Network analysis methods for smart inspection in the transport domain

Bruin, G.J. de

1 Introduction

Transportation supports our modern global economy like never before. Millions of vehicles, such as ships, planes, trains, and trucks, allow for truly worldwide trade [180], for most humans increasing welfare to levels previously thought to be unreachable [169]. However, the global transportation system also has its challenges; several dangers may come with the modern way of transporting goods and people, such as (1) environmental pollution, (2) culpable accidents, and (3) labor exploitation [88]. Reducing the severe risks involved is of utmost importance. Policy makers have recognized the need to limit transportation risks; therefore, national laws and international treaties have been developed to make transportation as safe and clean as possible [55]. The mere existence of laws and treaties does not immediately eliminate all of these dangers because vehicle operators may choose not to comply with legislation. Therefore, government inspectors periodically check vehicles to ensure compliance. Examples of noncompliant dangerous behavior include lack of safety training and disregarding rest periods (dangerous to humans) or lack of waste treatment (dangerous to, e.g., the environment and wildlife). Inspectorates have the job of ensuring compliance in the transport domain.

In the Netherlands, it is the responsibility of the Human Environment and Transport Inspectorate, in Dutch "Inspectie Leefomgeving en Transport (ILT)", to inspect vehicles and their operators. The inspectorate monitors 160 different policy issues and takes enforcement action when necessary [90]. Examples of issues are (1) the quality of fuel used in vehicles, (2) working conditions for transport personnel, and (3) illegal dumping of waste. Well-functioning inspectorates make a country a healthier, happier, cleaner, more prosperous, and safer place to live [136].

The remaining part of this introductory chapter is structured as follows. We start by exploring smart vehicle inspection in Section 1.1. At the end of this section, we introduce our contribution in the form of automated techniques that help ensure smart vehicle inspection. In Section 1.2, we introduce networks, a powerful model for achieving this task. Section 1.3 dives into one specific representation of a network where temporal

information is available, i.e., the temporal network. After that, in Section 1.4, we investigate the prediction of new network links as an approach to better understand the network's dynamics. Subsequently, in Section 1.5, we focus on the characteristics of the data used throughout the thesis, being transport networks. In Section 1.6, we provide the problem statement and research questions. The research methodology is presented in Section 1.7. Finally, an overview of the thesis is provided together with our contributions in Section 1.8.

1.1 Smart vehicle inspection

A major challenge for inspectorates is achieving maximum compliance towards legislation with finite inspection capacity [136]. For example, the cargo shipping industry is responsible for around 80% of global trade movements [190]. Historically, shipping inspectorates selected a random sample from all ships entering a port, such that all ships have an equal probability of being inspected. As a result, an inspectorate with a limited number of inspectors would only sporadically encounter noncompliant behavior at a ship, assuming that noncompliance is rare. Hence, ship owners might think there is no need to comply with legislation because noncompliant behavior is unlikely to be noticed. It can result in neglecting safety procedures and, therewith, more dangerous behavior.

Many inspectorates are limited in the number of inspectors they can employ. In the Netherlands, the Netherlands Shipping Inspectorate (NSI, part of ILT) can only inspect twelve ships per week [87, 146], while over 500 merchant ships arrive weekly in the port of Rotterdam alone [160]. Therefore, many inspectorates (including Netherlands Shipping Inspectorate (NSI)) are looking for innovative methods to maximize compliance and thereby minimize riskful behavior. One way of doing so is by improving the assessment procedure of vehicles for inspection so that more time can be spent on noncompliant vehicles. Traditionally, rule-based systems are considered to this end. The rules in these systems are based on expert knowledge. In this work, we consider the use of historical data to obtain better assessments of vehicles. Inspectorates performing data-driven assessment for inspections, as defined in Definition 1, are more likely to find noncompliant vehicles and are thus more effective in detecting dangerous behavior [45, 86, 137]. In this thesis, the terms "inspectorate" and "inspections" will be used solely in the context of the inspection of vehicles and their operators.

> **Definition 1.** *Data-driven assessment for inspection* is the process that uses (historical) data to determine what entities are likely associated with noncompliant behavior and thus need an inspection.

Taking the assessment procedure of vehicles for inspection one step further means that we not only make data-driven assessments (which may still involve human decisions) but require the assessment to be done in a so-called *smart* way, as detailed in Definition 2.

> **Definition 2.** *Smart inspection* is performed when a data-driven approach is taken to assess vehicles likely associated with noncompliant behavior in an accurate, automated, fair, and interpretable way.

Doing smart inspection ensures that vehicle owners are motivated to comply with legislation because they know that noncompliance will likely result in inspections and subsequent fines or legal consequences.

We briefly explain what we consider an (1) accurate, (2) automated, (3) fair, and (4) interpretable assessment in this paragraph and describe the last three aspects in more detail in the following subsections. While an accurate assessment is a logical consequence of an adequately performing machine learning model, the other three aspects deserve further elaboration.

- An *accurate* assessment is an assessment that closely matches the true outcome.
- An *automated* assessment is performed without human intervention and can automatically adjust to new data.
- A *fair* assessment does not discriminate towards sensitive characteristics.
- In an *interpretable* assessment, the entire approach, including how it arrives at an assessment, is clear to humans.

