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

Simone Paradiso[1*][0009-0007-4653-1789], Apostolos Ziakopoulos[2][0000-0001-6252-6743] and George Yannis[3][0000-0002-2196-2335]

[1] National Technical University of Athens, Athens, Greece
*simone_paradiso@mail.ntua.gr
[2] National Technical University of Athens, Athens, Greece
apziak@central.ntua.gr
[3] National Technical University of Athens, Athens, Greece
geyannis@central.ntua.gr

Abstract. A road network can be represented as a graph, and telematics data can be spatially aggregated onto it, enabling the capture of driving behaviors across the network. Consequently, Graph Neural Networks (GNNs) provide a robust framework for analyzing such data and enabling network-level inference. While extensive research has focused on nodes, working with edges, being non-point entities, presents additional challenges. This study adopts a dual graph approach, converting edges into nodes, to perform node embedding by using a GNN model on the main central area of Athens, integrating geometric and telematics data capturing both the spatial structure of the road network and driving behavior patterns. To partition the network, K-Prototypes was first applied to raw edge attributes to handle mixed variable types, while K-Means was applied to the numerical embeddings generated by the GNN. Clustering GNN-generated embeddings with K-Means outperformed K-Prototypes, showing stronger separation and more reliable partitioning of road networks, which can support proactive safety planning and infrastructure management.

Keywords: Road Safety, Telematics Data, Graph Neural Network, Dual Graph, Edge Embeddings.

1 Introduction

Road safety remains a critical global issue with an estimated 1.19 million road traffic deaths in 2021 and road crashes being the leading cause of death of children and youth [1]. Traditional statistical methods and emerging machine learning techniques have provided empirical foundations to identify risk factors and develop data-driven safety policies for road safety [2]. Nevertheless, telematics data, defined as digital data collected through the integration of telecommunications and information systems, is increasingly being considered due to the scarcity of crash data and the relative ease of its collection [3,4]. Furthermore, road safety is influenced by geographic context and is therefore closely linked to spatial analysis, which enables the identification of spatial patterns in road safety Key Performance Indicators (KPIs).

With the advent of big data, Machine Learning (ML) and Deep Learning (DL) techniques have gained prominence in spatial road safety analysis [5]. DL involves Artificial Neural Networks (ANN), which enables the model to capture complex patterns and relationships within the data [6]. The ANN architecture has evolved through the years leading to the development of Graph Neural Networks (GNN) extending ANNs to graph-structured data [7], which can be used to generate node representations from a graph capturing their structural or relational properties, preserving the context from neighboring nodes. This process, known as embedding, helps DL models to process the nodes more effectively for tasks like classification or clustering [8]. While most research in graph representation learning traditionally focus on nodes [9], edges can also serve as primary elements. This study adopts a dual graph approach, transforming edges into nodes to extend node-based methods to non-point network components, enabling the application of GNN models to perform node embedding tasks that correspond to edge-based analysis in the original road network.

A clustering algorithm was initially applied to the raw edge features using K-Prototypes, which handles mixed data types. The results were compared to those from K-Means clustering applied to the generated embeddings. The silhouette score, computed using Gower's distance for K-Prototypes, was evaluated across multiple values of K, but consistently yielded poor results. The same metric indicated a well-defined partition when applying K-Means to the generated embeddings. Consequently, the better-performing method was selected to produce a meaningful road network partition, thereby enabling a more accurate proactive and robust framework for road safety monitoring.

2 Data Collection and Preparation

The present work is based on analyzing telematics sensor data collected through a smartphone application developed by OSeven Telematics [10], that records driver behavior data [11] stored in compliance with Greek and European data protection regulations [12]. The dataset consists of over 13,000 anonymized different trips within a central area of the Athens metropolitan region, collected over the last four months of 2024 at a frequency of 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. These metrics are generated by proprietary ML algorithms developed by OSeven, whose accuracy has been validated against On-Board Diagnostics (OBD) data, on-road tests and literature benchmarks. Although the dataset comes from smartphone users of a private application, which may introduce demographic and trip-related biases, the large number of trips helps mitigate, though not fully eliminate, these limitations.

A graph of nodes and edges, based on the coverage of available telematics data, was extracted from OpenStreetMap (OSM), a free, open-source global map. 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. At this stage, the graph includes all edges enriched with telematics data and

features from OSM, along with the corresponding start and end nodes characterized by telematics data, without filtering by trip count. Segments with very few trips are treated as noise. This is illustrated in Figure 1.

