Mining and Learning from Railway Safety Data with Graphs and Tensors

Investigator Name: Jia Chen
Title: Assistant Professor of Teaching
Department: Electrical and Computer Engineering
University: University of California Riverside

Investigator Name: Evangelos Papalexakis
Title: Associate Professor
Department: Computer Science and Engineering
University: University of California Riverside

A Report on Research Sponsored by

University Transportation Center for Railway Safety (UTCRS)

University of California Riverside

September 2024

Technical Report Documentation Page

1. Report No. UTCRS-UCR-O2CY23	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Mining and Learning from Railway Safety Data with Graphs and Tensors		5. Report Date September 15, 2024
		6. Performing Organization Code UTCRS-UCR
7. Author(s) Jia Chen and Evangelos Papalexakis		8. Performing Organization Report No. UTCRS-UCR-O2CY23
9. Performing Organization Name and Address University Transportation Center for Railway Safety (UTCRS) University of California Riverside (UCR) 900 University Ave. Riverside, CA 92521		10. Work Unit No. (TRAIS)
		11. Contract or Grant No. 69A3552348340
12. Sponsoring Agency Name and Address U.S. Department of Transportation (USDOT) University Transportation Centers Program 1200 New Jersey Ave. SE Washington, DC, 20590		13. Type of Report and Period Covered Project Report June 1, 2023 – August 31, 2024
		14. Sponsoring Agency Code USDOT UTC Program
15. Supplementary Notes		
16. Abstract Railway systems are very complex pieces of cyberinfrastructure, interfacing with a number of transportation agents and other pieces of cyberinfrastructure. For instance, a railway crossing includes interactions between the railway system and a traffic intersection. Such a rich ecosystem of interactions among heterogeneous agents poses fascinating research challenges in modeling railway systems with data and conducting data-driven railway crossing safety assessment. In this project we leverage and extend powerful tensor and graph mining methods which can extract "needles in the haystack" within the abundance of collected data and produce actionable insights to better understand emerging accident patterns from historical data, identify underlying similarities in such patterns.		
17. Key Words Safety Analysis, Data Mining, Data Science	18. Distribution Statement This report is available for download from https://www.utrgv.edu/railwaysafety/research/operations/index.htm	

19. Security Classification (of this report) None	20. Security Classification (of this page) None	21. No. of Pages 15	22. Price

Table of Contents

List of Figures

List of Tables

None.

List of Abbreviations

AI	Artificial Intelligence
CIKM	ACM International Conference on Information and Knowledge Management
CPD	Canonical Polyadic Decomposition
FRA	Federal Railroad Administration
NeAT	Neural Additive Tensor Decomposition
t-SNE	t-distributed Stochastic Neighbor Embedding
USDOT	U.S. Department of Transportation
UTCRS	University Transportation Center for Railway Safety

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

Acknowledgements

The authors wish to acknowledge the University Transportation Center for Railway Safety (UTCRS) for funding this project under the USDOT UTC Program Grant No 69A3552348340.

1. Introduction

Railway systems are very complex pieces of cyberinfrastructure, interfacing with a number of transportation agents and other pieces of cyberinfrastructure. For instance, a railway crossing includes interactions between the railway system and a traffic intersection. Such a rich ecosystem of interactions among heterogeneous agents poses fascinating research challenges in modeling railway systems with data and conducting data-driven railway crossing safety assessment. In this project we leverage and extend powerful tensor and graph mining methods which can extract "needles in the haystack" within the abundance of collected data and produce actionable insights to stakeholders in order to better understand emerging accident patterns from historical data, identify underlying similarities in such patterns, towards ultimately reducing the number of accidents.

Given different entities such as rail companies or railway crossings and data collected for those entities (such as sensor measurements, video recordings, and public USDOT accident reports), we seek to identify hidden patterns that emerge from the data which are not visible by mere inspection of the vast raw data sources and can be turned into actionable insights that can inform policy by understanding emerging patterns of behavior in this complex ecosystem.
In this report we outline a number of technical advancements that were achieved as part of this project with respect to graph mining and tensor methods for railway safety applications.

