



Universiteit
Leiden
The Netherlands

Network analysis methods for smart inspection in the transport domain

Bruin, G.J. de

Citation

Bruin, G. J. de. (2023, November 16). *Network analysis methods for smart inspection in the transport domain*. SIKS Dissertation Series. Retrieved from <https://hdl.handle.net/1887/3656981>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/3656981>

Note: To cite this publication please use the final published version (if applicable).

Network Analysis Methods
for Smart Inspection
in the Transport Domain

Gerrit Jan de Bruin



Human Environment and Transport
Inspectorate
*Ministry of Infrastructure
and Water Management*



Universiteit
Leiden

The work in this book was funded by the Ministry of Infrastructure and Water Management through Leiden University. The author used facilities at the Leiden Centre of Data Science, the Leiden Institute of Advanced Computer Science, and the Innovation and Data Lab of the Human Environment and Transport Inspectorate (part of the Ministry of Infrastructure and Water Management).



SIKS Dissertation Series No. 2023-21

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Copyright ©2023 Gerrit Jan de Bruin

Without written permission of the author, no part of this thesis may be reproduced, stored, or published in any form.

Typesetting & Figures: $\text{\LaTeX}$, Gephi, `diagrams.net`, `Matplotlib`, and `Seaborn`.

All software used is open-source.

Open access: <https://scholarlypublications.universiteitleiden.nl/>

Cover design: Wolters Vormgeving

Printed by: Gildeprint, Enschede, the Netherlands, <https://www.gildeprint.nl/>

ISBN: 978-94-6419-967-3

Network Analysis Methods for Smart Inspection in the Transport Domain

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op donderdag 16 november 2023
klokke 13.45 uur

door

Gerrit Jan de Bruin
geboren te Amersfoort
in 1993

Promotor:

Prof.dr. H.J. van den Herik

Copromotores:

Dr. F.W. Takes

Dr. C.J. Veenman

Promotiecommissie:

Prof.dr. H.C.M. Kleijn

Prof.dr. T.H.W. Bäck

Dr. W.A. Kusters

Prof.dr. J.N. Kok (Universiteit Twente)

Prof.dr. N.V. Litvak (Universiteit Twente)

Preface

This thesis is part of an extensive collaboration between the Dutch Ministry of Infrastructure and Water Management (I&W) and Leiden University. My first personal encounter with I&W was during my master's study in Analytical Chemistry. I was determined to work on a subject with societal relevance, leading me to Jasper van Vliet, who worked at the Inspectie Leefomgeving en Transport (ILT), part of I&W. The aim of my master thesis was to conduct efficient compliance monitoring of cargo ship's fuel by using the information from chemical sensors. The thesis made it to a letter to the parliament [88, 89]. Afterward, Jasper invited me to participate in a new Ph.D. project of the Ministry. In this project, the ambition is to arrive at intelligence-led vehicle inspection by risk assessments. Two research directions were launched to explore the risk assessment of vehicles: (1) the application of machine learning techniques and (2) the application of network science techniques.

During my Ph.D. research, many developments occurred related to the topics of the thesis. I would like to mention two specific events that have had an impact on my research. The first one is the introduction of the General Data Protection Regulation (GDPR) in 2018. This law requires the use of transparent models that allow for an explanation of the results achieved. The new law led to increased awareness of the importance of fair data and fair models. The second event is the upset of the Dutch childcare benefits ("De Toeslagenaffaire"). In 2019 it became painfully clear how things can go wrong when authorities are (1) relying on biased data and (2) using models that are not validated fairly. I will address the two points (biases and non-validated models) in my thesis, although I work with non-personal data. They are in particular relevant for the proposed procedure to implement a smart inspection of cargo ships in Chapter 6.

Working both at Leiden University and at the ILT allowed me to interact with the wonderful world of academia and to stay in close contact with a governmental organization that makes a big impact by ensuring safe transportation and reducing the environmental pollution in the Netherlands. This enriching combination has helped me to create this thesis, for which I am grateful.

Gerrit Jan de Bruin, Utrecht, March 8th, 2023

Contents

Preface	vii
Contents	xi
List of Abbreviations	xv
List of Definitions	xvii
List of Figures	xix
List of Tables	xxi
1 Introduction	1
1.1 Smart vehicle inspection	2
1.2 Networks	6
1.2.1 Assortativity	7
1.2.2 Clustering	8
1.2.3 Community structure	9
1.2.4 GC, sparseness, small-world, and scale-free properties	9
1.3 Temporal networks	10
1.4 Link prediction	11
1.5 Transport networks	13
1.6 Problem statement and research questions	15
1.7 Research methodology	17
1.8 Thesis overview and contributions	17

