Learning with Graphs: Networks Embedding Algorithms

Luis Valenzuela
Dennis Mejicanos
Julio Guzmán
Supervisor: Vincent Gauthier

TELECOM SudParis

INSTITUT POLYTECHNIQUE DE PARIS

Outline

1. Motivation
2. Problem Statement
3. Deep Walk
4. Node2Vec
5. GCN
6. Experimental Results
7. Conclusions
8. References

Motivation

In order to apply machine learning models for graphs, we need to transform the vertices of a graph into a vector form. **By doing that we can take advantage of all ML algorithms (such as deep learning approaches) already developed to perform learning on graphs.**

We are going to compare three graph embedding techniques: DeepWalk, Node2Vec and Graph Convolutional Networks.

Problem Statement

Given three graph datasets: Facebook Large Page-Page Network, Github Social Network, and Cora, we will test and compare different graph embedding techniques: DeepWalk, Node2Vec on the Facebook y GitHub datasets, and Node2Vec and and Graph Convolutional Networks (GCN) on the CORA dataset.

We perform semi-supervised learning in order to guess features of nodes in a graph.

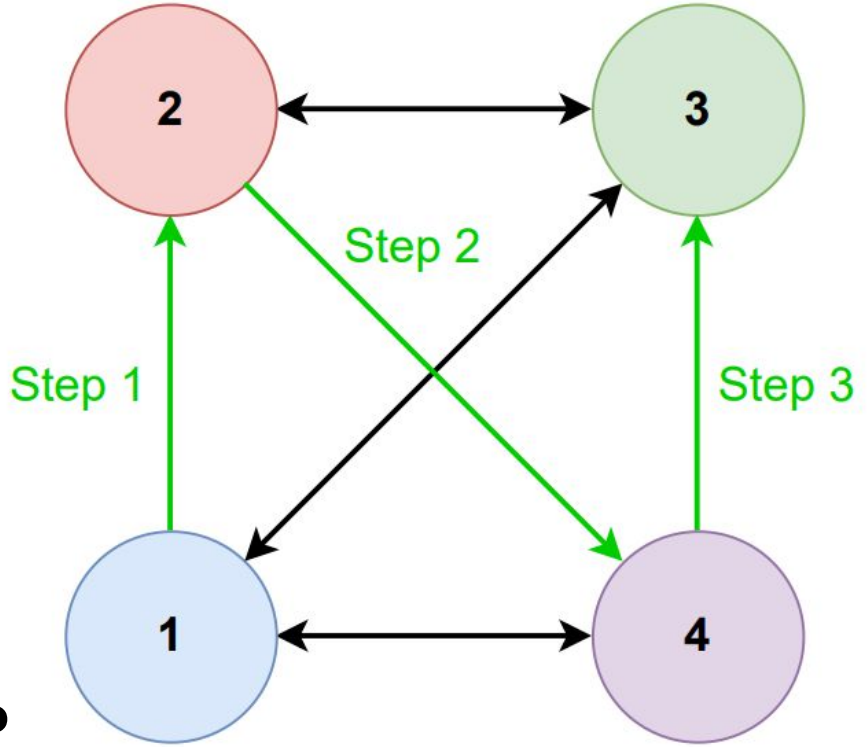
1.
Deep Walk

DeepWalk

- Run through the graph randomly
- Walks per node
- Walk Length

Walk Set: [''1'', ''2'', ''3'', ''4'']

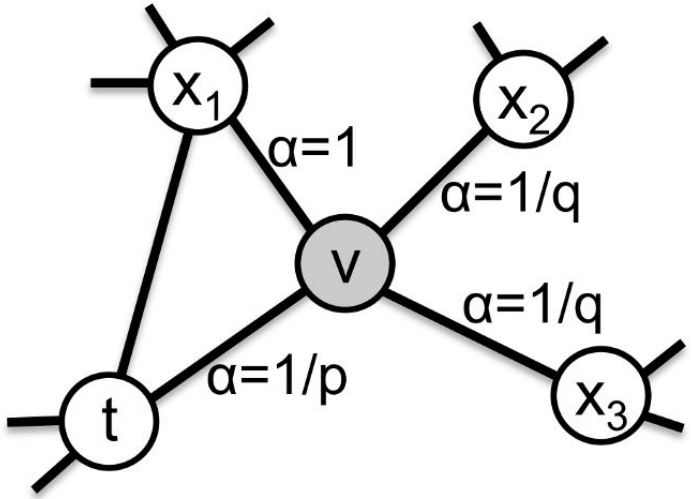
Deep Learning classification model to predict the context: Skip-gram Model



2.
Node2Vec

Node2Vec

- Biased Random Walker
  - P: Return Hyperparameter
  - Q: Inout Hyperparameter

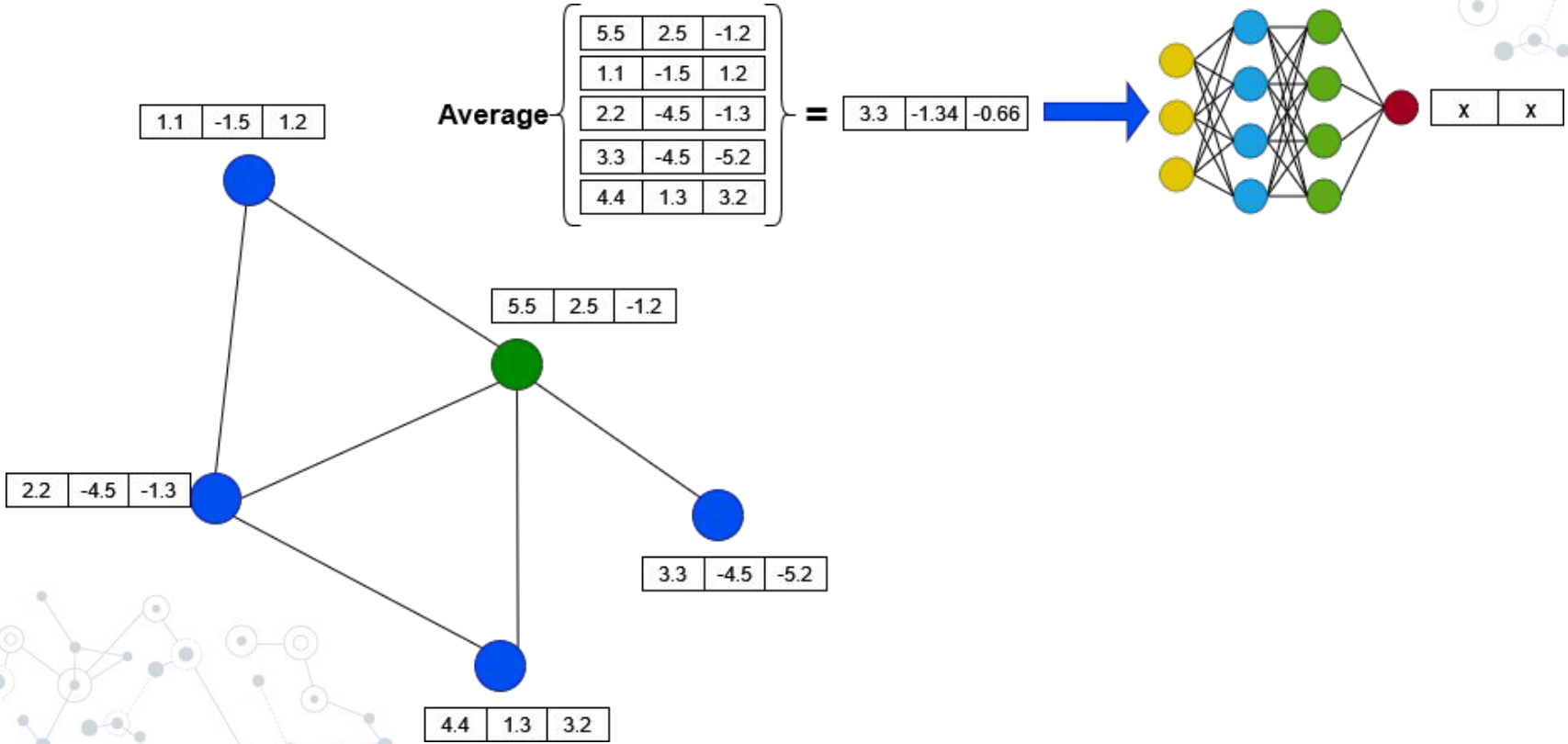


Grover, A. (2016). Node2Vec: Scalable Feature Learning for Networks. [Figure]. Recovered from: https://arxiv.org/pdf/1607.00653.pdf

3.

