Road Network Analysis of Leeds, UK

[*]**Akshay Saraf**
K23039901
King's College London,
London, UK
k23039901@kcl.ac.uk

[*]**Ameesh Arya**
K23114893
King's College London,
London, UK
k23114893@kcl.ac.uk

[*]**Pattarin Urapevatcharewan**
K23036414
King's College London,
London, UK
k23036414@kcl.ac.uk

[*]**Yue Ma**
K23031223
King's College London,
London, UK
k23031223@kcl.ac.uk

ABSTRACT

This report undertakes a comprehensive examination of the road network and associated events in Leeds city centre, UK. Initially, we delve into the spatial characteristics of the road network, emphasizing efficiency and planarity. Subsequently, our investigation extends to the distribution of road accidents and their correlations across interconnected roads. Leveraging Voronoi diagrams, we propose marathon route planning strategies geared towards enhancing citizen participation. Finally, adopting the W3C PROV standard, we unravel event provenance, complemented by PageRank computation and embedding technique evaluation.

INTRODUCTION

The analysis in this report mainly has four tasks, including creating the road network in Leeds, visualising and analysing the point pattern of road accidents, finding 42 km marathon paths, and representing road events by the W3C PROV provenance data model. Methods used in this study mainly come from network analysis and spatial data analysis. Drawing upon datasets from OpenStreetMap and the UK government, our analysis delves into various facets of urban dynamics. The report intends to investigate and analyse spatial patterns using a comprehensive strategy that includes data collecting, pre-processing, network visualisation, analysis, and planarity assessment. The study also investigates the visualisation and analysis of accident patterns, using quadrat, heatmap, and network visualisation methods to collect and comprehend accident occurrences. It also looks into how Voronoi diagrams can be used to identify key regions of interest and find marathon paths, with a focus on the selection and visualisation of these diagrams. Furthermore, the paper also investigates advanced mechanisms such as TransE, PROV, and PageRank for provenance data modelling and network analysis, emphasising the significance of embedding approaches in understanding spatial behaviour and accident patterns. The results of these analyses are then discussed, shedding light on the spatial dynamics of transportation accidents and urban spatial behaviour. Through these endeavors, our study furnishes valuable insights into urban road dynamics, thus offering tangible contributions to the realms of urban planning and safety management.

RELATED WORK

Boeing and Geoff[1] introduced the current research status and limitations of street network analysis and explained the potential of "OSMnx" library to address these problems. The analysis carried out in the lab week 6 [8] created the road network of Soho, London via "OSMnx" and "Folium" libraries, and then checked the planarity and built a node Voronoi graph. Lab in week 8 [9] introduced a "Spaghetti" library for analysing point patterns on a spatial network. It investigated the point pattern of crime data in Soho, London by snapping the point data to the road network, and calculating K-function and Moran's I, similar analysis has been performed in this research.

The study by Büchel et al. [4] aligns closely with our task of planning marathon routes in urban environments. While their framework focuses on individual recreational running routes, the principles of personalized route recommendation and consideration of user preferences can inform our approach to organizing marathon routes that maximize citizen participation and satisfaction. Their approach utilizes user-generated data, such as historical running routes and preferences, along with environmental factors and safety considerations, to generate route recommendations.By leveraging similar methodologies and considerations, we can adapt their framework to the context of organizing simultaneous marathons in different parts of the city.

The tutorial authored by Wasit Limprasert [3] provided comprehensive guidance on the Data Preparation, Creation of

[*]All members contributed to this work equally.

CHI'20, April 25–30, 2020, Honolulu, HI, USA

ISBN 978-1-4503-6708-0/20/04…$15.00

DOI: https://doi.org/10.1145/3313831.XXXXXXX

the TriplesFactory, and the construction of knowledge graph embeddings using PyKEEN. PyKEEN is a Python library designed for training and evaluating knowledge graph embedding models, including TransE, RotatE, and GCN. The implementation in this work draws inspiration from the methodologies outlined in Lab 9 [10] and Lab 10 [7], which introduced the PyKEEN Pipeline along with its requisite parameters. Additionally, the PageRank algorithm, as introduced in Lab 10, is implemented using the NetworkX PageRank method, which is integrated seamlessly for analyzing the network's centrality.

METHODS

Approach for Spatial Networks Creation and Planarity

Data Collection and Pre-Processing

Road Traffic Accident (RTA) data for Leeds, UK from 2016 to 2019 will be loaded from the UK Government's website (https://data.gov.uk/dataset/6efe5505-941f-45bf-b576-4c1e09b579a1/road-traffic-incidents). The OpenStreetMap (OSM) data will be utilized to extract the road network of Leeds, UK. Data will be further filtered to solely include driving roadways, leaving out walking paths and private roads. This will be achieved by utilising the OSMnx library, which facilitates the extraction and analysis of road networks from OSM data. Coordinate Reference System (CRS) of the data will be converted from British National Grid (CRS 'EPSG:27700') to WGS 84 (CRS 'EPSG:4326') coordinate system, which is a global coordinate system used by the Global Positioning System (GPS). A GeometryArray of Point geometries will be created from the 'Grid Ref: Easting' and 'Grid Ref: Northing' columns along with the latitude and longitude coordinates, which will help in the interpretation and manipulation of spatial data according to its geographic context i.e. for Leeds, UK location with the help of GeoPandas library.

