Geodesic and Neural Features for Link Prediction in the COVID-19 Biomedical Knowledge Graph

Lucas Hurley McCabe
Department of Mathematics, The George Washington University

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Introduction

The volume of new research publications is prohibitively vast, motivating scientific workflows driven by literature-based discovery. We analyze the COVID-19 bibliometric knowledge graph generated by ORNL, which represents biomedical concept relations extracted from 435 681 publications [1]. We compare three machine learning pipelines for automated hypothesis generation.

A concept-concept edge relation is *consistent* with a publication if both concepts are neighbors of a publication in the publication-concept network. We generate G_{TC} from all concept-concept edge relations whose latest consistent publication was prior to June 2020. Edges in G_{TC} are positive training pairs. The remaining edges are withheld as positive samples for validation (positive validation pairs).

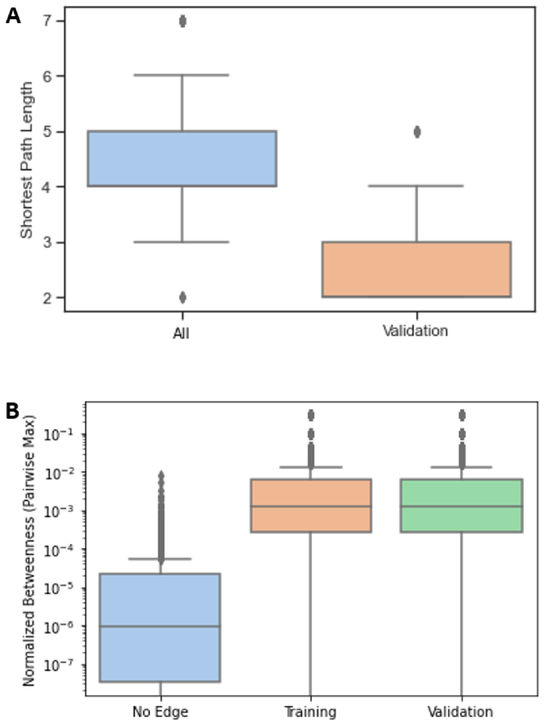

Figure 1: **(A)** Distributions of distances between all pairs of nodes in G_{TC} (All) and positive validation pairs (Validation). **(B)** Distribution of pairwise max normalized betweenness centrality in node pairs who have no edge in training or validation (No Edge), positive training pairs (Training), and positive validation pairs (Validation).

