Hierarchical Clustering on Graph Embeddings: A Scalable Approach to Risky Intersections

Simone Paradiso

IVORY

PhD Candidate and Researcher

oseven

Together with:
George Yannis & Apostolos Ziakopoulos

Department of Transportation Planning and Engineering
National Technical University of Athens

20th International Road Safety on Five Continents (RS5C 2025)

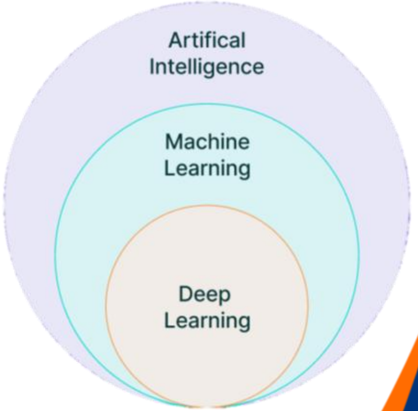
3-5 September 2025, Leeds, UK

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

Introduction

- Road crashes claim **1.3 million lives annually**, the leading cause of death for those **under 29** and among the top 10 globally.
- IVORY framework.
  - European Union's Horizon Europe research and innovation programme Marie Skłodowska-Curie Industrial Doctorates (grant No 101119590).
  - It develops fair and explainable **Artificial Intelligence (AI)** to analyze driver behavior, predict crashes, and enhance road safety while sharing knowledge.
  - DC9 focuses on creating an AI framework to analyze **road safety KPIs**, predicts crashes, and evaluates the **scalability of models** primarily across spatial.
- Traditional crash prediction relies on econometrics; now enhanced by **Machine Learning (ML)** & **Deep Learning (DL)**, with **Graph Neural Networks (GNNs)** extend DL to graph-structured data.

