Social Link Prediction in Mobility Data

Allan Frederick
University of Texas at Austin
Machine Learning on Real World Networks
allanfrederick1224@utexas.edu

David Johns
University of Texas at Austin
Machine Learning on Real World Networks
djohns442@gmail.com

Abstract

This project explores a social network leveraged from human mobility and points of interest data, where potential social interactions will be identified using link prediction. We investigate whether human mobility data can provide valuable insights into social interaction patterns and the formation of social networks. Two main networks are constructed: first, a scale-free person-to-person network connecting individuals based on shared location information, where edge weight is determined by time overlap. Second, a people-to-places network is also established, connecting people to points of interests (POIs) which are geographically mapped. A two-layer GCN and a two-layer GraphSAGE model are implemented to perform link prediction on both the people-to-people and people-to-places networks in order to predict the likelihood of interactions between people and where they might meet.

1. Introduction

This project will attempt to leverage location-based data, specifically human mobility and points of interest, to gain insight into who will interact with whom, and the time and place of these interactions. Understanding the nature of this particular network is valuable in public health and safety like virus spread prediction, city planning, and traffic monitoring [7]. Link prediction can be an effective technique in modeling how information will propagate in a given network, as it is used to identify unknown edges or edges that might appear in the future. We will evaluate the role which human mobility plays in the formation of social networks, and how it can be used to make inferences about the nature of interactions and predictions about future interactions.

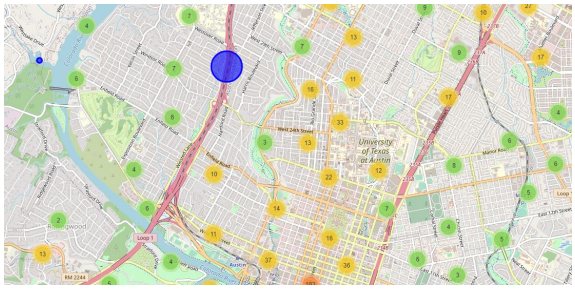


Figure 1: Illustration of POI and mobility data from one day in Austin, TX

2. Previous Work

Link prediction is a relatively new field, beginning in 2003 with the common neighbor counting approach [3]. Since then, many approaches to link prediction have been investigated for social network inferences, including supervised random walks [1] and deep learning [2]. Other approaches for social-link prediction have focused on multilayer networks, including both Foursquare and Twitter data, and have also devised unique similarity functions for link prediction [6]. Some attempts at using traditional machine learning classification methods on POI and person feature vectors have also been attempted [5].

These works have have mostly focused on spatial information for prediction. A recent article explored link prediction particularly using both spatial and temporal information to develop a ”multiview matching network”, which creates representations of users through location, time, and relation [8]. Results suggest that including spatiotemporal aspects improves link prediction.

Our approach will focus on utilizing several aspects from previous works such as biased random walk strategies for structural embeddings and graph neural networks. This will allow us to perform link predictions considering a two-pronged network approach built from spatiotemporal data.

3. Approach

We employ a dual strategy for link prediction, creating both people-to-people and people-to-places networks. Node embeddings from these networks serve as input for link-prediction models. In the people-to-people network, individuals are linked based on shared location information, with nodes representing individuals and links indicating shared locations at the same hour. Link weights reflect the normalized overlap time, ranging from 0 to 1, signifying the extent of shared time. Deeper interactions are differentiated by varying weights. Our assumption is that interactions themselves are defined by individuals sharing the same environment, such as breathing the same air but not necessarily communicating.

The location network links individuals and venues, connecting individuals to the checked-in venues. Nodes represent both individuals and venues, and weights are assigned to edges using SAG scores. This helps to depict strong links between individuals and places. We assume that people's data points involve multiple venue visits, creating overlap for community detection and mobility analysis. The people-to-people network provides insights into interactions based on frequency and duration. The places network reveals where interactions occur and highlights sub-communities within locations.