Leeds Road Transport Accident Area Selection

Based on over a period of four years worth of Leeds road transport accident data, a one-square-kilometer area around the city centre with the highest concentration of incidents will be chosen, ensuring that 300 or more accidents occurred across several years. The geographical coordinates of this area will be obtained through spatial analysis for future investigation.

Network Visualisation

The road network will be taken from OSM data for a specific area of Leeds as discussed in subsection 3.1.2, with a concentration on driving roads, and then will be visualised using red nodes (intersections). This network visualisation will help to analyse the structure and connection of the road network, which is critical for understanding road accident trends and assessing road safety in the area.

Network Analysis and Circuitry

Analysing a spatial network involves looking at several aspects to understand its structure and functioning. The number of nodes and edges, which are critical for comprehending the network's size and connectedness, will be investigated. Densities such as node, edge, intersection, and street density will reveal information about the distribution and concentration of elements in the network. Additionally, lengths will be examined, including total and average edge and street lengths, to better comprehend the scale and dispersion of the network. The spatial diameter and average circuity are indicators of the network's reach and intricacy, reflecting how far apart the farthest nodes are and the average path length between them. The average circuitry of the network will be estimated in order to evaluate the efficiency of using roads in the selected location. These evaluations, together with others, will provide a full perspective of the network's structure, allowing for insights into its road transport accidents in Leeds.

Network Planarity

The planarity of the network will be assessed using OSMnx's planarity metrics. Planarity is the geometric property of a graph in which all of its subgraphs are planar. The examination will consist of recognising and analysing non-planar subgraphs, as well as providing examples from Leeds roads networks to support the conclusion.

Visualization and Analysis of Accident Patterns

Quadrat, Heatmap and Network with Accidents Captured

Visualising the spatial distribution of road accidents has 2 main aspects, which are visualising based on the whole study region and visualising based on the road network. For the first aspect, quadrat and heatmap are good methods to reflect the spatial cluster of accidents and detect the hot spots. Quadrats can be seen as windows splitting the whole area into small cells and figures in each cell represent the number of accidents within the cell. Moreover, Quadrat is an elementary basis for generating a heatmap, which uses different colour bands to represent the strength and pattern of the data.

On the other hand, road accidents are prone to have a spatial pattern on the road network. Spaghetti library in Python can help to capture the accident points to the nearest position of the road network based on the ntw.snapobservations() method, and then conduct the study of road network analysis.

K-function and Moran'I

K-function and Moran's I are methods in spatial data analysis. K-function can assess whether the accidents are more or less clustered compared with a completely random distribution (CSR). Moran's I is a kind of index for detecting spatial autocorrelation and the direction, we can check the p-value of Moran's I to detect the significance of spatial autocorrelation and check the direction by the sign of the Moran's I value. Therefore, We can use the K-function and Moran's I to analyse the correlation between a high number of accidents on one road and a high number on connecting roads. If the value of Moran's I is positive and the p-value is less than 0.05, positive spatial autocorrelation exists. The higher the value, the stronger the spatial autocorrelation. On the contrary, a negative value indicates negative spatial autocorrelation. If the value is close to 0 and the p-value is greater than 0.05, it means that we can not reject the null hypothesis that the point data follow spatial randomness.

When we investigate whether accidents happen nearer to intersections or partway along roads, we can calculate the distance of every accident to its nearest intersection, and then calculate the ratio of the distance and the road length. If the ratio of most accidents is concentrated around 0, road accidents are

likely to happen near the intersection. On the contrary, if most of the ratios fall in the interval around 0.5, road accidents are likely to happen along the roads. Spaghetti API with the method of dist_to_vertex is suitable to study the question. And KDE (Kernel Density Estimation) is a good method to assess and visualise the probability density, and we can observe the concentration pattern of the ratios intuitively.

Voronoi Diagrams and Finding Marathon Paths

The first step in organizing parallel marathons across the city of Leeds involves strategically selecting the initial set of seed points. These seed points serve as the starting locations for dividing the city into various areas, ensuring accessibility and participation for citizens. The selection process incorporates multiple criteria to optimize the overall organization of the marathons.