European Commission
Marie Skłodowska-Curie Actions



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

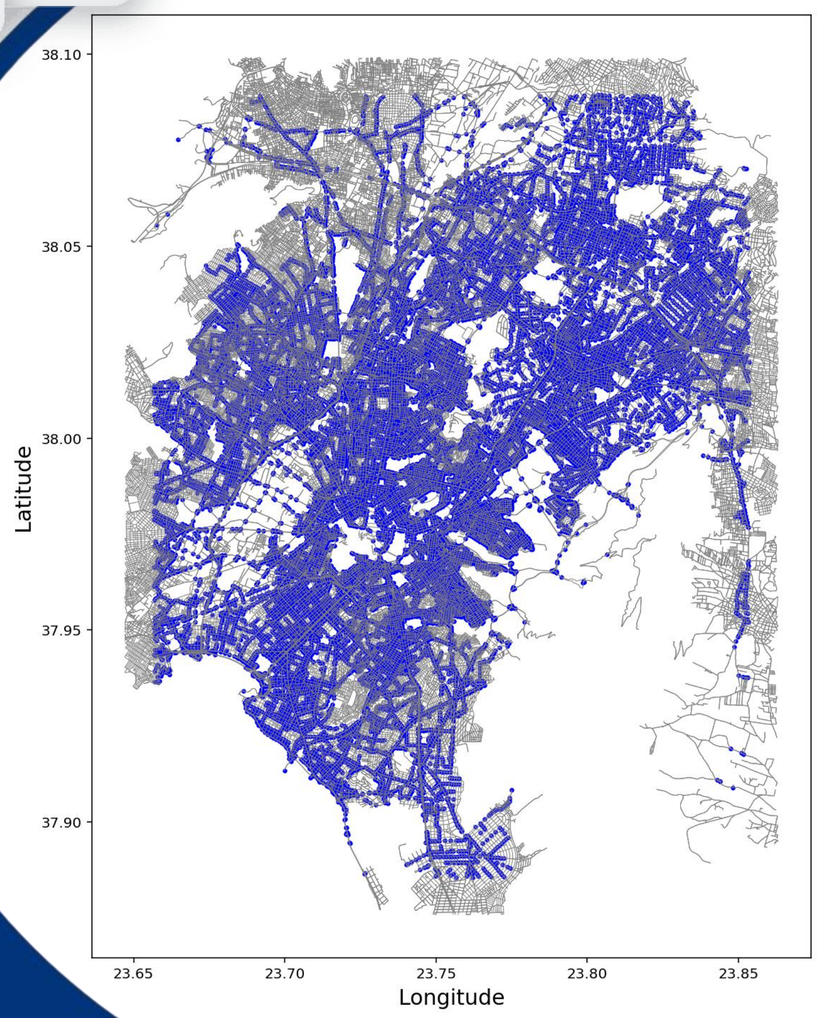
IVORY oseven

Data Sources

- OSeven Telematics provided **telematics data** collected via smartphone hardware sensors, anonymized and compliant with Greek and European personal **data protection** regulations (GDPR).
    - Raw data are processed by **proprietary machine learning** algorithms.
    - Reliability is validated against OBD data, on-road tests, simulators, and literature benchmarks.
    - Selected **features** from the provided **preprocessed dataset**: geographic coordinates, smoothened speed, and binary flags for hazardous behaviors.
- OpenStreetMap is a free, editable **global map** created by volunteers and released under an open-content license.

GDPR COMPLIANT

OpenStreetMap



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

IVORY oseven

Telematics Aggregation

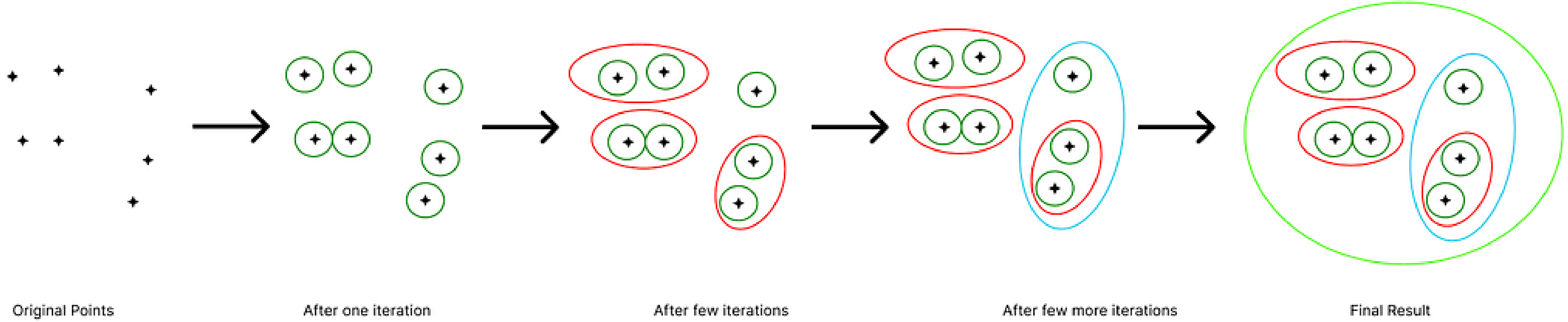
- ➢ Based on the telematics data, a coordinate bounding box was defined and used to extract a structured graph from OpenStreetMap via the OSMnx Python library.
- ➢ From the graph, node and edge features were stored in two different datasets.
- ➢ Telematics features were aggregated to OSM nodes using summation or averaging within a 50-meter buffer, coherently with existing literature.
- ➢ Each raw telematics point was matched to its closest edge, and features were aggregated per edge.

Features	Description
Street_Count	Number of streets connected to the intersection
SmoothenedSpeed	Average speed of vehicles near the node
SpeedingFlag	Count of speeding events near the node
Mobile_usage	Number of instances of phone usage near the node
Harsh_acc	Number of harsh acceleration events near the node
Harsh_brk	Number of harsh braking events near the node
Event_intensity	Average intensity of harsh events near the node
Trips_count	Number of trips recorded near the node

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

Agglomerative Hierarchical Clustering

- Clustering is an unsupervised machine learning technique to group similar items.
- An Agglomerative Hierarchical Clustering is an approach aiming to find a hierarchy of data point groups.
  - The method builds on a criterion to determine the distance between clusters, called Linkage Method, which has an impact on the shape and size of the clusters.
  - The Cophenetic Correlation Coefficient (CCC) is used to assess the clustering performance, defining how well the algorithm preserves the pairwise distances of the original data points.



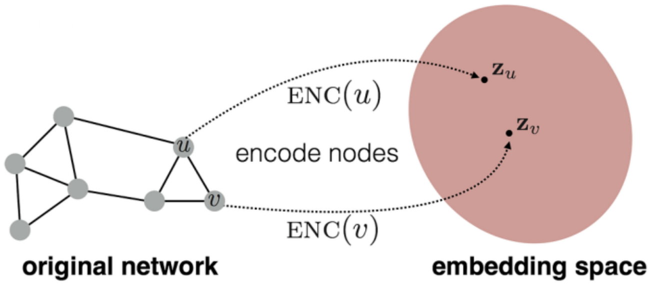
  - The output is a dendrogram displaying the hierarchical relationship between different sets of data.