Automated assessment

Ideally, vehicle assessment for inspection should be performed in an *automated* manner, considering many vehicles in a limited time, with little time-consuming human intervention. This moves away from the classically considered rule-based approach, in which solely human intelligence is used. In the current work, we consider *machine learning* methods for the assessment process. Machine learning is the process of learning (or equivalently, training) a model from examples of data represented by characteristic features [17, 76, 137]. The learned model can then make predictions about new (unseen) data. Features refer to characteristic properties of the examples provided to the machine learning algorithm. Engineering these features is an essential step in machine learning and can significantly affect the performance of a model. In the case of transport vehicles, features include vehicle characteristics, such as country of registration or maximal transport capacity. A machine learning model should be validated (Definition 3) and tested (Definition 4) to make sure to assess its performance.

> **Definition 3.** *Model validation* is the process of evaluating the performance and reliability of (possibly multiple) models on unseen data to select the best-performing model [76].

> **Definition 4.** *Model testing* is the final process of evaluating the performance of a model on unseen data after the model is fully trained [76].

Model validation and testing are done by dividing the examples into disjoint sets of data, usually the (1) train, (2) validation, and (3) test set [76]. A machine learning algorithm

then uses (1) the train set to learn the model, (2) the validation set to perform model validation, and (3) the test set to do model testing.

The power of machine learning models lies in their ability to easily identify trends and patterns in the data that are too complex for humans to find. Moreover, and especially useful in our setting, machine learning models can handle more vehicles than humans. The assessment of vehicles for inspection is thus ideally performed by a machine learning model.

Fair assessment

Assessment of vehicles should be performed in a fair manner (called fair assessment) to prevent discriminatory use of sensitive features. In our setting, sensitive features are properties of vehicles that the model should not consider as features for the model because of, e.g., legal restrictions or ethical considerations. An example specific to the transport domain is the registration country of a vehicle. There are at least two reasons why it is undesirable to use the registration country.

First, some countries are subject to more rigorous inspections than others. Therefore, historical data can be biased toward certain registration countries. This bias can occur when the inspection process is not standardized across all countries, leading to unequal levels of scrutiny. Second, vehicle operators themselves can initiate changes in the country of registration, thereby influencing the assessment outcome [131].

Now that we have established that one or more sensitive features can be present in the data, we mention two ways to limit the use such of sensitive information and thus arrive at a more fair model for the assessment of vehicles.

The first way is to disregard any sensitive information altogether. A clear advantage is that the sensitive information itself cannot be used to make a prediction. A significant drawback is that the sensitive information may correlate with non-sensitive information, resulting in the indirect use of sensitive information [13]. The second way is to use models that can produce fair assessments by special treatment of sensitive information, further detailed in Definition 5 below.

> **Definition 5.** A *fair model* produces assessments that do not discriminate towards characteristics of the example that are deemed sensitive.

Fair models minimize the negative outcome for sensitive groups by so-called decorrelation of assessments with sensitive information [75, 96, 157, 214]. Decorrelation is the reduction of correlation between sensitive information and the predicted outcome of a model. Moreover, recently developed methods allow users to tune the fairness-performance trade-off by controlling the level of decorrelation with the sensitive information. As such, these models can prevent sensitive information from being used in the assessment from being exploited, ensuring similar outcomes for the sensitive and non-sensitive groups. They urge careful consideration of the balance between performance on the one hand

and the restricted use of sensitive information on the other hand. In Chapter 6, we use a fair model and describe how fairness can be quantified.

The country of registration is not the only feature that may be deemed sensitive. In general, vehicle operators can manipulate static administrative information to arrive at a more favorable risk assessment. Examples of *static administrative* information include insurance company, vehicle's type, size, and construction year. In contrast, *behavioral* information, which is dynamic in nature, is more resilient to this type of manipulation. A good example of behavioral information used in our work is spatiotemporal information about the itineraries of vehicles. Expectedly, behavioral information, and not administrative information, is more indicative of riskful behavior.

Multiple ways to take behavioral information of a vehicle over time exist, such as time series analysis [73] and reinforcement learning [184]. Our work explores the use of *networks* (see Definition 7). Multidisciplinary studies repeatedly show that network-driven approaches can often reveal otherwise hidden complex patterns and properties that signal meaningful phenomena in the real world [6, 10, 21, 129, 183]. In this work, networks enable us to explicitly model vehicle relations, considering interactions between these vehicles as part of the national or global transportation system [164, 208]. In Section 1.2, we further explore and define the necessary network concepts and properties relevant to the transport domain.

Interpretable assessment

A challenge of most commonly used machine learning models is that their predictions are difficult to understand. It can be hard for humans to comprehend how multiple factors affect the inner workings of machine learning methods. As a result, people may perceive limited transparency [121]. Governmental organizations that motivate how they make their decisions and what data underpins these decisions (called interpretable assessment) are more trusted by society [198]. Hence, *interpretable* machine learning models (Definition 6) should be preferred in the inspection domain [122].

> **Definition 6.** An *interpretable model* is a model that allows humans to understand (1) what procedures were followed to make the model, (2) the inner workings of the model, and (3) how the model arrives at its predictions [8, 120, 123].

Our contribution towards smart vehicle inspection

While clearly within reach, full implementation of smart inspection has yet to be achieved. In the case of the ILT, a desire to become more data-driven has been expressed; however, this needs to be sufficiently translated into inspection practice. This thesis examines how data on vehicle behavior can be leveraged to better understand contemporary problems in the transport domain, focusing on the smart inspection of vehicles (Definition 2). In

particular, we model vehicle behavior by making use of networks. In addition to addressing several fundamental problems related to the analysis of networks, we use networks modeling vehicle behavior in machine learning approaches for the accurate, automated, and fair assessment of vehicles. By doing so, we (1) provide a novel approach toward the assessment of vehicles for smart inspection and (2) obtain a better understanding of the dynamics of the global transportation system. Ultimately, our findings will contribute to a safer and healthier environment [136].