2. Graph Mining of Public Accident Reports

Graphs are very powerful data structures that capture relationships (also called edges) between entities (also called vertices or nodes). In addition to lending themselves to intuitive visualizations, graphs capture important patterns that may not be visible when looking at the raw data. For instance, public USDOT accident reports, as captured in the publicly available dataset "US Highway Railroad Crossing Accident" [1] contain a lot of important information that can potentially allow us to understand mechanisms that influence different accidents and, thus, inform policy on how to improve the safety of railway crossings.

For the purposes of this project, we conducted a proof-of-concept study which was published at the Joint Rail Conference 2024 [2,3] where we demonstrate the power of graph mining in extracting actionable insights from public accident reports. In constructing a graph from raw

accident reports, we chose to set the railway companies as the entities/nodes of interest and connect them with edges in a way that captures the frequency of co-occurrence of accidents that those two companies have been involved in, based on the same locality and for a certain period of time.

By constructing the aforementioned graph, companies who tend to have similar accident patterns in similar locations will be connected with each other, and densely connected parts of that graph can point to groups of companies who share the same accident behavior. In order to extract those groups, we conduct Spectral Clustering, a graph mining technique that is able to group or cluster the nodes/entities of the graph into a set of disjoint clusters, where within each cluster nodes have high degree of connectivity with each other.

Our analysis [2] reveals interesting patterns both in terms of the geographic and temporal distribution of discovered clusters which can shed light into different factors that may contribute to an observed pattern of accidents. We see this work as a stepping-stone towards providing valuable insights for policymakers, railway companies, and public safety officials, advocating a shift toward data-driven analysis and predictive modeling.

3. Tensor Mining of Public Accident Reports

A natural next step in the direction of the exploratory analysis and pattern extraction that we developed in Section 1 is Tensor Mining. Tensors are multi-dimensional data structures that can represent very high-dimensional and heterogeneous datasets. A graph, such as the one in Sec. 2, in fact, can be seen as a 2-mode tensor (or a matrix). Thus, tensors allow us to express more complex relations among different entities or variables of interest. For example, we can have a 3-mode tensor which involves variables such as “Company Name”, “County”, and “Date”, and will essentially capture the number of accidents for each rail company within a given county over time. Such a tensor can be mined for patterns in a way that companies that have similar spatio-temporal accident behavior are clustered together, in our running example. One of the most popular ways in which we can do so is through the so-called Canonical Polyadic Decomposition (CPD) [4], which expresses a tensor into a sum of simple elements (called rank-one components), each one of which corresponding to a co-cluster (i.e., a cluster that involves all dimensions of the data) in the data. In our running example, a rank-one component would give us a co-cluster (i.e., a subset) of rail companies, counties, and points in time that have high similarity.

Given the number of available variables/dimensions in the USDOT public accident report data [1], tensor mining can be an extremely valuable tool in identifying needles in the proverbial haystack of the vast amounts of reports, and identifying emerging patterns across different variables of interest which are not visible by mere inspection or simple correlation analysis. In this Section, we describe how tensor mining can be applied in this novel railway safety application. In constructing tensor datasets from raw data, the challenge is that even though virtually any structured dataset (such as the USDOT accident reports [1]) can be readily and trivially seen as a tensor (i.e., a multi-dimensional matrix), we are particularly interested in tensors that can be mined for interpretable patterns [5], because tensor methods require data to have certain structure, typically so-called low-rank or low-dimensional structure. As a result, a first step here is to provide a method for identifying and forming as many well-structured tensors as possible from the raw data.

A major challenge in judging the quality of the structure in a given tensor dataset is that, since we are conducting exploratory and completely unsupervised analysis, there is no ground truth or golden standard baseline that we can use to evaluate our findings. However, not all hope is lost, since we can rely on so-called intrinsic measures of structure goodness, i.e., measures of structure goodness that capture numerically whether the structure within the data adheres to various reasonable definitions of "good structure". In our case, goodness of structure is measured by how well the data adhere to the CPD tensor model, as well as how "clusterable" are the data within the tensor. Thus, we are going to evaluate different automatically generated tensor datasets from the raw report data using a number of intrinsic measures of structure quality and then identify datasets which score high-enough on all measures.

