



"Ss. Cyril and Methodius" University in Skopje
**FACULTY OF COMPUTER
SCIENCE AND ENGINEERING**

Network Analysis and Node Classification

Mentor:
Sonja Gievska

Student:
Jovana Markovska 181009

March, 2022

Table of contents

Introduction	3
Dataset	3
Network Analysis	4
Cora graph (Real graph)	4
Erdős-Renyi Random Graph	5
Watts Strogatz Random Graph (Small World)	6
Evaluation	7
Node Classification using Graph Embeddings	8
Evaluation	8
Node Classification using GraphSAGE	9
Evaluation	9

Introduction

This document contains an evaluation of the methods that will be used in order to perform the following tasks on the Cora dataset:

1. Network Analysis
2. Node Classification using Graph Embeddings
3. Node Classification using GraphSAGE

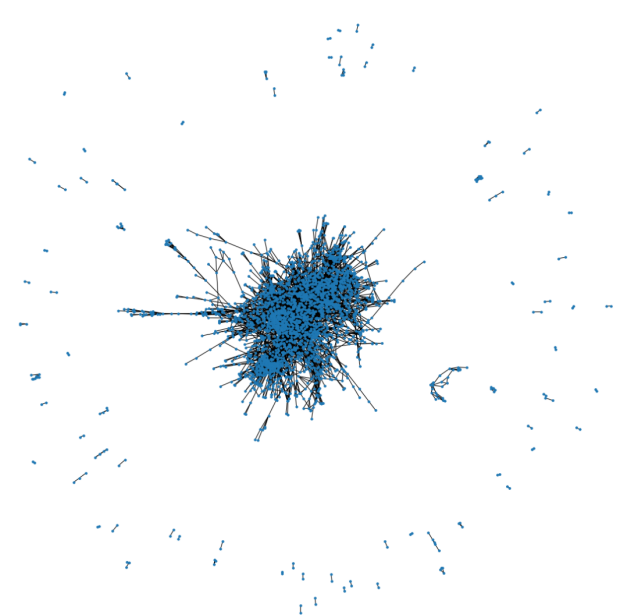
Dataset

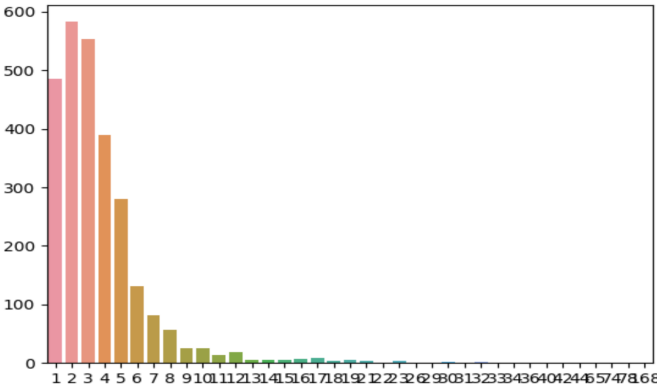
The Cora dataset consists of 2708 scientific publications classified into one of seven classes. The citation network consists of 5429 links. Each publication in the dataset is described by a 0/1-valued word vector indicating the absence/presence of the corresponding word from the dictionary. The dictionary consists of 1433 unique words.

Network Analysis

Cora graph (Real graph)

