# Network analysis methods for smart inspection in the transport domain

Bruin, G.J. de

# 5 Understanding behavioral patterns in truck co-driving networks

This chapter consists of two distinct research steps. The first step explores methods for *detecting communities* within truck co-driving networks. The second step investigates methods for understanding the relations of these communities with assortativity (cf. Definition 8). These steps allow us to better understand the behavioral patterns in truck co-driving networks. Understanding how to stimulate co-driving in turn may help to reduce traffic congestion and optimize fuel usage as a result of reduced aerodynamic drag.

The driving force behind edges in the truck co-driving network is analyzed in terms of assortativity. Moreover, we aim to understand the community structure of the truck co-driving network. We propose a novel metric, *the average maximal community assortativity metric*, to arrive at an understanding of the network community structure through assortativity.

The current chapter builds on the insights gained in the previous chapter, where we focused on assessing the evolution of co-driving networks. In this chapter, we address Research question 4, which reads as follows.

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

The current chapter corresponds to the following publication:
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

## 5.1 Truck co-driving network

In this chapter, we use network approaches to investigate what attributes lead to a group of truck drivers showing co-driving behavior. To do so, we use (1) network community detection [58] as well as (2) various metrics related to assortativity (also known as mixing patterns, see [126]).

We analyze a *unique* dataset gathered over one year, detailing the presence of at least two million trucks in the Netherlands (see Subsection 5.3.1 for a description of the data). We investigate the spatiotemporal data as a so-called *co-driving network*, wherein the nodes represent trucks (cf. Chapter 4). Trucks that are *co-driving* are observed at the same location within a very short time window. Those pairs of co-driving trucks that occur a certain number of times (e.g., more than once) are defined as *systematic co-driving trucks* (see Definition 12). In the co-driving network, the edges represent this systematic co-driving behavior. We will explain the construction of this network in Section 5.3.

The results of this work contribute to topics related to understanding human behavior, autonomous driving, and environmental sustainability. Using network metrics, we aim to derive *what attributes* may influence the decision of truck drivers to drive together systematically. The findings can be helpful for research on innovative forms of transportation, such as autonomous driving. We mention two possible benefits: (1) co-driving trucks can save up to 15% on fuel due to reduced aerodynamic drag [188] and (2) co-driving trucks reduce traffic congestion. It highlights the potential environmental implications of understanding co-driving behavior.

The co-driving network turned out to have at least three properties that are often encountered in real-world networks. First, the network has a significant *Giant Component* (GC, see item 4 in Subsection 1.2.4), which contains 37,858 nodes (trucks) and the majority of the co-driving links of the network. Second, the *average shortest path length* (cf. item 6 in Subsection 1.2.4) in the network is around nine edges, which, given a large number of nodes, is relatively tiny and hints at a small-world-like structure [119]. Third, our co-driving network is *scale-free* (cf. item 7 in Subsection 1.2.4), i.e., the degree distribution follows a power law [10]. We also investigate to what extent the network has a highly modular structure (cf. Subsection 1.2.3), meaning that a clear partitioning into communities exists.

As we will note in Subsection 5.3.1, we have access to additional node attributes (see Subsection 1.2.1). It allows us to study assortativity (Definition 8), which (1) enables insights into what attributes contribute to the network structure and (2) more importantly, explains co-driving behavior. Subsequently, we will use the node attributes to comprehend the communities better. With this knowledge, we aim to understand how local groups of co-driving trucks emerge and contribute to the global network structure. Furthermore, the proposed approach for understanding community detection results using assortativity is broadly applicable in other networks, providing a methodological contribution to the field.

The remainder of this chapter is organized as follows. After discussing related work in Section 5.2, Section 5.3 explains how the network was constructed from the raw data. Section 5.4 is concerned with the proposed approach and techniques to understand the network structure. Then, Section 5.5 provides details on the results obtained. Conclusions and suggestions for future work are provided in Section 5.6.

## 5.2 Related work on understanding behavioral patterns from networks

We start with an important contribution by Barrat and Cattuto [14], in which face-to-face contacts were recorded for twenty-second intervals using measurement infrastructure at several social settings. One of the results was that aggregated network topology and temporal behavioral properties are strongly related.

Second, Barrat and Cattuto showed that community detection could make a sensible partitioning of the network that was explainable by node attributes. Our study employs a similar approach, where the network topology and community structure are explained by the properties of the individual nodes and their assortative linking patterns.