1.2 Networks

We deem networks to be a suitable data model to capture complex patterns in the behavior of vehicles, with the ability also to capture temporal aspects (as further discussed in Section 1.3). We start by defining networks and related concepts and subsequently discuss seven commonly observed properties of networks useful for understanding the data modeled by these networks. These properties are leveraged in the data-driven approach taken in this thesis toward the accurate, automated, fair, and interpretable assessment of vehicles, i.e., smart inspection.

The field of research that, in a general sense, concerns itself with methods for discovering knowledge from real-world systems modeled as networks is referred to as *network science* [10]. We define a network in Definition 7.

> **Definition 7.** A *network* is a set of entities called nodes combined with a set of edges (or, equivalently, links) that connect pairs of nodes.

Nodes connected by an edge are said to be *adjacent* and are also called *neighbors*. For the remainder of the introduction, we assume that the edges in the network are *undirected* and *unweighted*. Some concepts slightly change when considering *directed* edges; this will be explained in the relevant chapters where needed. A node's *degree* is its number of neighbors. Nodes with a large degree are also called *hubs* and are often deemed to have a central role in the network. Two nodes are connected when there exists a *path* between these nodes; a path is a sequence of edges linking a series of nodes.

A *component* is a subset of nodes and edges for which it holds that (1) there is a path between all pairs of nodes in the component and (2) it is not part of any larger component. A network can consist of multiple components. With respect to the components, we introduce three new concepts. First, we frequently analyze the largest component of a network, commonly referred to as the *Giant Component (GC)*. Second, in a component, the *shortest path* is a path which uses a minimum number of edges to connect a pair of nodes. The length of the shortest path (called *distance*) equals the number of nodes involved minus one. Thus, two adjacent nodes have a distance of one $(2-1)$ to each other. Third, the *diameter* is the maximum distance between any pair of nodes in a component.

Social networks

A typical type of network often investigated is the *social* network, which is studied in many different disciplines, such as psychology, sociology, and mathematics. A node marks a person in these networks, while an edge indicates (for example) acquaintance. A figurative sketch of a so-called *ego network* of person D is given in Figure 1.1. An ego network consists of the individual node, its immediate neighbors and the edges connecting those neighbors. Node D has a degree of six and is part of three *triangles* (1: nodes A, B, D, 2: nodes B, C, D, and 3: D, E, G). A triangle is formed when three nodes are fully connected, i.e., have three edges between them. Like in many real-world networks, the ego network shown in Figure 1.1 contains extra contextual *node attributes*. In the figure, gender or profession (both indicated by the outfit) are examples of node attributes. The type of acquaintance (work, sport, or housemate) is considered an *edge* attribute. We can identify many types of networks in the real world, including information networks [108, 139], the aforementioned social networks [51, 119, 173, 174], technical networks [148], and transport networks [16, 94]. In the latter type of network, nodes are vehicles. We will explain this type of network in detail in Section 1.5.

In Subsections 1.2.1 to 1.2.4, we discuss seven common concepts to get a comprehensive understanding of networks. These concepts will prove relevant in the remainder of the thesis. Although we explain these concepts using examples from social networks, the discussed measures can be applied to any network.

1.2.1 Assortativity

The first concept is assortativity, defined in Definition 8.

> **Definition 8.** *Assortativity* refers to the inclination of nodes to link with other nodes that share similar (or dissimilar) characteristics [127].

The numeric value of assortativity is equal to the correlation coefficient (i.e., the Pearson coefficient) of the characteristics of linked nodes. Positive values for this measure indicate that neighboring nodes share a similar characteristic [10]. A value of zero indicates that there is no assortativity. Negative values indicate that nodes share dissimilar characteristics.

In social networks, the *degree* assortativity is usually observed. A positive value indicates that most people are acquainted with people with a similar number of friends [130]. It is well-known that celebrities often befriend other celebrities (i.e., the hubs in social networks). More specifically, marriages often occur between famous people, much more often than we would expect based on chance alone [10]. Generally, this strong degree assortativity is present in many more types of networks. We show a small example of assortativity in Figure 1.2; the two nodes with the highest degree (indicated in black) link to other nodes with a relatively large degree. In turn, the nodes with the smallest

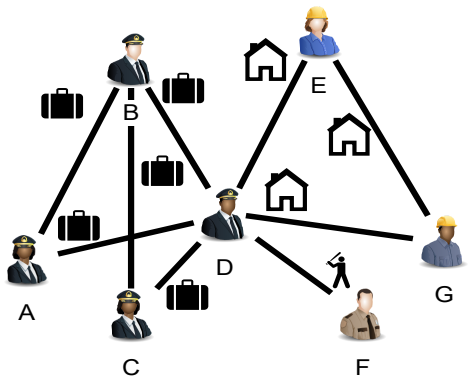

Figure 1.1: An example of an ego network (here part of a social network).