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

IVORY oseven

Introduction to GNNs

- GNNs learn compact **vector representations** of nodes that capture structural roles, **neighborhood context**, and node features.
- They encode graph-structured data by leveraging **topological relationships** rather than flattening the graph into vectors.



- Advancements include **Graph Convolutional Networks** (GCN) using convolution operations, and **Graph Attention Networks** (GAT) applying **attention mechanisms**.
- A simple neural network with two GAT layers was defined.
  - Leveraging **attention coefficients** to weigh neighbors differently and to incorporate both **edge features** and **neighboring nodes**.
  - Using a **multi-head attention** to stabilize training and improve accuracy.
  - Trained within a self-supervised framework, using a self-designed **contrastive loss function** inspired by literature in this field.

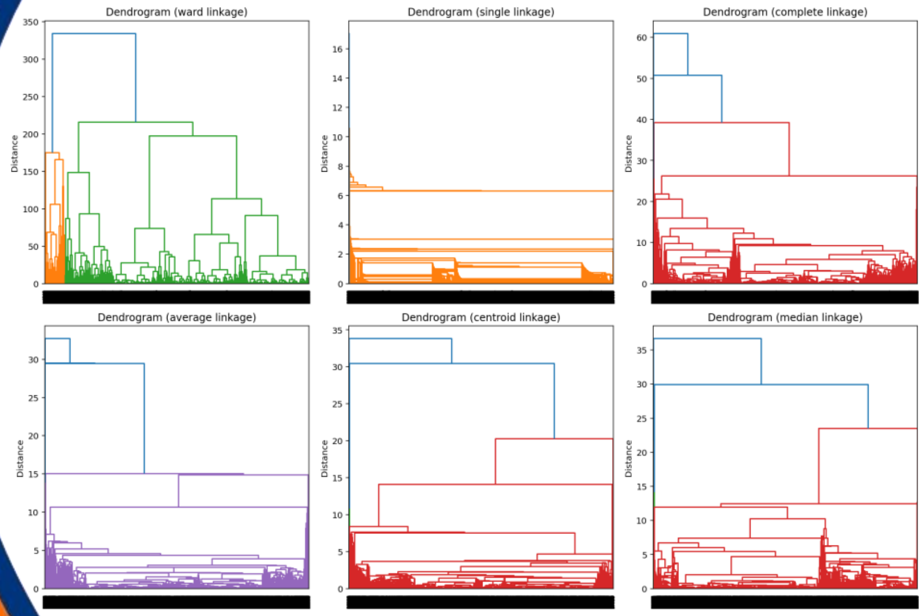
This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101119590

IVORY oseven

Dendrogram Analysis of Raw Features

- A **dendrogram** displays the step-by-step grouping of data points, where the **height** of each merge shows how far apart clusters are when they merge.
  - At each merge, if the **linkage height ≤ threshold**, the clusters are colored the same.
  - In SciPy, the default color threshold is **70% of the maximum merge height**.
  - If most of the dendrogram is **one color**, the data forms one large cluster with a few smaller or outlier groups standing apart at that level.

Linkage	CCC
Ward	0.616
Single	0.652
Complete	0.682
Average	0.931
Centroid	0.916
Median	0.499



- **Ward's** method achieves a **moderate CCC** and more clearly **separates** dense data points.
  - While it sacrifices **exact pairwise distances** to form clusters, different stem lengths in the dendrogram still show **meaningful separation**.