Following network construction, an initial Node2Vec embedding was performed for each network. A training loop was implemented over Node2Vec to optimize the embeddings for facilitating the downstream link-prediction tasks; the more optimized the Node2Vec embeddings, the more efficient training is for the link-prediction models. After Node2Vec embeddings, the networks were split into positive and negative graphs, with each graph containing the same set of nodes as the original graph. The links in the positive graphs represented existing edges in the original graph, while links in the negative graphs represented unconnected node pairs in the original network. The original networks were split into training, validation, and testing subgraphs, where validation and testing were both ten percent of the entire network. For the people-to-places network, data from the first 3 weeks was used as training data and the last week as used as testing data. The people-to-people network training, evaluation, and testing data was split randomly. The Node2Vec embeddings for the training subgraph was used as the initial input features to the link-prediction models.

Using the Deep Graph Library, we constructed a two-layer GCN and a two-layer GraphSAGE model to implement link prediction on each network for each month (July and August). Models were implemented for each month separately in order to capture shorter-term information that could otherwise be lost by including both months in the training. The link prediction models were learned by computing dot-product similarity between node pair embeddings for both the positive and negative graphs of the training subgraph. The link-prediction models were trained over 100 epochs with a binary cross entropy loss function. The final output resulted in a positive and negative link score for each node pair. A high positive score means a high likelihood of a link existing, which would correspond to a low negative score for that pair of nodes. A complete workflow of the pipeline is illustrated in 2.

With learned node embeddings as input features to the GCN and GraphSAGE models, we were able to effectively predict the existence of links in both the people-to-people and people-to-places networks.

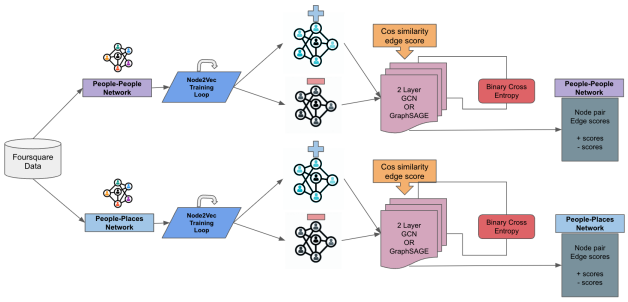


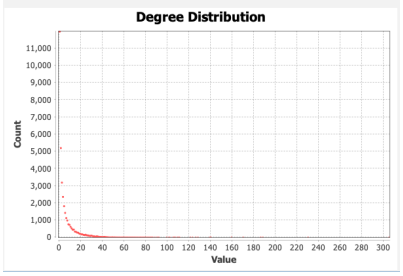
Figure 2: Link-prediction pipeline. Two different networks are constructed from Foursquare data, each network undergoes Node2Vec training loop for initial embeddings. Then the networks are split into positive and negative graphs and input into the GNNs for training. Outputs are the edge scores predicting existence of links.

4. Experimental Setup and Results

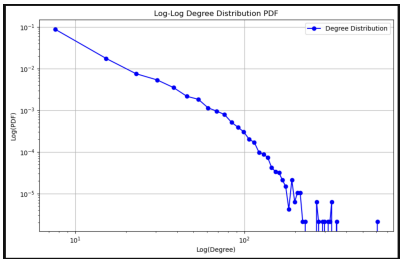
4.1 Network construction

Preliminary analysis of the complete Austin people-to-people network reveals a power law degree distribution, as seen in Fig 3. Individuals with high connections are rare (hubs) while most individuals have a low connection degree. This suggests that the nature of these interactions is scale-free, which is to be expected given that social networks tend to be scale-free in nature. In terms of demographics, the most prominent groups are females aged 30-40 and males aged 30-45. This demographic corresponds to SAG scores ranging 100-150, demonstrated by the SAG scores PDF in Fig. 4.

We can also see the power law demonstrated in the people-to-places network as seen in Fig 5. This implies the venues represent a scale-free network and suggests the emergence of clear hubs, such as the airport and grocery stores.