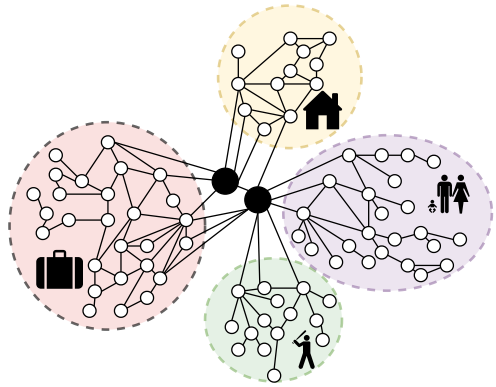

Figure 1.2: Community structure of a social network.

degree mainly connect to nodes with low degrees. Social networks generally also show assortativity in terms of age and race [22].

In contrast, in degree *disassortative* networks, nodes with a high degree are more likely linked to nodes with a lower degree. An example of a degree disassortative network is the topology of the internet [132], where hubs (servers, also called autonomous systems) frequently link to low-degree nodes (individual machines). The degree disassortative nature of such a network has consequences for how resilient the internet is towards failures. A breakdown of a limited number of hubs can prove disruptive to the overall connectivity of the network. Exactly this happened in recent times; major outages of the internet occurred in November 2020 [68], July 2021 [199], and June 2022 [182], because some hubs failed.

In Chapter 2, the measure of degree assortativity is used to characterize the structure of networks. Moreover, assortativity is used in Chapter 5 to understand truck behavior.

1.2.2 Clustering

The second concept commonly observed in real-world networks is *clustering*. Particularly in social networks, people tend to organize themselves in tightly knitted groups, so-called cliques [80, 197]. The *clustering coefficient* quantifies the extent to which clustering is present.

In particular, the *node clustering coefficient* quantifies the fraction of triangles that exist compared to how many triangles could exist between a node's neighbors. For example, in Figure 1.1, we observe that node D is part of three triangles. When all neighbors of node D are connected, there are $6 \cdot 5/2 = 15$ triangles, meaning that the local clustering coefficient of node D is $3/15 = 0.2$.

In Chapters 2, 5 and 6, the clustering coefficient is used to characterize the structure of networks.

1.2.3 Community structure

The third concept is that numerous networks possess a clear *community* structure. Communities are groups of nodes that are densely linked amongst each other but sparsely linked with other communities [64]. While defined above purely based on network structure, it is repeatedly observed that communities correspond to nodes sharing some property in real-world settings. In the social network depicted in Figure 1.2, each community is indicated by a colored area and a pictogram indicating the correspondence with the social groupings by household, interest, neighborhood, or profession.

Many methods for *community detection* exist [58, 76, 106]. In community detection, the goal is to optimally split the network into communities. When contextual information is known, it may be utilized to find the right communities. However, often we wish to find communities in an automated way by only using the network topology. A way to achieve this is by optimization of the so-called *modularity* measure [19, 186], which is typically computed as the difference between the actual number of edges within a community and the expected number of edges within the communities assuming the connections between the nodes were randomly created [25]. Suppose the optimization process obtains a high modularity value. In that case, nodes within the discovered communities are more connected to each other than they are to nodes in other communities. Hence, a strong community structure is likely present. Likewise, when a low modularity value is obtained, this often indicates that the network does not have a strong community structure [128].

In Chapter 5, assortativity (as discussed in Subsection 1.2.1) is used to understand the community structure of a network.

1.2.4 Giant Component, sparseness, small-world, and scale-free properties

We continue with four more common concepts frequently observed in real-world networks. These concepts (numbered 4, 5, 6, and 7) relate to the macro-scale of a network, meaning they can only be observed when considering the overall structure of a network.

4. **Large Giant Component.** Real-world networks often exhibit a Giant Component (GC) spanning the vast majority of all nodes. Throughout this work, we frequently use the GC to ensure all nodes are connected.
5. **Sparseness.** Real-world networks are typically *sparse*, meaning that from all the pairs of nodes that could be linked, relatively few links exist [10]. The sparseness of links in networks has implications for predicting new links, which we will discuss further in Section 1.4.
6. **Small-world.** Small-world networks are networks where nodes can typically reach each other using a shortest path of small length [119, 197]. The *average path length* or *average distance* of a component is equal to the average length of a *shortest* path (i.e., the distance) between all pairs of nodes [10]. We characterize networks using

the average shortest path length in the GC in Chapters 4 and 5. The GC in small-world networks tend to have relatively low average distances, even if the overall component is large in terms of the number of nodes and edges [6]. The significance of small-world networks is that they can provide efficient communication between distant nodes while maintaining local connectivity and resilience to node failures.

7. **Scale-freeness.** Many real-world networks are believed to be scale-free, which means many nodes have a relatively low degree, and few nodes have a very high degree. The degrees of nodes in a scale-free network thus lack a characteristic scale, making the degree distribution "scale-free" [195]. Therefore, the notion of scale-free networks is closely related to the presence of hubs. There is some controversy [82, 92] as to whether scale-free networks occur frequently [11, 12, 15, 195] or not [26]. Part of this discussion can be traced back to how closely the degree distribution resembles a power law distribution, lognormal distribution, or other types of skewed distributions [26]. Some scholars consider real-world networks universally scale-free, regardless of the domain of the network and the identity of the nodes [12]. The scale-free structure of many networks also has implications for predicting new links, which we will discuss further in Section 1.4.

1.3 Temporal networks

So far, we have assumed that networks are static, meaning we assume that *all* edges exist at some point in time. Real-world networks usually evolve and are therefore better modeled by a *temporal* network, which we defined in Definition 9.

> **Definition 9.** A *temporal network* is a network in which the edges are associated with a timestamp or time interval [38, 83].

