**Representation Learning on Graphs using Node2vec**

Pankaj RAjoria

[p.rajoria@tu-braunschweig.de](mailto:p.rajoria@tu-braunschweig.de)

MASTER OF DATA SCIENCE, TECHNISCHE UNIVERSITÄT BRAUNSCHWEIG

Contents

[Introduction 2](#_Toc140756065)

[Graph Learning 2](#_Toc140756066)

[Representation Learning 3](#_Toc140756067)

[Node Embedding 3](#_Toc140756068)

[Pairwise similarity function 4](#_Toc140756069)

[Encoder 4](#_Toc140756070)

[Decoder 4](#_Toc140756071)

[Loss function 4](#_Toc140756072)

[Graph Embedding 4](#_Toc140756073)

[Random Walks 4](#_Toc140756074)

[Node2Vec 5](#_Toc140756075)

[Biased Random Walks 6](#_Toc140756076)

[Search strategies 6](#_Toc140756077)

[Breadth-first Sampling (BFS) 6](#_Toc140756078)

[Depth-first Sampling (DFS) 6](#_Toc140756079)

[Search bias α 6](#_Toc140756080)

[Return parameter, p 6](#_Toc140756081)

[In-out parameter, q 6](#_Toc140756082)

[The node2vec algorithm 6](#_Toc140756083)

[Implementation 6](#_Toc140756084)

[Results 6](#_Toc140756085)

[References 7](#_Toc140756086)

# Introduction

Graphs are a versatile and powerful data structure used to represent relationships between entities. They consist of nodes, which represent individual units, and edges, which represent connections between nodes. A graph can be denoted as a pair **(V, E)**, where V represents the set of nodes in the graph, and E represents the set of edges indicating the connections between the nodes. Typically, each edge in E is defined as an unordered pair of nodes **(u, v)**, signifying a relationship between nodes u and v.

Graphs serve as a fundamental model for capturing and analysing various real-world scenarios that involve entities and their interactions. In the field of machine learning, graphs are particularly useful for representing structured data. Techniques like representation learning are employed to extract and encode the structural information into low-dimensional vectors, which then serve as features for various machine learning tasks. This enables the discovery of patterns, predictions, classifications, and recommendations based on the relationships between nodes.

One of the most commonly used representations of graphs is the adjacency matrix. An adjacency matrix **Aij** is a square matrix that provides a compact way to represent the connections between nodes in a graph. The rows and columns of the matrix correspond to the nodes in the graph, and the entries of the matrix indicate whether there is an edge between two nodes.

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | 0 | **1** | **1** |
| 2 | **1** | 0 | 0 |
| 3 | **1** | 0 | 0 |

**3**

**1**

**2**

## Graph Learning

Graph learning is a branch of machine learning that focuses on understanding and analysing structured data represented as graphs. In this context, the entities in the data are represented as nodes, and the relationships between them are represented as edges connecting the nodes. The goal of graph learning algorithms is to extract valuable insights and patterns from these interconnected data points. Graph learning techniques have diverse applications across various domains. For instance, in social network analysis, graph learning can be employed to classify users based on their behaviour or predict user preferences and interests. By leveraging the graph structure of social connections, graph learning algorithms can uncover hidden patterns and make accurate predictions.

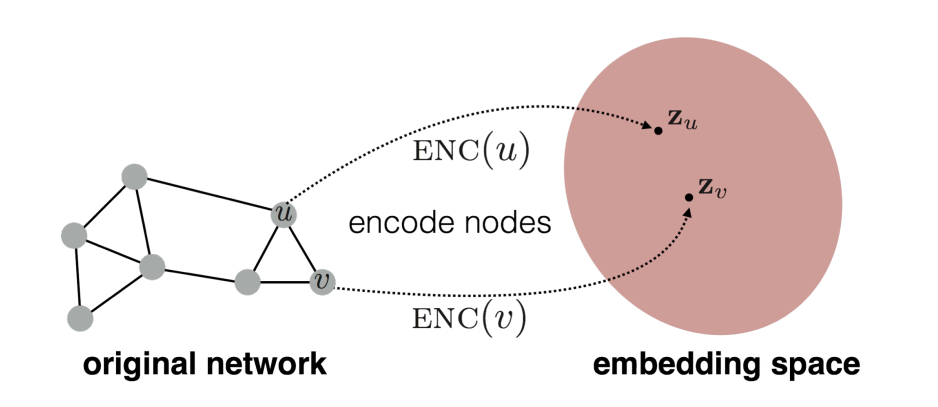
In recommendation systems, graph learning plays a crucial role in link prediction. It enables the suggestion of new connections between users and items based on their past interactions, facilitating personalized recommendations. By learning from the graph structure, graph learning algorithms can effectively capture the collaborative filtering effect and enhance recommendation accuracy. In the field of bioinformatics, graph learning is used to analyse biological networks and predict protein functions. By representing proteins as nodes and interactions between proteins as edges, graph learning algorithms can classify proteins based on their functional properties and infer functions for unannotated proteins.

Graph learning also finds applications in knowledge graphs, where entities and their relationships are represented as a large-scale graph. By learning meaningful representations from the knowledge graph, graph learning algorithms can enable semantic search, question answering, and knowledge graph completion. These applications contribute to enhancing the understanding and utilization of vast amounts of interconnected knowledge

## Representation Learning

Representation learning on graphs, also known as graph representation learning, refers to the process of extracting meaningful and low-dimensional representations (embeddings) from graph-structured data. The goal is to encode the structural information of the graph into numerical vectors that can be used as features for various machine learning tasks.