Third, in a more recent study, Kassarnig *et al.* [95] handed over a thousand phones to students who agreed to have their communication and spatiotemporal activities traced. The work showed that network metrics (such as academic performance of peers, centrality, and the fraction of low and high-performing peers) are more informative indicators of university performance than node attributes indicating an individual's characteristics such as personality, class attendance, and the Facebook activity level. It underpins the value of network metrics compared to classical data aggregates.

Fourth, research by da Cunha and Gonçalves [43] on the Brazilian Federal Police criminal intelligence network used network science techniques to uncover behavioral patterns amongst criminals. Similar to our data, their network also featured a significant Giant Component (GC) and a degree distribution that follows a power law. Their observed low density and high average shortest path length were explained as "no trust among thieves". Additionally, Cunha and Gonçalves showed that their GC had a highly modular structure, which was explained by the necessity of (1) being *efficient in running criminal activities* within the group while (2) at the same time also being *obscure to the outside world*.

Throughout this chapter, we will employ *community detection* and *node attributes* in a way comparable to those in the works mentioned above, aiming to extract behavioral insights. To the best of our knowledge, the work of this chapter is the first to investigate the phenomenon of truck co-driving using network science methods and techniques.

## 5.3 Network construction

This section explains the network construction. We start with the characteristics of the data in Subsection 5.3.1. In Subsection 5.3.2, we describe how we selected systematic co-driving events. We continue with the co-driving network and its node attributes in Subsection 5.3.3. Subsection 5.3.4 reports two validation metrics to confirm that we selected the right value of a parameter. Finally, Subsection 5.3.5 details a regional co-driving network and its additional node attributes.

### 5.3.1 Truck observation data

The data is obtained from an ANPR system[1]. The Dutch Infrastructure and Water Management Ministry maintains the system. The data contains over 16,000,000 observations of trucks passing at a measurement system. These systems are situated at evenly distributed locations in the Netherlands. For each observation, the following data was available:

- license plate (serving as a unique identifier);
- location $\ell$ (either one of seventeen highway locations);
- lane $h$, indicating which of the (at most two) lanes the truck was in;
- speed $v$ (in $\mathrm{km\,h^{-1}}$);
- timestamp $t$ at a $10\,\mathrm{ms}$ resolution; and
- *country* (using the ISO-2 country code).

We note that a slightly different dataset was used compared to Chapter 4. In this chapter, we retained observations of all locations and used only data available when performing the calculations in 2018.

We briefly mention two insights from the truck observation data. First, the frequency distribution of how often each distinct truck (identified by its license plate) is measured is given in Figure 5.1a. The horizontal axis denotes the number of measurements per truck and the vertical axis indicates the corresponding probability. The distribution is highly skewed to the lower values, meaning that most trucks are only measured a few times. There appears to be a truncated power law present. Second, the interval distribution between two successive measurements of the same truck at the same location is shown in Figure 5.1b. It demonstrates how most trucks that return have a diurnal pattern, visible from the peaks at multiples of $24\,\mathrm{h}$. Similarly, a weekly pattern is present. This figure indicates that most individual trucks have regular driving patterns.

### 5.3.2 Selection of systematic co-driving events

In the co-driving network, nodes are trucks, and edges represent systematically co-driving trucks Definition 12. We follow the same selection procedure as used in Subsection 4.4.3.

[1]See https://international.fhwa.dot.gov/pubs/pl07028 for details on this system.

We employ the following three criteria to determine which truck pairs are systematically co-driving together.

1. Trucks $a$ and $b$ should be at the same place, i.e., their location is identical, so $\ell_a = \ell_b$ (a *co-occurrence*).
2. Moreover, the *co-driving* trucks should be so in a time window of at most $\Delta t_{\text{max}}$, so $|t_a - t_b| \leq \Delta t_{\text{max}}$.
3. Finally, *systematically co-driving* trucks are those co-driving trucks $(a, b) \in E$ occurring at least $\Theta > 0$ times.

Thus, to derive the co-driving network, we must set parameters $\Delta t_{\text{max}}$ and $\Theta$.