(a) Degree count



(b) Log-Log scale

Figure 3: People-People network degree distribution in, Austin, TX

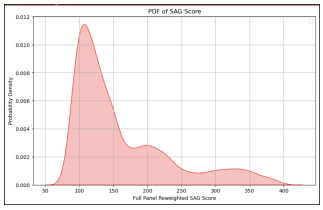


Figure 4: PDF of the SAG Score

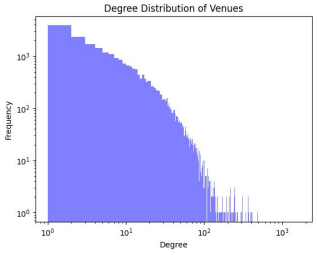


Figure 5: Degree distribution of Austin venues in July 2020

4.2 Network embedding and learning

The performance of the Node2Vec embeddings is demonstrated in figure 7. The training loss decreases over the number of epochs, indicating that the embeddings are being learned and optimized. However, the minimized loss was found to be around 1.1, suggesting that there is still room for further optimization. Yet, performance is decent enough to use as input into a link-prediction model because the embeddings undergo further optimization in the GNNs.

The test performance of the GCN model for the person-to-person July network is illustrated in 8. In addition, the test performance of the GraphSAGE model for the person-to-places July network is illustrated in 9. A complete comparison of all the models for each network can be seen in the tables 1 and 2. The people-to-people GCN and GraphSAGE models perform quite well in comparison to random chance. GraphSAGE outperforms the GCN for the month of July, while the GCN outperforms GraphSAGE for the month of August. The people-to-places models also perform reasonable well above chance level (indicated by the blue diagonal line) seen in 9. GraphSAGE outperforms GCN for both July and August in the people-to-places networks.

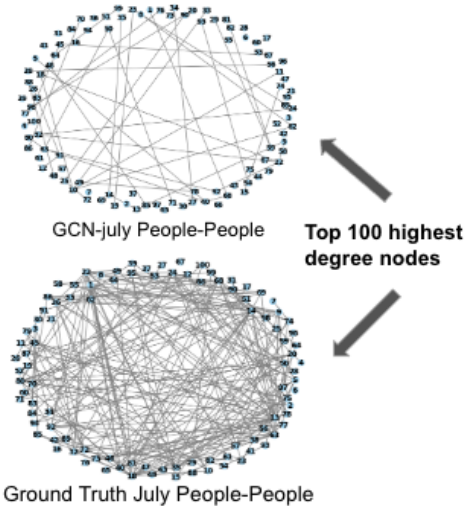


Figure 6: Predcited node pairs vs. ground truth node pairs

Model	Train Loss	Val AUC	Test AUC
GCN-july	0.473	0.905	0.901
GraphSAGE-july	0.470	0.933	0.932
GCN-aug	0.491	0.902	0.901
GraphSAGE-aug	0.516	0.896	0.896

Table 1: People-people model performances.

Model	Train Loss	Val AUC	Test AUC
GCN-july	0.502	0.912	0.909
GraphSAGE-july	0.498	0.923	0.921
GCN-aug	0.512	0.897	0.889
GraphSAGE-aug	0.507	0.905	0.906

Table 2: People-to-places model performances.

Figure 6 highlights the main differences between the predicted node pairs and the ground truth node pairs of the

top 100 highest degree nodes of the test network. The ground truth mobility network clearly demonstrates higher interconnectivity between the highest ranking node pairs, as compared to the predicted links. It is important to note that despite this large contrast, the performances of the prediction models were still decent. Figure 6 only illustrates 100 nodes, whereas the entire network comprises of thousands of other nodes such that the plots of the predicted network vs ground truth network become virtually indistinguishable.

