Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

Link Prediction for Automated Scientific Hypothesis Generation

Pooja Ganesh
Advisor - Brian Nord
Co-advisor - Ashia Livaudais, Aleksandra Ciprijanovic

FCSI Presentation - Deepskies Laboratory

11 August 2023

DEEP SKIES

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

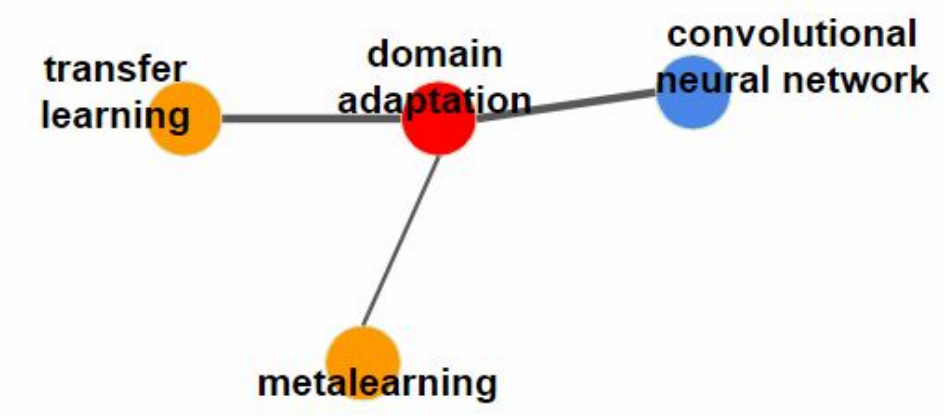
The HypoGen Project

Motivation

Can we accelerate scientific discovery using automated hypothesis generation, a machine learning method to make predictions about possible links across scientific literature?

Solution

Exploring link prediction for scientific concepts.



Graph Data Structure

Pipeline - From Data Preprocessing to Model Assessment

1

Graph Data

- Edge removal for training data
- Creation of edge features and target variable for unconnected node pairs
- Prevention data leakage by careful splitting
- Datasets - Cora, Science4cast

2

Node2vec

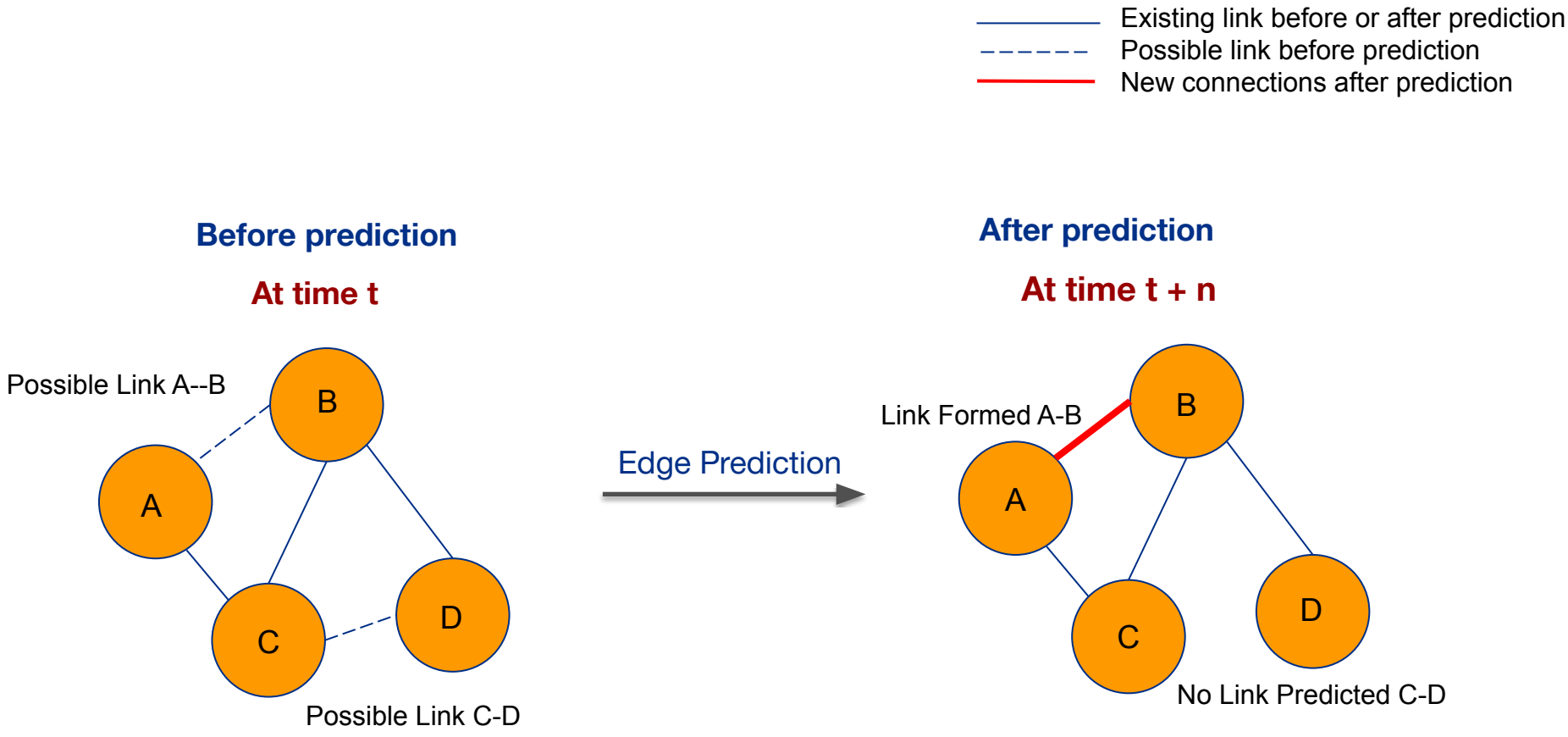
- Vector Representation for nodes
- Representation of network topology
- Input for neural network model additional to features

