Computing traffic accident high-risk locations using graph analytics

Iyke Maduako¹² · Elijah Ebinne¹ · Victus Uzodinma¹ · Chukwuma Okolie³ · Emmanuel Chiemelu¹

Abstract Several studies on traffic accident hot spots have adopted spatial statistics and Geographic Information Systems where spatial point patterns are modelled solely on spatial dependence without consideration of the temporal dependence of the events. This could lead to under-estimation or over-estimation of results because of the temporal aggregation of the events to an absolute time point. Furthermore, traffic accidents are usually considered as events occurring randomly in two-dimensional geographic space. However, traffic accidents are network-constrained events. In this study, we adopt the connectivity of graph on a network space approach that identifies accident high risk-locations based on space–time-varying connectivity between traffic accident events and the road network geometry. A simple but extensible traffic accident space time-varying graph (STVG) model is developed and implemented using traffic accident data from 2010 to 2015 for Brevard County in Florida, United States. Traffic accident high risk-locations were identified and ranked in space and time using time-dependent degree centrality and PageRank centrality graph metrics respectively through time-incremental graph queries. In the analysis, traffic accident patterns were discovered based on network graph analytics. Our findings offer a new and efficient approach for identifying, ranking and profiling accident-prone areas in space and time at different scales.

Keywords Road network · Traffic accident hot spots · Space–time-varying graph data model · Degree centrality · PageRank centrality

1 Introduction

Identification, ranking and profiling of traffic accident high-risk locations have been a major focus of traffic accident analysis by researchers, transport authorities and traffic accident departments [1–3]. These agencies are particularly interested in understanding how traffic accidents occur in terms of location (where) and time (when), and the pattern of occurrence over time and space. Several studies have utilized different methods based on spatial point pattern and Geographic Information Systems (GIS) to analyze traffic accident hot spots and hazards. These spatial statistical and GIS methods include, the Moran’s I [4], Getis-Ord [5], Kernel Density Estimation (KDE) [6], Full Bayes hierarchical model [7], Quasi-Poisson model [8] and recently Network KDE (NETKDE) [9]. However, the challenge is that these methods are based on spatial statistics where spatial point patterns are analyzed only based on the spatial constraints of the accident locations with no account of the temporal constraints of these events. This leads to either under or over-estimation of the analysis results based on the aggregation of the events to an absolute time point. Therefore, a method that analyses traffic accident hot spots (high risk-locations) from the perspective of connectivity in space and time between road network topology (streets and intersections) and traffic...
accident events is required to gain more insights on the spatio-temporal pattern of accident high risk-locations.

The last two decades have produced vast studies and literature on hot spots and spatial pattern analysis of traffic accidents. Traffic accidents analysis has mainly been studied based on spatial statistical methodologies. This literature review brings as a background, some of the spatial and statistical methods recently used to analyze traffic accidents. Local spatial statistics methods, the Moran’s I and Getis-Ord statistics [10, 11] have been popularly used while other studies have adopted methods such as Kernel Density Estimation (KDE) [12, 13], Network KDE (NETKDE) [14, 15], Quasi-Poisson model [8], Full Bayes hierarchical model [16–18], Neuro-Fuzzy approach [19, 20] and even Neural Networks [21–25].

Moran’s I (MI) is a local spatial statistical method that is used in traffic accident analysis to measure the spatial dependence of accident locations and can also be used to examine the density of their spatial pattern, and how dispersed or randomly distributed the cluster patterns are. The combination of MI and Getis-Ord statistic have been very effective in furthering the understanding of the processes that lead to spatial dependency [10]. A group of high index Getis-Ord values (z-score and p-values) represent hot spots while the low values are low incident areas. These methods have been used by [1, 26–31] for spatial pattern, hot spot detection and severity index of traffic accidents. KDE is one of the most used methods in traffic accident density estimation to examine first order attributes of spatial point dispersion/distribution patterns [32]. Its use has been showcased in the determination of traffic hot spots for example, in measuring spatial concentration and pattern of road accidents [33], wildlife and vehicle accident spatial analysis [34], traffic accident spatial cluster (hot spots) analysis [26], determination of pedestrian crash areas [35], detection of accident hot spots with combined KDE and Poisson function [26], cyclist’s traffic hazard density estimation [36], classification of road accidents [37], traffic accident spatial pattern analysis with combined KDE, nearest neighbor distance and K function [38]. While KDE has long been used in traffic accident analysis, some authors suggest that it has some limitations based on the premise that KDE’s analysis is always constrained to the planar 2-D Euclidean space and not the network space. Therefore KDE’s density estimation is based on Euclidean distance instead of the network distance. This has given rise to the development of Network-based KDE (NETKDE) by Xie and Yan [39] to estimate traffic accident density over the network space. NETKDE has been utilized in traffic accident spatial density estimation over the network space [14, 40–43]. The introduction of NETKDE showed some useful effort to examine spatial point patterns over the road network geometry, however, traffic accident analysis based on this model focuses on 1-D network space (spatial dimension) with no account of time (temporal dimension).

Our approach, therefore, differs from the previous approaches in the literature such that we identify and rank accident high risk-locations in space and time within a network based on the connectivity between accident events, road network geometry, and time using a graph data model. Graph data models present fundamental data structure for capturing and analyzing connectivity between events, objects and time. While graphs have been used for decades to analyze interactions in network sciences, their usage to analyze network-constrained events such as traffic accidents in a transit network remains to be investigated. In this paper, we use a graph data model for identification, ranking and analysis of traffic accident hot spot patterns in space and time at different spatial and temporal granularities.

The main contribution of this paper lies in the development of a graph data model that can support the computation of graph metrics to identify and rank traffic accident high-risk locations as well as the discovery of the space-time varying pattern at different space and time scales. We propose two graph metrics to understand and quantify the connectivity that exists between road network geometry and traffic accident events, facilitating the analysis of accident activity levels at road segments and intersections, accident activity ranking and overall accident event pattern over time.