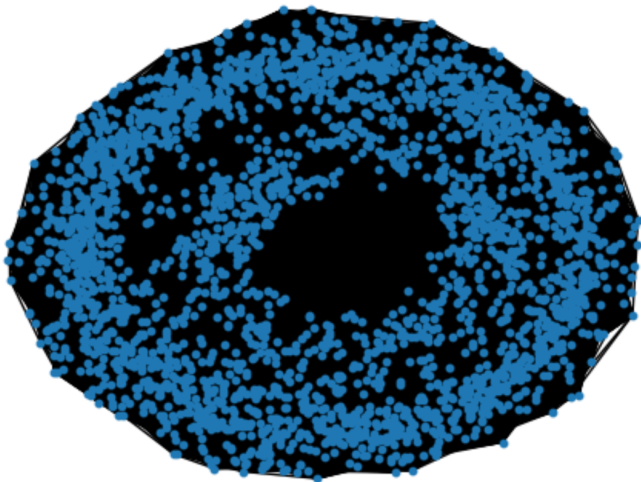
Visualization of the Cora graph

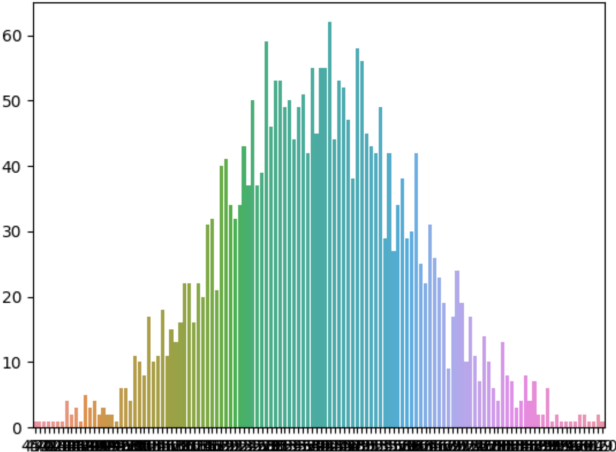


	Cora graph
Degree distribution	
Connected components	78
Diameter	19
Average clustering coefficient	0.2406732985019372

Erdős-Renyi Random Graph

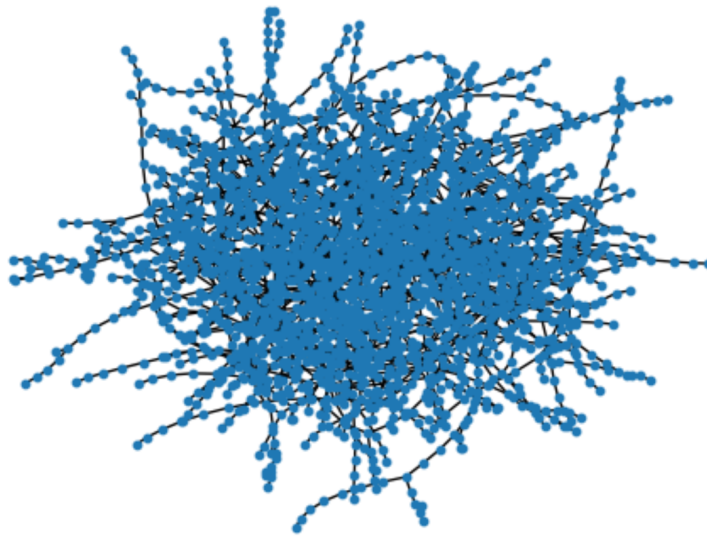
Visualization of the Erdős-Renyi Random Graph

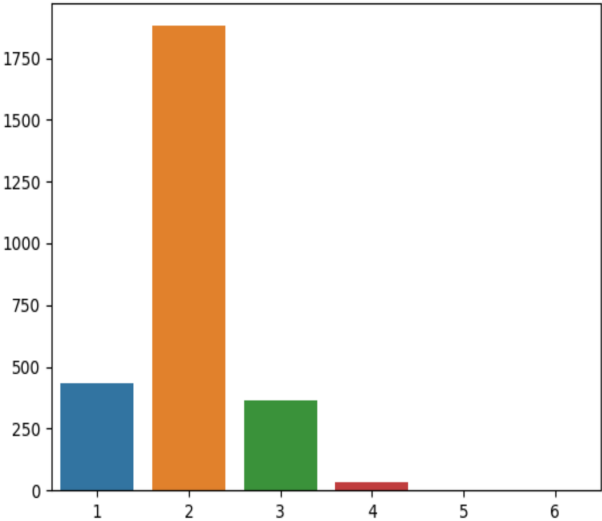


	Erdős-Renyi graph
Degree distribution	
Connected components	1
Diameter	2
Average clustering coefficient	0.19950958783982486

Watts Strogatz Random Graph (Small World)

Visualization of Watts Strogatz Random Graph



	Watts Strogatz graph														
Degree distribution	 <table><thead><tr><th>Degree</th><th>Frequency</th></tr></thead><tbody><tr><td>1</td><td>450</td></tr><tr><td>2</td><td>1800</td></tr><tr><td>3</td><td>350</td></tr><tr><td>4</td><td>50</td></tr><tr><td>5</td><td>10</td></tr><tr><td>6</td><td>5</td></tr></tbody></table>	Degree	Frequency	1	450	2	1800	3	350	4	50	5	10	6	5
Degree	Frequency														
1	450														
2	1800														
3	350														
4	50														
5	10														
6	5														
Connected components	5														
Diameter	227														
Average clustering coefficient	0.0														