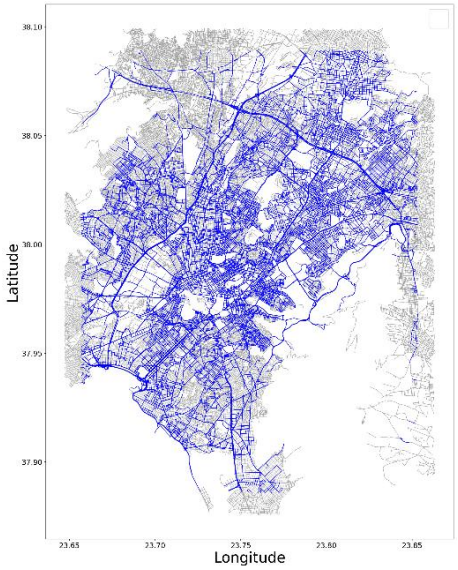

Fig. 1. Road Network with Telematics-Characterized Nodes and Edges

After aggregating telematics data per node and per edge, features including SpeedingFlag, mobile_usage, harsh_acc, harsh_brk, and event_intensity were normalized relative to exposure, with per-node values expressed per trip to account for differences in traffic volume, and per-edge values expressed per meter of road per trip to account for both traffic volume and segment length. Edge features are presented in Table 1.

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

Node features used for the dual node embedding task are the same as in Table 1 except length, oneway, and RoadType are excluded, and street_count is added. Normalization uses trip_counts only, not road length.

3 Methods

This study partitions a road network using features from telematics and OSM. It combines clustering methods, K-Means and K-Prototypes, evaluated via silhouette analysis, and leverages a Graph Neural Network (GNN) to generate edge embeddings for clustering. Each method is described in detail below.

3.1 Clustering with K-Means and K-Prototypes

The present study employs the K-Means algorithm described in detail in [13], which is an unsupervised machine learning technique that groups data points into clusters based on their similarity. K-means requires the number of clusters as input, which can be guided either by domain knowledge or by validation metrics such as the silhouette score, for example selecting the optimal K that maximizes the silhouette score across a range of values. The silhouette score [14] ranges from -1 to 1, and it is calculated for each data point in the dataset. The average score provides information about whether the clusters are well-separated or overlapping.

K-Prototypes extends the K-means algorithm to categorical or binary domains and domains with mixed values [15], while the Gower's distance combines numeric and categorical variables into a single dissimilarity metric [16] and it can be used to compute the Gower's matrix, which allows the calculation of the silhouette score based on the cluster labels obtained from K-Prototypes.

3.2 GNN Node Embedding

A GNN is an extension of existing ANNs suitable for data represented in graph domains [7]. In this work, the GNN model may be used to perform a node embedding task aiming to encode each node as a numeric vector, capturing its graph position and the structure of its local neighborhood context [17]. Since this study focuses on road network edges, a dual graph was constructed to transform the edge embedding task into a node embedding problem. Two key steps are as follows:

1. If two edges in the original graph share a common node, their corresponding nodes in the dual graph are connected by an edge. The features of the common node are assigned as features to the edge in the dual graph.
2. Each edge in the original graph becomes a node in the dual graph, carrying the original edge features as node features in the dual graph.

A GNN layer updates each node's representation by using a learnable function $f_{\emptyset}$, which is typically a neural network and an aggregation function that aggregates neighbor features. The updated node representation is computed as:

$$h_i' = f_{\emptyset}\left(h_i, AGGREGATE\left(\{h_j \mid j \in \mathcal{N}_i\}\right)\right) \quad (1)$$

Here, $\mathcal{N}_i$ denotes the neighbors of node i (including i), while h_i and h_i' represent its current and updated feature vectors. This study uses Graph Attention Networks (GAT) which weight neighbors via an attention mechanism α [18], which is a single-layer feedforward ANN, parametrized by a vector $\bar{a}$. The coefficient between nodes i and j is:

$$\alpha_{ij} = \frac{e^{(LeakyReLU(\vec{a}^T[W\vec{h_i}||W\vec{h_j}]))}}{\sum_{k\in\mathcal{N}_i} e^{(LeakyReLU(\vec{a}^T[W\vec{h_i}||W\vec{h_k}]))}} \quad (2)$$

Where || is the concatenation operation. The attention coefficients weigh neighbor features to create each node's updated representation, which can then pass through an activation function to produce the new node representation $h_i{}'$.

4 Results and Discussion

As first step of the analysis, the numeric edge and node features were scaled. Afterwards, the silhouette score was computed using the Gower's matrix based on cluster labels from each K-Prototypes run. Scores ranged from 0.07 to 0.33 across different K values, indicating weak clustering performance which would limit the reliability of the findings.