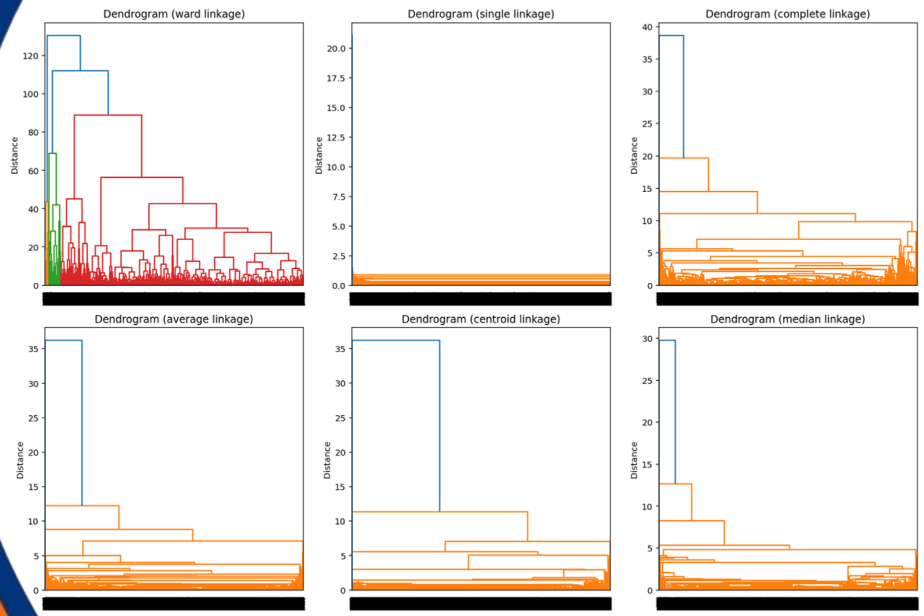
This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101119590

IVORY oseven

Dendrogram Analysis of Embeddings

- Most dendrograms show one color, meaning the data forms a single large, dense cluster at low linkage heights.
  - The single method exhibits an emphasized chaining effect, indicating a continuous structure in the data, expected from the embedding process, further highlighted by high CCC.
- Ward's method achieves a good CCC and produces clearer cluster separation.
  - One large cluster with two smaller groups emerge.
- The Ward's method using embeddings (instead of raw features) achieved a higher CCC.
  - For analysis, cluster labels were mapped back to the raw features and averaged.

Linkage	CCC
Ward	0.690
Single	0.792
Complete	0.746
Average	0.930
Centroid	0.925
Median	0.553



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

IVORY oseven

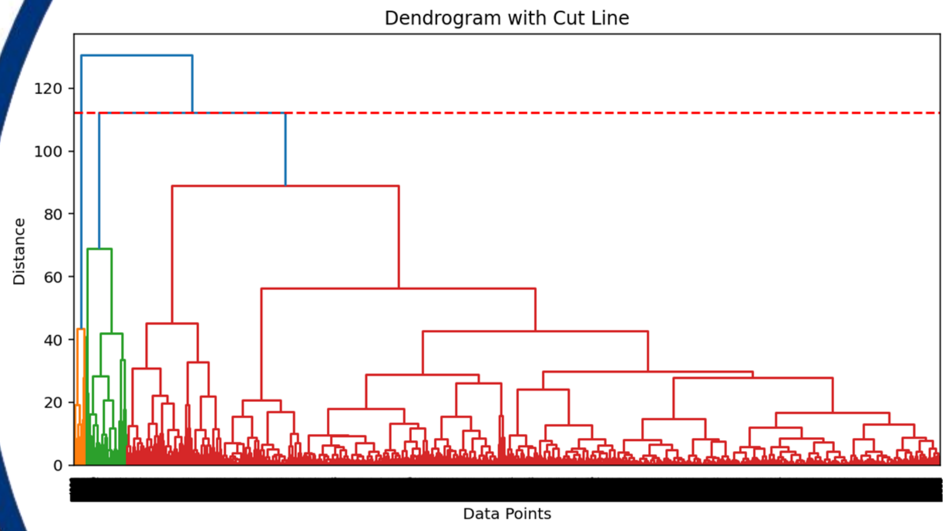
Discussion: Key Insights (1/2)

➢ The dendrogram encourages separation in 2 clusters:

i. One risky cluster displaying increased speeding events, mobile usage, and more frequent harsh events and trips.

ii. The other displays lower-risk traits, though with a comparable average speed.

