

Your Original and Relevant Course Project Title

Social Network Analysis for Computer Scientists — Course paper

Chenyu Shi

s3500063@umail.leidenuniv.nl

LIACS, Leiden University

Leiden, Netherlands

Shupe Li

s3430863@umail.leidenuniv.nl

LIACS, Leiden University

Leiden, Netherlands

ABSTRACT

KEYWORDS

node2vec, GCN, graph embeddings, social network analysis, network science

ACM Reference Format:

Chenyu Shi and Shupe Li. 2022. Your Original and Relevant Course Project Title: Social Network Analysis for Computer Scientists — Course paper. In *Proceedings of Social Network Analysis for Computer Scientists Course 2022 (SNACS '22)*. ACM, New York, NY, USA, 2 pages.

1 INTRODUCTION

Graphs are mathematical objects that can model complex relationships on non-Euclidean space. They are widely used in multiple domains such as molecular structure modelling, social network analysis, recommender systems, etc. To leverage the information contained in graphs, it is essential to develop efficient techniques for representing graph-structured data numerically.

Traditional statistical and machine learning methods are designed for extracting features from structured data on Euclidean space. For example, principal component analysis (PCA), uniform manifold approximation and projection (UMAP), and t-distributed stochastic neighbor embedding (T-SNE) are common techniques to reduce dimensions and capture features of data. Although these methods have achieved satisfactory performance on structured data, they are hard to be generalized to graph-structured data, because they highly depend on properties of Euclidean space.

This challenge led to the development of techniques specifically for graph-based representation. There are two main types of these techniques: shallow embedding method and deep embedding method [4]. Shallow embedding methods use shallow encoder functions to map the original graph structure onto a Euclidean space and obtain the embedding matrix. If data is labelled, we can apply supervised learning algorithms, e.g. label propagation, to extract embeddings later used in a supervised task. However, labels are not available or only partly available in most cases, where we need unsupervised learning or semi-supervised learning to distill information about graph structure. These methods can be divided into distance-based method and outer product method further [4]. Generally, distance-based methods select a metric function that indicates distances between any pairs of nodes and optimize the

function to generate embeddings. Representative distance-based methods include multi-dimensional scaling and laplacian eigenmaps. Outer product-based methods use matrix operations to evaluate the similarity between nodes. Most of early studies in graph embedding field adopt matrix factorization to reduce dimensionality of data while preserve the structure information [1]. Another mainstream outer product-based method is inspired by the development in natural language processing. Existing research generalizes the skip-gram word embedding framework to capture the graph embeddings, which has been proved to be efficient on many graph related tasks [5][6]. Node2vec [2], one of algorithms addressed in this paper, is also a variation of skip-gram-based method.

Node2vec is a semi-supervised algorithm whose goal is learning features from networks [2]. It transforms the graph embedding learning into a maximum likelihood optimization problem in a similar way to skip-gram architecture of word embedding learning. Likelihood calculation requires a clear definition of the neighborhood. Textual data has the intrinsic semantic order that can be naturally employed as word neighborhoods. However, graph-structured data has no explicit neighborhoods. Node2vec introduces the idea of the second-order random walk into graph neighborhood sampling strategy. The emphasis of node2vec model is easy to switch between breadth-first sampling (BFS) and depth-first sampling (DFS) by adjusting hyperparameters. Moreover, its computational complexity is less than classical BFS and DFS strategies. Because of its efficiency and great performance on graph embedding learning task, node2vec is an ideal choice among shallow embedding methods.

In recent years, a lot of studies have focused more on deep embedding method rather than the shallow one. Deep embedding method usually refers to algorithms that learn graph features via graph neural networks (GNN). The GNN is a class of artificial neural networks constructed for graph-structured data. Inspired by the success of convolutional neural networks (CNN) on grid data, many GNN architectures have been proposed to generalize the convolution operation on graphs. In this paper, we mainly focus on a method called graph convolutional networks (GCN) [3]. GCNs defines the graph convolution based on the graph Laplacian spectrum. It has achieved state-of-the-art performance on common graph related tasks, such as node classification, link prediction, etc.

We propose a novel method to extract embeddings from graphs in this paper. Our method is based on node2vec and GCNs. Motivated by the concept of meta learning, we regard the embeddings returned by node2vec as the meta information for GCNs. This prior knowledge helps to improve the quality of final graph embeddings, and therefore enhances the model performance in various tasks.

The rest of the paper is organized as follows. Section 2 reviews related works on graph embedding learning methods. We illustrate

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SNACS '22, Master CS, Fall 2022, Leiden, the Netherlands

© 2022 Copyright held by the owner/author(s).

necessary notations, formula, and assumptions in Section 3. And then we describe approaches in detail in Section 4. Section 5 is an introduction to five open source data sets we use in the project. After that, we present our experimental set-up and results in Section 6. The paper ends with a conclusion section.

2 RELATED WORK

3 PRELIMINARIES

4 APPROACH

5 DATA

6 EXPERIMENTS

7 CONCLUSION

ACKNOWLEDGMENTS

REFERENCES

- [1] HongYun Cai, Vincent W. Zheng, and Kevin Chen-Chuan Chang. 2018. A Comprehensive Survey of Graph Embedding: Problems, Techniques, and Applications.

- IEEE Transactions on Knowledge and Data Engineering* 30, 9 (2018), 1616–1637. <https://doi.org/10.1109/TKDE.2018.2807452>
- [2] Aditya Grover and Jure Leskovec. 2016. Node2vec: Scalable Feature Learning for Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*.
- [3] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations*.
- [4] Kevin P. Murphy. 2022. *Probabilistic Machine Learning: An Introduction*. MIT Press.
- [5] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. DeepWalk: Online Learning of Social Representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*.
- [6] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. LINE: Large-Scale Information Network Embedding. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*.