Graph Convolutional Networks (GCN)

GCN


1.1 -1.5 1.2
Average
5.5 2.5 -1.2
1.1 -1.5 1.2
2.2 -4.5 -1.3
3.3 -4.5 -5.2
4.4 1.3 3.2
=
3.3 -1.34 -0.66
X X
5.5 2.5 -1.2
2.2 -4.5 -1.3
3.3 -4.5 -5.2
4.4 1.3 3.2

4. Experimental Results

4.1

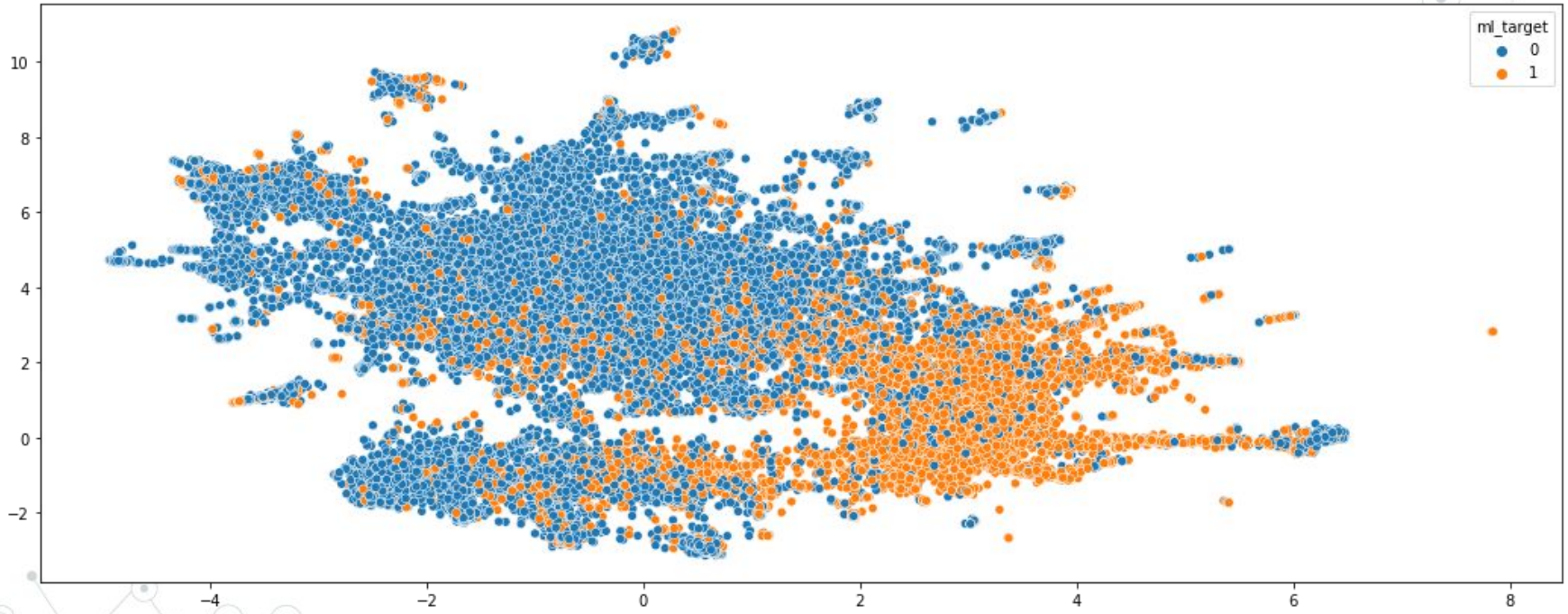
Github Dataset

Github Dataset

- Nodes are developers who have at least 10 repositories on github.
- Edges are mutual follower relationships between the developers.
- Features are extracted based on the location, employer and e-mail address.
- Nodes: 37,700: 27,961 Web Dev, 9,739 ML Dev.
- Edges: 289,003.
- Predict if a GitHub user is web or machine learning developer.

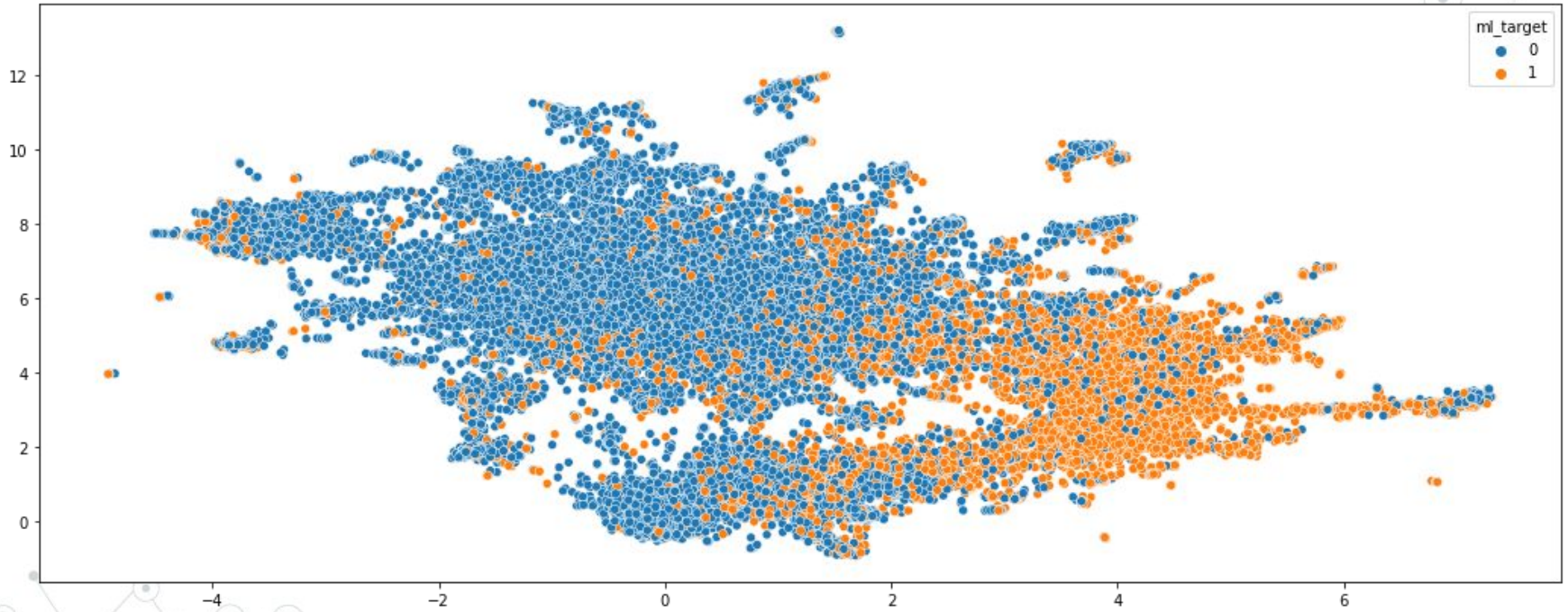
Source: https://github.com/benedekrozemberczki/datasets