2	Supervised link prediction in large-scale temporal networks	23
2.1	Link prediction	24
2.2	Related work on link prediction	26
2.3	Preliminaries	27
2.3.1	Notation	27
2.3.2	Real-world network properties and their measures	28
2.3.3	The goal of a supervised link prediction model	28
2.4	Chapter research methodology	29
2.4.1	Features	29
2.4.2	Supervised link prediction	33
2.5	Data and the statistics used	33
2.6	Experiments	35
2.6.1	Experimental setup	36
2.6.2	Improving prediction with temporal information	36
2.6.3	Link prediction performance and networks structure	40
2.6.4	Enhancement of performance with past event aggregation	45
2.6.5	Node- and edge-centered link prediction	45
2.7	Chapter conclusions and outlook	47
3	Performance of split strategies in link prediction	49
3.1	Machine learning methods on networks	50
3.2	Related work on validation of link prediction models	51
3.3	Chapter research methodology	52
3.3.1	Link prediction	52
3.3.2	Splitting strategies	53
3.3.3	Features	55
3.3.4	Tree-based gradient boost classifier	56
3.3.5	Performance metric: Average Precision	57
3.4	Properties of the six temporal networks	57
3.5	Experimental setup	58
3.5.1	Distance selection	59
3.5.2	Time intervals	59
3.5.3	Training and testing	59
3.5.4	Improved performance	59
3.5.5	Robustness checks	60
3.6	Results of the two different splitting strategies	60
3.7	Chapter conclusion and outlook	62

4	Understanding dynamics of truck co-driving networks	65
4.1	Co-driving network	66
4.2	Relevant related work on dynamics in networks	67
4.3	Truck mobility data	68
4.4	The co-driving network	68
4.4.1	Procedure to obtain intentional co-driving events	69
4.4.2	Determining maximal time interval between co-driving trucks	69
4.4.3	Network construction	70
4.4.4	Network statistics	71
4.5	Chapter research methodology	72
4.5.1	Link prediction	72
4.5.2	Features	73
4.5.3	Classifier	75
4.5.4	Class imbalance	75
4.6	Experimental setup and results	76
4.6.1	Experimental parameter setup	76
4.6.2	Results	76
4.7	Chapter conclusions and outlook	77
5	Understanding behavioral patterns in truck co-driving networks	79
5.1	Truck co-driving network	80
5.2	Related work on understanding behavioral patterns from networks	81
5.3	Network construction	82
5.3.1	Truck observation data	82
5.3.2	Selection of systematic co-driving events	82
5.3.3	Co-driving network and node attributes	83
5.3.4	Two validation metrics	83
5.3.5	Regional co-driving network	85
5.4	Chapter research methodology	86
5.4.1	Understanding co-driving behavior by assortativity	86
5.4.2	Understanding co-driving behavior by community structure	86
5.5	Analysis of co-driving behavior	87
5.5.1	Network statistics	87
5.5.2	Assortativity	89
5.5.3	Average maximal community assortativity	89
5.6	Chapter conclusion	91

6	Fair automated assessment of noncompliance in cargo ship networks	93
6.1	Smart cargo ship inspection	94
6.2	Related work on ship risk profile	96
6.3	Cargo shipping data	97
6.3.1	Port calls	98
6.3.2	Inspections	98
6.3.3	Merging port calls and inspections	98
6.4	Chapter research methodology	100
6.4.1	Cargo ship network	100
6.4.2	Feature engineering	101
6.4.3	Fair random forest classifier	103
6.4.4	Performance measures	104
6.4.5	Fairness measures	104
6.5	Results	105
6.5.1	Experimental setup	105
6.5.2	Cargo ship network	106
6.5.3	Performance of the baseline ship risk profile	106
6.5.4	Performance of the random forest classifier	109
6.5.5	Performance of the fair random forest classifier	110
6.5.6	The effect of the orthogonality parameter	110
6.5.7	The effect of the orthogonality and threshold quantile together	112
6.6	Discussion on limitations	112
6.7	Chapter conclusions	113
7	Conclusions	117
7.1	Answers to the research questions	117
7.2	Answer to the problem statement	119
7.3	Future research directions	120
	References	123
	Summary	141
	Samenvatting	145
	Curriculum Vitae	151
	Publications	153
	Acknowledgments	157
	SIKS dissertation series	161