3

Graph Convolutional Network Model

- Link Prediction for unconnected node pair
- Performance Metrics AUC

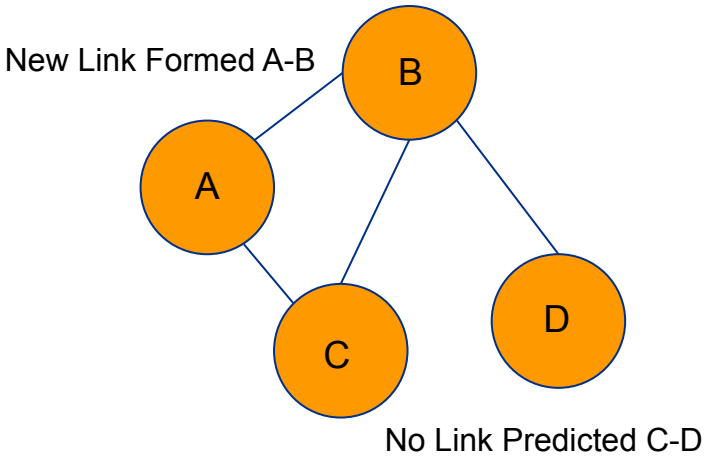
Graph Data and Link Prediction (Graphical Structure)



- We need to use graphs at two different instances of time to extract the target variable.
- But in real life, we only have one initial graph.

Graph Data and Link Prediction (Feature Set View)