Criteria used for Seeds Selection

1. **Even Distribution:** One crucial aspect is the even distribution of the seed points across the city. This ensures that each marathon is geographically dispersed, allowing citizens from different neighborhoods to participate conveniently.
2. **Accident Prone Areas:** Prioritizing safety, the selection process considers areas with the least history of accidents. By avoiding regions with frequent accidents, we aim to minimize potential risks for marathon participants.
3. **Proximity to Public Transportation:** Accessibility is key to maximizing citizen participation. Seed points are strategically chosen to be near public transportation hubs, facilitating easy access for participants from various parts of the city.

To implement the selection process, the following steps are undertaken:

- **Accident Data Analysis:** Utilizing historical accident data, areas with the least occurrences of accidents are identified using geographic information systems (GIS) techniques.
- **Public Transportation Analysis:** Public transportation networks are analyzed to identify central locations with easy access to bus stops, train stations, or other transit options.
- **Spatial Distribution Analysis:** Employing spatial analysis techniques, the city is divided into evenly distributed areas.
- **Multi-Criteria Evaluation:** A multi-criteria evaluation approach is adopted to weigh and combine the above criteria, ensuring a balanced selection of seed points that align with the overarching objectives of safety, accessibility, and inclusivity.

Visualization of Voronoi Diagram

To visualize the cells yielded by the selection of seed points, we employ a Voronoi diagram. The Voronoi diagram partitions the space into regions based on the proximity to the seed points, providing a clear representation of the areas allocated to each marathon.

The decision to use a node network Voronoi diagram is based on its alignment with the problem requirements and the nature of the seed points. Since the seed points represent specific locations in the city, and the objective is to allocate marathon cells around these points, a node network Voronoi diagram provides a clear and intuitive representation of the resulting regions. Additionally, this type of diagram facilitates easy interpretation and analysis of the spatial distribution of marathon cells, making it the most suitable choice for this problem.

Finding Marathon Paths

We explore different strategies to find paths for marathon cells that are exactly 42 km long and finish at the starting point.

We initially selected seed points by dividing Leeds into 10 equal areas and choosing locations within each area with the least accidents. This approach aimed to balance safety and accessibility for marathon participants.

Options for Increasing Path Availability

To address the challenge of finding suitable paths for all cells, we propose the following alternative approach:

- **Increase Seed Points:** Adding more seed points across the city increases the coverage area and allows for a wider range of potential paths to be explored within each cell.
- **Grid-based Exploration:** We adopt a grid-based approach to systematically explore the surrounding area of each seed point. By dividing the search area into smaller grid cells, we can methodically examine each cell to identify paths that fulfill the length and return-to-start requirements.
- **Explore Alternative Criteria:** We considered using Centroids of evenly distributed areas and proximity to public transportation for seeds selection. Exploring alternative criteria can provide new insights into suitable locations for marathon cells.
- **Combining Criteria:** we combined the first approach (evenly distributed and accident-prone areas) with the proximity to public transportation criterion. This hybrid approach aimed to increase the number of seed points while still prioritizing safety and accessibility.

TransE, PROV, and PageRank Mechanisms

Provenance Data Model

To represent the provenance of important events in the road network of Leeds, we employed the PROV Python library and defined namespaces for entities, activities, and agents involved in of events such as accidents, marathons, road maintenance, protests, and public transport strikes. Entities were defined to represent various aspects of each event, for example accident location, marathon route, and protest location and the public transport strike hub. Activities were defined to represent actions associated with each event, such as accident investigation, marathon event management, road closure planning and strike response. Agents were defined to represent individuals or organizations involved in each event, such as a police officer, marathon runner, protestor and the public transport strike organiser. We specified relationships between entities, activities, and agents using PROV concepts such as wasGeneratedBy, used, wasAttributedTo, wasAssociatedWith,

and these relationships capture the flow of information, dependencies, attributions, and connections between different elements involved in the events.

Page Rank - Provenance Network

In general, PageRank is used to identify the most important and central nodes where nodes with higher PageRank values are considered more important or influential within the network. Where as in a provenance network, PageRank can be used to find the most important events, agents that would be the most influential and entities that might be most frequently used. The NetworkX library was used to compute the PageRank value for all the nodes of the Provenance Network.

TransE, RotatE and GCN Embeddings

For this task, we trained and evaluated three different types of embeddings: TransE, RotatE, and GCN (Graph Convolutional Network) embeddings. TransE is a translational embedding model that represents entities and relations in a low-dimensional space while RotatE is an extension of TransE that introduces rotational transformations to capture symmetric relations in the embedding space. GCN on the other hand utilizes graph convolutional neural networks to learn embeddings directly from the graph structure. These embeddings were trained using the CoDExMedium dataset, which contains knowledge graph triples, and then evaluated on the provenance network representing important events in the road network of Leeds. The goal was to assess how well these embeddings capture the underlying structure and relationships within the provenance network, thereby providing insights into the significance and influence of different entities, activities, and agents involved in road network events. Mean Reciprocal Rank (MRR) and Hits@10 were used to evaluate the performance of the embeddings.