In Figure 10 and Figure 11 we have some predictions shown in comparison to the ground truth. These predictions are from the last week of July 2020. On the left we have a map with the predictions from the people-to-places network, while on the right is the ground truth data. The numbers inside the circles represent the number of people that have visited that area. The more orange the circle, the more people have visited that location. As you can see the predictions somewhat closely represent the ground truth data. For example, the most popular destination, the airport, was predicted with 77% precision and 84% recall

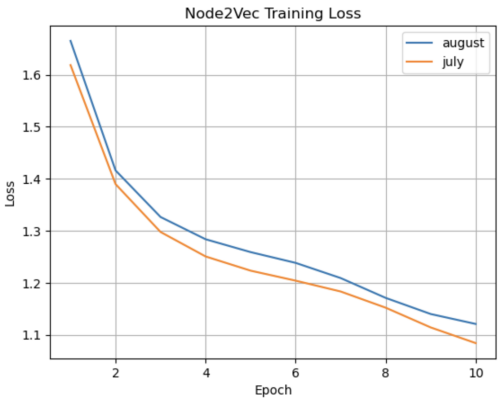


Figure 7: Training loss of Node2Vec embeddings on people-people network for both July and August

These results suggest that it is possible for us to accurately predict who will interact with whom, as well as where people will go in the future. It is possible to use these two predictions together to then predict if two people will interact and where they will interact. One current limitation of our approach is the separation between the two systems. We must consider that just because two people are likely to visit the same location and are likely to interact with each other does not necessarily mean they will interact at that location. This subtlety is somewhat absent in our approach. It could, however, be solved by adding links between people in the people-to-places network. It is important to note another limitation, that the link predictions on each network are being performed on static networks and do not include any temporal aspect in model learning. This limitation could potentially be handled by extending the link prediction task

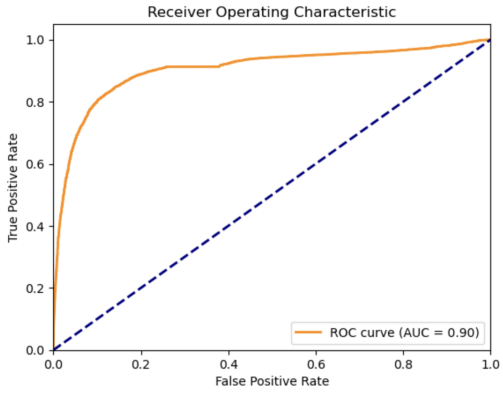


Figure 8: ROC AUC of GCN link-prediction test performance on people-people network July network

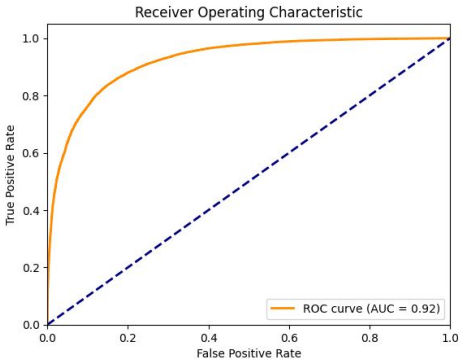


Figure 9: ROC AUC of GraphSAGE link-prediction test performance on people-to-places network July network

into an edge classification that would distinguish an edge among two or three time bins (morning,afternoon,evening). This could be helpful in predicting when a link may occur.