Github Dataset - DeepWalk Embedding



Walks per node = 10, walk length = 80

Github Dataset - Node2Vec Embedding



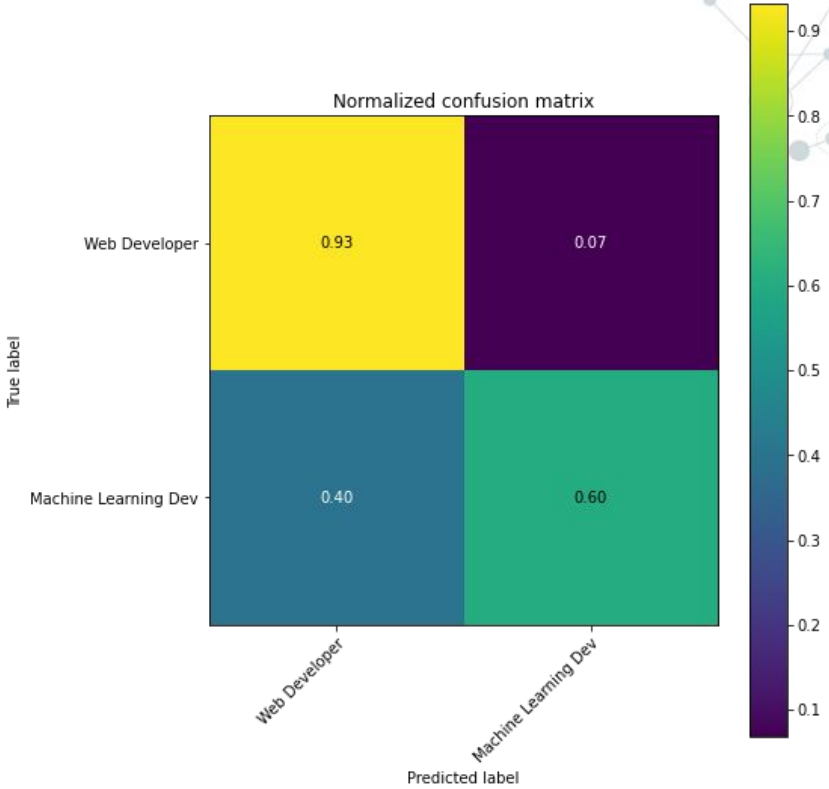
Walks per node = 10, walk length = 80, p = 0.25, q = 0.25

