A Dual Graph Framework for Edge Embedding and Clustering in Road Safety Analysis

Simone Paradiso[12]*, Apostolos Ziakopoulos[1] and George Yannis[1]

[1]National Technical University of Athens, Department of Transportation Planning and Engineering, 5 Iroon Polytechneiou Street, 15773, Athens, Greece

[2]OSeven Single Member Private Company, 27B Chaimanta Street, 15234 Chalandri, Greece

TRA TRANSPORT RESEARCH ARENA BUDAPEST 18-21/05/26

Introduction

- **Road safety** is a critical global issue – over **1 million fatalities** annually.
- **Spatial context matters** – road safety is closely linked to geography.
- **Graph Neural Networks** (GNNs):
  - Extend neural networks to **graph-structured data**.
  - Generate node embeddings capturing structural and relational properties along with neighboring context (**Embedding**).
- **Dual-graph** approach:
  - Transforms edges into nodes.
  - Enables node-based methods for edge-level analysis in road networks.
- **Clustering algorithms** on**:**
  - **Raw edge features**:
    - **K-Prototypes** to handle mixed data types.
    - Silhouette scores (via Gower's distance) were consistently low across K values.
  - **Edge embeddings**:
    - **K-Means** on numeric features.
    - Silhouette scores indicated well-defined partitions.
- Clustering on road networks improves significantly when using GNN-generated embeddings.
  - More **coherent** and **meaningful partitions**.

Data Collection and Preparation

- **Data sources**:
  - **Telematics sensor data** collected through a smartphone application developed by **OSeven Telematics** in compliance with GDPR.
    - **13,000+ anonymized trips** within the central area of the Athens metropolitan region, collected over the **last four months of 2024** at sampling rate **1 Hz**.
    - The data includes trip coordinates along with **speed data**, **periodic binary flags denoting the presence of harsh events and their intensity**, **speeding** or **mobile use**, generated by proprietary ML algorithms developed by OSeven.
  - **OpenStreetMap** (OSM) for graph extraction based on telematics coverage.
- Telematics were **aggregated** to the OSM entities by summing or averaging, depending on the feature's nature.
  - The data were matched to the **nearest edge** and aggregated per edge.
  - **For nodes**, only observations within a **buffer of 50 m** and on the **connected edges to the nodes** were used to compute node features, helping prevent excessive overlaps.
- The **graph** includes all edges enriched with telematics data and features from OSM, along with start and end nodes characterized by telematics data (Figure 1).

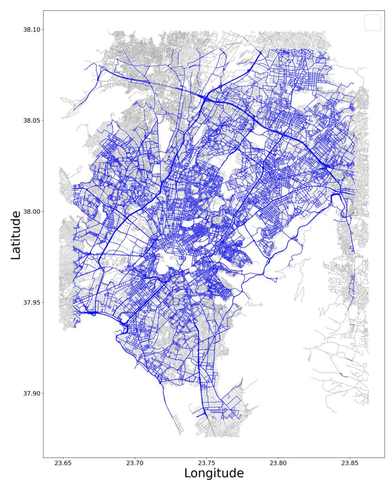

Figure 1 Road Network with Telematics-Characterized Nodes and Edges

- **Edge features** like SpeedingFlag, mobile_usage, harsh_acc, harsh_brk, and event_intensity were **exposure-normalized**, with edge values expressed **per meter per trip** to account for traffic and segment length (**Table 1**).
- **Node features** match Table 1 except that length, oneway, and RoadType are excluded, street_count is added, and **normalization uses trip counts only**.

Table 1 Edge Features

Feature	Description
Oneway	Binary indicator of travel direction along the road segment
RoadType	Classification of the road segment (e.g., service road, urban, rural)
SmoothenedSpeed	Average vehicle speed along the road segment
SpeedingFlag_per_vkt	Number of speeding events per meter of road per trip
Mobile_usage_per_vkt	Number of phone usage events per meter of road per trip
Harsh_acc_per_vkt	Number of harsh acceleration events per meter of road per trip
Harsh_brk_per_vkt	Number of harsh braking events per meter of road per trip
Event_intensity_per_vkt	Average intensity of harsh events per meter of road per trip

Methods

- **Clustering** is an unsupervised machine learning method that **groups data points** by similarity.
  - **K-Means** requires **specifying the number of clusters**, chosen via domain knowledge or validation metrics like the **Silhouette** score (−1 to 1) indicating cluster separation.
  - **K-Prototypes** extends K-Means to handle categorical, binary or **mixed data**. While **Gower's distance** combines these variable types into a single dissimilarity metric to compute silhouette scores.
- A **GNN** extends neural networks to perform **node embedding** task encoding each node as a numeric vector capturing its position and neighborhood context (Figure 2).
  - A **dual graph** converts edge embedding into a node embedding task.
  - GNN layer updates each **node's representation** by using a **learnable function** $f_\emptyset$ and **aggregated neighbor features**:

$$h_i' = f_\emptyset\left(h_i, AGGREGATE\left(\{h_j \mid j \in \mathcal{N}_i\}\right)\right)$$

where $\mathcal{N}_i$ denotes the neighbors of node i (including i), and h_i and h_i' are the current and updated node feature vectors.