RESULTS

Analysing Spatial Networks and Planarity

Leeds RTA Data Collection and Pre-Processing

Four CSV files containing road transport accidents in Leeds, a comprehensive dataset including 8,654 accidents in Leeds spanning the years 2016 to 2019, were imported. By including columns for longitude, latitude, and geometry in the Leeds RTA dataframe and conversion of the Coordinate Reference System (CRS) from British National Grid (EPSG:27700) to WGS 84 (EPSG:4326), the data's compatibility with global mapping systems was ensured. The utilisation of the GeoPandas library enabled the precise analysis and manipulation of spatial data in Leeds, UK. The absence of null values in the longitude and latitude columns served as evidence of the data's integrity and completeness, providing a solid foundation for future analysis and area selection.

Leeds Road Transport Accident Area Selection

The Leeds road transport accident area selection highlights the significance of identifying a specific geographical area within Leeds city center, characterized by a high concentration of road transport accidents over a period of four years. The coordinates (-1.5331415549020968, 53.801679858701405) were identified through an examination of accident data as the location with the highest number of accidents (13), surpassing all others and demonstrating a notable concentration of incidents; thus, they were designated as the focal point for subsequent analysis. The creation of a polygon around this central point, encompassing a 950 m^2 area, facilitates a focused study of accident patterns within this high-risk zone.

Figure 1. Selected area ploygon

The boundary of this polygon, defined by its coordinates (-1.547783, 53.7931442, -1.5186735, 53.8102162) is shown in Figure 1, was established to encapsulate the area with the highest number of accidents from 2016 to 2019, totaling 706 incidents as shown in Figure 2. The chosen coordinates and the polygon boundary serve as a foundation for focused analysis which allowed for a more in-depth examination of the factors contributing to the high number of accidents in this specific area, potentially revealing patterns and trends that could inform targeted interventions to improve road safety.

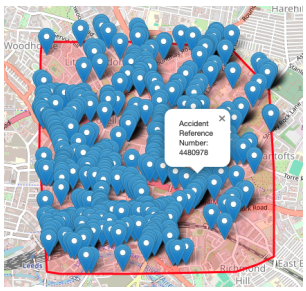

Figure 2. Leeds road transport accident spots in selected area

Leeds RTA Network Analysis, Circuitry, and Visualization

Leeds road transport accident network visualisation is shown in Figure 3. The Leeds RTA network has 735 nodes and 1,427 edges, representing a dense network of intersections and an intricate network of roadways, respectively. The number of nodes and edges is large, which presents both a strength and a difficulty in terms of traffic control and accident risk. The average degree refers to the average number of connections (edges) that each node possesses. The value 3.882993 is high, indicating a tightly connected network, which may result in increased traffic volumes and, potentially, more accidents. Total Edge Length (94,912.381 m) and Average Edge Length (66.51183 m) provide information on the total and average lengths of roads in the network, assisting in determining the network's scale and the possible influence of road conditions on traffic flow and safety. These numbers are high, showing Leeds' substantial road network, which is typical for a city of its size and importance.

Figure 3. Leeds road transport accident selected area network

The **spatial diameter** is the maximum shortest path distance between any two network nodes and is a measure of the network's total connectedness. A spatial diameter of 49 m is a low value, indicating a compact network that can result in more

efficient, manageable, and safer transport services. This is because compact networks reduce the distance between nodes, lowering transit times and distances, simplifying road management, and allowing for faster reaction to emergencies. The total and **average street lengths** provide information about the total and average lengths of streets in the network, which can be useful for planning and safety analysis. The values 64,369.607 m and 67.40273 m suggest a large total length of streets and a relatively long average street length, which is expected for a city like Leeds with a dense road network. **Node density** (213.125459 km²) indicates the number of intersections per unit area, suggesting a high concentration of points where roads intersect. **Intersection density** (168.180635 km²) further highlights the density of these intersections, indicating a complex network of road connections. **Edge density** (27,521.421501 km²) and **street density** (18,665.036821 km²) provide insights into the total length of roads and streets per unit area, highlighting Leeds' substantial transportation network. These high density indicate a more developed and robust transport infrastructure but it can also suggest potential issues such as larger traffic volumes and an increased chance of accidents due to the network's complexity and density, as demonstrated by the 706 accidents from over a period of four years. The above discussed metrics and other metrics along with their respective values are shown in Table 1 [5] [6].