The edges of a temporal network are consequently defined by (1) the source, (2) the target, and (3) an edge attribute containing temporal information on edge formation. In this work, we consider only temporal networks of which the edges are formed at a specific point in time, thus where the third edge attribute in Definition 9 is a timestamp. We do not consider any edge removal. Temporal networks allow for a more in-depth study of the growth mechanisms of a network [5, 52]. For example, a growth process known as preferential attachment can lead to the emergence of aforementioned scale-free networks. *Preferential attachment* is the process where new edges are preferentially linked with nodes that are hubs (i.e., have a high degree) at the time of edge formation [12]. The process will result in a feedback loop in which hubs increase their large degree even further, causing an increasingly skewed degree distribution. Generally, we differentiate between two types of temporal networks [81], viz. networks with (1) persistent relations and (2) discrete events.

First, we have temporal networks modeling *persistent relations* between the nodes. An example of such relations can be found in *acquaintance networks*, where an edge connects

two people if they are acquainted with each other in some way (such as friendship, kinship, or a professional relation) [119]. At most, one edge exists between two nodes in the network, and those edges are assumed to be present indefinitely, i.e., they appear but do not disappear.

Second, we can consider temporal networks modeling *discrete events* [133]. Multiple edges between a pair of nodes can exist, each with its associated timestamp. A *communication network* is an example of a temporal network containing discrete events. Like social networks, the nodes are people, but now the edges consist of communication events, such as calls or messages. Two persons can communicate often; thus, each edge has a distinct timestamp, and many edges may exist between the same two nodes.

1.4 Link prediction

An important task in network science is *link prediction*. It has numerous applications in real-world scenarios, such as spam mail detection in communication networks or friend recommendations in online social networks. The link prediction task is defined differently for varying purposes. In the broadest definition, the task is to predict which links exist between two nodes in a network. These links may be unobserved or even missing. In this definition, link prediction [114] can be employed on *static* networks (meaning no time information is present). Therefore, we call this task *missing* link prediction. However, in this work, we are interested in the *temporal* aspect of this task. Hence, we define link prediction in Definition 10.

> **Definition 10.** *Link prediction* is the task of predicting which links will appear in the future [10, 62].

Link prediction, as defined above, requires the use of temporal networks because links that appear later in time need to be known to train the model. Therefore, it is sometimes also called *temporal* link prediction.

Commonly, the link prediction task is formulated as a machine learning problem. The examples provided to the model consist of all pairs of nodes that are not adjacent in a current network snapshot, being the network consisting of all edges up to a certain point in time. The machine learning model aims to predict whether each currently unconnected pair of nodes is linked in a future network snapshot. Multiple types of features can be utilized to perform this task [54].

An example type is the *similarity-based* feature type, which considers how similar the surrounding network structure of two nodes is. Two typical similarity-based features are (1) the number of *unique* neighbors and (2) the number of *common* neighbors of both nodes. To explain the workings of these features, let us consider the well-known Zachary karate club social network [144, 213], depicted in Figure 1.3. It is a network of different karate club members, with links marking social interactions outside the club.

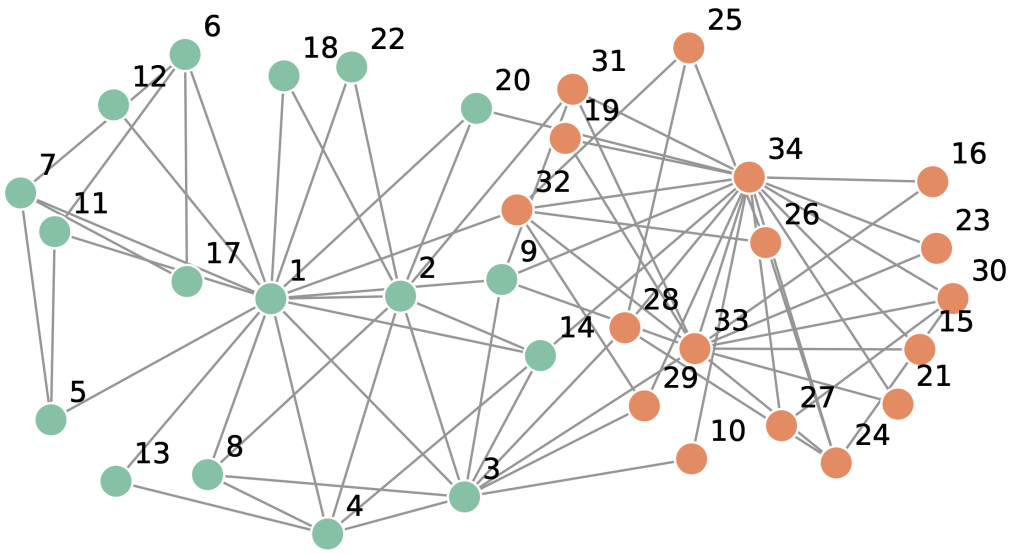

Figure 1.3: The well-known Zachary karate club social network.

Members 16 and 23 have a high similarity, as they both have a degree of two, and all their neighbors are shared. A clear advantage of the machine learning approach towards link prediction is that (1) multiple types of features (provided as input to the model) can be conveniently combined to arrive at a well-performing model [63] and (2) the approach is interpretable when simple topological network features are used [54].

In the remainder of this section, we mention two challenges of link prediction. These challenges will be addressed in Chapters 2 and 3.