Objectively, the developed space and time-varying graph (STVG) model [44] is used in this study to analyze traffic accidents hot spot patterns over the road network entities and across different time granularities. Graph metrics through graph queries are employed to discover traffic accident high-risk locations and analyze their patterns over 6 years (2010–2015) at different time granularities. The following analytical queries are computed based on graph metrics from the STVG graph database of traffic accident events: Which intersections and streets have the highest crash index? When and where do most crashes occur? How does the spatio-temporal pattern of fatalities differ from the spatio-temporal pattern of traffic accidents overall? What is the spatial pattern during weekdays and weekends, morning, afternoon and night times? When and where are the blackspots for crashes involving elderly drivers, teenagers, alcohol or weather related?

The remaining sections of this paper are organized as follows. Section 2 describes our proposed space–time-varying graph (STVG) model for traffic accidents. The graph metrics used to identify and rank the hot spots and temporal pattern analytics, and the practical implementation of the model in the Neo4j graph database are described.
in Sect. 3. The results of the analysis are presented in Sect. 4 and the conclusion follows in Sect. 5.

2 Methods

There are conceptual and implementation challenges in representing and analyzing the dynamics in traffic accident patterns and high risk-locations based on network connectivity. The conceptual and logical graph model of these events and the road network topology must assume a flexible structure that is defined in terms of relationships among the entities in both spaces (geographic and event spaces) of the entire network as a function of time. The modelling process thereby involves the logical representation of the interplay between the different perspectives of a traffic accident event network onto a single graph model. At the implementation level, the manner of storing the traffic accident data, and the adaptation of the conceptual graph model in a graph database have an impact on the analytical tasks needed for computing the analysis metrics. Identifying, ranking and analyzing patterns present in a traffic accident event network over time and space are fundamental analytical tasks needed to be supported by a graph database.

2.1 Space–time concept of the model

Conceptually, the space–time-varying graph (STVG) model of traffic accident events is a directed property graph model with two major key aspects and two network spaces. The aspects are the network elements (nodes and edges) and the time-tree (time instants). The network elements are spatio-temporal entities on two logical spaces (the geographic and event spaces) of the network. The spatial component of the graph model is a property in the nodes and edges of the network. Therefore, the STVG model can be defined and visualized as a graph, G made up of (N, E, T): where N represents the set of nodes (network vertices), E represents the set of all edges (relationships) and T represents the set of all time instants (from the time-tree).

The set of all nodes in the graph, G is represented as N(G), the set of all edges is represented as E(G) and the set of all time instants is represented as T(G). N(G) is made up of spatial nodes n, and spatio-temporal nodes nst, instances and non-spatio-temporal nodes n. E(G) is consisting of spatio-temporal edges est and non-spatio-temporal edges e. A spatial node in the graph is a node with spatial properties (such as spatial coordinates) among other properties that it has but with no temporal property; these are the street and intersection nodes (see Fig. 1). Non-spatio-temporal nodes are nodes with no spatial and temporal properties; these are the accident influencing factor nodes. While the spatio-temporal nodes are nodes with spatial and temporal properties in the graph model, that is, the traffic crash node. Spatio-temporal edges with the space–time relation “NEXT” are in the geographic and the event space. The edges can be weighted spatially by distance and temporally by time, and stop nodes. A spatial edge es ∈ E(G) is an ordered tuple, that is, es = (u, v, wt), where u, v ∈ N(G) are the source and target nodes, respectively and wt is the spatial weight of the edge (distance).

A temporal edge et ∈ E(G) is an ordered quintuple, et = (u, ta, v, tb, wt), where u, v ∈ N(G) are the source and target nodes, respectively, ta, tb ∈ T(G) are the source and target time instants, respectively and wt is the temporal weight of the edge (time). While the spatio-temporal edge est ∈ E(G) is an ordered sextuple est = (u, ta, v, tb, ws, wt) where u, v ∈ N(G) are the source and target nodes, respectively, ta, tb ∈ T(G) are the source and target time instants, respectively, ws is the spatial weight of the edge (distance) and wt is the temporal weight of the edge (time). Non-spatio-temporal edges are edges with no space–time relation and properties such as the “LOCATED_AT”, the “CAUSED_BY” and “HAPPENS_AT” edges. This is an edge, es ∈ E(G), that is, es = (u, v), where u, v ∈ N(G) are the source and target nodes, respectively.

2.2 Logical graph representation of the model

We propose a graph model as a data structure that consists of nodes (also known as vertices) and edges (also known as links). The reasons for presenting every element in our graph model as a node are, (i) to accommodate the space–time dynamics of every node in the network and (ii) to enable the computation of relevant graph metrics to determine high risk-locations and accident activity levels which are majorly nodal metrics. The nodes consist of entities that serve as the basis of the representation and edges act as relationships between these entities. A node can be an entity or a real-world feature that exists in a traffic accident event network such as the geographic features (streets and intersections). A node can also represent purely conceptual and abstract entities such as the traffic crash and the influencing factors. In our graph model, the edges are used to represent the directed relationships between the entities in the network.

2.3 Geographical and event spaces

In this research work, we propose the design of a graph model for the connectivity interplay between traffic accident events and geographic spaces on the road network topology where nodes and edges can be combined to form
complex retrieval criteria based on graph metrics such as centrality metrics. Our approach proposes the superimposition of two perspectives of the network using a single graph model. The first perspective is the relative view of the road network using the Geographical Space. In our case, the Geographical Space is a topological space based on the existence of geographical places and their neighborhood relationships such as left and right or being connected with. This represents the sequence of places (locations) on the road network, the traffic crashes are located at, sequentially and temporally as they occur. This sequence of occurrence is captured in both spaces of the model by the “NEXT” space–time relationship (edge). The second perspective is the relative view of the traffic crashes on the event space. The Event Space is an abstract space that records the sequence of traffic crashes as they occurred. These traffic crashes are recorded in the event space of the graph database sequentially and are connected by the “NEXT” space–time relation.