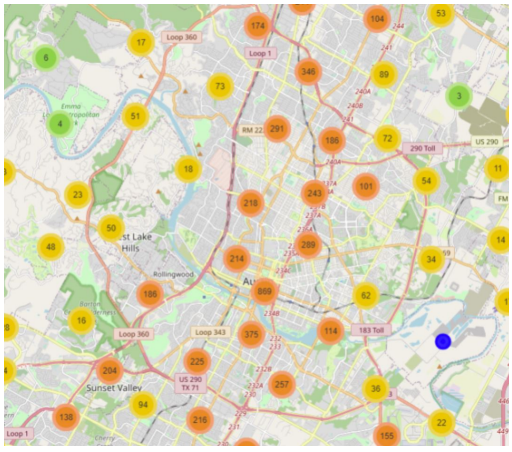


Figure 10: Predictions on people-to-places network on last week of July of 2020

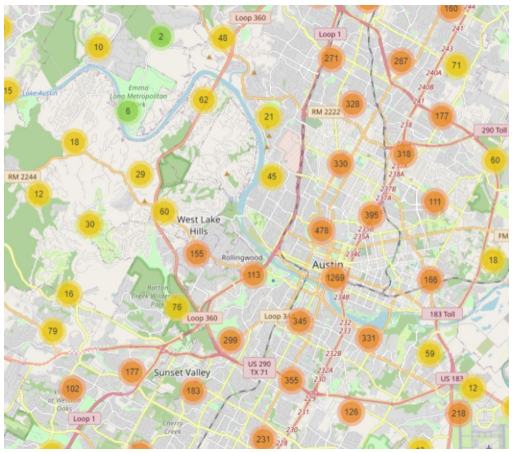


Figure 11: Ground truth of people-to-places network on last week of July of 2020

5. Conclusion and Future Work

This project aims to use the power of link prediction with human mobility and point of interest data to understand and predict social interactions, potentially influencing public health and safety. The social interactions exhibit scale-free properties, where the locations of their interactions can be also be geographically mapped. Node embeddings were created for both people-people and people-to-places networks, which were fed as input features into two-layer GCN and two-layer GraphSAGE models for each month of data. The models generally performed well for each type of network for each month. Thus, we have been able to mostly predict whom will interact with whom, and where these interactions may potentially occur. Future work would incorporate training on the the temporal aspects of each network, Foursquare user check-in times, in order to predict when an interaction might occur.

6. Contributions and Lessons Learned

Each team member worked on one prong of the dual-pronged approach. People-to-people network creation, visualizations, embedding, and link prediction model training was done by Allan. The people-to-places network creation, visualizations, embedding, and link prediction model training was done by David.

We learned how to create networks out of an appreciably large dataset, analyze the characteristics of these networks, perform embeddings on these networks, and construct link prediction models for each network with decent performance. This did not include the temporal aspect of the networks but we learned how to approach the ”who” and ”where” components of link prediction in a contact network. A significant component to approaching this project involved how to deal with such a large dataset, which is why we chose to implement two separate networks in parallel, to break the project down into more manageable data. Another significant component to this project building the networks themselves, and manipulating the network into being compatible for Node2Vec and Link prediction model training. Perhaps just as importantly, we learned how to improve our paper writing and presentation techniques to communicate more effectively.

References

[1] L. Backstrom and J. Leskovec. Supervised random walks: Predicting and recommending links in social networks. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM ’11, page 635–644, New York, NY, USA, 2011. Association for Computing Machinery.

[2] X. Li, N. Du, H. Li, K. Li, J. Gao, and A. Zhang. A deep learning approach to link prediction in dynamic networks. pages 289–297.

[3] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. 2003.

[4] A. Rossi, D. Barbosa, D. Firmani, A. Matinata, and P. Merialdo. Knowledge graph embedding for link prediction: A comparative analysis. 15(2):1–49, 2021.

[5] S. Scellato, A. Noulas, and C. Mascolo. Exploiting place features in link prediction on location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’11, page 1046–1054, New York, NY, USA, 2011. Association for Computing Machinery.

[6] R. Yang, C. Yang, X. Peng, and A. Rezaeipanah. A novel similarity measure of link prediction in multi-layer social networks based on reliable paths. *Concurrency and Computation: Practice and Experience*, 34(10):e6829, 2022.

[7] Y. Yang, N. V. Chawla, P. Basu, B. Prabhala, and T. La Porta. Link prediction in human mobility networks. pages 380–387, 2013.

[8] W. Zhang, X. Lai, and J. Wang. Social link inference via multiview matching network from spatiotemporal trajectories. *IEEE Transactions on Neural Networks and Learning Systems*, 34(4):1720–1731, 2023.