



UNIVERSITÀ  
DEGLI STUDI  
DI PADOVA



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**

<b>1 Dataset preprocessing</b>	<b>1</b>
<b>References</b>	<b>1</b>
<b>Contribution of Authors and AI Usage</b>	<b>2</b>

# 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 and the Twitch dataset in the proposal. The dataset from [1] has been used instead of the Pennsylvania dataset from the proposal. 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:  $|V| = 1,088,092$ ,  $|E| = 3,083,796$ , and Directed.

The Twitch dataset was replaced with the Deezer dataset from [2]. This was necessary due to the large number of edges of Twitch dataset, revealed unfeasible from the initial experiments. The Deezer dataset has the following properties:  $|V| = 143,884$ ,  $|E| = 846,915$ , and Undirected.

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] A. Dasgupta J. Leskovec K. Lang and M. Mahoney. “Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters”. In: *Internet Mathematics* 6.1 (2009), pp. 29–123. URL: <https://snap.stanford.edu/data/roadNet-PA.html>.
- [2] Benedek Rozemberczki et al. “GEMSEC: Graph Embedding with Self Clustering”. In: *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2019*. ACM. 2019, pp. 65–72. URL: <https://snap.stanford.edu/data/gemsec-Deezer.html>.

## 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.