Graph Structure



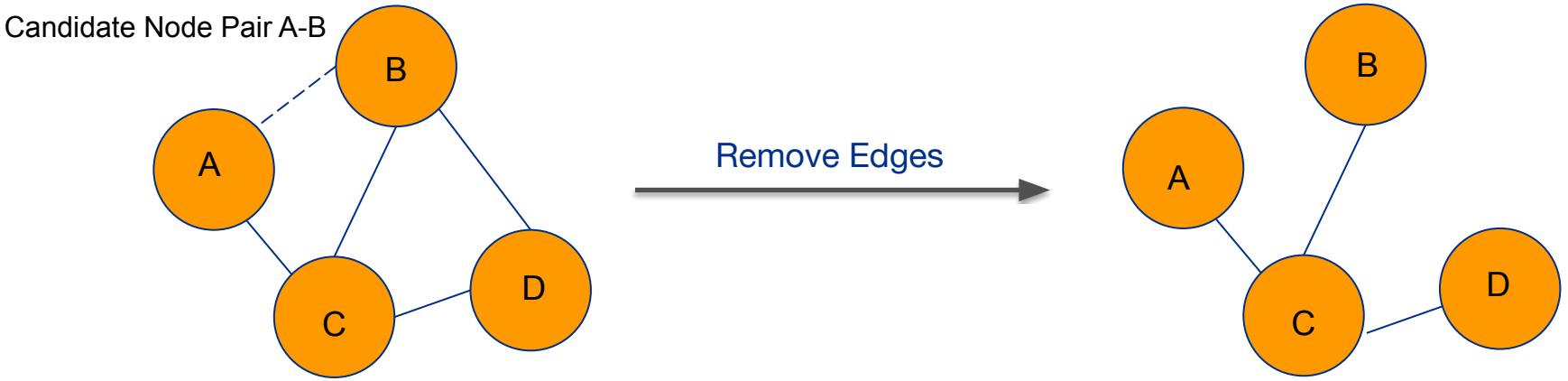
time: t + n

Feature Set

Features	Link (Target Variable)
A-B Node Pair	1
C-D Node Pair	0

- We need predictor and target variables to predict link between 2 unconnected nodes

Creating Training Data using Graph in present time t with one graph

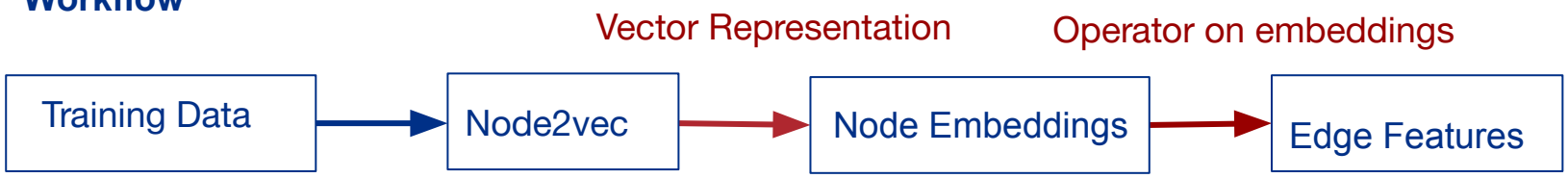


- Objective is to predict link existence between candidate node pair : A-B
- Feature creation for all the unconnected node pairs including the ones for which we have removed the edges.
- Usually imbalanced dataset, needs to be balanced to get proper evaluation

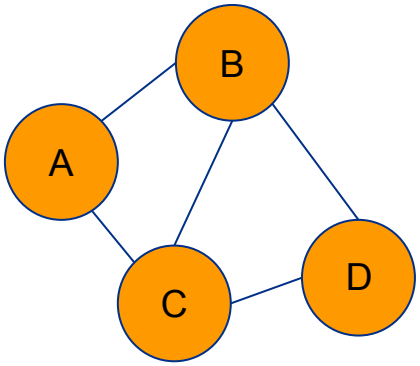
Features	Links (Target)
B-D	1 (removed)
A-D	0 (never existed)

Node2vec - Feature Extraction (Node Embeddings)

