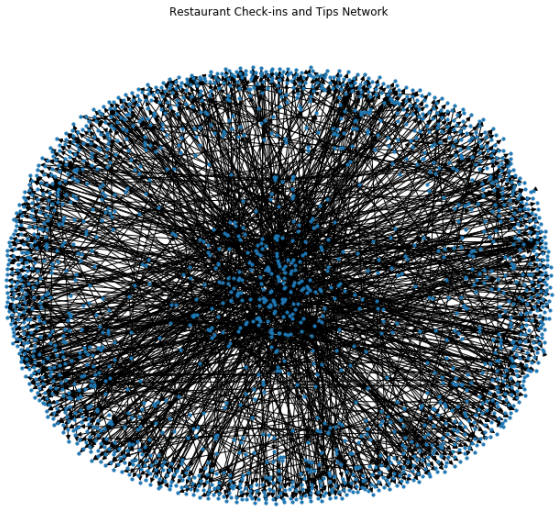

Figure 1: Restaurants Check-ins and Tips network

The degree distributions for people and venues are represented using log-log scale representations, which highlight connection patterns within the dataset. The user degree distribution shows that the majority of users have degrees between 1 and 8, which suggests a decentralised network structure where users with few connections are more common. This implies that there is a greater number of irregular users in the social network.

The degree distribution of the venue ranges from 0 to 22. Locations with a degree of 0 represent remote locations with no links to other locations, whereas locations with higher degrees—up to 22—indicate well-liked loca-

‡Electronic address: 202003045@daiict.ac.in

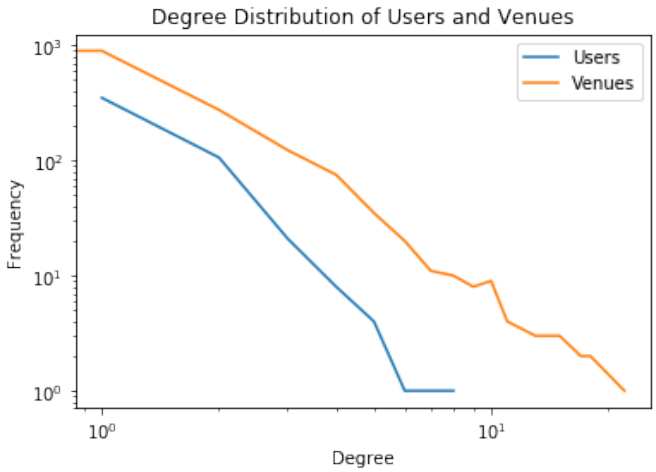

Figure 2: Degree distribution of User and Venue nodes on log-log scale

tions with large user bases. The different types of user-venue connections in the network are influenced by this variation in venue degrees.

IV. 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.

V. SCORING METHODS

To predict potential links in social networks, different scoring techniques were used to evaluate the probability of future connections between venues and users. Every technique takes a distinct approach when it comes to identifying the underlying patterns in the network.

A. Distance Score

For an edge connecting user x and venue y, the scoring mechanism represented by $\text{score}(x, y)$, is defined as the negative of the shortest distance path. Since edges with shorter distances are more likely to be established, the reasoning for using the negative distance as the score highlights this point.

In the implementation process, two sets are initialized: the source set S containing the user x and the destination set D having the venue y. In each iteration, the algorithm adds all neighbors of either set S or set D to their respective sets. The decision of which set to expand is based on the set with fewer neighbors at that particular iteration. This iterative process continues, gradually expanding the sets with neighboring nodes, and it aims to capture the potential connections by prioritizing the nodes with fewer neighbors, indicative of shorter distances in the underlying graph.

B. Common Neighbors:

The common neighbors scoring technique relies on the notion that a higher number of shared neighbors between two nodes increases the likelihood of a future connection. In the context of a unipartite graph, the common neighbor score ($\text{Score}(x, y)$) is given by the equation:

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

For our bipartite graph, there are two distinct methods to calculate the common neighbor's score:

- **Common Neighbors Score using User ($\textbf{Score}(x, y)_{\textbf{user}}$:**

$$\text{Score}(x, y)_{\text{user}} = \max_{xi \in \Gamma(y)} |\Gamma(x) \cap \Gamma(xi)|$$

  According to this scoring method, the connection between user x and venue y is strengthened if there exists another user who visits venue y and shares a significant number of visited venues with user x. This implies a higher probability that user x will visit the venue y.