Characteristics	Values
Nodes	735
Edges	1427
Average Degree	3.882993
Total Edge Length (m)	94912.381
Average Edge Length (m)	66.51183
Average Streets per Node	2.67483
Streets per Node Counts	0: 0, 1: 155, 2: 23, 3: 471, 4: 78, 5: 8
Streets per Node Proportions	0: 0.0, 1: 0.21088, 2: 0.03129, 3: 0.64081, 4: 0.10612, 5: 0.01088
Intersection Count	580
Total Street Length (m)	64369.607
Street Segment Count	955
Average Street Length (m)	67.40273
Average Circuity	1.04409
Self Loop Proportion	0.001047
Node Density (km²)	213.125459
Intersection Density (km²)	168.180635
Edge Density (km²)	27521.421501
Street Density (km²)	18665.036821
Number of Connected Components	1
Number of Nodes in Connected Components Subgraph	735
Number of Strongly Connected Components	31
Spatial Diameter (m)	49
Number of Nodes in Largest Strongly Connected Component	664

Table 1. Leeds Road Transport Accident Network Characteristics

The **average circuity** value of 1.04409 in the Leeds road transport network shows a fairly efficient network, with the shortest network distance about matching the Euclidean distance between origin-destination pairs. Circuity is a measure of a transportation network's efficiency, with 1 indicating the most efficient network. An average circuity of 1.2 is common [2], implying that Leeds' network is more efficient than the average. This efficiency is critical for analysing a transport network's overall performance since it represents the network's ability to reduce travel distance and time. Cities with lower average circuity values may have more direct routes to main sites of interest, resulting in shorter travel times and potentially reduced traffic congestion. Cities with higher average circuity values, on the other hand, may have more complicated road networks with more intersections and turns, resulting in longer travel times and increased traffic congestion. The circuity value can also be used to determine travel distances, showing that Leeds' road network is reasonably efficient in this respect as well.

Leeds RTA Network Planarity

The planarity of a road transport network is how closely it resembles a flat plane. In the context of Leeds, a city recognised for its complicated topography and the presence of elevated constructions such as bridges and underpasses that connect various areas of the city, the network is not planar, which means it does not lie flat, but rather has a three-dimensional layout. Planarity is an important aspect in transportation network analysis since it affects traffic flow, safety, and overall network efficiency. Presence of Kuratowski subgraphs as shown in Figure 4 also confirms that the selected area is non planar with edges intersecting not only at their ends, but also edges cross one another. Figure 5 and Figure 6 shows an example of underpass and a bridge on A64(M) in Leeds inner ring road motorway on the selected area in subsection 4.1.2, further confirming selected Leeds area is non planar.

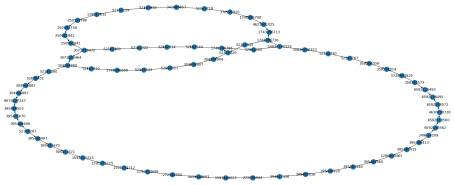

Figure 4. Kuratowski subgraphs of Leeds RTA selected area network

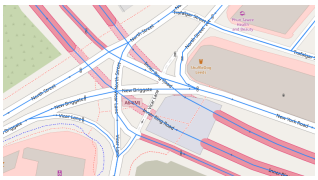

Figure 5. Map of underpass and bridges in Leeds selected area road network on A64(M)

Figure 6. Example underpass and bridge on A64(M) in Leeds selected area road network

Visualization and Analysis of Accident Patterns

Quadrat and Heatmap

Through the quadrat counts on 7, we know most of the accidents in the study region happened in southwest part. Then we convert to the heatmap8 of road accidents in this study region, we can intuitively see the red area mainly concentrate around southwest corner, which aligns with the results of quadrat counts. The roads of the red areas seems denser and main intersections are included.

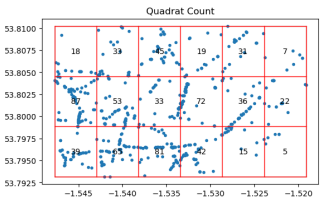

Figure 7. Qudrat Counts of Road accidents

Through visualising road accidents on spaghetti road network, we can compare the observed accidents 9 and the snapped accidents 10 on road network of the study region. From the observed accidents 9, there are some accident points occur on the blank areas but not on the network. This actually make sense that when convert to the heatmap8, we can see these accidents are indeed happened on the road. This is because the road network we extracted is the "driving" network and some roads only for walking are not included. Then we snap the accidents on the driving network 10, we can see accidents are happened almost every segment of the road.

Figure 8. Heatmap and the Road Network of the Study Region

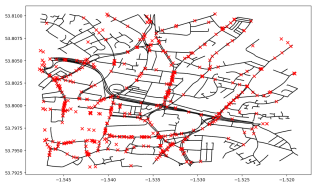

Figure 9. Observed Accidents of Spaghetti Network

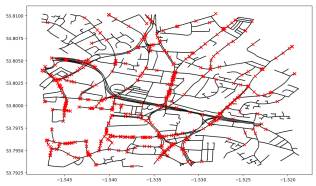

Figure 10. Snapped Accidents of Spaghetti Network

Spatial Characteristics of Road Accidents

After calculating the K-function, we can then plot the values and confidence intervals 11. At all distances, the line of the observed values is on the above of the lines of confidence intervals, which indicates that the accident points are more clustered than would be expected. Then we look at the plot Moran's I 13, the value of Moran's I is 0.113 represented by the red vertical line, while the expected Moran's I is -0.00026 represented by the black line. Most of the figures are concentrated around the expected Moran's I, and the p-value is less than 0.05. The results suggest that there is significant evidence to support that spatial autocorrelation exists, and road accidents are not randomly distributed.