The steps of our proposed method are:

1. Start with an initial set of variables of interest. Here we can include all dimensions, however, this can generate an explosive number of candidates, so it is best to judiciously identify a reasonable subset of variables, such as railroad company name, county, date, and variables relating to road conditions or maintenance, in order to narrow down the search.
2. Given the above variables of interest, generate in a brute force manner all possible tensors that can involve a subset of those variables. For simplicity we can also constrain here the number of modes/dimensions in the tensor to a fixed number, e.g., three. This will generate all possible subsets of size three from the initial set.

3. For each tensor generated in step #2, we need to measure a number of structure quality scores. In particular, we measure:
    a. Fit: For a user-defined range of number of components, measure how well a CPD decomposition approximates the data. This can be a number [0,1], with 1 being perfect fit. We choose the maximum fit along that range.
    b. Core Consistency [5]: This is a score that measures how well the data adhere to a CPD model by using concepts behind the definition of the model. A very negative number here indicates poor modeling, so we eliminate negative entries, and produce a number between [0,100], where 100 is perfect structure, and values in the range of 50-80 indicate noisy but acceptable structure. We calculate this over a range of components and keep the maximum attained.
    c. Silhouette [7]: This score judges how well a CPD decomposition clusters an entity (e.g., rail company) when forced to produce hard clustering assignments by choosing the maximizing component per instance of that entity [8]. This number is in the range of [-1,1] with -1 indicating poor clustering and 1 indicating perfect clustering. We also calculate this over a range of components and keep the maximum.
4. Given a triplet of scores for each automatically generated tensor dataset, we filter out datasets that do not match certain thresholds per dimension. The datasets which match the criteria (e.g., Fit > 0.8, Core Consistency > 50, and Silhouette > 0.2) are output as candidates for tensor mining, since they are more likely to have high-quality structure within.

Figure 1 shows an example from the USDOT accident report data where we identify a subset of the automatically generated tensors with high quality structure. This result is quite encouraging because it indicates that there is a good number of high-quality datasets that can be automatically generated from the raw accident reports, which can be used to derive actionable insights. In future work we will consider scalability of our proposed method, since currently in order to score each tensor dataset we have to explicitly compute the CPD decomposition multiple times. In recent work [6], we have shown that we can build self-supervised machine learning models in order to learn similar values of interest for tensor data, and we would like to build upon that work in order to accelerate our current proposed method.

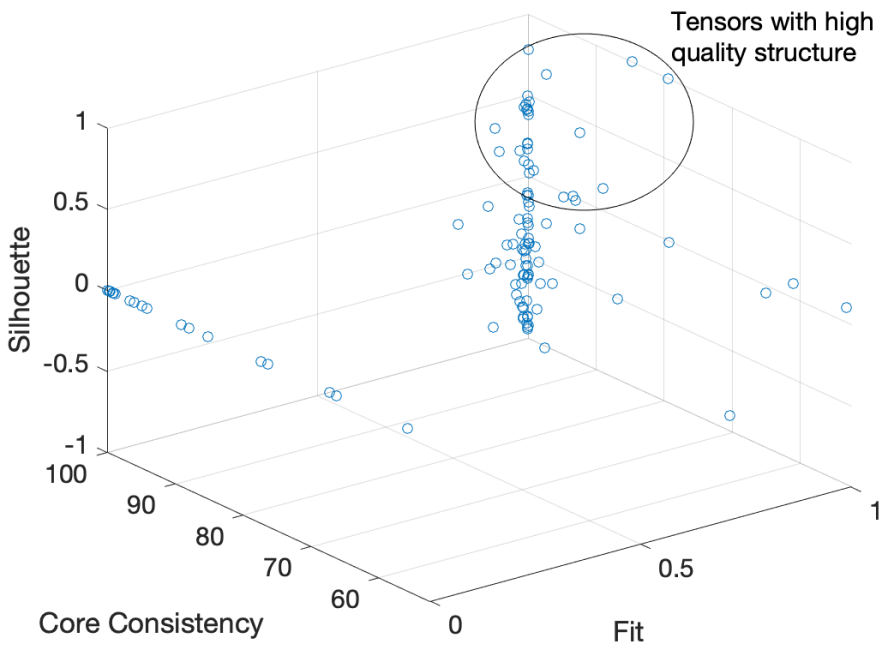

Figure 1: Scoring automatically generated tensor datasets with respect to various quality of structure measures. The circled portion scores reasonably high in all three measures.