The first challenge is that most works in the literature do not consider temporal information associated with the network's edges. Thereby, they ignore the evolution of the network observed so far. Using time-aware measures can improve prediction, but it ignores an essential dichotomous aspect of many temporal networks, namely that two types of temporal networks exist. The two types were discussed earlier in Section 1.3: (1) networks where edges are *persistent relations* and (2) networks where edges mark *discrete events*. We recall that temporal networks with discrete events may contain multiple edges between nodes, each having its own timestamp. This type of temporal network allows the evolution of edges between a pair of nodes to be exploited in the link prediction task.

In Chapter 2, we will show that we can improve link prediction performance when accounting for these discrete events.

The second challenge is that the validation (Definition 3) and the testing (Definition 4) of link prediction models are two nontrivial tasks often overlooked in existing work. It is only after applying proper model validation and testing that we may have sufficient confidence in applying a model in the real world. In particular, validation and testing

can identify *overfitting*, see Definition 11, which causes a too-optimistic performance estimation and, therefore, in machine learning, the well-known warning is that overfitting will reduce the validity of a study.

> **Definition 11.** *Overfitting* happens when a model matches the training data too closely, and the model is not working well on new, unseen data [167].

Obtaining a hold-out validation and using a general test set is impossible for network data because network data is, by definition, "related". Returning to the Zachary karate club social network in Figure 1.3, it is generally agreed that the nodes can be divided into two communities, which is indicated by the color of the node. A rigorous approach would be to sample the green nodes for model learning and the red ones for model validation when applying missing link prediction on the network. However, the ego networks of some nodes, in particular, node 9 and 10, are severely altered when such a sampling step is performed. These alterations could happen at a large scale in real-world networks, making the resulting link prediction model unusable. It is even more problematic for measures based on distance that use more global information beyond a node's direct neighborhood.

In Chapter 3, we explore two different splitting strategies in an attempt to discover how to perform adequate model validation on a collection of real-world temporal networks.

1.5 Transport networks

As explained in Section 1.1, our research examines how we can leverage transport networks to better understand vehicle behavior. We distinguish two different types of network data used throughout this work, being (1) *co-driving trucks* (Definition 12) in Chapters 4 and 5 and (2) *cargo ship networks* (Definition 13) in Chapter 6. Both datasets have national or even international coverage and systematically record nearly all vehicles for a specific period and location. Big datasets like these allow for a complete overview of all transport of that specific type. In both cases, the study of the temporal network aspects allows us to understand the behavior of the trucks (and ships) in relation to all other trucks (and ships). Below, we briefly describe (1) these two transport networks and (2) what we seek to understand from them.

Truck co-driving networks

The Ministry of I&W gathers movements of trucks by Automatic Number-Plate Recognition (ANPR) systems; see Figure 1.4. The systems monitor any vehicle that passes, although some data may be missing, for example, because of misread license plates or avoidance of the cameras. Subsequently, it registers details such as license plate, country of registration, hazardous substances, length, weight, speed, and of course, the time of registration. By

Figure 1.4: The Weigh-In-Motion system.

exploring the data, we aim to learn and better understand what factors influence the trucks to do co-driving, an activity which we define in Definition 12 (and detail further in Chapter 4).

> **Definition 12.** *Co-driving* is the process where two trucks are observed at the same location within a very short time window. *Systematic* co-driving occurs when two trucks drive together frequently (e.g., more than once).

To investigate the process of truck co-driving, we consider the so-called truck co-driving network. We construct a temporal network from all systematic co-driving events by considering every truck as a node, linking two trucks when they show systematic co-driving behavior. Each temporal edge is thus characterized by the two trucks it links and the time period of the systematic co-driving event. The location of the co-driving activity that occurred is included as a spatial edge attribute. This spatiotemporal network is a particular extension of the temporal network, as both time and spatial information are available.

The truck co-driving networks have our interest for two reasons. First, we are interested in the properties of the co-driving network and the comparison with networks from other domains (e.g., social networks). Second, we want to know what communities of trucks are present in the network and what factors contribute to the formation of these communities. Chapters 4 and 5 provide a complete account of our research using the co-driving trucks dataset.

Ultimately, we mention two societal advantages of understanding truck co-driving behavior. First, understanding truck co-driving behavior can help reduce traffic congestion [187]. Moreover, co-driving and therewith platooning trucks can optimize fuel usage because of the aerodynamic drag reduction.

Cargo ship network

The second set of data comprises all port calls of sea-going cargo ships in Europe, including the times of entrance into and departure from the port. These are collected from each port's administrative systems. All inspectorates have access to the same dataset in Europe, and the data can thus be used for smart ship inspection. We capture the behavior of the ships in relation to other ships by considering this dataset as a network. Deriving features from this network allows us to incorporate more information in a machine learning model than we would otherwise capture from the static data. The construction of a so-called cargo ship network (Definition 13) allows for extracting meaningful information for a machine learning model identifying noncompliant behavior of ships.

> **Definition 13.** The *cargo ship network* is a temporal network of all movements of cargo ships between ports [94]. The departure and arrival ports and the time of departure characterize edges.

This network is spatiotemporal as well. The edges have a temporal attribute indicating when the movement occurred. Unlike the truck co-driving network, each node is associated with a location. Relevant to our setting is that the inspectorate keeps records of all ships where noncompliances have been found, which can be used as node attributes and ultimate labels in a machine learning model. The entire study of the cargo ship network is presented in Chapter 6.

