Vivian Feng

AP Lang Period 1

Science Essay First Draft

December 5, 2022

Word Count: 1173

Graph Embeddings and Link Prediction

One of my favorite novels is *The Lord of the Rings*. While the novel's text is an excellent literary work, its format is not conveniently parsable for computer analysis. For example, it’s evident that Frodo and Gandalf are friends, but how separated are Frodo and a character from an earlier era, such as Fëanor? Both of them play a critical role in *The Lord of the Rings* mythology, but it's difficult to quantify their degree of separation just by reading the book. However, by using a network of hyperlinks between the different character pages on *The Lord of the Rings* fan site tolkiengateway.net as a proxy for direct interactions between characters, their narrative closeness is easily visualized using a scatterplot. Through the Computer Systems Lab, I am implementing and running various link prediction algorithms on *The Lord of the Rings* hyperlink graph and other networks to better understand the properties of networks hidden in plain sight.

A network, or graph, is a set of nodes linked together by connections known as edges. The link prediction problem aims to find unobserved edges in incomplete networks (Chen et al., 2022). Though rarely noticed, network link prediction is present everywhere. Product recommendations on online shopping websites, such as Amazon, use networks of customer purchase history to predict what shoppers might be interested in buying next. Netflix uses networks of watch history to suggest new shows for users to watch, and Spotify does the same, but with music (Schrage, 2022).

The simplest class of link prediction algorithms is the node similarity heuristic. Node similarity heuristics assign a score to pairs of nodes, where a higher score indicates a greater likelihood of an edge. The scores are calculated from the properties of the nodes, such as the number of common neighbors, in which common neighbors are shared adjacent nodes (Chen et al., 2022). As Liu et al. (2023) note, while effective in many cases, similarity heuristics make assumptions about networks that are not universal. For instance, in a friend network, a higher number of mutual friends (common neighbors) between two people (nodes) suggests that two people are more likely to know each other. However, in protein interaction networks, two proteins that share many mutual interaction proteins are not necessarily more likely to interact with each other (Liu et al., 2023).

Recently developed algorithms using machine learning and artificial intelligence techniques ameliorate many of the issues faced by node similarity heuristics (Liu et al., 2023). Directly performing computations on large networks, which can have millions of edges, is slow and impractical. To improve algorithm efficiency, researchers attempt to map networks to lower-dimension representations in the form of coordinates, equivalently termed coordinate vectors. These coordinate vectors are known as embeddings. Embeddings also represent graphs in a more standardized fashion. They convert graphs into a format that is conveniently inputted into neural network architectures, a class of powerful mathematical modeling tools.

Researchers use random walks as a way of obtaining a representative sample of graph characteristics. Random walks are a random sequence of nodes where each node in the sequence is a neighbor of the preceding one. Grover and Leskovec (2016) developed the random-walk-based node embedding technique Node2Vec. It treats random walks as “sentences” and nodes as “words.” Then, these sequences are fed into a neural network framework originally developed for representing words as coordinate vectors, called the Skip-gram architecture, to generate a simpler representation of the original network. The embeddings are then used as inputs to a classifier model. Current research analyzing networks frequently uses Node2Vec as an experimental baseline.

An alternate method of graph embedding generation uses a graph neural network. Each node is assigned an initial representation coordinate, and the graph neural network progressively updates these representations by aggregating its neighbors' coordinates into a single vector and using the aggregation to update the node representation (Liu et al., 2023). Likewise, these embeddings can be chained with other architectures to build link prediction models.

Node embeddings' usefulness is not limited to link prediction. Other tasks such as edge and node classification, wherein a model tries to predict which group a node or edge belongs in, also use embeddings. However, most embedding methods focus on the topological structures of networks. For classification tasks, focusing only on graph structure neglects the predictive power of other node properties. Bielak et al. (2023) propose the method AttrE2Vec, which combines edges of a random walk and feature vector describing nonstructural properties of an edge using an encoder to generate the embedding vector representing the edge. The lead author Piotr Bielak (personal communication, November 20, 2022) explains that while AttrE2Vec currently cannot be used for link prediction, it could serve as the foundation for improved link prediction models.

Applications of link prediction are field-agnostic. Link prediction algorithms that rely only on the structure of a graph can be ported easily from one field to another. For instance, Grover and Leskovec (2016) originally used Node2Vec on social networks, but Kim et al. (2022) used the algorithm on smart home device networks to equal effect. Random walk sampling is widely used. Predicting edges in protein-protein interaction networks and predicting edges in an

Internet of Things (smart home devices) application networks both rely on the use of random walks to represent graph features for use in graph embeddings (Kim et al., 2022; Nasiri et al., 2021).

The field-agnostic nature of link prediction algorithms inspired me to test link prediction algorithms on *The Lord of the Rings* hyperlink network. To ensure my study covered a broad range of link prediction algorithm use cases, I also included a dolphin pod interaction network, an ecological communities network, a road network, and an airplane flight network in my dataset. Each of these graphs has around a few thousand edges. I am testing some node similarity heuristics and Node2Vec on this dataset. To test an algorithm, I partition a network's list of edges into a "training" and "test" set. Then, I build an incomplete graph from the training set and try and identify edges in the set of nonexistent edges of the training set that fall in the test set. After obtaining Node2Vec embeddings, I project these coordinates into the 2D plane and build a scatterplot to visualize the nodes. Furthermore, I will use the embeddings to train a logistic regression link prediction model. So far, I have found node similarity heuristics to be fairly robust, consistent with the results of earlier research.

Link prediction provides hints of the direction of the next major scientific advancement. For example, the algorithm proposed by Nasiri et al. (2021) for identifying protein-protein interactions can be used to uncover new metabolic pathways. The study of link prediction on Internet of Things networks by Kim et al. (2022) provides insights into user behavior that can guide product design. The state of link prediction technology is in constant flux, and my research into the efficacy of established algorithms on novel graphs, such as *The Lord of the Rings* hyperlink graph, continues to expand the scope of this technology.

References

Bielak, P., Kajdanowicz, T., & Chawla, N. V. (2022). AttrE2vec: Unsupervised attributed edge representation learning. *Information Sciences*, *592*, 82-96. https://doi.org/10.1016/j.ins.2022.01.048

Chen, Y.-L., Hsiao, C.-H., & Wu, C.-C. (2022). An ensemble model for link prediction based on graph embedding. *Decision Support Systems*, *157*, 113753. https://doi.org/10.1016/j.dss.2022.113753

Grover, A., & Leskovec, J. (2016). Node2vec: Scalable feature learning for networks. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 855-864. https://doi.org/10.1145/2939672.2939754

Kim, S., Suh, Y., & Lee, H. (2022). What IoT devices and applications should be connected? Predicting user behaviors of IoT services with Node2vec embedding. *Information Processing & Management*, *59*(2), 102869. https://doi.org/10.1016/j.ipm.2022.102869

Liu, X., Li, X., Fiumara, G., & De Meo, P. (2023). Link prediction approach combined graph neural network with capsule network. *Expert Systems With Applications*, *212*, 118737. https://doi.org/10.1016/j.eswa.2022.118737

Nasiri, E., Berahmand, K., Rostami, M., & Dabiri, M. (2021). A novel link prediction algorithm for protein-protein interaction networks by attributed graph embedding. *Computers in Biology and Medicine*, *137*, 104772. https://doi.org/10.1016/j.compbiomed.2021.104772

Schrage, M. (2022, April 27). *The recommender revolution*. MIT Technology Review. Retrieved November 29, 2022, from https://www.technologyreview.com/2022/04/27/1048517/the-recommender-revolution/