Workflow



About node2vec

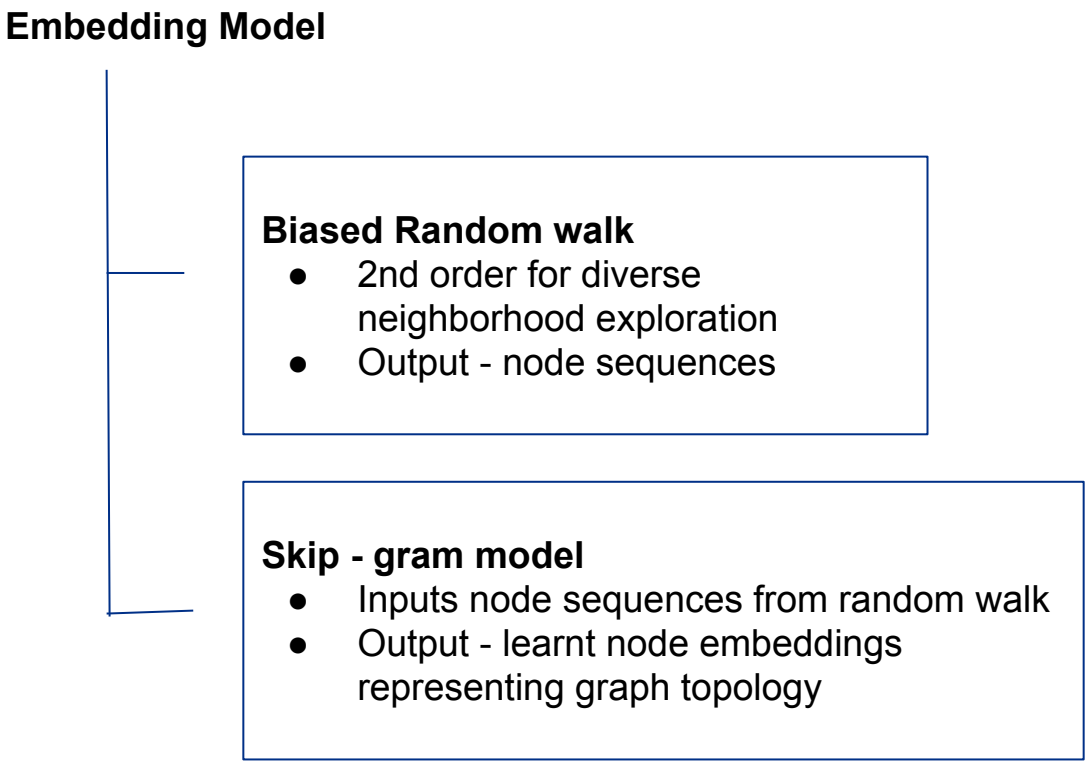


- Mapping each node in graph with a vector sequence interpretable my a neural network model
- Numerical representation of the graph
- Dimensions size for each node = n (default = 128)

Nodes	Embeddings/Features
A	[0.1, 2.85, 3.25,....n]
B	[0.03, 4.95, 0.5,....n]
C	[1.02, 8.85, 4.05,...n]
D	[0.12, 0.25, 6.9,.....n]

Node2vec - Feature Extraction (A peek inside the module)

- Two major steps that happen - biased random walk, followed by training a skip-gram model.



Science4Cast Benchmark Dataset

Motivation for benchmark - vast dataset with concepts as node to facilitate hypothesis generation
Problem Statement - originally used to predict new links between concepts n years into the future

Inside the dataset

construction

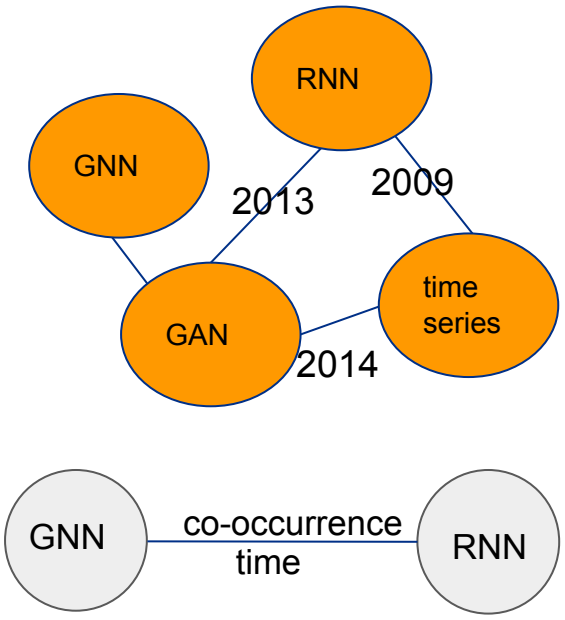
- Concept keywords extracted from **titles** and **abstracts** of scientific papers
- **Concept examples** - graph NN, recurrent NN
- Papers between 1990 and 2014
- Unweighted edges

data

- 64,000 concept nodes
- 18,000,000 edges
- Multiple edges can exist between two nodes