Github Dataset Results

DeepWalk

F1 Score: 0.8407

Walks per node = 10, walk Length = 80



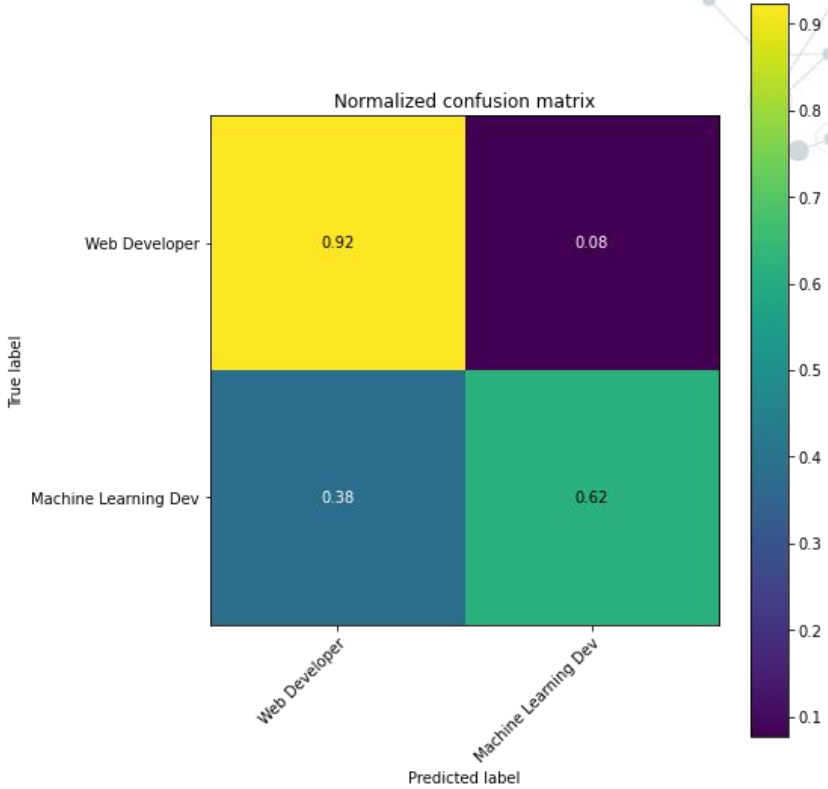
Github Dataset Results

Node2Vec

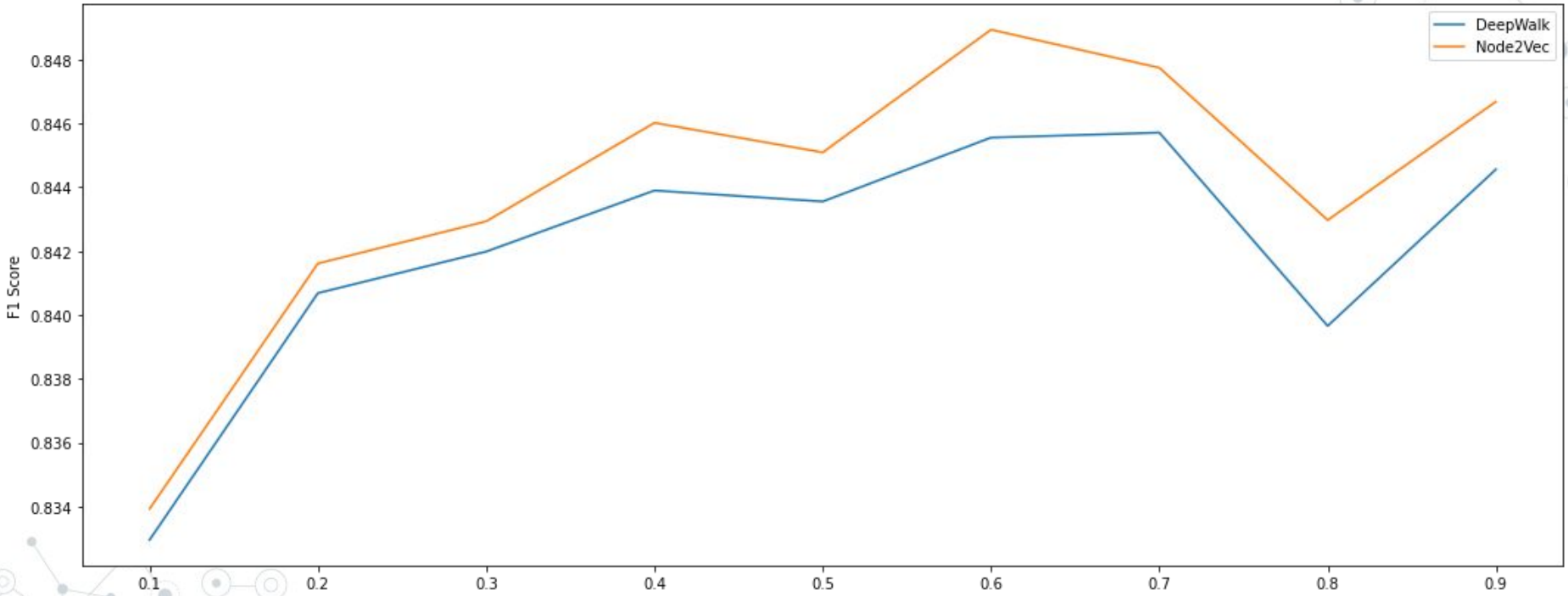
F1 Score: 0.8448

Walks per node = 10, walk Length = 80

p = 0.25, q = 0.25



Github Dataset Results



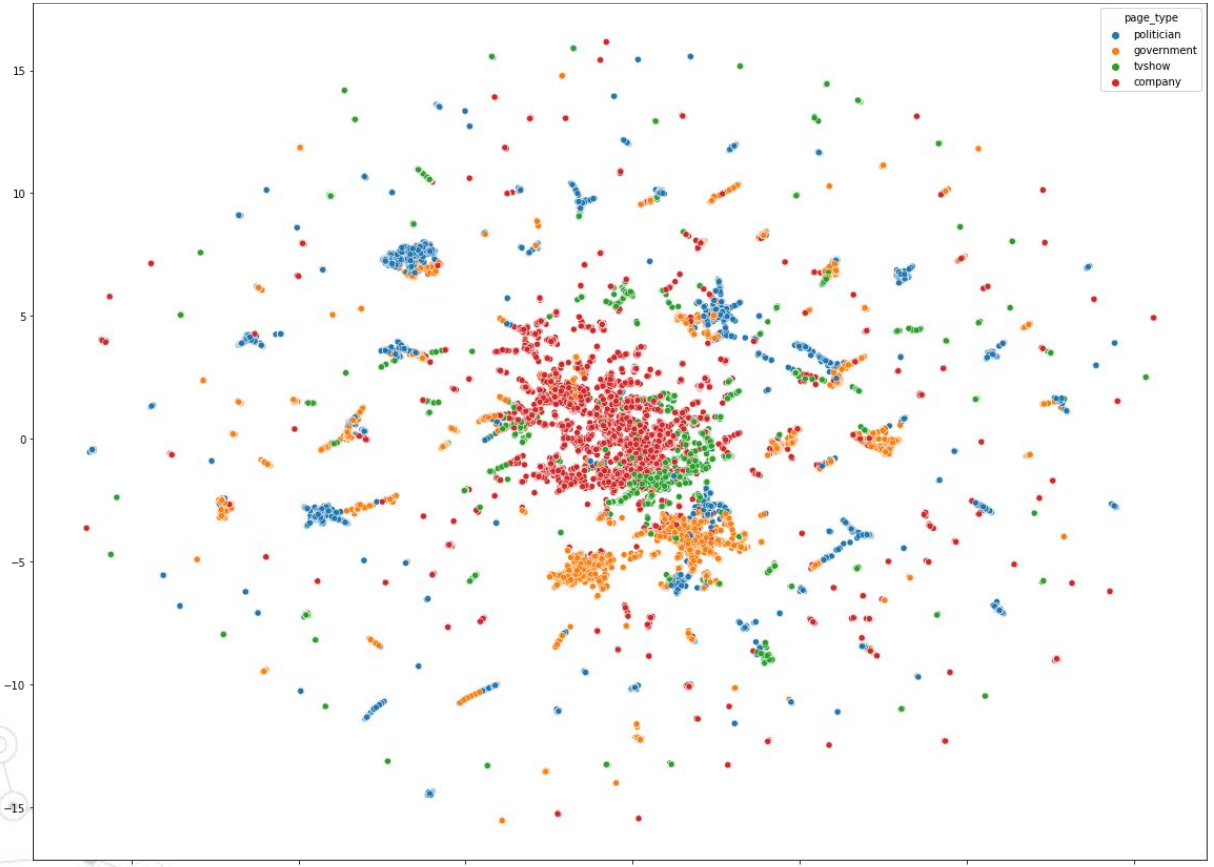
Source Code: https://github.com/julioegb/DeepWalk_vs_Node2Vec

4.2
Facebook Large Page-Page Network

Facebook Large Page-Page Network

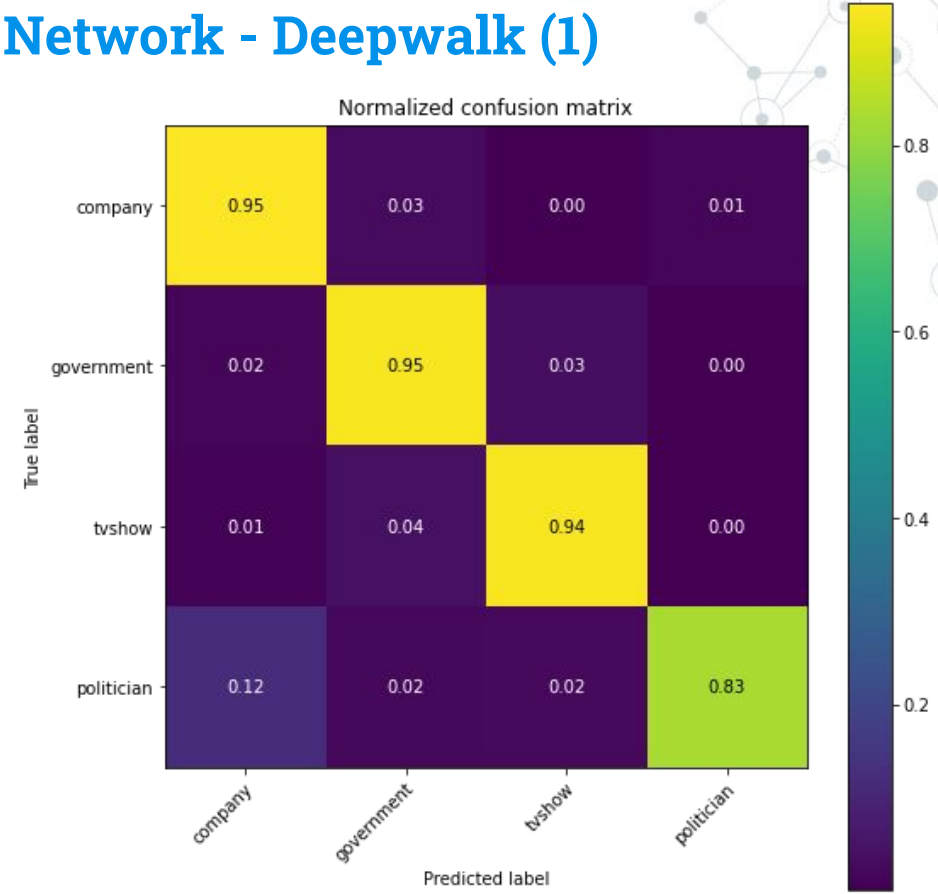
- Nodes: 22,470
- Edges: 171,002
- Categories:
  - Politicians
  - Governmental Organizations
  - Television Shows
  - Companies
- Features are the site descriptions that summarize the purpose of the site.