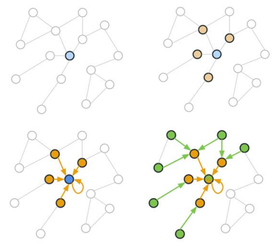

Figure 2 How GNNs Learn: Information Sharing

  - This study uses Graph Attention Networks (**GAT**) which aggregates neighbor feature through a **weighted average**, weighing neighbors via an **attention mechanism** α**.**
  - The model is trained in a **self-supervised way** with a graph contrastive loss. For each node i, the loss is defined as:

$$\mathcal{L}_i = -\log\left(\frac{\sum_{j\in P_i} e^{\frac{S_{ij}^{+}}{\tau}}}{\sum_{j\in P_i} e^{\frac{S_{ij}^{+}}{\tau}} + \sum_{j\in N_i} e^{\frac{S_{ij}^{-}}{\tau}}}\right)$$

where S_{ij} denotes the node similarity between and τ is a temperature parameter.

The final loss averages $\mathcal{L}_i$ over all nodes, pulling **representations of neighboring nodes** closer together and pushing **non-neighbors** apart in the **embedding space**.

Comparing Clustering Approaches

- Numeric edge and node features were scaled.
- **K-Prototypes**:
  - Silhouette score computed using the Gower's matrix based on cluster labels from several K-Prototypes run.
  - Scores ranged from 0.07 to 0.33 across different K values, indicating **weak clustering performance**.
- **GNN + K-Means**:
  - **GNN architecture** defined as follows:
    - Two GAT layers, each followed by a ReLU activation function.
    - Multi-head attention with three heads.
    - Shallow FNN layer to preprocess binary features before GAT input.
    - Trained on the dual graph for 10 epochs using PyG NeighborLoader and Adam optimizer.
  - **Learned embeddings clustered** using K-Means for multiple K values and evaluated with silhouette scores.
  - The **score peaked at 0.92 (K=2)**, indicating a well-defined partition consisting of two road groups (Figure 2).

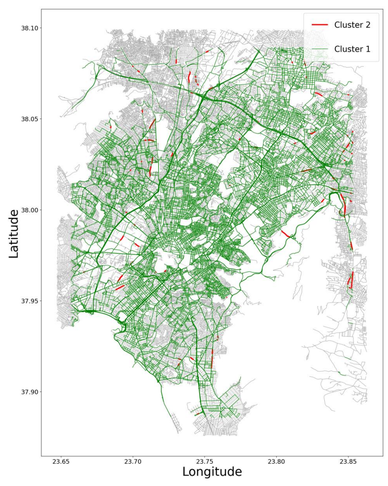

Figure 3 Road Network with Telematics-Characterized Nodes and Edges

Key Findings and Discussions

- Cluster **interpretation** approach:
  - **Embedding abstraction** inherently limits **interpretability**.
  - **Cluster labels** were mapped back to the **raw dataset**, averaging numerical features and using the mode for categorical/binary features to create representative cluster profiles.

Table 2 Cluster Profiles

Features	Cluster 1 (43,378 edges; 99.30%)	Cluster 2 (1305 edges; 0.70%)
Oneway	1	1
RoadType	Urban	Rural
SmoothenedSpeed	26.47	49.63
SpeedingFlag_per_vkt	0.0008	0.032
Mobile_usage_per_vkt	0.005	0.003
Harsh_acc_per_vkt	0.0001	0
Harsh_brk_per_vkt	0.0001	0.0003
Event_intensity_per_vkt	0.0003	0.0006

- Table 2 presents the cluster profiles. **Cluster 1** represents:
  - Roads associated with **higher average speeding** and speeding-flag values and with a **predominantly rural environment**, compared to Cluster 1.
  - Roads with more **frequent** and **intense harsh braking**, suggesting that drivers tend to drive faster and brake more abruptly.
    - **Low numerical magnitude** of cluster means reflects exposure-normalized scaling and does not influence clustering (features were standardized).
- While **Cluster 2** represents:
  - Most roads in the network showing lower or comparable average values across all telematics measures, suggesting **more moderate driving behavior** across them.
- The discussed approach:
  - Improves **clustering quality**, yielding more accurate and meaningful network partitions.
  - Enables **the use of K-Means** by working exclusively with numerical data.
  - By constructing a **dual graph**, the method leverages the well-established framework of **node embedding tasks** to generate edge embeddings.
- While crash data were not used, these **behavioral patterns** may be indicative of conditions associated with road safety risk and can inform targeted, data-driven **interventions** using edge-based **spatial analysis**.

Conclusions and Limitations

- Linking **driver behavior** with the **spatial context** of road networks provides **valuable insights** for spatial driving behavior patterns.
- Generating edge embeddings with a GNN and clustering them via K-Means produces a clear road-network partition, supporting **proactive risk assessment** and **targeted interventions**.
- **Future work**:
  - Exploring alternative **clustering algorithms** and different **GNN architectures** could offer further insights into graph partitioning.
  - Incorporating **macroscopic traffic features** or additional telematics variables may enhance the framework's **effectiveness in real-world applications**.
  - **Segmenting** the analysis by **urban versus rural environment** could provide additional insights, as well as a distinction between **one-way versus bi-directional roads**.

Acknowledgements

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101119590

Contact Information

Simone Paradiso

Department of Transportation Planning and Engineering

5, Iroon Polytechneiou str., GR-15773, Zografou, Athens

Tel: + 210.772.1575

Email: simone_paradiso@mail.ntua.gr

Web: https://linktr.ee/simone_paradiso