1.6 Problem statement and research questions

This section will describe our problem statement and research questions. As explained at the beginning of this introduction, smart inspections (Definition 2) are essential to ensuring a healthy and clean environment. In our work, we consider four aspects (see Section 1.1) of smart inspection and aim to handle them. It lead us to the following problem statement.

> **Problem statement:** *How can network science methods leverage behavioral data for smart inspection of vehicles?*

We subdivide the problem statement into five research questions. The first two questions address fundamental network science challenges, and the last three address more applied questions in the transportation domain. Below we describe the background and rationale behind this subdivision, i.e., our research strategy.

It has previously been observed that not all networks perform similarly in the missing link prediction task (e.g., [63]). In addition, literature so far has not extensively dealt with the relation between *network structure* and *performance* in the (temporal) link prediction task (Definition 10). These two observations leads us to Research question 1.

Research question 1: *What is the relation between network structure and model performance in link prediction?*

Let us now turn to the validation of link prediction models. A common approach to model validation (Definition 3) and testing (Definition 4) on tabular data is to use a hold-out set, i.e., a separate test set to evaluate the model's performance. Such a hold-out set is impossible to obtain for network data because all data are inherently related (see Section 1.4). If the hold-out criterion is not met, it can result in overfitting (Definition 11). We therefore formulate Research question 2 as follows.

Research question 2: *How can we obtain accurate estimates of the performance of link prediction models by using adequate splits into train, validation and test sets?*

Having posed our research questions addressing fundamental network science challenges, we now consider the research questions addressing smart vehicle inspection.

Our exploration of smart vehicle inspection starts by considering the case of the co-driving of trucks. We want to learn what factors contribute to truck co-driving, for reasons explained in Section 1.5. We do so by exploration of the co-driving network, arriving at Research question 3.

Research question 3: *How do network structure and vehicle attributes relate to co-driving behavior?*

We continue with the analysis of the truck co-driving network. For the inspectorate, it is interesting to understand (1) which groups of truck operators show frequent co-driving behavior and (2) what brings the truck operators in these groups together. When inspectorates want to change the behavior of truck operators, they can target specific communities via targeted communication. The question is which community detection model (and what parameter setting) yields the best partitioning into communities to do so. We explore the use of node attribute information to find such an optimal partitioning in Research question 4.

Research question 4: *How can node attribute information be exploited to automatically create a good partitioning of a co-driving network into communities?*

Finally, we proceed to the smart cargo ship inspection. We use information from the cargo ship network to improve the fair assessment of cargo ships for inspections, allowing us to answer Research question 5.

Research question 5: *How can ship behavior be utilized to enable smart inspection of cargo ships?*

Answering these five research questions allows us to deepen our understanding of machine learning methods on network data. The other way around, it improves our understanding of the effects of information on connectivity and relatedness of individual entities, i.e., the network aspect, on machine learning tasks. In turn, this knowledge can

improve the understanding of the behavior of different vehicles. Ultimately, it may enable smart inspection of vehicles, thereby maximizing the impact of new regulations for a sustainable planet.

1.7 Research methodology

We answer the five research questions by the following research methodology, consisting of six phases:

1. We establish the *context* of the question at hand.
2. We collect *relevant literature*.
3. We establish *preliminaries* and set up experiments.
4. We determine what *data* is available and what properties does this data possesses.
5. We *report and discuss* the findings of the experiments.
6. We provide a *conclusion* and suggest *future work*.

Answering the five research questions allows us to formulate an answer to the problem statement in Chapter 7.

1.8 Thesis overview and contributions

Below, we first provide an overview of the thesis and then indicate which research questions are answered within each chapter and what methodology was used.

We can differentiate three topics that our research covers: (1) machine learning, (2) network science, and (3) smart vehicle inspection. Each chapter relates to at least two topics. In Figure 1.5, we present a diagram we coin as a “ranked classification diagram”. It is a Venn diagram with each chapter assigned to one of the three topics above. The ranking aspect comes from the following; a chapter is more closely related to a topic when put nearer the corresponding circle.