It is crucial to note that the geographical and event spaces have complementary characteristics which are evident in their many similarities and important differences. For example, each node in the event space has an equivalent node in the geographical space. This connection between the two spaces is represented by the “LOCATED_AT” relationship. In terms of similarities, they are both non-metric spaces representing the connectivity between geographical places and abstract traffic crash entities. Connectivity is a central characteristic that is being used as the basis of the representation of nodes and edges in both Geographical and Event Spaces. Irrespective of the space that is represented in the graph model, a temporal relationship (HAPPENS_AT) is used to connect all nodes to a time tree. HAPPENS_AT is crucial for representing the time dimension in the graph model that is needed in time-varying graph queries and metrics for the analysis.

One important difference between the Geographical Space and Event Space is that numerical weights are only given to the edges of the Event Space. The edge weights are non-negative numbers that represent the time elapsed between two consecutive crash events. This is used to record the difference in time between successive crash events which can be used in the computation of weighted edge-metrics such as shortest paths between events.

The traffic accident influencing factors are also modeled as abstract entities and stored as separate nodes in the graph database. This makes it easier for traffic accidents to be classified in terms of influencing factors using graph queries and metrics. Each traffic crash can have one or more influencing factors which are connected to the event space by the relationship “CAUSED_BY” edge.

2.4 Overview of the graph data model

Figure 1 provides a pictorial view of our proposed graph model. It depicts the meta-graph and conceptual representation of the entities in the model as implemented in the graph database.
In the proposed model, the time dimension is not represented by a timestamp property associated with the nodes or edges of the Geographical and Event Spaces, but instead it becomes a relationship between the nodes in the network and the time tree. The time tree is a hierarchical and multi-level temporal indexing structure that represents the natural levels of a timestamp [45]. It is comprised of a “Root” node that links all the “Year” nodes, and in turn, each “Year” node is linked to the twelve “Month” nodes of a year, and so on. Figure 2 illustrates this hierarchical structure where the top or lower-level nodes are connected through the “CONTAINS” relationship, meanwhile nodes situated at the same level of the time tree are sequentially connected through the “NEXT” relationship.

Fig. 2 An example of the hierarchical indexing structure of the time tree [46]
This hierarchical temporal indexing structure which is a component of the model improves the processing speed of time-dependent graph queries. Essentially, a time-dependent query arrives at the time-instant node or the range of time-instant nodes of the query. Then it traverses to all the linked events or nodes without scanning through all the nodes in the large graph.

2.5 Graph metrics for analysis

There are many graph metrics such as connectedness, shortest and longest shortest paths, degree, betweenness and PageRank centralities as well as network diameter and density. However, to identify and rank traffic accident high-risk locations and accident activity patterns over time and space, degree and PageRank centrality metrics are the major focus in the study. These metrics are implemented and retrieved through incremental graph queries to the graph database. The incremental queries create projects subgraph footprints (SF) defined at time-points or time-intervals (STVGt1, STVGt2, STVGtn, STVGti, ……STVGtn).

2.5.1 Degree centrality

Centrality measures give the relative measure of activity level a node has in the graph [45]. The Degree of a node is simply the amount of connectivity (the in-degrees and out-degrees) that it has with the other nodes in the graph. The reputation of a node in the graph is enhanced if it has a high number of in-edges and out-edges. In this paper, the degree centrality algorithm is utilized to identify traffic accident high-risk locations on streets and at intersections. Degree centrality is the simplest graph metric to compute but the most popular and extensively used in network sciences. In our graph model, which is a directed graph, there are two-degree centrality measures for a node, the in-degree which is the number of connections that point inward to the node, and the out-degree which is the number of connections that originate from the node and point outward to other nodes. The formula below states the simple definition of a degree centrality in-degree or out-degree. The actual degree centrality of a node in our graph is gotten from the summation of the in-degree and out-degree measures of the node.

By definition:

Degree centrality \( C_D(i) = \sum_{j=1}^{n} a_{ij} \),

where, element \( a_{ij} = 1 \) (if a direct edge exists between nodes i and j) and \( a_{ij} = 0 \) (if there is no edge).

For example, in this case study, degree centrality is used in different time windows and granularities to retrieve from the STVG database, intersections and street segments with a high occurrence of traffic accidents across time. High-risk intersections and street segments are derived from high degree centrality and PageRank centrality measures based on a simple ‘high’ and ‘low’ classification from the average. The average degree centrality is equal to the summation of all centrality measures of the intersections or street segments in the study area divided by the total number of intersections or street segments.

2.5.2 PageRank centrality

PageRank was originally adopted by Google [47] for ranking search results in the web network. PageRank is a variant of a more advanced view of centrality known as eigenvector centrality which permits connections in the network to have a variable value [48]. The results from PageRank centrality analysis most times correlates with that of degree centrality as both measure the level of importance of a node based on connectivity. However, with PageRank, nodes are regarded as being more reputable if there are more incoming edges than outgoing edges, and the nodes that link to it are also important. In this paper, PageRank is used to rank accident high-risk locations over time and space.

The PageRank score of a node in STVG, \( n \in STVG \) was computed by aggregating the stationary probabilities of all the incident edges on the node.

PageRank of a node i, is defined as

\[
PR(i) = \frac{d}{n} + (1 - d) \sum_{j \neq \in N} PR(j) / \text{OutDegree(j)},
\]

where d is the damping factor (ranging from 0 to 1), n is the total number of nodes and OutDegree is the number of outgoing edges from node j.

Incremental queries are used to implement these metrics at specific time-points and intervals on the projected subgraph footprints to evaluate traffic accident high-risk locations, activity level and ranking on the road network topology.