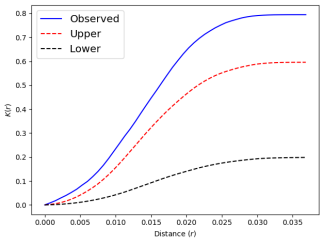

Figure 11. K-function

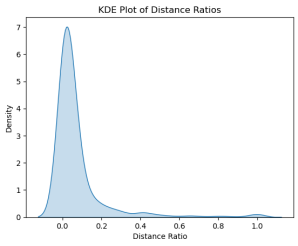

Figure 12. KDE (Distance Ratios)

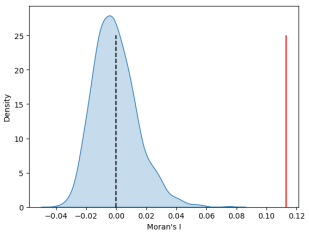

Figure 13. Moran's I

The average distance of accident points to their nearest intersections is really short, and most of the figures are concentrated around 0 in the KDE plot 12. These results indicate that accidents are more likely to happen near the intersection rather than along the road. However, there are some values of ratio greater than 0.05, this may be because the edge of the study region cut off some roads and these edges only connect to one node.

Evaluating Marathon Paths

Voronoi diagram

Initially, we selected 10 seed points distributed across Leeds based on criteria emphasizing areas with low accident rates and geographic diversity. These seed points were intended to serve as the basis for generating the Voronoi diagram. Using

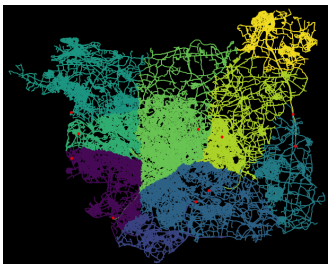

Figure 14. Voronoi Diagram

the selected seed points, we generated a Voronoi diagram, as in Figure 14, to partition the city into regions based on proximity to each seed point. The resulting Voronoi cells represented the areas allocated for organising marathon routes.

The Voronoi cells effectively covered the entire city and exhibited good connectivity, ensuring that each area was accessible for organizing marathons.

Marathon Paths result

1. Less Accident Prone and Evenly Distributed Criteria
    - No path meeting the stringent 42 km length requirement was found, but six paths within a 150-meter tolerance were identified, offering viable marathon route options.
2. Increase in Seed Number in the first Criteria
    - Only one path within the tolerance range was located despite increasing the seed count from 10 to 50.
3. Centroids of Evenly Distributed Areas Criteria
    - No path of exact 42 km length was found, yet two paths within the tolerance range emerged.
4. Proximity to Public Transportation Criteria
    - No exact-length paths were discovered, but three within-tolerance paths were identified, enhancing accessibility.
5. Increase in Seed Number for Public Transportation Criteria
    - No exact-length paths were discovered, though four tolerance-range paths were found with 20 seeds.
6. Combining the first and the fifth Criteria

- Combining the criteria resulted in identifying areas suitable for organizing ten simultaneous marathon paths within tolerance, providing a plethora of viable options for marathon routes.

The visualization as shown in Figure 15 depicts both the cells and the found paths for organising 10 simultaneous marathon events. It provides a comprehensive overview of the spatial distribution and organization of marathon routes within the city. The cells, represented in 10 colors, delineate distinct

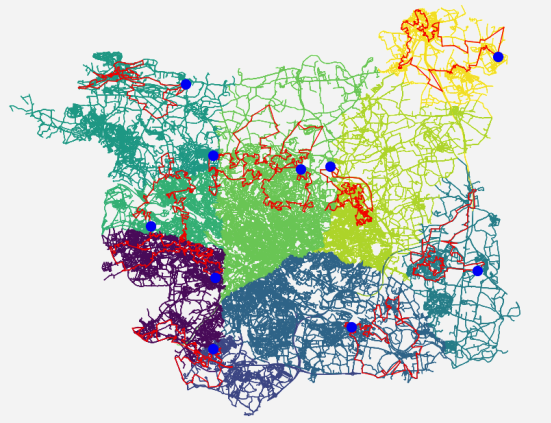

Figure 15. 10 simultaneous marathon paths

areas across the urban landscape, each serving as a potential location for marathon events. These cells are strategically positioned to ensure geographic diversity and accessibility, with careful consideration given to factors such as proximity to public transportation and low accident rates.