Next, we will demonstrate some indicative results from tensor analysis. For this particular tensor, we chose 10 variables of interest:

{'RailroadName',

'ReportYear','CountyName','StateName','HighwayUser','HighwayUserPosition','EquipmentType', 'Visibility','WeatherCondition', 'HighwayUserAction'}

resulting a 10-mode tensor. We should point out here that in the literature of tensor mining, one rarely encounters tensors with more than 4-5 modes, and as the number of modes (or order) of the tensor grows, computational complexity increases as well and some models and computations become prohibitive. Thankfully, the CPD decomposition scales well with the number of modes, and we are able to mine patterns from such high-order tensors using this model.

In this particular tensor we formed, we are expecting to extract co-clusters that relate all 10 dimensions with each other, giving us very fine-grained information about rail companies and their spatio-temporal behavior (over counties and time) as well as different parameters of a given accident. Given the exploratory nature of the proposed analysis, it is unfortunately very hard to

assign quantitative measures of success, since success is measured by the ability of the proposed method to uncover hidden patterns. To that end, we provide a qualitative assessment of the proposed method by observing whether it is able to extract coherent patterns, aiming for a rather objective evaluation of the patterns extracted and avoiding highly-subjective notions of interestingness and importance, which can vary depending on the stakeholder who is consuming the provided patterns.

Figure 2 demonstrates the above point. In particular, we show a 2-dimensional projection of the rail company representation learned by the CPD decomposition using t-SNE. Furthermore, we have colored company cluster assignments with respect to CPD maximum component membership (left subfigure) and by using k-means clustering on that space. (right subfigure) It is clear that our proposed CPD maximum component membership clustering identifies nearly perfectly all coherent structure, whereas traditional clustering fails to do so.

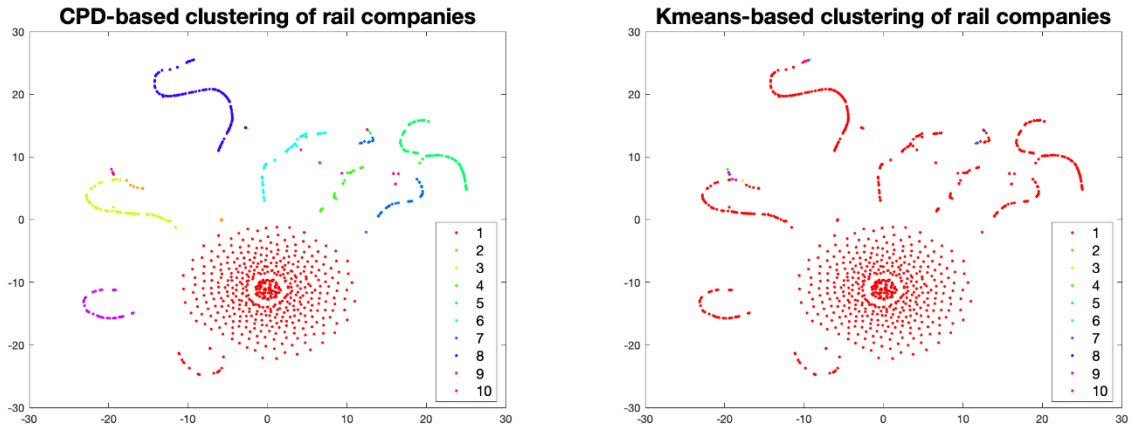

Figure 2: Left: Clustering rail companies using our proposed tensor method. Same color indicates same cluster membership. We observe that even though the clusters are not necessarily linear, our proposed method groups coherent regions of the space successfully. Right: Clustering of companies using traditional K-means clustering, which fails in capturing the non-linear patterns in the space.

We are planning to submit this work for publication to a relevant conference or workshop during Fall 2024.