3 Datasets and data processing

The used case traffic accident dataset for this study was obtained from the free-access database of the Florida Department of Highway Safety and Motor Vehicles (FHSMV) from 2010 to 2015. The coverage area of this dataset is Brevard County. Road network, boundary dataset and other ancillary spatial datasets were obtained from the Brevard County property appraiser free public dataset. The traffic accident dataset was downloaded as a shapefile that contained the 13 attributes as described in Table 1 and a
total of 1,048,575 tuples. The tuples did not have any information about their spatial locations such as streets and intersections. Therefore we had to carry out data-pre-processing and spatial contextualization.

### 3.1 Data pre-processing and spatial contextualization

Data pre-processing and spatial contextualization were vital to: (i) add spatial context to the traffic accident datasets such as street name and intersect ID for seamless integration and connectivity at the database level, and (ii) to decompose the intersections and streets in the entire Brevard County road network vector dataset into discrete vector units, that is points (for intersection) and lixels (for street segments). These processes were carried out using some python scripts based on arcpy functions which involve an automated pipeline of geoprocessing as described in Fig. 3 and the following steps.

#### Step 1: Computation of Intersections

This involved a set of python functions coded in arcpy to: (i) merge streets and roads of the same unique name that are disconnected in the road network dataset into single streets and roads, (ii) a spatial search function is used to search for points where two or more streets or roads are intersected, (iii) these intersected points are automatically extracted into a list with unique IDs and a concatenation of the street names that formed the intersection, and (iv) duplicate points are filtered, removed and the final list of points is saved as a vector layer called intersections.

#### Step 2: Computation of Lixels

Each street or road segment is divided into a basic linear unit of 50 m length. These linear units of segments are called Lixels as defined by Xie and Yan [39]. This is to facilitate a systematic selection of regularly spaced locations along the street network to enhance connectivity to discrete street nodes in the graph model.

#### Step 3: Connectivity Neighborhood Creation

This step creates a 15-m connectivity radius around each street lixel and intersection point. This is used to associate the traffic accident events to their respective streets and intersections within 15 m radius of the accident epicenter. These connectivity neighborhoods are converted to a vector feature layer and each search neighborhood is labeled by the name of its associated streets lixel of intersection. This 15-m connectivity neighborhood worked optimally for this study area even at ramps and crossroads, however this could be dynamically selected as we did. We iteratively tested different radii ranging from 10 to 50 m. The selected 15 m connectivity radius gave us the balance between connectivity precision and coverage. We also acknowledge that this might not be the optimal radius for a different study area with its peculiarities in road structures. This should be selected iteratively to arrive at an optimal connectivity radius.

#### Step 4. Overlay Traffic Accident and Connectivity Neighborhood

A spatial overlay function is called up from the arcpy python package to associate each traffic accident tuple to its respective connectivity neighborhood.

**Fig. 3** Data pre-processing and contextualization workflow

---

**Table 1** Attributes of the traffic accident dataset

| Attribute     | Description          |
|---------------|----------------------|
| Crash_DT      | Crash date           |
| Crash_TM      | Crash time           |
| Age           | Age of the driver involved |
| Crash_HOD     | Crash hour of the day |
| Crash_DOW     | Crash day of the week |
| Crash_MOY     | Crash month of the year |
| Crash_Y       | Crash year           |
| Crash_WK      | Crash week number of the year |
| Fatalities    | Number of fatalities recorded |
| Injury        | Injury recorded      |
| Alcohol_Related | Yes/no                |
| Distraction_Related | Yes/no               |
| Weather_Condition | Clear/cloudy/rain    |
Step 5: Geo-enrich the Traffic Accident tuple Each traffic accident tuple is geo-enriched by adding as an extra attribute of this tuple, the label of the associated connectivity neighborhood which is either a street name or an intersection ID. Finally, the spatially contextualized traffic accident dataset is realized to fit the graph model described in Sect. 3.3.

Step 6: Temporal Sequence of Occurrence The last preprocessing carried out on the contextualized traffic accident dataset is the sequential arrangement of events based on the temporal sequence of occurrence. An additional column was added to the CSV file of the dataset called “Sequence”. This enables us to create the “NEXT” space–time relation both in the event space and geographic space in the graph database.

3.2 Neo4j graph database

The implementation of our graph model was done using the Neo4j graph database management system. Neo4j is a popular native graph database with applications in fraud detection, real-time recommendation engines, and network operations. It uses the Cypher query language which enables ad-hoc queries and user-defined functions (UDFs).

One important characteristic of the Neo4j database is the creation of the edges of the network at the same time as the nodes are being created, rather than creating them at query time through JOIN operations. Online resource 1 shows an example node creation cypher statement to load the traffic accident tuples into the Neo4j database. After the nodes have been created, the edges can be created in the database. Online resource 2 shows an example of the edges created between the nodes of the Geographical Space and Event Space. A total of 1,137,695 nodes and 344,765 edges were created in a few seconds in the Neo4j database. The nodes were sequentially linked to the time tree from the highest to the lowest level of time granularity as shown in Online resource 3. This Cypher code was used to create the sequential “HAPPEN_AT” edge between the Traffic accident nodes and the corresponding time tree leaf nodes. In total, 109,844 edges were created for “HAPPEN_AT” edges within the database.

3.3 Cypher queries

Cypher is a graph query language that allows for efficient querying, updating and analysis of graph properties and metrics. Graph metrics (e.g., PageRank and degree centralities), and other UDFs are easily encoded within cypher queries. The computation of the graph metrics used in this study was coded in cypher with a pipeline of query functions. Online resource 4 illustrates this pipeline for a PageRank cypher query. Through the Cypher query manager, cypher queries are used to project the part of the graph valid for the specified time window in-memory. Appropriate graph metrics are then implemented on the projected graph to retrieve analytical values needed for traffic accident high-risk location analysis. Online resource 5 shows the time-incremental Degree-Centrality query.