Overlaying the cells are the identified paths, depicted in red, which represent viable marathon routes within each designated area. These paths, meticulously plotted based on a combination of criteria including even distribution, safety, and accessibility, offer a multitude of options for organizing marathon events simultaneously across various parts of the city. Additionally, each blue marker denotes the start and end nodes of the marathon paths, serving as pivotal points for participants' journey through the designated areas.

Together, the visualization of cells and found paths provides valuable insights into the planning and execution of marathon events, facilitating informed decision-making to optimize participant engagement, safety, and overall success of the initiatives.

Exploring TransE, PROV, and PageRank

Provenance Data Model

The provenance model created is illustrated in the Figure 16 and it represents the events on the Leeds Road Network. using the W3C PROV provenance data model standard. It provides a visual depiction of entities, activities, and agents involved in various events such as accidents, marathons, road maintenance, protests, and public transport strikes. The relationships between these elements capture the flow of information, dependencies, attributions, associations, and derivations within and across different events.

Figure 16. Leeds - Provenance Data Model

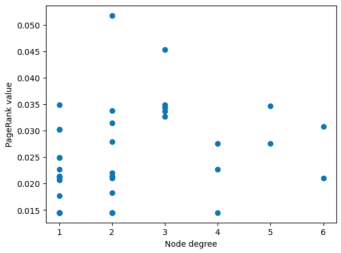

Figure 17. PageRank of the Provenance Network - Scatter Plot

Page Rank - Provenance Network

The computed PageRank values for all the nodes of the provenance network are visualized using a scatter plot as seen in Figure 17, where the x-axis represents the node degree (number of connections) and the y-axis represents the PageRank value. Additionally, a color-coded network diagram is generated as seen in Figure 18, where nodes are colored based on their PageRank values, with a color bar indicating the range of values. From the plots, it is identified that the most important nodes are ProtestOrganizer (Agent), RoutePlanning (Activity), StrikePlanning (Activity). It makes sense because these three are the most important parts of the 3 of the 5 events we have analysed.

TransE, RotatE and GCN Embeddings

After training the TransE embedding with CodexMedium Dataset and testing it on the provenance network, the Mean Reciprocal Rank (MRR) was 0.0005 and Hits@10 was 0,0 after hyperparameter tuning. 50 Embedding Dimensions with a learning rate

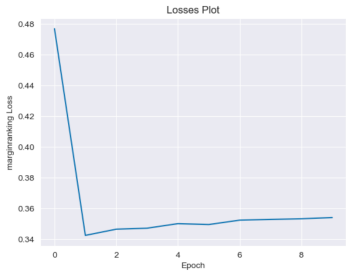

Figure 19. Losses plot for TransE embedding during training on CodeXMedium dataset

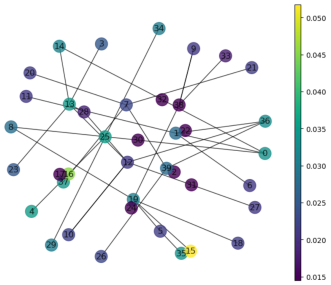

Figure 18. PageRank of the Provenance Network - Network Diagram

of 0.01 gave the best model. Figure 19 shows the losses visualized during training.

Similarly for the RotatE embedding, after hyperparameter tuning, MRR was 0.002, while Hits@10 was 0.013. The best parameters were 75 Embedding Dimensions with a Learning Rate of 0.01.

GCN embeddings were trained and evaluated using the same approach as TransE and RotatE. MRR was found to be 0.0003, while Hits@10 was 0.0 after Hyperparameter Tuning. The best parameters were 100 embedding dimensions and 0.01 learning rate.

DISCUSSION

The examination of the Leeds road transport accident network provides valuable insights into the network's characteristics, efficiency, circuity, and planarity. The network's compact spatial diameter and high average street length indicate that it is well-suited for efficient transport services, which could reduce travel times and distances. However, the high node, intersection, and edge densities suggest a complex network that may lead to higher traffic flow and accident risks. The average circuity value of 1.04409 indicates that the network is relatively efficient, with the shortest network distance closely matching the Euclidean distance between origin-destination pairs. This efficiency is crucial for minimising travel distance and time, which may be particularly especially useful in a city like Leeds with a dense road network. The network's non-planarity, as demonstrated by the presence of Kuratowski subgraphs and presence of underpass and bridges constructions, suggests that the network's complexity could contribute to increased accident risks. This complexity could also have an influence on the network's efficiency, while these structures can reduce travel times by providing direct routes, they may also lead to longer routes in some cases as the need for longer routes to navigate around underpass and bridges which could increase travel times and distances thus further impacting network efficiency and accident risks.