Facebook Large Page-Page Network - Deepwalk



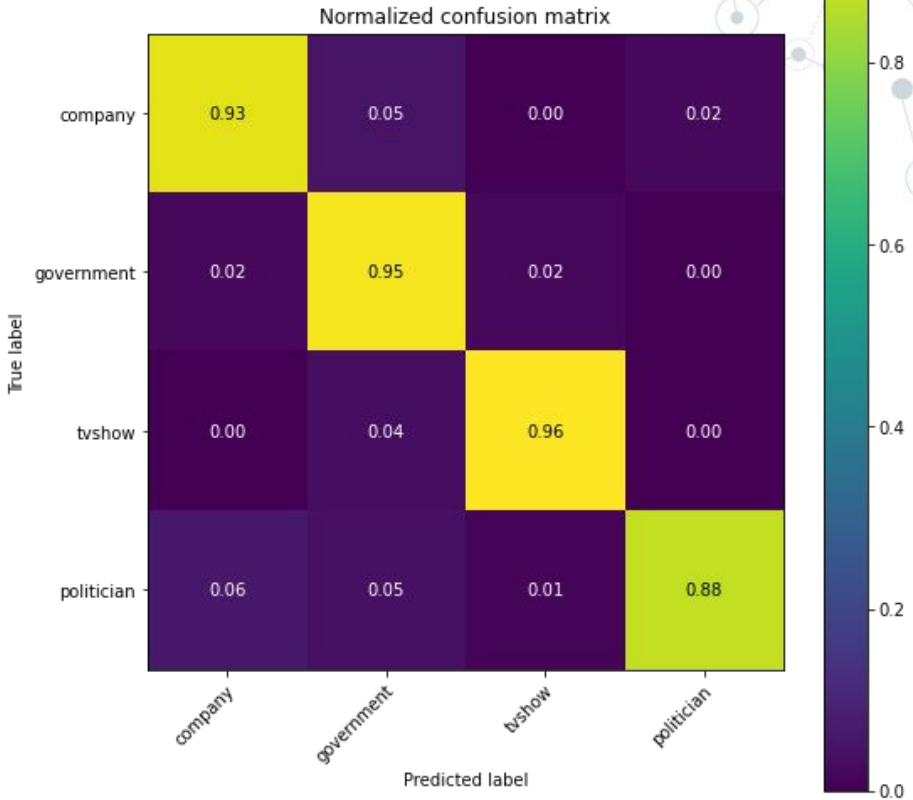
Facebook Large Page-Page Network - Deepwalk (1)

- Deep Walk
  - Random Forest
  - Not Balanced
  - Accuracy = 0.931

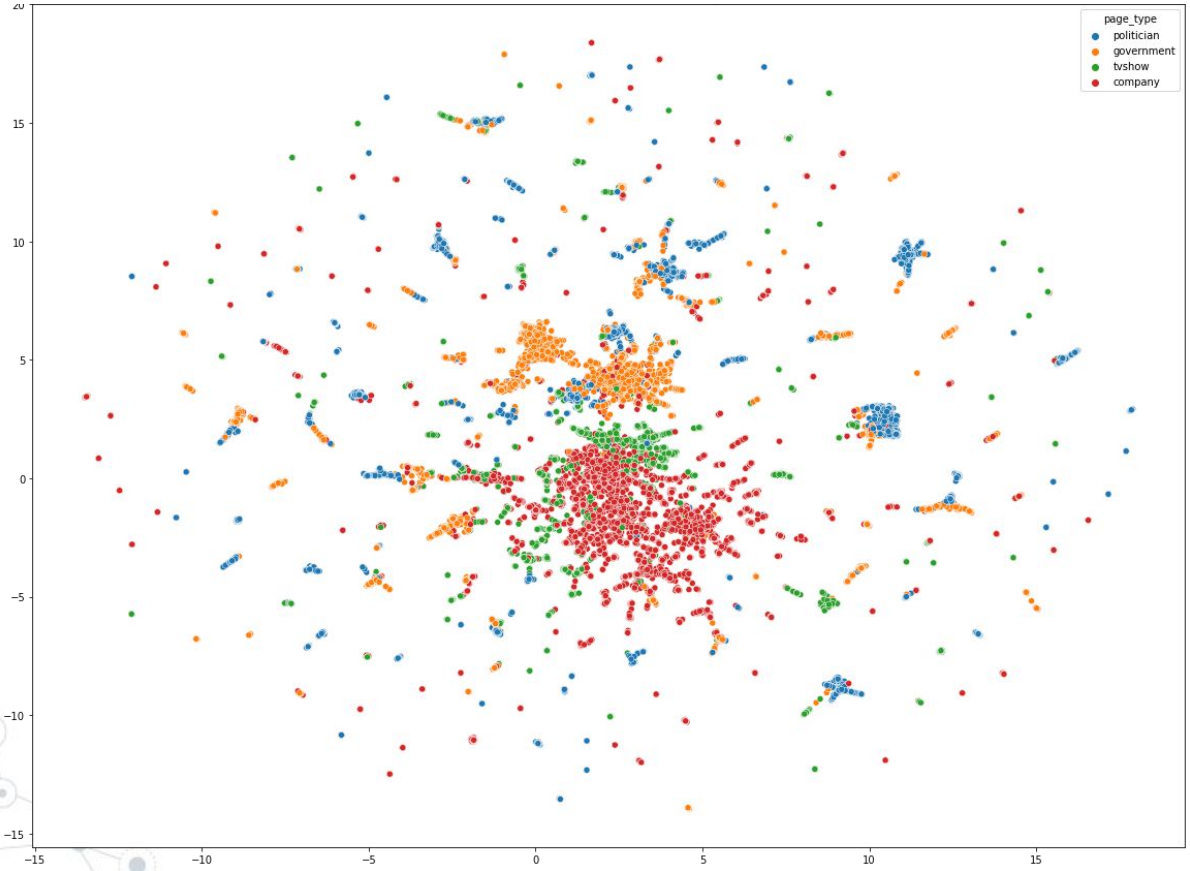


Facebook Large Page-Page Network - Deepwalk (2)