4 Results and discussion

Identification, ranking and profiling of traffic accident high-risk locations have always been the major concerns of most transport authorities and traffic accident departments. They also want to be able to carry out this analysis at varying spatial and temporal scales efficiently with simple queries and fewer complexities. Current analysis methods based on GIS and spatial statistics are not only modeled without temporal context in them but can be very challenging and technically complex for a traffic agent to process and make sense of the analytical results. Graph metrics such as degree and PageRank can be implemented in simple graph queries to the database with temporal and spatial context in the queries. The results are retrieved faster and are easier to understand than in the case of any GIS and statistical analytics tools.

The analysis and results below, present the use of degree and PageRank graph metric queries as shown in Table 2 to answer the following questions: Which intersections and streets have the high crash index? When and where do most crashes occur? How does the spatio-temporal pattern of fatalities differ from the spatio-temporal pattern of traffic accidents overall? What is the spatio-temporal pattern during weekdays and weekends, morning, afternoon and night times? etc. Table 3 shows the top-20 intersections with the highest crash activity based on degree centrality graph metric while Table 4 shows the top-20 street segments with the highest crash activity and Ranking based on degree centrality and PageRank graph metrics respectively.

In this analysis, the degree centrality graph measure is very efficient in identifying traffic accident high-risk locations both on the road network segments (streets) and on intersections only based on network connectivity between the geographic space and the event space. Figure 4 shows the graph visualization of these interactions. Our analysis identified that intersections with ID 1886 and 12639 on the Ramp and I 95 roads (Fig. 5) with degrees of 3565 and 3355 respectively shown in Table 5 have the highest traffic crash activity over the 6-year period.

I 95 and Babcock roads (Fig. 4) are the most traffic accident-prone road segments with degree values of 7018 and 6203 respectively within a period of 6 years.

However, as the degree of centrality employs in-edges and out-edges for its computation, PageRank gives more
relevance to the in-coming edges and the importance of the nodes that link to the node in question. PageRank values as shown in Table 4 discovers similar results in terms of traffic accident streets ranking over a period of 6 years. However, the top 20 intersections with the highest degree centrality measures are not ranked top in the PageRank analysis shown in Table 5. This is because PageRank does not only consider the connectivity between the intersections and the traffic crash events but also factors in the degree of importance of that intersection in the entire network. That is, the number of roads or street segments that have a direct connectivity with any intersection in question. The highly ranked intersections have more segment connections than others. For example, the top 10 ranked spots in this analysis have at least 4 connecting segments. PageRank metric also considers some intersections very important because they are high-risk locations and have direct connections to high-risk street segments in the network.

In Fig. 6, the number of accidents that take place yearly on these roads especially I 95 and Babcock Street maintained a record high throughout these 6 years as well as in every month of the year. However, the analysis shows that a street as small as Babcock Street has maintained the highest degree of traffic accidents since 2013 consistently. Based on these metrics it becomes easy for a traffic accident agent to recommend Babcock Street for traffic accident profiling as it has maintained the highest degree of accidents since 2013 on a yearly and monthly basis. The intersection with ID 1886 on Ramp Road has maintained the top-accident prone intersection over the 6 years, except in 2014, when the intersection with ID 12639 on I 95 road hit a record high over all other intersections in the network. This suggests that the spot was probably in a bad shape in 2014. Coincidentally, no accident was recorded on this intersection, 12639 in August and September throughout these 6 years. Interestingly, the months of October, November and December are relatively the months with

---

**Table 2** Example Queries to compute and rank average crash index of intersections

```sql
MATCH (cr: Crashes)-[r]-(int:Intersections)
RETURN int.IntersectionID
int.IntersectionName
Size((int)-[r]-()) AS degree
ORDER BY degree DESC LIMIT 20
```

**Table 3** Top-20 intersections with the highest crash activity based on degree centrality graph metric

| Intersection Name                          | Intersection ID | Degree Centrality |
|-------------------------------------------|----------------|-------------------|
| Ramp&Ramp                                 | 1886           | 3565              |
| I 95&I 95                                 | 12639          | 3355              |
| Highway 1&Turnaround&Highway 1            | 1817           | 1585              |
| Ramp&I 95&I 95                            | 14472          | 1099              |
| Highway 1&Highway 1&Turnaround            | 12764          | 763               |
| Rockledge&Rockledge&Turnaround            | 23264          | 659               |
| Highway A1A&Highway A1A                   | 20243          | 561               |
| Wickham&Wickham                           | 25863          | 547               |
| Ramp&Ramp                                 | 3485           | 531               |
| Washington&Turnaround&Washington          | 13047          | 469               |
| Minton&Palm Bay&Minton&Palm Bay            | 25495          | 391               |
| Wickham&Sarno&Sarno&Wickham               | 22107          | 301               |
| Eau Gallie&Eau Gallie&Harbor City&Harbor City | 23012     | 276               |
| Minton&Emerson&Emerson&Minton              | 23851          | 269               |
| Turnaround&Minton&Minton                   | 13591          | 256               |
| Eau Gallie&Eau Gallie                      | 1821           | 249               |
| Atlantic&Atlantic&Fourth&Fourth            | 22580          | 247               |
| Eau Gallie&Eau Gallie&South Patrick&Riverside | 24088     | 218               |
| Turnaround&Courtenay&Courtenay            | 14833          | 216               |

---

**Table 4** Example Queries to compute and rank average crash index of intersections

```sql
MATCH (int:Intersections)
WITH collect(int) AS nodes
CALL apoc.algo.pageRank(nodes) YIELD node, score
RETURN node. IntersectionID, node.IntersectionName, score
ORDER BY score DESC LIMIT 20
```
the lowest degree of traffic crashes both on the road segments and intersections. Figure 7 shows daily analysis which depicts consistent peaks on Mondays and Fridays for both on the segments and intersections. However, the two most high-risk accident intersections 1886 and 12639 maintained the average highest on a daily basis similar to the Babcock and I 95 roads. These traffic accident peaks are recorded mainly during peak hours of the day as shown in the hourly charts. The highest peaks for the crashes are recorded between the hours of 3 pm and 7 pm daily. Time-varying PageRank analysis did not show a significant difference from the overall PageRank average described in Tables 3 and 4.