Based on the study region and driving road network, accidents are clustered in the southwest part of the area. Transportation hubs and intersections are more prone to have road accidents, and Burmantofts Street, Regent Street and York Street are the most accident-prone road sections. According to the analysis above, we can conclude that road accidents are not randomly distributed because there is significant spatial Autocorrelation exists. Moreover, road accidents are clustered denser than expected from the results of the K-function, which means that a high number of accidents may be correlated with a high number of connecting roads. However, there are other factors affecting the spatial patterns of road accidents. For example, two-way lanes and parallel lanes are more prone to road accidents. In addition, the design of traffic lights and the travel habits of the citizens are also influential factors.

We reflect on the outcomes of the evaluation of marathon paths using various criteria. The findings reveal both the strengths and limitations of each criterion in identifying viable routes for marathon events. While none of the individual criteria met the stringent 42 km exact length requirement, the combined approach proved to be a more promising strategy, yielding ten simultaneous marathon paths within tolerance. This suggests that leveraging multiple criteria, such as geographic diversity, safety, and accessibility, can lead to more comprehensive and effective marathon route planning. Moving forward, there are several avenues for future research and improvement. Refining the criteria used for selecting marathon paths could enhance the accuracy and reliability of the results. Additionally, optimizing the path-finding algorithms and incorporating additional factors, such as terrain elevation and participant demographics, could further improve the selection process.

The use of the W3C PROV provenance data model standard provides a structured means of documenting key events in Leeds' road network. By delineating entities, activities, and agents involved in each event and articulating their relationships, this model enables a comprehensive understanding of event dynamics and their ramifications. It facilitates event lifecycle tracking, cause-and-effect analysis, pattern recognition, and informed decision-making for road network management. Through PageRank analysis, central entities such as ProtestOrganizer, RoutePlanning, and Strike-Planning were identified, informing resource allocation and decision prioritization. Training and evaluating TransE, RotatE, and GCN models on the CodexMedium dataset revealed varied performance. Both TransE and RotatE embeddings exhibited moderate improvements post-hyperparameter tuning, with RotatE marginally outperforming TransE in MRR and Hits@10. However, their effectiveness in capturing the nuanced relationships within the provenance network remains limited, as evidenced by their relatively low performance metrics. Despite their potential for knowledge representation, the embeddings' ability to encapsulate the complexities of the provenance network is constrained. Future research endeavors may focus on refining embedding models and incorporating additional contextual information to enhance the modeling of Leeds' road network dynamics.

CONCLUSION

In conclusion, this paper provides a complete examination of the road network and associated events in Leeds city centre, UK, with a focus on spatial characteristics, accident distribution, marathon route planning, and event provenance. The analysis discovered that the road network, while highly efficient and planar in terms of spatial diameter and average street length, is complex due to high node, intersection, and edge densities, which may enhance accident risk. Accident patterns exhibited strong spatial autocorrelation, with higher clustering in the southwest, notably near transportation hubs and intersections. The study of marathon pathways using Voronoi diagrams and multiple parameters revealed the possibility of increasing Leeds citizen engagement through efficient route planning. Furthermore, using the W3C PROV standard for event provenance in conjunction with PageRank calculation and embedding methodologies provided insights into event lifecycles and decision-making for road network management. These findings highlight the importance of spatial dynamics, accident patterns, and event provenance in urban planning and safety management.

REFERENCES

[1] Geoff Boeing. 2017. OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, environment and urban systems* 65 (2017), 126–139.

[2] Jie Huang and David M. Levinson. 2015. Circuity in urban transit networks. *Journal of Transport Geography* 48 (2015), 145–153. DOI:http://dx.doi.org/https://doi.org/10.1016/j.jtrangeo.2015.09.004

[3] Wasit Limprasert. 2023. Tutorial: Knowledge Graph Embedding with PyKEEN. Medium. (2023). TriplesFactory, PyKeen embeddings.

[4] Benedikt Loepp and Jürgen Ziegler. 2018. Recommending Running Routes: Framework and Demonstrator.

[5] Albert Merono Penuela. 2024a. Network Data Analysis: Topic 5: Spatial and Planar networks. Microsoft PowerPoint. (2024). Spatial and Planar networks.

[6] Albert Merono Penuela. 2024b. Network Data Analysis: Topic 6: Spatial networks in context. Microsoft PowerPoint. (2024). Spatial networks in context.

[7] Albert Merono Penuela. 2024c. Network Data Analysis: Week 10: W3C PROV Model and PageRank. Google colab. (2024). W3C PROV Model, PageRank.

[8] Albert Merono Penuela. 2024d. Network Data Analysis: Week 6: Spatial network. Google colab. (2024). Spatial network creating, Planarity, Voronoi.

[9] Albert Merono Penuela. 2024e. Network Data Analysis: Week 8: Spatial networks in context. Google colab. (2024). Spaghetti network, K-function.

[10] Albert Merono Penuela. 2024f. Network Data Analysis: Week 9: Knowledge Graph Embeddings. Google colab. (2024). Knowledge Graph Embeddings.

APPENDIX

GitHub Link: https://github.kcl.ac.uk/K23114893/NetworkX