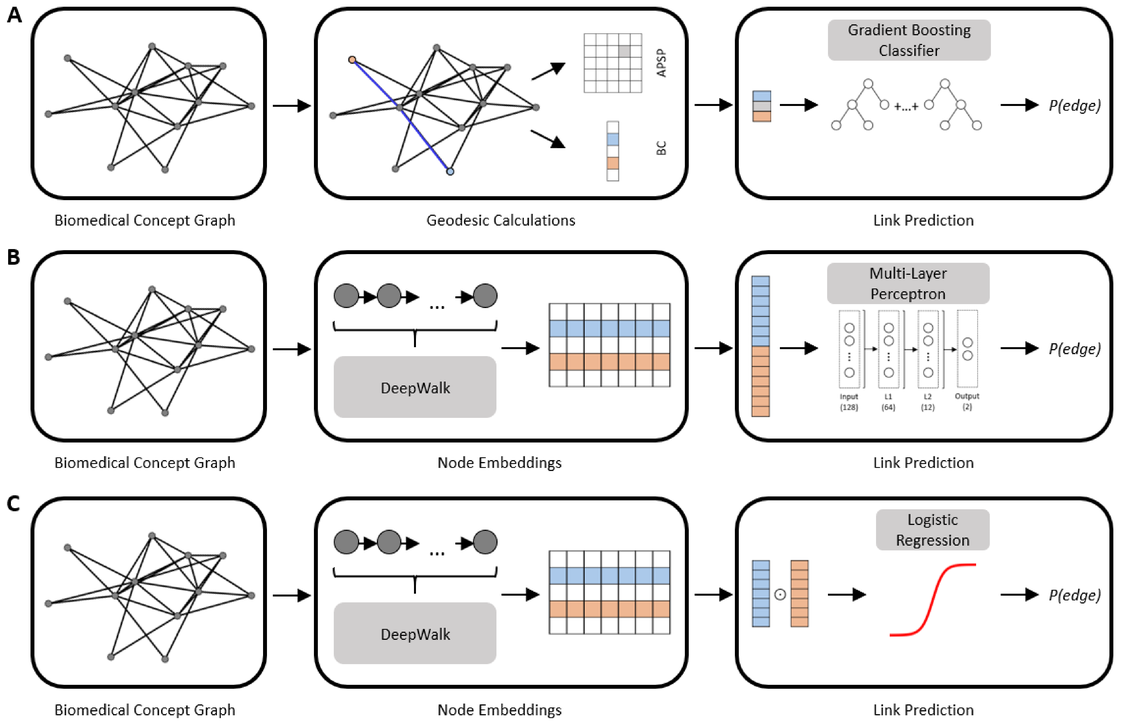

Figure 2: Illustration of link prediction pipelines: **(A)** *Geodesic+GBC*, **(B)** *DeepWalk+MLP*, and **(C)** *DeepWalk+LR*.

Methodology

We implement and assess three link prediction pipelines for automated hypothesis generation:

- *Geodesic+GBC*: All-pairs shortest paths distances and betweenness centrality metrics are calculated based on G_{TC}. For each hypothesized edge, a gradient boosting classifier estimates the probability of the edge based on the betweenness centrality of each node and the length of the shortest path between them (Figure 2A).
- DeepWalk node embeddings are computed based on G_{TC} [2]. We construct two embedding-based pipelines:
  - *DeepWalk+MLP* : For each hypothesized edge, a multi-layer perceptron estimates the probability of the edge based on the embeddings of each node (Figure 2B).
  - *DeepWalk+LR* : For each hypothesized edge, we use the Hadamard operator to generate a combined representation of the edge's nodes' feature vectors. We estimate the probability of the edge using this combined representation via logistic regression (Figure 2C).

While shortest path length is an informationally-dense feature for link prediction (Figure 1A), it provides several practical challenges. First, since the generation of training data warrants sampling positive edges and removing them from the graph, only a small number of positive samples can be used without excessively sparsifying the source data. For this reason, we limit the positive training data to 34 668 positive edge samples (10% of G_{TC}'s edges).

Such a procedure may make the graph disconnected, resulting in incalculable shortest path lengths. Where edge removal results in disconnected vertices in the training data, we replace the shortest path length with the effective diameter [3].

Results and Discussion

We evaluate performance on a class-balanced validation set, considering classification accuracy (ACC), area under a receiver operating characteristic curve (AUC), and average precision score (AP), as outlined in the table below.

Pipeline	ACC	AUC	AP
DeepWalk+MLP	70.0	77.6	76.4
DeepWalk+LR	74.5	82.1	83.7
Geodesic+GBC	**96.8**	**99.4**	**99.5**

Geodesic statistics are exceedingly effective features for link prediction, but their advantage comes at significant computational and memory costs for large graphs. Additionally, as the knowledge graph grows over time, re-computation of features must be repeated from scratch, making the time-consuming task of computing geodesic features unscalable. As such, the more moderately performant ***DeepWalk+LR*** pipeline may still be better-suited for a production environment. For the purpose of guiding scientific experimentation, however, these computations may be infrequent, making the less scalable but higher-precision ***Geodesic+GBC*** pipeline more attractive.

Acknowledgements

Support for DOI 10.13139/OLCF/1782714 dataset is provided by the U.S. Department of Energy, Project smcdc21_challenge2 under Contract DE-AC05-00OR22725. Project smcdc21_challenge2 used resources of the Oak Ridge Leadership Computing Facility at Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

Citations

[1] D. Herrmannova, R. Kannan, S.-H. Lim, and T. E. Potok. Covid-19 Knowledge Graph – Dataset for SMCDC 2021 Challenge 2. Technical report, Oak Ridge National Lab (ORNL), Oak Ridge, TN (United States), 2021.

[2] B. Perozzi, R. Al-Rfou, and S. Skiena. DeepWalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 701–710, 2014.

[3] J. Kunegis. *Handbook of Network Analysis*. 2021.