4. Neural Tensor Decomposition

As we have demonstrated in Sec. 2, tensor decomposition, and particularly the CPD model, are very powerful and expressive in terms of identifying hidden patterns/co-clusters in the data. However, recent advances in tensor methods have brought about the concept of neural tensor decompositions, tensor decompositions that leverage neural networks in order to express highly complex and non-linear patterns in the data which classical methods such as CPD may not be able to. State-of-the-art neural tensor methods have been shown to successfully do so, however, the price to pay is that they forfeit the notion of pattern interpretability that classical models like CPD enjoy, where each component of the decomposition corresponds to a single interpretable pattern/co-cluster in the data.

We have developed NeAT [9], a neural tensor model that starts from first principles and judiciously extends the classical CPD model by introducing independent neural networks for each extracted component. By doing so, each component still stands for an important pattern in the data, while the neural network that is assigned to it allows it to express non-linear patterns. Figure 3 shows pictorially how NeAT readily extends CPD, and thus, inheriting some of its interpretability properties, and how this contrasts to existing neural tensor methods.

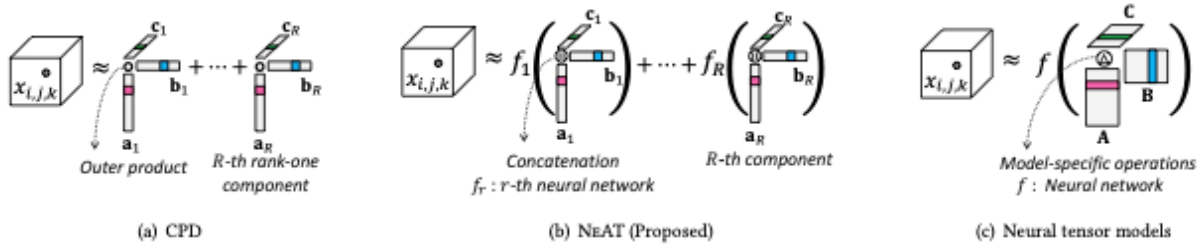

Figure 3: The proposed NeAT method judiciously extends CPD by introducing separate neural networks per component.

This work [9] will appear at the 33rd ACM International Conference on Information and Knowledge Management (CIKM) 2024, a top-tier data science conference.

5. Developing a Tensor-based Framework for Extracting Insights from Rail Crossing Videos

Going beyond public accident reports, in collaboration with PI Khattak from University of Nebraska-Lincoln, we set out to investigate how we can derive actionable insights from video recordings of rail crossings. Video is a very rich modality and essentially captures behavioral patterns of that particular crossing over the course of time. Even though it is possible for an analyst to go over the video and understand different risk factors or other patterns of interest for a single intersection, it becomes a humanly-impossible task when the goal is to understand how different crossings may "behave" similarly over different periods of time, especially when that behavior is time-varying (e.g., two crossings may be very similar during rush hour but may exhibit drastically different risk factors during night time).

In this work we propose to develop a tensor-based framework that can do exactly that: given video recordings of rail crossings in given locale over time, identify groups of crossings that exhibit similar visual behavior (thus potentially exhibit similar risk factors) over a certain period of time. In order to determine similarity of behavior, we are going to rely on 3D Convolutional Neural Networks which are used to process and classify video data in order to compute vector representations of video segments per crossing. We, thus, assume that similarity in that video embedding space correlates with visual similarity, and we are going to restrict the time windows over which that similarity is computed in order to make sure that we eliminate non-stationarity issues.