Evaluation

Degree distribution	<p>The degree distribution plots of above mentioned graphs show us that we can find similarity in the distributions of the Real and Small World Graph. The most common node degree in both graphs is the degree of 2.</p> <p>On the other hand, Erdős-Renyi Graph degree distribution differs as a result of the number of connected components which is only one.</p>
Connected components	<p>Like an assortative network, the Real Graph high degree nodes tend to stick together. The Small World and the Erdős-Renyi Graph differ from the Real graph in terms of connected components, which can be seen on the visualised graphs. The Erdős-Renyi Graph has only 1 connected component which makes it far more different than the Real graph.</p>
Diameter	<p>In terms of the diameter, the Real graph has a diameter of 19, Erdős-Renyi Graph has a diameter of 2, Small World has a diameter of 227. All three differ from one another.</p>
Average clustering coefficient	<p>The average clustering coefficient of the Small World graph is 0.0 which means that this graph has the lowest degree of transitivity. This can be perceived when looking at the string-like structure of the graph itself. On the contrary, regarding the average clustering coefficient, it can be concluded that the Real graph and the Erdős-Renyi Graph are somewhat more similar. We can see dense core and more sparse peripheries which shows us an assortative networks with disassortative hubs.</p>

Node Classification using Graph Embeddings

This model is using 50-dimensional vectors, with 0.000001 values for both L1 and L2 regularization. It contains two hidden layers with the size of 100 and 50. The subjects are encoded using Ordinal Encoder and the classifier is a Random Forest Classifier.

```
sdne = SDNE(d=50, beta=5, alpha=1, nu1=0.000001, nu2=0.000001, K=2,  
            n_units=[100, 50], n_iter=50, xeta=0.01, n_batch=500)
```

Evaluation

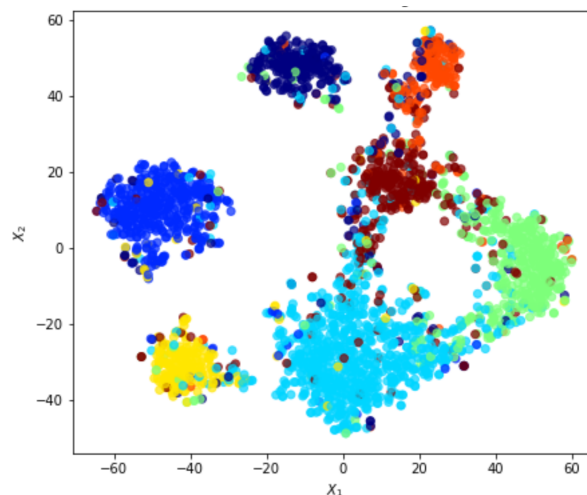
SDNE test set metrics	
Accuracy score:	0.25830258302583026
F1-micro:	0.25830258302583026
F1-macro:	0.13218615813078952

This model does not take into account all of the node features resulting in a not-so-high accuracy. If we were to use all of the node features, the accuracy would increase respectively.

Node Classification using GraphSAGE

The best GraphSAGE model with the highest accuracy of 0.8171 is a combination of 2 layers with 32-dimensional hidden node features at each layer and one fully connected layer with the same unit size as train_targets, followed by a softmax activation function.

In comparison to the previous method used for node classification - SDNE, the accuracy of GraphSAGE is significantly higher, which makes it better.



TSNE visualization of GraphSAGE embeddings for Cora dataset

Evaluation

	Predicted	True
31336	Probabilistic_Methods	Neural_Networks
1061127	Rule_Learning	Rule_Learning
1106406	Reinforcement_Learning	Reinforcement_Learning
13195	Reinforcement_Learning	Reinforcement_Learning
37879	Probabilistic_Methods	Probabilistic_Methods
1126012	Reinforcement_Learning	Probabilistic_Methods
1107140	Reinforcement_Learning	Theory
1102850	Probabilistic_Methods	Neural_Networks
31349	Probabilistic_Methods	Neural_Networks
1106418	Theory	Theory

Ratio between predicted categories and true categories

GraphSAGE test set metrics	
Accuracy score:	0.8153618906942393
F1-micro:	0.8153618906942393
F1-macro:	0.8009958616552086