- **Common Neighbors Score using Venue ($\textbf{Score}(x, y)_{\textbf{venue}}$:**

$$\text{Score}(x, y)_{\text{venue}} = \max_{yi \in \Gamma(x)} |\Gamma(y) \cap \Gamma(yi)|$$

  In this scoring method, the connection between user x and venue y is influenced by the existence of another venue that attracts a substantial number of the same users as venue y. If user x visits this common venue, there is a higher chance that user x will also visit venue y.

Here, $\Gamma(x)$ represents the neighbors of user x, and $\Gamma(y)$ represents the neighbors of user y.

C. Tip-based Prediction:

In order to forecast possible connections between users and venues in a bipartite network, the tip-based Prediction approach makes use of the content of user-provided tips or remarks. 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 have an effect on how recall and precision are given off.

D. Preferential Attachment-based Prediction:

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 in new connections.

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

E. Community Detection Score:

In order to evaluate the possibility of links, this scoring technique integrates community discovery methods. It is anticipated that nodes in the same community are more likely to be linked.

Identifying communities and calculating the likelihood of a link between nodes in the same community are the mathematical tasks involved in the Community Detection Score (CDS(x,y)).

F. k-Nearest Neighbors (k-NN):

Link prediction uses k-Nearest Neighbours (k-NN) because of its capacity to evaluate node similarity inside a social network. k-NN facilitates the identification of possible connections between users and venues by calculating the closeness of nodes based on shared attributes. The technique predicts links by analysing a node's k-nearest neighbours and using the similarity measure. In this particular implementation, the k-NN model is trained to identify comparable venues based on user check-in patterns, with the check-in counts acting as a feature vector. By utilising the network's intrinsic structure, this method enables k-NN to efficiently identify and forecast significant connections.

VI. RESULTS

Table I summarises the outcomes of the link prediction techniques. For each of the scoring methods—Negative Shortest Distance Score, User Common Neighbours Score, Venue Common Neighbours Score, Tip-based Prediction, Preferential Attachment Prediction, Community Detection Score, and k-nearest Neighbours (k-NN)—the table displays the percentages of Precision and Recall.

Table I: EER (in %) for corresponding features and classifiers

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

With a very high recall of 83.72%, k-NN is remarkably effective at identifying a significant portion of true positive cases. The corresponding Precision, at 2.87%, is comparatively lower, indicating a trade-off where the algorithm produces a significant number of false positives in addition to capturing a large number of actual positive linkages. This demonstrates how k-NNs naturally balance precision and recall, with emphasising precision frequently coming at the expense of recall.

The Common Neighbours scoring techniques yield scores for both User and Venue that are notable. The User Common Neighbours Score illustrates the importance of common connections among users in link prediction, with a Precision of 34.03% and a Recall of 8.85%. In the same way, the Venue Common Neighbours Score highlights the importance of common venues in link prediction with a Precision of 54.74% and a Recall of 13.80%. These findings support the hypothesis that users are more likely to make connections in the future if they frequent comparable venues or have more connections in common. The predictive effectiveness of these scoring techniques is significantly enhanced by shared connections, which is highlighted by Common Neighbours Scores as an important factor to take into account when forecasting relationships.

VII. CONCLUSION

The comparison of several link prediction techniques concludes by highlighting the small trade-offs between recall and precision, with each method displaying unique shortcomings and strengths. The k-NN method is distinguished by its high recall, which highlights its effectiveness in detecting possible connections. As the model produces a significant amount of false positives, precisely balancing this recall is a challenging task. Moreover, common connections are essential for utilising

local network architecture for link prediction, which is another benefit of Common Neighbours scoring methods for users and venues.

In order to maximise on the complementing advantages of various strategies, future research may investigate hybrid models that combine numerous scoring methods. Link prediction models may also be more accurate if dynamic features of the network are examined, taking into account temporal fluctuations and changing user preferences. Deep learning architectures and other sophisticated machine learning methods may provide fresh perspectives on identifying complex patterns in social networks. By offering a substantial comprehension of the link prediction landscape, the research opens up new avenues for methodological investigation and improvement in the ever-evolving field of social network analysis.

References

[1] D. Yang, D. Zhang, Z. Yu, and Z. Yu, "Fine-grained preference-aware location search leveraging crowdsourced digital footprints from lbsns," in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, ACM, 2013, pp. 479–488.

[2] D. Yang, D. Zhang, Z. Yu, and Z. Wang, "A sentiment-enhanced personalized location recommendation system," in *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, ACM, 2013, pp. 119–128.