Features	Cluster 1 Mean Values (31,409 nodes)	Cluster 2 Mean Values (467 nodes)
Street_Count	3.36	3.24
SmoothenedSpeed	27.08	31.26
SpeedingFlag	4.59	34.92
Mobile_usage	10.59	99.19
Harsh_acc	0.36	4.16
Harsh_brk	0.31	3.36
Event_intensity	0.37	1.37
Trips_count	25.53	230.16



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

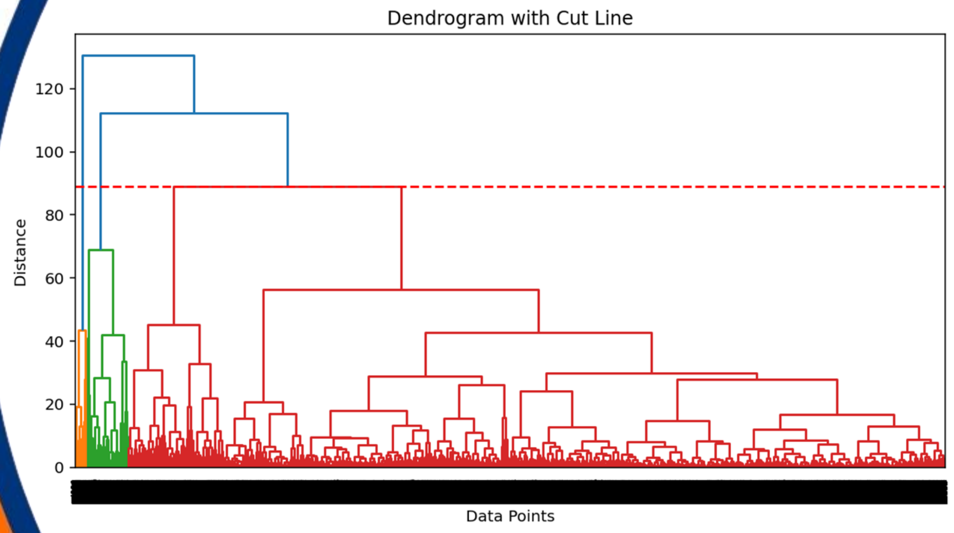
Discussion: Key Insights (2/2)

➢ Downwards, the number of clusters increased to 3, resulting in the previous safer cluster being split into two subgroups:

i. These new clusters show similar levels of harsh events and mobile usage, both lower than the original risky cluster.

ii. However, the smaller of the two has the highest average speed, making it risky in terms of speed, while the original risky cluster remains higher in harsh events and phone use.

Features	Cluster 1 Mean Values (29,937 nodes)	Cluster 2 Mean Values (1,472 nodes)	Cluster 3 Mean Values (467 nodes)
Street_Count	3.37	3.19	3.24
SmoothenedSpeed	25.99	49.33	31.26
SpeedingFlag	3.07	35.31	34.92
Mobile_usage	10.62	10.02	99.19
Harsh_acc	0.36	0.47	4.16
Harsh_brk	0.30	0.56	3.36
Event_intensity	0.36	0.62	1.37
Trips_count	24.55	45.5	230.16



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

IVORY oseven

Potential applications

➢ The work aims to provide node-based insights, informing on where to focus safety efforts and resources to improve overall traffic road safety.

- Risky cluster areas can be targeted for interventions to enhance road safety, such as infrastructure improvements, awareness campaigns, or enforcement measures.
- Insurers can use this clustering to define risk profiles by identifying patterns of risky or safe driving behavior, enabling insurers to offer more accurate, location-based pricing and targeted advice.
- The hierarchical structure enables finer granularity for more targeted interventions. Lower in the hierarchy, clusters may reflect different risk types, such as speed-related or harsh event and phone-related behavior.

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

IVORY oseven

Conclusions

- Graph-based representations improve the understanding of complex road safety data by leveraging node features, topology, and edge attributes to enhance clustering performance.
- The resulting hierarchy enables clearer cluster selection via dendrograms and supports scalable analysis for deeper road safety insights. Dividing the safer cluster reveals alternative risk perspectives, highlighting the value of multi-level clustering for capturing nuanced behaviors.
- Future directions may include testing different GNN architectures and linkage methods.
- Incorporating traffic, temporal, and contextual features, such as rural vs. urban classification, could enhance real-world relevance.
- However, hierarchical clustering remains computationally intensive, requiring significant resources, particularly with large-scale graphs.

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

Hierarchical Clustering on Graph Embeddings: A Scalable Approach to Risky Intersections

Simone Paradiso

IVORY

PhD Candidate and Researcher

oseven

Together with:
George Yannis & Apostolos Ziakopoulos

Department of Transportation Planning and Engineering
National Technical University of Athens

ROAD SAFETY RS5C ON FIVE CONTINENTS

20th International Road Safety on Five Continents (RS5C 2025)
3-5 September 2025, Leeds, UK

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