Università
degli Studi
di Padova

Dipartimento
di Ingegneria
dell'Informazione

INFORMATION ENGINEERING DEPARTMENT

MASTER'S DEGREE IN COMPUTER ENGINEERING

APEROL from Networks: Analyzing Pipeline and Embedding Representations for Optimized Learning (from Networks)

Professor
Fabio Vandin

Students
Marco Annunziata, 2160851
Silvia Mondin, 2141201
Sveva Turola, 2160852

ACADEMIC YEAR 2025-2026

Contents

1 Dataset preprocessing

Before proceeding to a discussion of the preprocessing of the datasets, it is necessary to inform readers that changes have been made to the Pennsylvania dataset in the proposal. The dataset from [1] has been used instead. The decision to change the dataset was made to have a directed graph, instead of an undirected one, so to have more uniformity among road networks. The main properties of the new dataset are the following: $|V| = 1088092, |E| = 3083796$, and Directed.

All of the datasets utilized in this study were subjected to a unification of format through a preprocessing procedure. The datasets were in various file formats and contained different types of information. Initially, all the datasets were manually converted to CSV files. Subsequently, a Python script was implemented to automate the preprocessing of the datasets. To ensure a uniform data structure, all datasets were converted into CSV files comprising three columns: `u`, `v`, and `weight`. The columns `u` and `v` represent the source and target nodes of an edge, respectively. In the case of an undirected dataset, each edge between two vertices, denoted by (u, v), is represented by two directed edges, namely (u, v) and (v, u). The column `weight` represents the weight of the edge. In the case of an unweighted dataset, a uniform weight of 1.0 was assigned to all edges. Finally, if the node identifications were not numeric, they were mapped to numeric consecutive IDs starting from 0.

References

[1] Jure Leskovec and Andrej Krevl. *SNAP Datasets: Stanford Large Network Dataset Collection.* `http://snap.stanford.edu/data`. June 2014.

Contribution of Authors and AI Usage

The contributions for this project are:

- Datasets preprocessing: Silvia Mondin (implementation, writing);

The following AI tools were used:

- Copilot was used to suggest code and report snippets.
- DeepL was used to correct the grammar and syntax of the text.