- Deep Walk
  - Random Forest
  - Balanced
  - Accuracy = 0.937

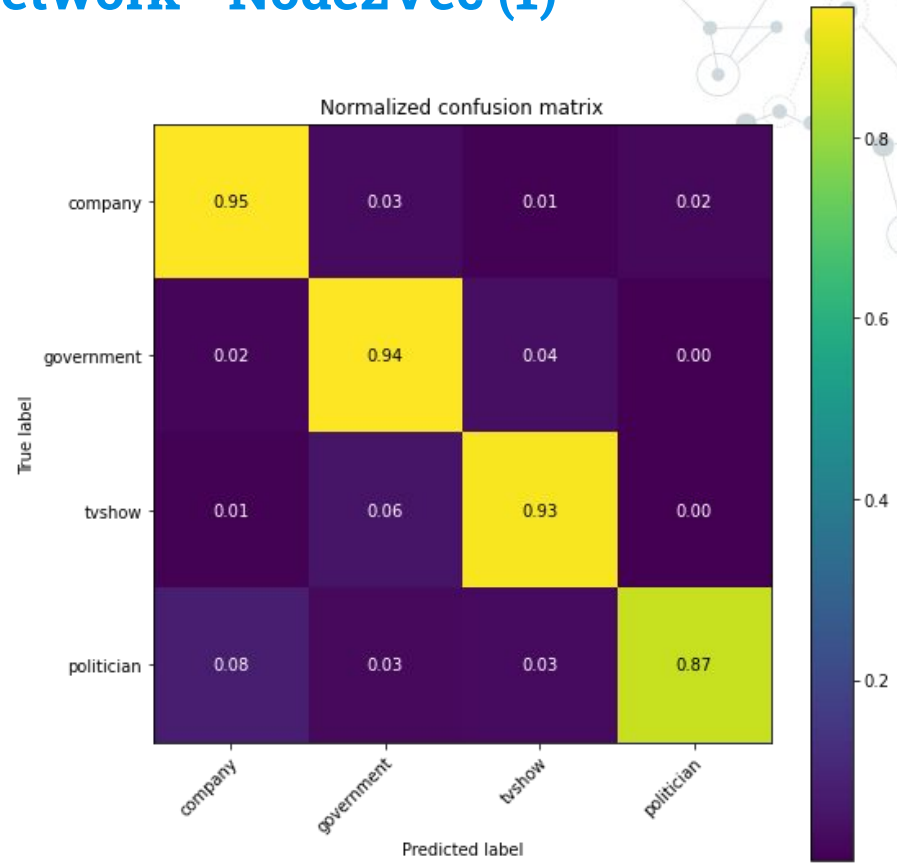


Facebook Large Page-Page Network - Node2Vec



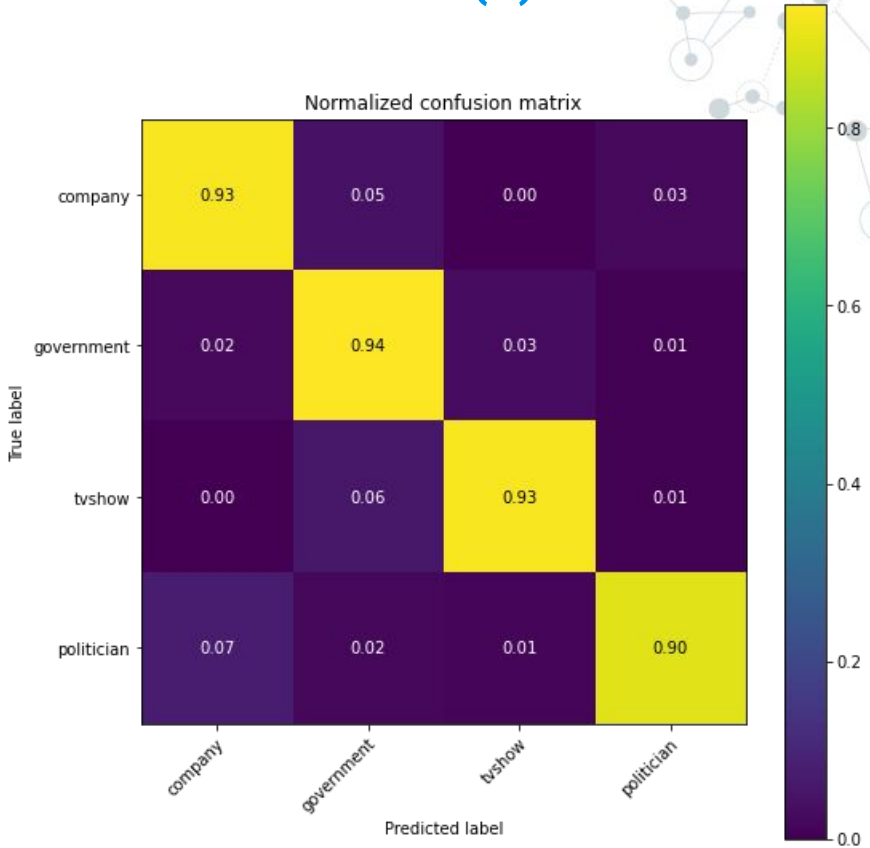
Facebook Large Page-Page Network - Node2Vec (1)

- Node2Vec
  - P = 10
  - Q = 0.1
  - Random Forest
  - Not Balanced
  - Accuracy = 0.931

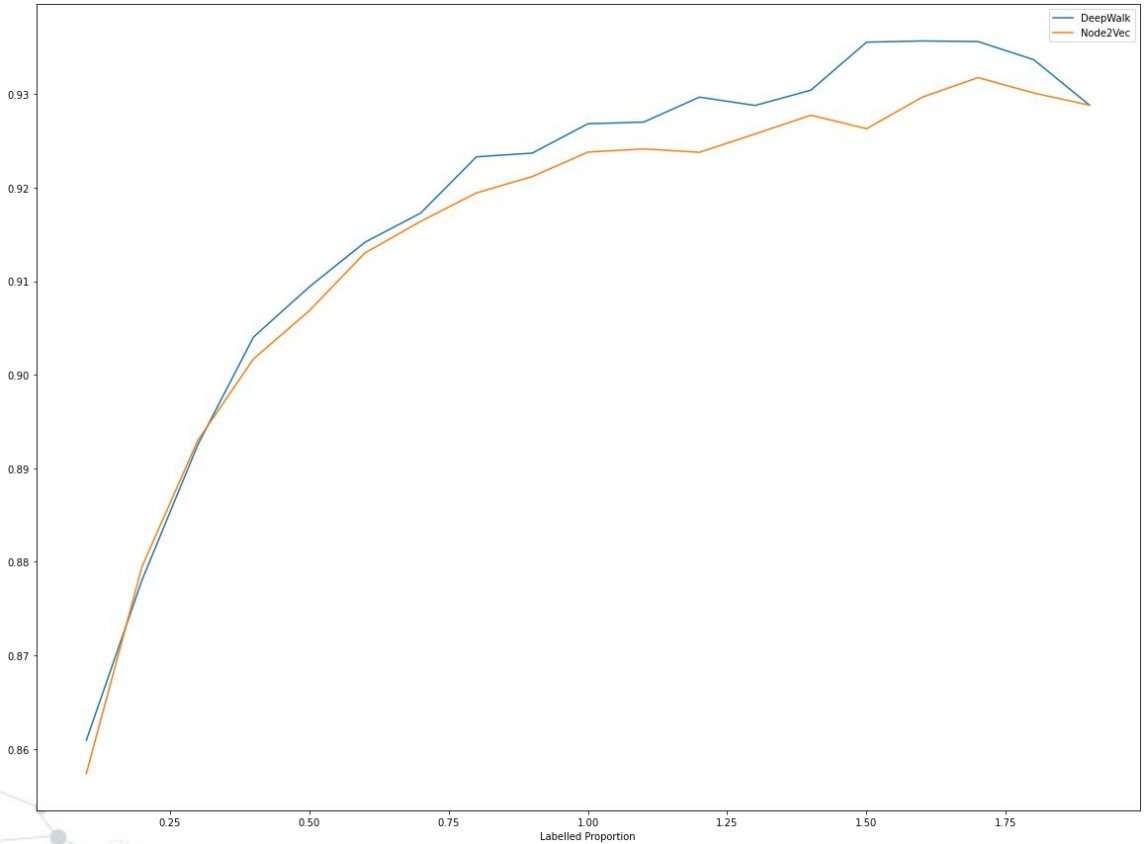


Facebook Large Page-Page Network - Node2Vec (2)

- Node2Vec
  - P = 10
  - Q = 0.1
  - Random Forest
  - Balanced
  - Accuracy = 0.928