Figure 8 shows that most accidents in intersection 1886 were not fatal as the intersection did not appear in the top 20 high traffic accident fatality charts. However, the intersection, 13462 on Highway 1 which ranks third in the degree of accident occurrence has recorded the highest fatalities in Brevard County from 2010 to 2015. I95 maintained the highest in terms of the degree of accidents, ranking and fatality rate.

5 Conclusion
While previous research on spatio-temporal traffic accident pattern analysis have mostly been based on GIS-based methods and spatial point pattern analysis, in this paper, we...
have taken the perspective of space–time interactions in the entire network space of road network topology and traffic accident events. In other words, this study has focused on discovering traffic accident patterns in space and time based on graph analytics of the network that emerges from the space–time-varying relationship between road networks and traffic accident events.

This paper describes a simple space time-varying graph (STVG) model of traffic accidents that is generic and scalable. The STVG model provides us with the capability to identify, rank and profile traffic accident high-risk locations in space and study their pattern over time. The developed STVG model was implemented in the Neo4j graph database with inbuilt graph algorithms. The time-tree is a hierarchical temporal indexing structure that provides the ability to analyse these patterns at different space and time scales. Data spatial contextualization became very necessary so that the traffic accident dataset could fit in properly in the space time-varying graph model. This was carried out through a pipeline of spatial data preprocessing.

Traffic accident blackspots and segments were identified and ranked in space and time using time-dependent degree and PageRank centrality graph metrics. Degree centrality is simple and effective in discovering these high-risk locations and segments in space and time while PageRank metric ranking adopted a more complex procedure. The PageRank of the street segments correlates with the degree centrality analysis but does not correlate with degree centrality analysis of the intersections. PageRank metric requires some modification to suit this type of network. However, the degree centrality metric proved to be the most suitable.

---

**Fig. 5** Some streets and intersections shown on the road network topology; I 95 and Babcock streets in red lines while the intersections are circled in red

---

**Table 5** Top-20 intersections with the highest crash activity ranking based on PageRank graph metric

| Intersection name                                      | Intersection ID | Pagerank score |
|--------------------------------------------------------|-----------------|----------------|
| Wickham&Sarno&Sarno&Wickham                           | 13,047          | 14.06875       |
| Eau Gallie&Eau Gallie&Harbor City&Harbor City         | 23,012          | 13.00625       |
| Minton&Emerson&Emerson&Minton                          | 23,851          | 12.5175        |
| Eau Gallie&Eau Gallie&South Patrick&Riverside          | 24,088          | 10.35          |
| Harbor City&Harbor City&Strawbridge&Strawbridge        | 24,586          | 9.90375        |
| Interchange&San Filippo&Malabar&Malabar                | 22,399          | 9.33           |
| Aurora&Wickham&Wickham&Aurora                          | 23,190          | 8.92625        |
| Wickham&Lake Washington&Wickham&Lake Washington        | 21,677          | 7.77875        |
| Barnes&Murrell&Murrell&Murrell&Barnes                   | 25,766          | 7.5875         |
| Minton&Palm Bay&Minton&Palm Bay                         | 25,495          | 6.81111        |
| Minton&Palm Bay&Minton&Palm Bay                         | 25,523          | 6.81111        |
| Hibiscus&Hibiscus&Babcock&Babcock&Babcock              | 25,758          | 6.71625        |
| Harbor City&Sarno&Harbor City&Harbor City              | 22,271          | 6.67375        |
| Garden&A Max Brewer Memorial&Washington&Washington     | 22,938          | 6.27           |
| Merritt Island&Newfound Harbor&Merritt Island&Merritt Island | 25,736   | 5.675          |
| Wickham&Parkway&Wickham&Wickham&Parkway                | 25,761          | 5.675          |
| Cypress&Eau Gallie&Cypress&Montreal&Eau Gallie          | 23,255          | 5.6325         |
| St Lucie&Atlantic&Atlantic&Atlantic&St Lucie           | 25,863          | 5.56875        |
| Atlantic&Atlantic&Turnaround&Marion                    | 25,371          | 5.52625        |
| Otterbein&Clearlake&Clearlake                           | 6745            | 5.48375        |
Fig. 6 Graphical illustration showing the variations in traffic accident degree scores, yearly at the a top four intersections, and b top four streets; and monthly at c top four intersections, and d top four streets.