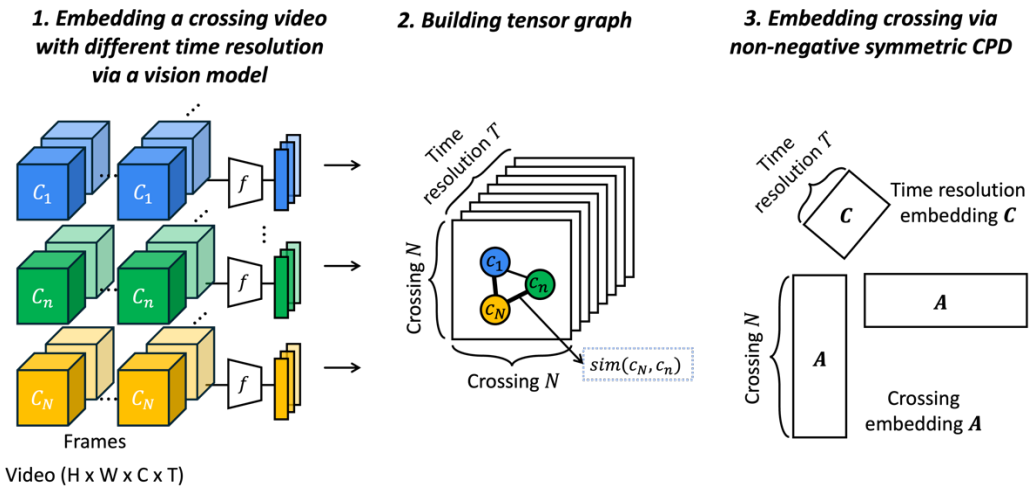

Figure 4: Overview of our proposed framework for mining rail crossing video

Figure 4 shows an outline of our framework: Given a railway video dataset, we consider each crossing as a node in a graph and segment video data into intervals (1 min, 30 min, 1 hr) to create different temporal resolutions. For each segment, we compute vector embeddings using a pre-trained video model. We then calculate the similarity between these embeddings for each time interval, resulting in a temporal tensor graph (crossing, crossing, time). We apply nonnegative symmetric CPD to this tensor graph to identify crossings with similar behaviors.

Currently, we have developed an end-to-end system and we are in close collaboration with UTCRS partners in order to apply our framework to real data and identify meaningful behaviors. We are planning to submit a paper detailing the framework and indicative results to a relevant conference or workshop in the Fall.

6. References

[1] US Department of Transportation. "US Highway Rail Grade Crossing Accident Dataset." https://www.kaggle.com/datasets/yogidsba/us-highway-railgrade-crossing-accident?resource= download. Data retrieved from US DOT.

[2] Villalobos, Ethan, Hector Lugo, Biqian Cheng, Miguel Gutierrez, Constantine Tarawneh, Ping Xu, Jia Chen, and Evangelos E. Papalexakis. "Spectral Clustering in Railway Crossing Accidents Analysis." In ASME/IEEE Joint Rail Conference, vol. 87776, p. V001T05A011. American Society of Mechanical Engineers, 2024.

[3] Villalobos, Ethan, Constantine Tarawneh, Jia Chen, Evangelos E. Papalexakis, and Ping Xu. "Kernel Ridge Regression in Predicting Railway Crossing Accidents." In ASME/IEEE Joint Rail Conference, vol. 87776, p. V001T05A013. American Society of Mechanical Engineers, 2024.

[4] Sidiropoulos, Nicholas D., Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E. Papalexakis, and Christos Faloutsos. "Tensor decomposition for signal processing and machine learning." IEEE Transactions on signal processing 65, no. 13 (2017): 3551-3582.

[5] Papalexakis, Evangelos E. "Automatic unsupervised tensor mining with quality assessment." In Proceedings of the 2016 SIAM International Conference on Data Mining, pp. 711-719. Society for Industrial and Applied Mathematics, 2016.

[6] Shiao, William and Papalexakis, Evangelos E. “FRAPPE: Fast Rank Approximation with Explainable Features for Tensors”, To Appear at Springer Data Mining and Knowledge Discovery 2024

[7] Rousseeuw, Peter J. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." Journal of computational and applied mathematics 20 (1987): 53-65.

[8] Gujral, Ekta, and Evangelos E. Papalexakis. "Smacd: Semi-supervised multi-aspect community detection." In Proceedings of the 2018 SIAM International Conference on Data Mining, pp. 702-710. Society for Industrial and Applied Mathematics, 2018.

[9] Ahn, Dawon and Saini, Uday Singh and Papalexakis, Evangelos E. and Payani, Ali “Neural additive tensor decomposition for sparse tensors,” in 33rd ACM International Conference on Information and Knowledge Management. ACM, 2024.