Deep Walk vs Node2Vec

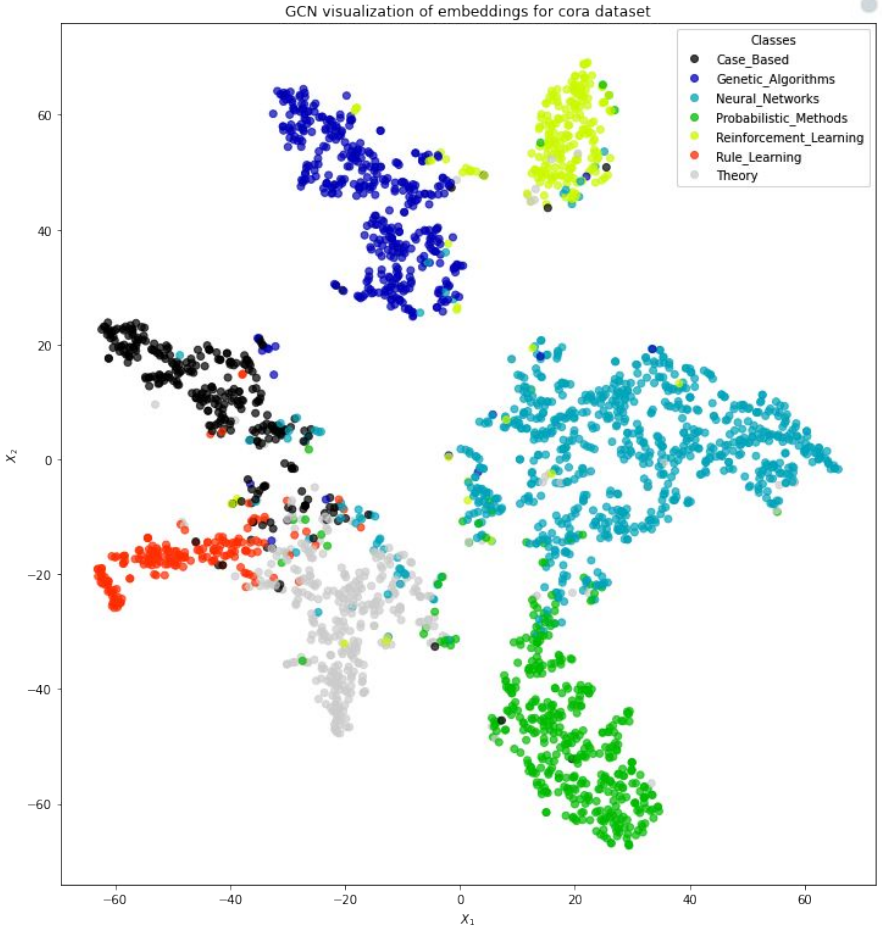


4.3
Cora Dataset

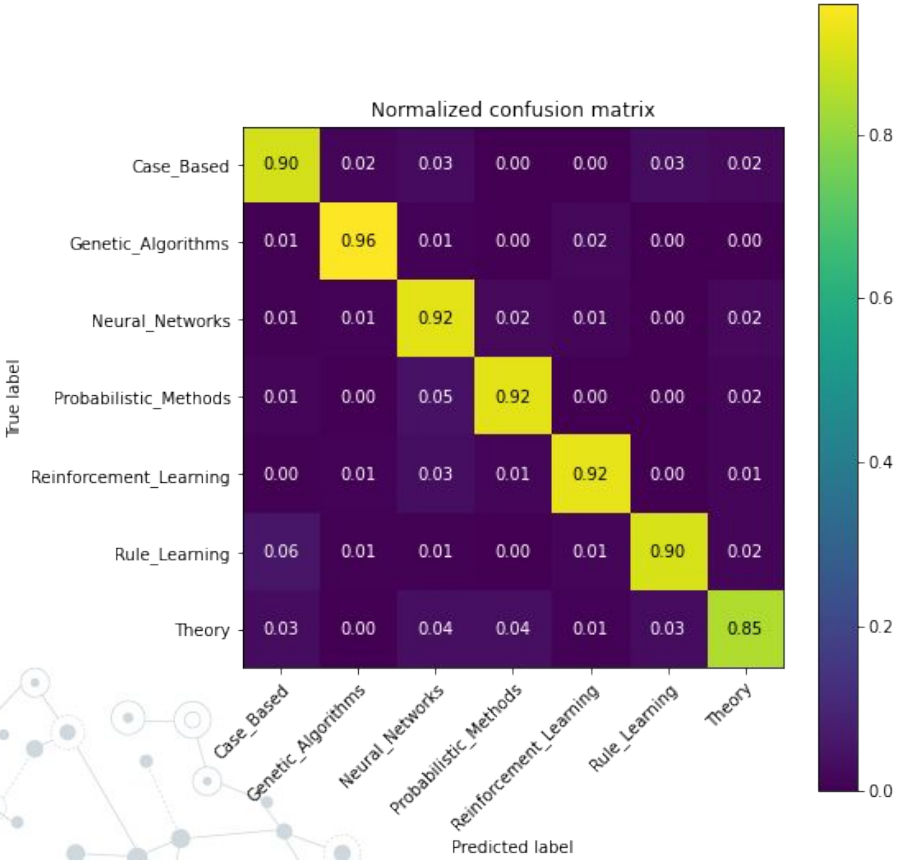
Cora Dataset

- Nodes: 2,780
- Edges: 5,429
- Categories:
  - Neural_Networks
  - Probabilistic_Methods
  - Genetic_Algorithms
  - Theory
  - Case_Based
  - Reinforcement_Learning
  - Rule_Learning