Fig. 7 Graphical illustration showing the variations in traffic accident degree scores, daily at the a top four intersections, and b top four streets; and hourly at c top four intersections, and d top four streets.
Further analysis will incorporate more of the non-spatial and temporal entities of the network such as the influencing traffic accident factors and fatality in the analysis. Further analysis can answer questions such as when and where are the blackspots for crashes involving elderly drivers, teenagers, alcohol or weather related. Also, a comparative analysis between the graph-based approach and the use of NETKDE to compute traffic accident hotspots in space and time is desirable, however, we have to find a way to incorporate temporal dimension into NETKDE. As mentioned in Sect. 1, one of the main advantages of using an STVG model is the ability to carry out time-varying analysis of the hotspots in a dynamic network. For lack of temporal dimension in NETKDE, we were not able to carry out a comparative analysis of our approach with NETKDE in this study.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Moons, E., Brijs, T., & Wets, G. (2009). Improving Moran’s index to identify hot spots in traffic safety. Geocomputation and Urban Planning. https://doi.org/10.1007/978-3-540-89930-3_7
2. Bíl, M., Andrášik, R., & Sedoník, J. (2019). A detailed spatiotemporal analysis of traffic crash hotspots. Applied Geography, 107, 82–90. https://doi.org/10.1016/j.apgeog.2019.04.008
3. Le, K. G., Liu, P., & Lin, L. (2020). Determining the road traffic accident hotspots using GIS-based temporal-spatial statistical analytic techniques in Hanoi Vietnam. Geo-spatial Information Science, 23(2), 153–164. https://doi.org/10.1080/10095020.2019.1683437
4. Moran, P. A. (1948). The interpretation of statistical maps. Journal, Source Statistical, Royal Series, Society, 38(2), 189–203.
5. Getis, A., & Ord, J. K. (1992). The Analysis of Spatial association by use of distance statistics. Geographical Analysis, 24(3), 189–206. https://doi.org/10.1111/j.1538-4632.1992.tb00261.x
6. Hashimoto, S., Yoshiki, S., Saeki, R., Mimura, Y., Ando, R., & Nanba, S. (2016). Development and application of traffic accident density estimation models using kernel density estimation. Journal of Traffic and Transportation Engineering, 3(3), 262–270. https://doi.org/10.1016/j.jtte.2016.01.005
7. Caliendo, C., Guida, M., Postiglione, F., et al. (2021). A Bayesian bivariate hierarchical model with correlated parameters for the analysis of road crashes in Italian tunnels. Statistical Methods and Applications. https://doi.org/10.1007/s10260-021-00567-5
8. Ma, L., Yan, X., & Qiao, W. (2014). A quasi-poisson approach on modeling accident hazard index for urban road segments. Discrete Dynamics in Nature and Society. https://doi.org/10.1155/2014/489052
9. Tang, L., Kan, Z., Zhang, X., Sun, F., Yang, X., & Li, Q. (2016). A network Kernel Density Estimation for linear features in space–time analysis of big trace data. International Journal of Geographical Information Science, 30(9), 1717–1737. https://doi.org/10.1080/13658816.2015.1119279
10. Satria, R., & Castro, M. (2016). GIS Tools for Analyzing Accidents and Road Design: A Review. Transportation Research Procedia, 18, 242–247. https://doi.org/10.1016/j.trpro.2016.12.033
11. Zhang, Z., Ming, Y., & Song, G. (2019). Identify road clusters with high-frequency crashes using spatial data mining approach. Applied Sciences, 9(24), 5282. https://doi.org/10.3390/app9245282
12. Okabe, A., Satoh, T., & Sugihara, K. (2009). A kernel density estimation method for networks, its computational method and a GIS-based tool. *International Journal of Geographical Information Science, 23*(1), 7–32. https://doi.org/10.1080/13658810802475491

13. Thakali, L., Kwon, T. J., & Fu, L. (2015). Identification of crash hotspots using kernel density estimation and kriging methods: A comparison. *Journal of Modern Transportation, 23*, 93–106. https://doi.org/10.1007/s40534-015-0068-0

14. Xie, Z., & Yan, J. (2013). Detecting traffic accident clusters with network kernel density estimation and local spatial statistics: An integrated approach. *Journal of Transportation Geography, 31*, 64–71. https://doi.org/10.1016/j.jtrangeo.2013.05.009

15. Mohaymany, A. S., Shahri, M., & Mirbagheri, B. (2013). GIS-based method for detecting high-crash-risk road segments using network kernel density estimation. *Geo-spatial Information Science, 16*(2), 113–119. https://doi.org/10.1007/s100502013.766396

16. Aguero-Valverde, J., & Jovanis, P. P. (2006). Spatial analysis of fatal and injury crashes in Pennsylvania. *Accident Analysis and Prevention, 38*(3), 618–625. https://doi.org/10.1016/j.aap.2005.12.006

17. Zeng, Q., Sun, J., & Wen, H. (2017). Bayesian Hierarchical Modeling Monthly Crash Counts on Freeway Segments with Temporal Correlation. *Journal of Advanced Transportation, 2017*(8p). https://doi.org/10.1155/2017/5391054

18. Fawcett, L., Thorpe, N., Matthews, J., & Kremer, K. (2017). A novel Bayesian hierarchical model for road safety hotspot prediction. *Accident Analysis & Prevention, https://doi.org/10.1016/j.aap.2016.11.021*

19. Hosseinilou, M. H., & Sohrabi, M. (2009). Predicting and identifying traffic hot spots applying neuro-fuzzy systems in intercity roads. *International Journal of Environmental Science and Technology, 6*, 309–314. https://doi.org/10.1007/BF03327634

20. Effati, M., Rajabi, M. A., Samadzadegan, F., & Shabani, S. (2014). Transportation A geospatial neuro-fuzzy approach for identification of hazardous zones in regional transportation corridors. *International Journal of Civil Engineering, 12*(3), 289–303.

21. Mussone, L., Bassani, M., & Masci, P. (2017). Analysis of factors affecting the severity of crashes in urban road intersections. *Accident Analysis and Prevention, 103*, 112–122. https://doi.org/10.1016/j.aap.2017.04.007

22. García de Soto, B., Bumbacher, A., Deublein, M., & Adey, B. T. (2018). Predicting road traffic accidents using artificial neural network models. *Infrastructure Asset Management, 5*(4), 132–144. https://doi.org/10.1680/jinam.17.00028

23. Pradhan, B., & Ibrahim, S. M. (2020). Review of traffic accident predictions with neural networks. In B. Pradhan & M. Ibrahim Sameen (Eds.), *Laser scanning systems in highway and safety assessment. Advances in science, technology & innovation*. Springer.