List of Abbreviations

AA	Adamic-Adar.....	24, 29, 30, 32
Aminer	Arnetminer.....	35
ANPR	Automatic Number-Plate Recognition.....	13, 68, 73, 82
AP	Average Precision.....	57, 60, 61
AUC	Area Under the Receiver Operating Characteristic Curve.....	33, 36, 38, 45, 57, 76, 103–105, 108–110, 113
CN	Common Neighbors.....	24, 29, 30, 32, 56
DBLP	Digital Bibliography & Library Project.....	xix, 30, 31, 34, 37, 39, 43, 44
FOC	“Flag Of Convenience”.....	98
GC	Giant Component.....	xx, 6, 9, 10, 57, 71, 72, 80, 81, 83, 84, 87–89, 91, 106, 107
GDPR	General Data Protection Regulation.....	vii
GPS	Global Positioning System.....	66
HPLP	High-Performance Link Prediction.....	55
I&W	Infrastructure and Water Management.....	vii, 13, 82
ILT	Inspectie Leefomgeving en Transport (English: Human Environment and Transport Inspectorate).....	vii, viii, 1, 2, 5
IMO	International Maritime Organization.....	98
JC	Jaccard Coefficient.....	24, 30, 32
KONECT	Koblenz NETwork CollecTion.....	35
MoU	Memorandum of Understanding.....	94, 98, 113
NIR	New Inspection Regime.....	94–96
NL	the Netherlands.....	85
NN	Number of Neighbors.....	56
NSI	Netherlands Shipping Inspectorate.....	2
PA	Preferential Attachment.....	24, 30, 32
PF	PropFlow.....	56
PSC	Port State Control.....	94

RDW	Rijksdienst voor het Wegverkeer (English: Netherlands Vehicle Authority)	85
ROC	Receiver Operating Characteristic	xv, xx, 33, 76, 77
SCAFF	Splitting Criterion Area under the curve for Fairness	103, 106
SNAP	Stanford Network Analysis Project	35
SP	Shortest Paths	56
WiFi	Wireless Fidelity	66
WTF	Weighted Temporal Features	32
XGBoost	eXtreme Gradient Boosting	36, 38, 39, 56, 59

List of Definitions

1	Data-driven assessment for inspection	2
2	Smart inspection	3
3	Model validation	3
4	Model testing	3
5	Fair model	4
6	Interpretable model	5
7	Network	6
8	Assortativity	7
9	Temporal network	10
10	Link prediction	11
11	Overfitting	13
12	Co-driving	14
13	Cargo ship network	15

List of Figures

1.1	An example of an ego network	8
1.2	Community structure of a social network	8
1.3	The well-known Zachary karate club social network	12
1.4	The Weigh-In-Motion system	14
1.5	The relation between this thesis's three topics and chapters	18
2.1	Mapping of three weighting functions for the DBLP network	31
2.2	Number of nodes and edges of the 26 temporal networks	35
2.3	Link prediction performance of the 26 temporal networks	38
2.4	Correlations between network properties and performance	41
2.5	Degree assortativity and link prediction performance	41
2.6	Link prediction performance with and without past event aggregation	46
2.7	Link prediction performance of node- and edge-centered features	46
3.1	Procedure to obtain instances for the binary link prediction	54
3.2	Two different strategies exist to obtain disjoint and independent sets	54
a	Random split	54
b	Temporal split	54
3.3	Precision-recall curves of the AskUbuntu network for robustness checks	61

4.1	Monthly variations in truck registrations	68
4.2	Histogram of number of registrations per truck	68
4.3	Frequency distribution of Δt for both intentional and random co-driving	70
4.4	Number of trucks driving between a pair of co-driving trucks	70
4.5	Degree and strength distribution of co-driving cargo truck network	72
a	Degree distribution	72
b	Strength distribution	72
4.6	The ROC curve of the random forest link prediction classifier	77
4.7	The Gini feature importance of the various feature sets	77
5.1	Summary statistics of cargo truck data	84
a	Probability distribution of the number of measurements per truck	84
b	Time interval that truck re-occur	84
5.2	Statistics of the co-driving cargo truck network	84
a	Number of nodes in the network and its GC	84
b	Number of edges in the network and its GC	84
c	Density of the network and its GC	84
d	Diameter and average shortest path length in the GC	84
5.3	Validation metrics for establishing systematic co-driving	85
a	Distributions of the difference in speed Δv	85
b	The fraction of trucks driving on different lanes	85
5.4	Degree and weight distribution of the co-driving cargo truck network	88
5.5	Properties of the communities	90
6.1	Fraction of ships being noncompliant per country	99
6.2	Number of ships registered to each country	99
6.3	Fraction of ships for each flag state being noncompliant	100
6.4	The considered cargo ship network	107
6.5	Confusion matrices	108
6.6	Performance and fairness of proposed ship selection classifier	111
6.7	Fairness measures evaluated on the proposed classifier	112

List of Tables

2.1	Summary statistics of the 26 temporal networks	34
2.2	Link prediction performance of the 26 temporal networks	37
2.3	Link prediction performance with past event aggregation	39
2.4	Degree assortativity of networks after rewiring	43
2.5	Link prediction performance after rewiring	44
3.1	Summary statistics of the six temporal networks	57
3.2	Link prediction performances for different split strategies	61
4.1	Nine statistical properties of the co-driving cargo truck network	71
4.2	The features of the link prediction classifier and their importance	74
4.3	Overview of available truck information	74
5.1	Statistics of the full and regional co-driving cargo truck networks	88
5.2	Calculated assortativities of the full and regional truck co-driving network . . .	90
6.1	Summary statistics of considered cargo ship networks	107
6.2	Performance and fairness measures for the different models	108