At this stage, a GNN architecture was defined with two GAT layers, each followed by a ReLU activation function and multi-head attention with three heads. The binary inputs were first processed through a shallow ANN with LeakyReLU activation, to enable learning a non-linear transformation of the binary features before passing them to the GAT input layer. The architecture was trained on the dual graph to generate edge embeddings over 10 epochs, using PyG's NeighborLoader and the Adam optimizer.

The model was trained in a self-supervised way with a contrastive loss inspired by prior work [19]. For each node i, the loss is defined as:

$$\mathcal{L}_i = -\log\left(\frac{\sum_{j\in P_i} e^{\frac{s_{ij}{}^+}{\tau}}}{\sum_{j\in P_i} e^{\frac{s_{ij}{}^+}{\tau}} + \sum_{j\in N_i} e^{\frac{s_{ij}{}^-}{\tau}}}\right) \quad (3)$$

The final loss is obtained by averaging $\mathcal{L}_i$ over all nodes. Here, S_{ij} denotes cosine similarity between node representations and τ is a temperature parameter that defines how strictly the loss treats the similarity scores. The objective uses softmax-normalized cosine similarity to increase similarity between node i and its neighbors P_i, while decreasing similarity to randomly sampled non-neighbors N_i, pulling the representations of neighboring nodes closer together and pushing the non-neighbors farther apart in the embedding space.

The generated embeddings were then clustered using K-Means across several values of K and evaluated using silhouette scores. The silhouette score reached a peak of 0.92 at $K = 2$, which is close to the maximum value of 1 and indicates a well-defined partition of the data. The resulting network partition is shown in Figure 2.

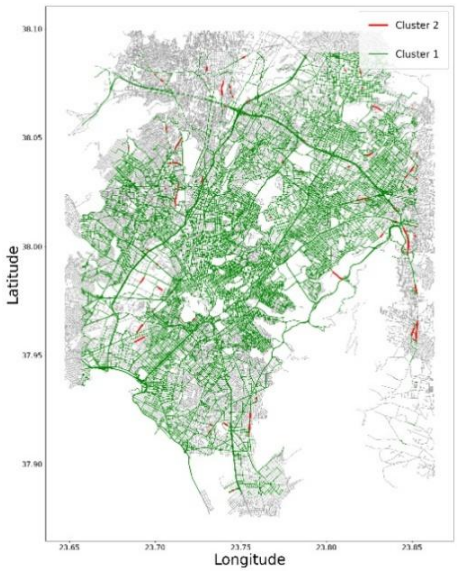

Fig. 2. Road Network Partitioning Based on GNN-Generated Edge Embeddings

Although the silhouette scores for K-Prototypes and K-Means are not directly comparable as they are computed on different feature spaces, the clustering structure in the embedding space yields a silhouette score close to 1, indicating a very well-defined partition, while the structure in the raw feature space produces a score near 0. This demonstrates a substantial improvement in road network partition when applying K-Means to the embeddings.

However, in this case interpretability is lost due to the abstraction introduced by the embedding process. To bridge the gap between interpretability and performance, cluster labels were mapped back to the raw feature dataset and all numerical features were averaged within clusters, while the mode was used for categorical and binary features, to obtain a representative profile of each cluster. And the results are shown in Table 2.

Table 2. Cluster Profiles Derived from Edge Embeddings

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

With a silhouette score of 0.92 the findings are reasonably reliable. Table 2 shows how the clustering detected a very small group exhibiting higher average values in terms of speeding and speeding flag and associated with a predominantly

rural environment. This group also exhibits more frequent and intense harsh braking per vehicle per road length, indicating a few roads where drivers tend to drive faster and brake more harshly. It is also worth noting that small numerical magnitude of the variables of cluster means reflects the exposure-normalized scale and does not affect clustering, which was performed on standardized features. The larger cluster, which encompasses most roads in the network, shows lower or comparable average values across all telematics measures, suggesting more moderate driving behavior.

The discussed novel approach not only improves clustering quality, yielding more accurate and meaningful network partitions, but also enables the use of K-Means by working exclusively with numerical data, instead of relying on mixed-type clustering methods like K-Prototypes. Additionally, by constructing a dual graph, the method leverages the well-established framework of node embedding tasks to generate edge embeddings.

While the road network partition is defined purely by observed driving patterns without crash data, the behavioral differences identified could be considered indicative of conditions associated with road safety risk. These patterns may help target interventions on roads where aggressive driving behaviors are more pronounced, highlighting the potential of edge-based spatial analysis to support proactive, data-driven road safety strategies.