- That is a huge dataset!
- Computationally expensive due to input to the model is a matrix of size N X M, where N - nodes and M is embedding size! That is 64000 X 128 !!
- Node2vec is also expensive

our network/graph



Science4Cast Dataset Link Prediction -Benchmark

Training Data

- Split across years before 2014
- Example - graph for 2011, 2012, 2013, 2014

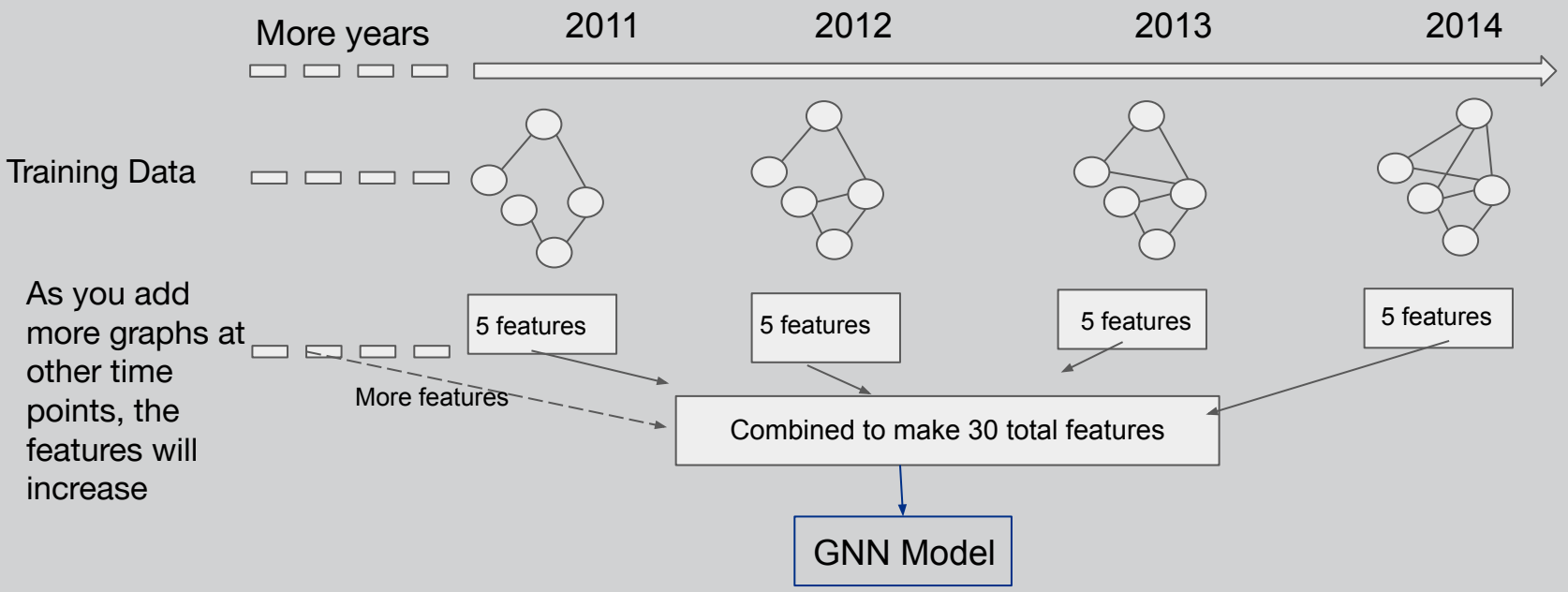
Feature Creation

- Handcrafted features for node pairs
- Example - degree of the node pairs in various years

GNN

- Simple 2 layer GNN
- **AUC train 0.74**
- **AUC test 0.82**

Workflow with visualization



Cora and Comparisons

- Less complicated structure of graph dataset
- Much smaller than Science4cast
- Paper citation network
- Undirected graph
- **Nodes - 2485 and Edges - 4689**

Dataset	Method	Features	AUC train	AUC test
Science4cast	2 layer GNN	Hand crafted (30)	0.74	0.82
Cora	Logistic Regression	Node Embeddings	0.86	0.88

Observations

- Recorded AUC for both test and train are higher for Cora dataset
- Reason - small dataset, use of node embeddings

Scope for Improvement for science4cast dataset

- Adding more layers in GNN
- Increasing the number of handcrafted features
- Adding node embeddings as model inputs to capture the graph topology better

Future work in HypoGen

- Exploring other link prediction algorithms
- Employing feature extraction methods other than node2vec to build edge features
- Exploring new ways to build better edge features
- Exploring larger, more complex datasets with predefined edge features and weights to strengthen link prediction algorithm.