24. Gutierrez-Osorio, C., & Pedraza, C. (2020). Modern data sources and techniques for analysis and forecast of road accidents: A review. *Journal of Traffic and Transportation Engineering, 7*(4), 432–446. https://doi.org/10.1016/j.jtte.2020.05.002

25. Moghadam Gilani, V. N., Hosseinian, S. M., Ghasedi, M., & Nikookar, M. (2021). Data-Driven Urban Traffic Accident Analysis and Prediction Using Logit and Machine Learning-Based Pattern Recognition Models. *Mathematical Problems in Engineering, https://doi.org/10.1155/2021/9974219*

26. Erdogan, S., Yilmaz, I., Baybura, T., & Gullu, M. (2008). Geographical information systems aided traffic accident analysis system case study: City of Afyonkarahisar. *Accident Analysis and Prevention, 40*(1), 174–181. https://doi.org/10.1016/j.aap.2007.05.004

27. Truong, L. T., & Somenahalli, S. V. (2011). Using GIS to Identify Pedestrian-Vehicle Crash Hot Spots and Unsafe Bus Stops. *Journal of Public Transportation, 14*(1), 99–114. https://doi.org/10.5038/2375-0901.14.1.6

28. Kuo, P. F., Lord, D., & Walden, T. D. (2013). Using geographical information systems to organize police patrol routes effectively by grouping hotspots of crash and crime data. *Journal of Transportation Geography, 30*, 138–148. https://doi.org/10.1016/j.jtrangeo.2013.04.006

29. Rankavat, S., & Tiwari, G. (2013). Pedestrian Accident Analysis in Delhi using GIS. *Journal of the Eastern Asia Society for Transportation Studies, 10*, 1446–1457.

30. Tortum, A., & Atalay, A. (2015). Spatial analysis of road mortality rates in Turkey. *Transport, 168*(6), 532–542. https://doi.org/10.1016/j.jtrangeo.2014.00.029

31. Aghajani, M. A., Dezfulian, R. S., Arjoody, A. R., & Rezaei, M. (2017). Applying GIS to Identify the Spatial and Temporal Patterns of Road Accidents Using Spatial Statistics (case study: Ilam Province, Iran). *Transportation Research Procedia, 25*, 2131–2143. https://doi.org/10.1016/j.trpro.2017.05.0409

32. Gatrell, A. C., Bailey, T. C., Diggle, P. J., & Rowlingson, B. S. (1996). Point Spatial application pattern analysis geographical epidemiology. *Transactions of the Institute of British Geographers, 2*(1), 256–274.

33. Levine, N. E. D., Kim, K. E., & Nitz, L. H. (1995). Spatial analysis of Honolulu motor vehicle crashes: I. Spatial patterns. *Accident Analysis and Prevention, 27*(5), 663–674. https://doi.org/10.1016/0001-4575(95)00017-T

34. Diaz-Varela, E. R., Vázquez-Gonzalez, I., Marey-Pérez, M. F., & Álvarez-López, C. J. (2011). Assessing methods of mitigating wildlife-vehicle collisions by accident characterization and spatial analysis. *Transportation Research Part D: Transport and Environment, 16*(4), 281–287. https://doi.org/10.1016/j.trd.2011.01.002

35. Pulugurtha, S. S., Krishnakumar, V. K., & Nambisan, S. S. (2007). New methods to identify and rank high pedestrian crash zones: An illustration. *Accident Analysis and Prevention, 39*(4), 800–811. https://doi.org/10.1016/j.aap.2006.12.001

36. Delmelle, E. C., & Thill, J. C. (2008). Urban Bicyclists: Spatial Analysis of Adult and Youth Traffic Hazard Intensity. *Transportation Research Record: Journal of the Transportation Research Board, 2074*(1), 31–39. https://doi.org/10.3141/2074-04

37. Anderson, T. K. (2009). Kernel density estimation and K-means clustering to profile road accident hotspots. *Accident Analysis and Prevention, 41*(3), 359–364. https://doi.org/10.1016/j.aap.2008.12.014

38. Yalcin, G., & Duzgun, H. S. (2015). Spatial analysis of two-wheeled vehicles traffic crashes: Osmaniye in Turkey. *KSCE Journal of Civil Engineering, 19*(7), 2225–2232. https://doi.org/10.1007/s12205-015-0661-0

39. Xie, Z., & Yan, J. (2008). Kernel Density Estimation of traffic accidents in a network space. *Computers, Environment and Urban Systems, 32*(5), 396–406. https://doi.org/10.1016/j.compenvurbsys.2008.05.001

40. Okabe, A., Okunoki, K. I., & Shiode, S. (2006). The SANET toolbox: New methods for network spatial analysis. *Transactions in GIS, 10*(4), 535–550. https://doi.org/10.1111/j.1467-9671.2006.01011.x

41. Okabe, A., Yomono, H., & Kitamura, M. (1995). Statistical Analysis of the Distribution of Points on a Network. *Geographical Analysis, 27*(2), 152–175. https://doi.org/10.1111/j.1538-4632.1995.tb00341.x

42. Yamada, I., & Thill, J. C. (2004). Comparison of planar and network K-functions in traffic accident analysis. *Journal of Transport Geography, 12*(2), 149–158. https://doi.org/10.1016/j.jtrangeo.2003.10.006
43. Lu, Y., & Chen, X. (2007). On the false alarm of planar K-function when analyzing urban crime distributed along streets. *Social Science Research, 36*(2), 611–632. https://doi.org/10.1016/j.ssresearch.2006.05.003

44. Maduako, I., Wachowicz, M., & Hanson, T. (2019). STVG: An evolutionary graph framework for analyzing fast-evolving networks. *J Big Data*. https://doi.org/10.1186/s40537-019-0218-z

45. Maduako, I. D., Wachowicz, M., & Hanson, T. (2019). Transit performance assessment based on graph analytics. *Transportmetrica A: Transport Science, 15*(2), 1382–1401. https://doi.org/10.1080/23249935.2019.1596991

46. Ivarsson, T., Kollegger, A., Neubauer, P., Svensson, J., & Weber, J. (2014). *The Neo4j manual v1.3*. Neo-Technology.

47. Brin, S., & Page, L. (1998). The anatomy of a large scale hypertextual Web search engine. *Computer Networks and ISDN Systems, 30*(7), 107–117.

48. Hansen, D. L., Shneiderman, B., Smith, M. A., & Himelboim, I. (2020). Analyzing social media networks with NodeXL. *Insights from a connected world* (2nd ed., pp. 31–51). Elsevier.

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.