BITS :: Call for Abstracts 2022 - Oral communication

Type	Oral communication
Session	Algorithms for Bioinformatics
Title	Right Normalized Laplacian improves the performance of Node2Vec embedding for edge prediction in STRING PPI Graphs
All Authors	Cappelletti L(1,2), Taverni S(1,2), Fontana T(1,2), Joachimiak M(3), Reese J(3), Casiraghi E(1,2), Robinson P(4), Valentini G(1,2)

Affiliation

(1) AnacletoLab, Dipartimento di Informatica, Università degli Studi di Milano, Italy
(2) Laboratorio Nazionale InfoLife, CINI
(3) Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
(4) The Jackson Laboratory for Genomic Medicine, 10 Discovery Drive, Farmington 06032, CT, USA

Motivation

Several relevant problems in computational biology and network medicine can be modelled as edge prediction tasks in graphs, such as the in-silico prediction of protein interactions (using e.g., STRING (DOI: 10.1093/nar/gkaa1074) graphs) or the prediction of drug-disease associations (using e.g., CTD (DOI: 10.1093/nar/gkaa891) graphs). Popular edge-prediction approaches firstly compute node embeddings through Node2Vec (DOI: 10.1145/2939672.2939754) or Struc2vec (DOI: 10.1145/3097983.3098061) and then use the node embeddings to train a classifier model to predict the existence of edges.

Such methods either sample the graph edges directly or explore the graph through a weighted random walk. Unfortunately, these approaches are biased towards high degree nodes, which are more frequently sampled, while node embeddings of nodes with low degrees, underrepresented in the samples, lead to poor euclidean representation of the nodes, as shown in the analysis of CTD data in the attached figure.

As a result, the areas of the graph with low degree nodes, which often represent the most interesting part of the biological network under study, are poorly represented by the node embeddings, thus leading to poor prediction results.

Methods

In a Node2Vec second-order random walk, the node sampling is biased through the in-out and the return parameters that allow us to interpolate the walk between a breadth-first and depth-first search.

We propose a Right Normalized Laplacian (RNL) approach to further bias a Node2Vec random walk towards low degree nodes, i.e. we divide each edge weight by the inbound degree of the destination node. This approach makes the sampling of high and low degree nodes equiprobable, thus avoiding the bias toward high degree nodes.

In a Node2Vec second-order random walk, the node sampling is biased through the in-out and the return parameters that allow us to interpolate the walk between a breadth-first and depth-first search.

We propose a Right Normalized Laplacian (RNL) approach to further bias a Node2Vec random walk towards low degree nodes, i.e. we divide each edge weight by the inbound degree of the destination node. This approach makes the sampling of high and low degree nodes equiprobable, thus avoiding the bias toward high degree nodes.

Results

We evaluated the proposed approach by performing edge prediction tasks on eight weighted STRING PPI graphs: Homo Sapiens, Drosophila Melanogaster, Saccharomyces Cerevisiae, Mus Musculus, Sus Scrofa, Amanita Muscaria, Alligator Sinensis and Canis Lupus. Before executing the node embeddings, we applied an edge weight cut-off at 700 and dropped the remaining singleton nodes. To perform the experiments, we trained a simple Perceptron GNN using the CBOW Node2Vec embeddings obtained by using either the traditional or the proposed RNL approach on an edge prediction task.

We executed ten connected holdouts for each task with an 80/20% split of the positive edges.

We evaluated the model performance using three different imbalance ratios between existing and non-existing edges (1:1, 1:10, 1:100). The CBOW models were trained for 50 epochs, using negative sampling (10 negative nodes) and a 10-steps context window.

The figure shows that RNL can better represent low degree nodes, separating the different CTD node types even when their degree is low.

We built the entire experimental pipeline using the GraPE library (DOI: 10.48550/arXiv.2110.06196).

The average test F1 Score across the eight edge prediction tasks in STRING improved from 0.72 ± 0.068 to 0.79 ± 0.056, using RNL with respect to the traditional approach (see figure).

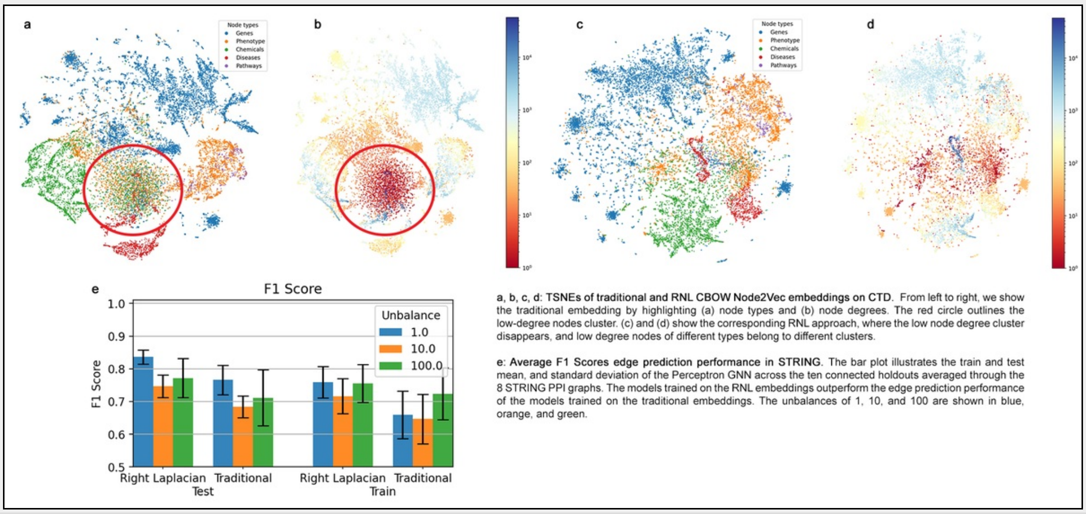
Even though the RNL approach severely changes the graph topology sampling, seemingly destroying the impact of a node's degree in the random walk, it does not prevent the model from learning the node topology. On the contrary, it leads to significantly better embedding and edge prediction results for low degree nodes.

Info

5 authors [Luca Cappelletti, Stefano Taverni, Tommaso Fontana, Elena Casiraghi, Giorgio Valentini] are members of the CINI Infolife laboratory

filename	right_laplacian_96dpi.png

Figure



a, b, c, d: TSNEs of traditional and RNL CBOW Node2Vec embeddings on CTD. From left to right, we show the traditional embedding by highlighting (a) node types and (b) node degrees. The red circle outlines the low-degree nodes cluster. (c) and (d) show the corresponding RNL approach, where the low node degree cluster disappears, and low degree nodes of different types belong to different clusters.

e: Average F1 Scores edge prediction performance in STRING. The bar plot illustrates the train and test mean, and standard deviation of the Perceptron GNN across the ten connected holdouts averaged through the 8 STRING PPI graphs. The models trained on the RNL embeddings outperform the edge prediction performance of the models trained on the traditional embeddings. The unbalances of 1, 10, and 100 are shown in blue, orange, and green.

Availability	https://github.com/LucaCappelletti94/right_laplacian_node2vec

Corresponding Author

Name, Surname	Giorgio, Valentini
Email	giorgio.valentini@unimi.it
Submitted on	29.04.2022