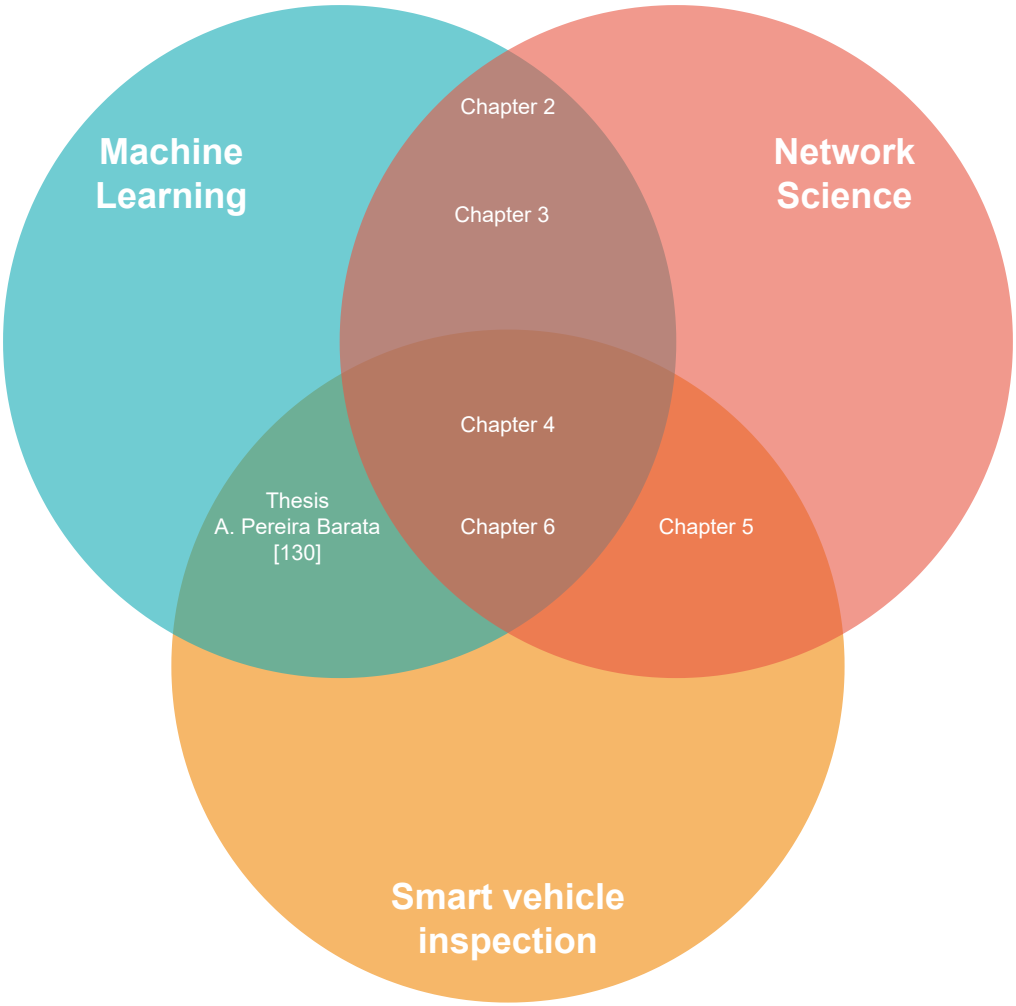

Figure 1.5: The relation between this thesis's three topics and chapters (indicated by a "ranked classification diagram").

Contributions

- In Chapter 2, we address Research question 1. It starts with the topics of machine learning and network science. A large corpus of publicly available temporal networks is gathered. Link prediction is applied to all of them, and the link prediction performance and properties of the temporal networks are systematically investigated. The content of this chapter is based on the work described in:

  G. J. de Bruin, C. J. Veenman, H. J. van den Herik, and F. W. Takes. „Supervised temporal link prediction in large-scale real-world networks." *Social Network Analysis and Mining* 11, 80 (2021). DOI: 10.1007/s13278-021-00787-3.
- Chapter 3 is devoted to Research question 2. Topics covered in this chapter are again machine learning and network science. The topic of smart vehicle inspection is not directly covered, but the evaluation of link prediction strategies is important when used in smart vehicle inspection. We also use a corpus of publicly available temporal networks gathered in this work. Different strategies for link prediction are assessed and evaluated. The content of this chapter is based on the work described in:

  G. J. de Bruin, C. J. Veenman, H. J. van den Herik, and F. W. Takes. „Experimental evaluation of train and test split strategies in link prediction." In: *Proceedings of the 9th International Conference on Complex Networks and Their Applications*. Studies in Computational Intelligence 994. Springer, 2021, pages 79–91. DOI: 10.1007/978-3-030-65351-4_7.
- Chapter 4 is answering Research question 3, thereby covering the topics of network science, machine learning, and smart vehicle inspection. The chapter considers the construction of the truck co-driving network. We analyze the properties of the network and apply link prediction to the network to understand the (social) processes underlying the co-driving behavior. The content of this chapter is based on the work described in:

  G. J. de Bruin, C. J. Veenman, H. J. van den Herik, and F. W. Takes. „Understanding dynamics of truck co-driving networks." In: *Proceedings of the 8th International Conference on Complex Networks and Their Applications*. Studies in Computational Intelligence 882. Springer, 2020, pages 140–151. DOI: 10.1007/978-3-030-36683-4_12.
- Chapter 5 addresses Research question 4. It covers the topics of network science and smart vehicle inspection. A new approach to community detection using assortativity is proposed and applied to the truck co-driving network. The content of this chapter is based on the work described in:

  G. J. de Bruin, C. J. Veenman, H. J. van den Herik, and F. W. Takes. „Understanding behavioral patterns in truck co-driving networks." In: *Proceedings of the 7th International Conference on Complex Networks and Their Applications*. Studies in Computational Intelligence 813. Springer, 2018, pages 223–235. DOI: 10.1007/978-3-030-05414-4_18.

- Chapter 6 provides an answer to Research question 5. It brings together all the topics: network science, machine learning, and smart vehicle inspection. We provide an approach to smart cargo ships inspection. A comprehensive analysis is made of the fairness and performance of the model. The content of this chapter is based on the work described in:

  G. J. de Bruin, A. Pereira Barata, C. J. Veenman, H. J. van den Herik, and F. W. Takes. „Fair automated assessment of non-compliance in cargo ship networks." *EPJ Data Science* 11, 13 (2022). DOI: 10.1140/epjds/s13688-022-00326-w.

Chapter 7 concludes the thesis with answers to the research questions and the problem statement. Possible future research directions are provided as well.

Cooperation

It deserves to be noted that the work by Antonio Pereira Barata was part of the same project as this thesis and thus is also concerned with machine learning and smart vehicle inspection [153, 157]. His work focused on methods for assessing the impact of missing data in the truck registration data (see Section 1.5), as well as machine learning methods for *tabular data*, whereas this thesis focuses on methods for better understanding *network data* in relation to smart vehicle inspection.