5 Conclusions

The current study proposes a novel approach for analyzing graph-structured data, aiming to identify an edge-based partition within a road network enriched with telematics data. This framework enables the detection of structurally and functionally coherent road segments, leveraging telematics data to reveal spatial patterns in human driving behavior. Applying a K-Prototypes algorithm directly on raw mixed-type features yields poor road network partitioning. By contrast, converting the edge embedding task into a node embedding problem via a dual graph and using GNN-generated embeddings as input to K-Means produces a well-defined, road-based network partitioning. This approach handles mixed variable types and lays the groundwork for a proactive risk assessment and informing targeted interventions on the road network.

Exploring alternative clustering algorithms and different GNN architectures could offer further insights into graph partitioning. Furthermore, 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 was not considered, which could provide additional insights, as well as a distinction between one-way versus bi-directional roads.

6 Acknowledgments

This research is based on work carried out within the IVORY project. The project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement No 10111959.

7 References

1. Global status report on road safety 2023, https://www.who.int/publications/i/item/9789240086517, last accessed 2026/03/13.
2. Skaug L, Nojoumian M, Dang N, Yap A.: Road Crash Analysis and Modeling: A Systematic Review of Methods, Data, and Emerging Technologies. Applied Sciences 15, 7115 (2025). https://doi.org/10.3390/app15137115
3. Ziakopoulos A, Tselentis D, Kontaxi A, Yannis G.: A critical overview of driver recording tools. J Safety Res. 72, 203–212 (2020). https://doi.org/10.1016/j.jsr.2019.12.021
4. Joshi M, Bamney A, Wang K, Zhao S, Ivan J, Jackson E.: Analyzing the Suitability of Vehicle Telematics Data as a Surrogate Safety Measure for Short-Term Crashes. Transportation Research Record 2679, 489–504 (2025). https://doi.org/10.1177/03611981241263341
5. Ziakopoulos A, Yannis G.: A review of spatial approaches in road safety. Accident Analysis & Prevention. 135, 105323 (2020). https://doi.org/10.1016/j.aap.2019.105323
6. LeCun Y, Bengio Y, Hinton G.: Deep learning. Nature 521, 436–444 (2015). https://doi.org/10.1038/nature14539
7. Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G.: The Graph Neural Network Model. IEEE Transactions on Neural Networks. 20, 61–80 (2009). https://doi.org/10.1109/TNN.2008.2005605
8. Xu M.: Understanding Graph Embedding Methods and Their Applications. SIAM Rev. Society for Industrial and Applied Mathematics 63, 825–853 (2021). https://doi.org/10.1137/20M1386062
9. Chen F, Wang Y-C, Wang B, Kuo C-CJ.: Graph representation learning: a survey. APSIPA Transactions on Signal and Information Processing 9, e15 (2020). https://doi.org/10.1017/ATSIP.2020.13
10. OSeven Homepage, https://oseven.io/, last accessed 2026/03/13.
11. Kontaxi A, Tzoutzoulis D-M, Ziakopoulos A, Yannis G.: Exploring speeding behavior using naturalistic car driving data from smartphones. Journal of Traffic and Transportation Engineering 10, 1162–1173 (2023). https://doi.org/10.1016/j.jtte.2023.07.007
12. Papadimitriou E, Argyropoulou A, Tselentis DI, Yannis G.: Analysis of driver behaviour through smartphone data: The case of mobile phone use while driving. Safety Science 119, 91–97 (2019). https://doi.org/10.1016/j.ssci.2019.05.059
13. Steinley Douglas.: K-means clustering: A half-century synthesis. British Journal of Mathematical and Statistical Psychology 59, 1–34 (2006). https://doi.org/10.1348/000711005X48266
14. Shahapure KR, Nicholas C.: Cluster Quality Analysis Using Silhouette Score. 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA). pp. 747–748 (2020). https://doi.org/10.1109/DSAA49011.2020.00096
15. Huang Z.: Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery 2, 283–304 (1998). https://doi.org/10.1023/A:1009769707641
16. Akay Ö, Yüksel G.: Clustering the mixed panel dataset using Gower's distance and k-prototypes algorithms. Communications in Statistics - Simulation and Computation. Taylor & Francis 47, 3031–3041 (2018). https://doi.org/10.1080/03610918.2017.1367806
17. Hamilton WL.: Graph Representation Learning. Morgan & Claypool Publishers (2020).
18. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y.: Graph Attention Networks. arXiv preprint arXiv:1710.10903 (2018). https://doi.org/10.48550/arXiv.1710.10903
19. Chen T, Kornblith S, Norouzi M, Hinton G.: A Simple Framework for Contrastive Learning of Visual Representations. arXiv preprint arXiv:2002.05709 (2020). https://doi.org/10.48550/arXiv.2002.05709