## Node Embedding

Node embedding aims to map each node in the graph to a low-dimensional vector. The objective is to ensure that nodes with similar structural roles or properties in the graph are represented by similar vectors. There are various algorithms and approaches for node embedding, such as DeepWalk, node2vec, GraphSAGE, and Graph Convolutional Networks (GCNs). These methods typically use random walks, neighbourhood aggregation, or graph convolution operations to capture the structural information and generate node embeddings.

Node embedding methods typically involve the following four methodological components:

1. Pairwise similarity function
2. Encoder
3. Decoder
4. Loss function

### Pairwise similarity function

### Encoder

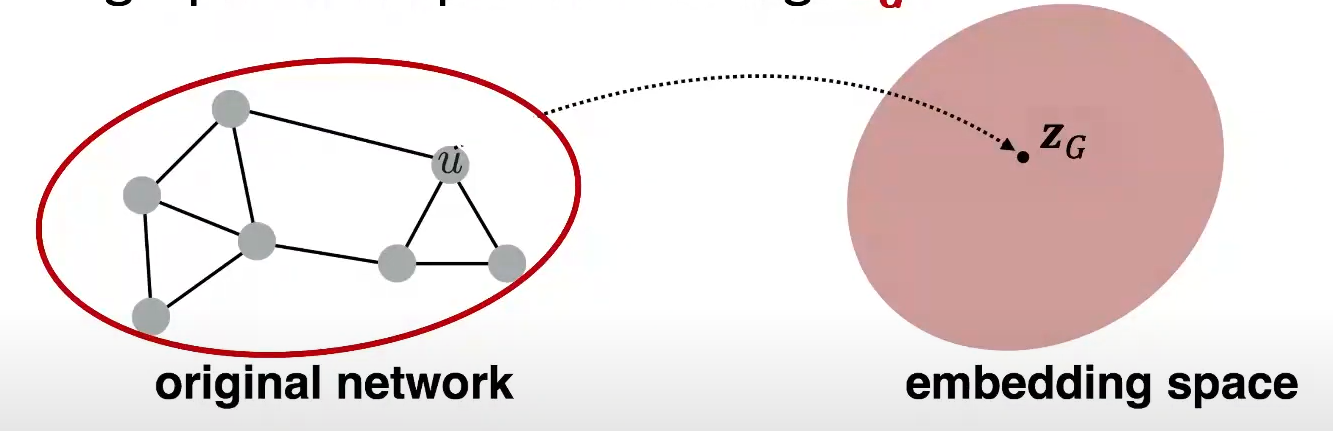
### Decoder

### Loss function

These four methodological components work together to learn node embeddings from graph data. The pairwise similarity function captures the notion of similarity between nodes, the encoder generates low-dimensional representations, the decoder reconstructs pairwise similarities, and the loss function guides the training process by evaluating the quality of the reconstructions.

## Graph Embedding

In addition to node embedding, some applications require capturing the overall graph-level information. Graph-level embedding aims to summarize the entire graph or subgraphs into a fixed-length **vector** representation. This can be useful for tasks such as graph classification or clustering. Techniques like Graph Isomorphism Networks (GIN) and Graph Attention Networks (GAT) can be used to aggregate node embeddings and generate graph-level embeddings.



# Random Walks

Random walks are used to explore and capture information from graph or network structures. Random walks involve taking sequences of steps on a graph, with each step determined by randomly selecting a neighbouring node. The goal is to generate sequences that capture the local neighbourhood structure of the graph. Starting with a graph representation, a random walk begins at a randomly chosen node. At each step, the walker moves to a neighbouring node based on a predefined probability distribution. By performing multiple steps, a sequence of visited nodes is generated, reflecting the graph's connectivity patterns. These sequences are then used to learn representations of the nodes or the entire graph, employing techniques such as Skip-gram, DeepWalk, or node2vec. The learned representations enable the capture of important structural information and similarities between nodes in a low-dimensional vector space, facilitating downstream tasks like node classification, link prediction, or community detection. Random walks in representation learning provide a means to explore and encode the local structure of graphs, enabling meaningful analysis and machine learning on graph data.

D

E

I

Random walk for a fixed length of 3, and from starting node H, is H🡪G🡪C🡪A

Node2Vec  
Node2vec is a semi-supervised algorithm designed for scalable feature learning in networks. It aims to learn feature representations for nodes in a network such that the neighbourhoods of nodes are preserved in a d-dimensional feature space. The algorithm accomplishes this by employing a custom graph-based objective function optimized using stochastic gradient descent (SGD) and drawing inspiration from prior work in natural language processing. At the core of node2vec is a 2nd order random walk approach, which is used to generate network neighbourhoods (i.e., samples) for nodes. The algorithm's key innovation lies in its flexible definition of a node's network neighbourhoods. By adopting a family of biased random walks, it efficiently explores diverse neighbourhoods for each node, enabling the learning of representations that organize nodes based on their network roles and community affiliations.

The flexibility of node2vec allows control over the search space through tuneable parameters, unlike previous rigid search procedures. This adaptability enables the algorithm to generalize well and model various equivalences observed in networks. Additionally, the parameters governing the search strategy have intuitive interpretations and can be learned in a semi-supervised fashion using a small fraction of labelled data. It extends the learned feature representations of individual nodes to pairs of nodes (edges) through simple binary operators. This characteristic makes it suitable for prediction tasks involving both nodes and edges in the network.

## Biased Random Walks

## Search strategies

### Breadth-first Sampling (BFS)

### Depth-first Sampling (DFS)

## Search bias α

### Return parameter, p

### In-out parameter, q

## The node2vec algorithm

# Implementation

# Results

# References

1. Hamilton, W. L. (n.d.). Graph Representation Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 14(3), 1-159. Morgan and Claypool.