Cora Dataset Classification GCN

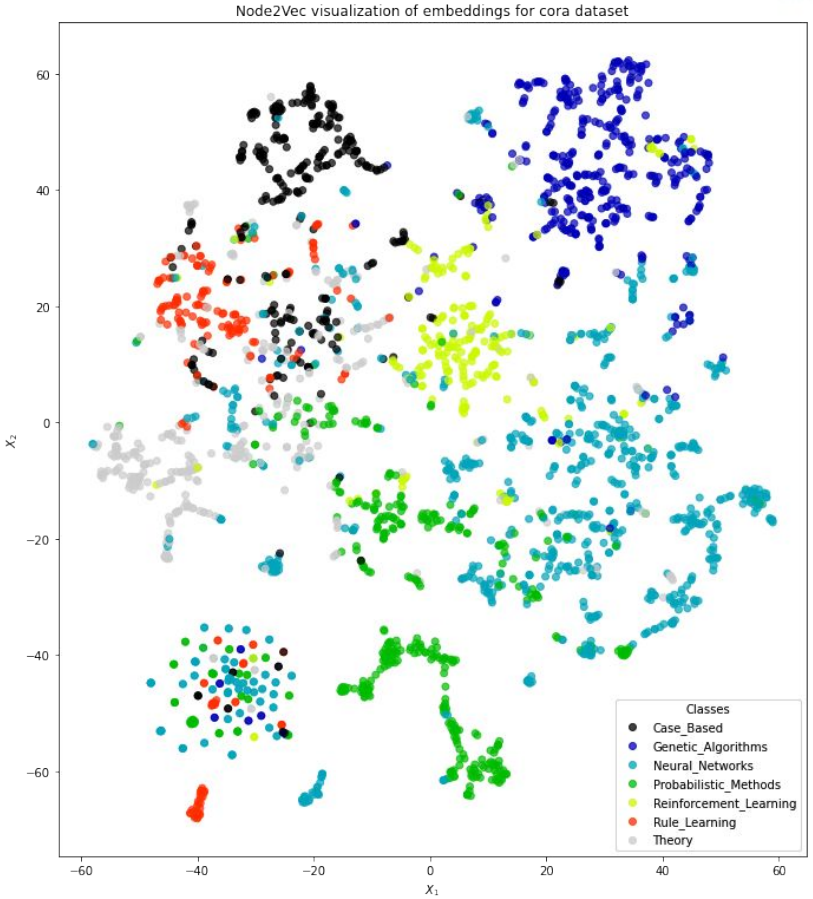


Cora Dataset - Classification - GCN



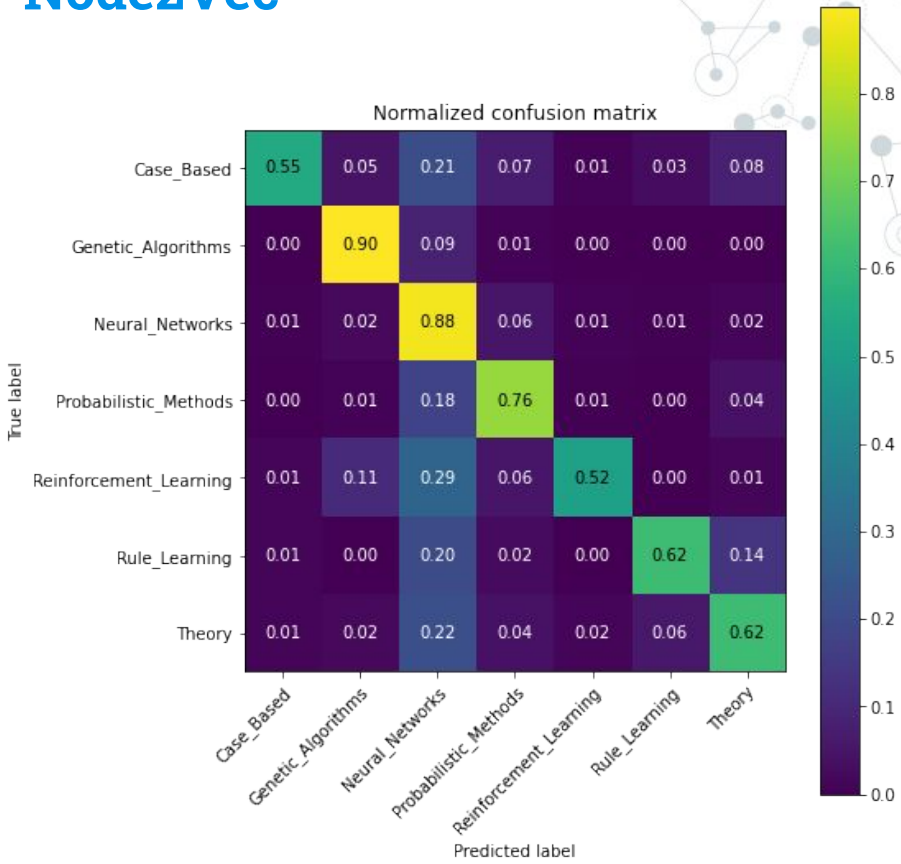
- GCN
  - Accuracy = 0.913

Cora Dataset Classification Node2Vec

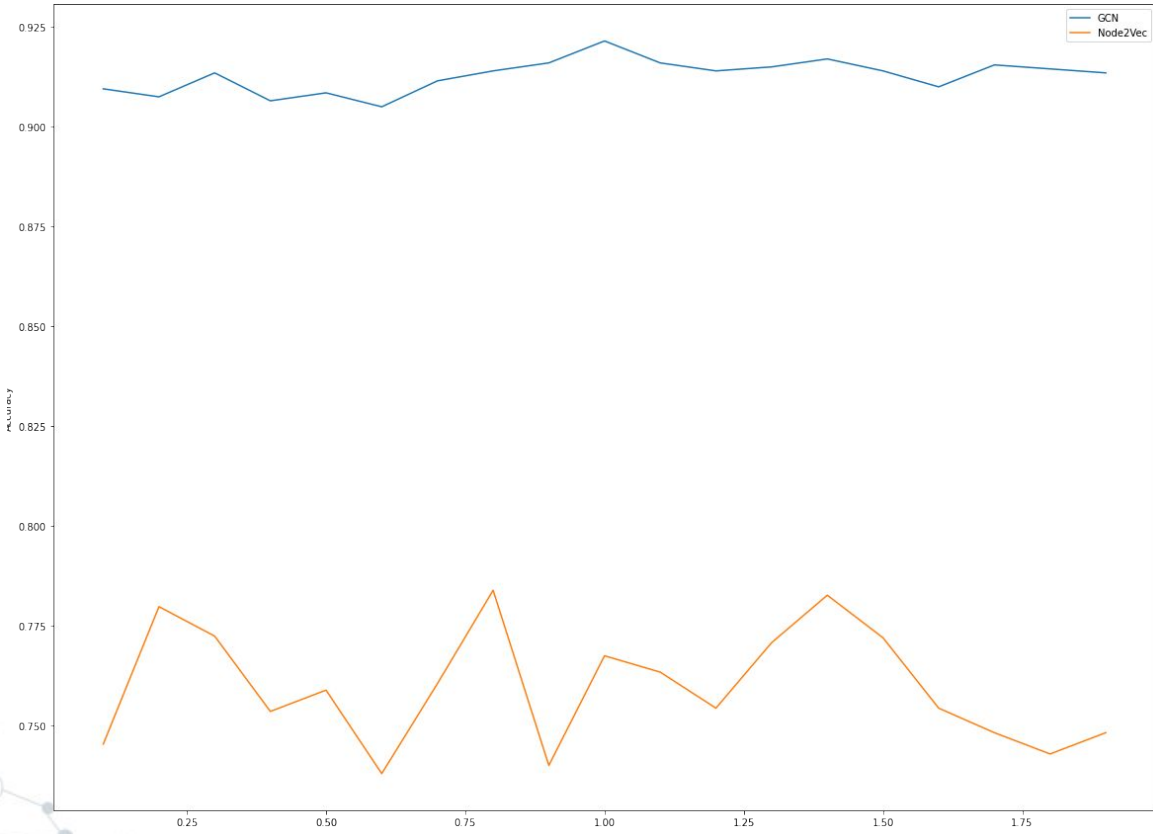


Cora Dataset - Classification - Node2Vec

- Node2Vec
  - P = 0.5
  - Q = 0.25
  - Random Forest
  - Accuracy = 0.748



GCN vs Node2Vec



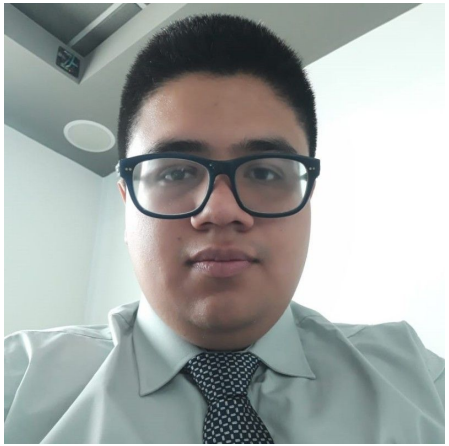
Conclusion

- Node2Vec is an improvement to DeepWalk. Graph Convolutional Networks is an improvement of Node-2Vec.
- For community detection it is better to use Node2Vec. Ranking improves compared to Deep Walk.
- For classification after comparing the different methods, GCN obtained a more precise accuracy

References

- A. Grover and J. Leskovec, “Node2vec: Scalable Feature Learning for Networks ,” Arxiv.org. [Online]. Available: https://arxiv.org/pdf/1607.00653.pdf. [Accessed: 09-Feb-2022].
- B. Perozzi, R. Al-Rfou, S. Skiena, “DeepWalk: Online Learning of Social Representations,” Arxiv.org. [Online]. Available: https://arxiv.org/abs/1403.6652. [Accessed: 09-Feb-2022].
- Leskovec, J. (s. f.). SNAP: Network datasets: Wikipedia Article Networks. SNAP: Stanford Network Analysis Project. https://snap.stanford.edu/data/facebook-large-page-page-network.html
- L. Valenzuela, “-research-project/research_project_luisvalenzuela.ipynb at main · lvalenzuelana/-research-project,” GitHub. [Online]. Available: https://github.com/lvalenzuelana/-Research-Project/blob/main/Research_Project_LuisValenzuela.ipynb. [Accessed: 09-Feb-2022].
- Relational.fit.cvut.cz. 2022. *Dataset*. [online] Available at: <https://relational.fit.cvut.cz/dataset/CORA> [Accessed 13 February 2022].
- Mejicanos, D., 2022. *Research/Research_GCN_Dennis_Mejicanos.ipynb at main · DennisM23/Research*. [online] GitHub. Available at: <https://github.com/DennisM23/Research/blob/main/Research_GCN_Dennis_Mejicanos.ipynb> [Accessed 13 February 2022].
- Jupyter Notebooks DeepWalk and Node2Vec Github Dataset. Available at: <https://github.com/julioegb/DeepWalk_vs_Node2Vec> [Accessed 13 February 2022].
- GitHub Social Network. *Dataset*. [online] Available at: <https://github.com/benedekrozemberczki/datasets> [Accessed 13 February 2022].

Our Team



Dennis Mejicanos

Telecommunications Engineer

dennis.mejicanos_garcia@telecom-sudparis.eu

Luis Valenzuela

Telecommunications Engineer

luis.valenzuela_navarro@telecom-sudparis.eu

Julio Guzmán

Computer Science Engineer

julio.guzman_barraza@telecom-sudparis.eu