We derive the right parameter setting for $\Delta t_{\text{max}}$ in a data-driven manner below. In Figure 5.2, network characteristics are shown for increasing values of $\Delta t_{\text{max}}$. Definitions of these metrics, all common in the field of network science, can be found in [10]. Recall that a high value for $\Delta t_{\text{max}}$ will increase the probability that a pair of co-occurring trucks is added by chance. Therefore we choose to keep the value relatively low, namely at $\Delta t_{\text{max}} = 8\,\text{s}$. At this value, the density of the resulting network is lowest, while the GC's size compared to the full network (in terms of both nodes and edges) has become stable. Other network metrics, such as the GC's diameter and average shortest path length, also stabilize around this value, as seen in Figure 5.2d.

We expect the probability that two trucks randomly co-drive twice is negligible. Therefore, we identify non-random and, thus, systematic co-driving by setting $\Theta = 2$.

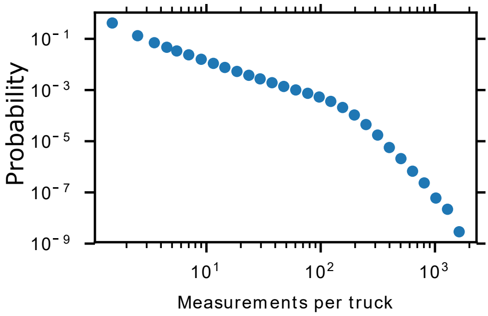
### 5.3.3 Co-driving network and node attributes

The co-driving network is an undirected weighted network $G = (V, E, w)$, where $V$ is the set of all trucks involved in a co-driving activity at least once. For a truck pair $(a, b) \in E$, the weight $w_{a,b}$ indicates the number of times the two trucks drove together. It should be greater than or equal to a certain threshold: $w_{a,b} \geq \Theta$. We furthermore consider four node attributes: (1) $country$, directly derived from the license plate; (2) $\tilde{v}$, the median truck speed; (3) $n_\ell$, the number of different locations where the truck was observed; and (4) $\ell_{\text{max}}$, the location where the truck was most frequently observed.

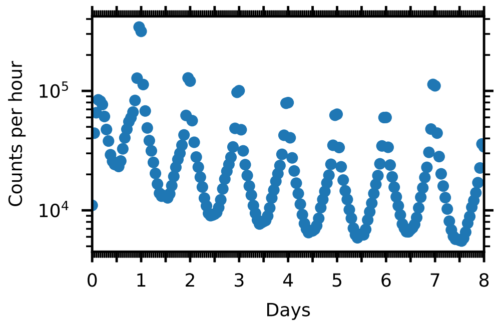
### 5.3.4 Two validation metrics

We validate our choice of $\Delta t_{\text{max}} = 8\,\text{s}$ by assessing whether two metrics from the raw truck measurement data differ when applied on two non-systematically ($w_{a,b} < 2$) co-driving truck pairs and two systematically ($w_{a,b} \geq 2$) co-driving truck pairs.

The first validation metric is $\Delta v$: the speed difference $|v_a - v_b|$ between two co-occurring trucks within $\Delta t_{\text{max}}$. We are inclined to assume that trucks that drive systematically together for longer distances would have a lower value of $\Delta v$ as their speed needs to be aligned. In Figure 5.3a, we observe that this is indeed the case. Here, the

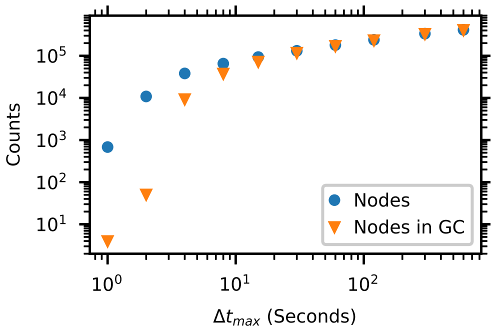



(a) Probability distribution of the number of measurements per truck.

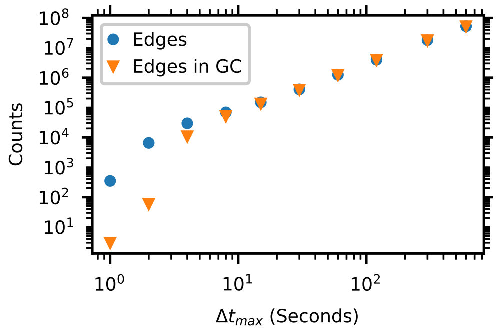



(b) Time interval that truck re-occur.

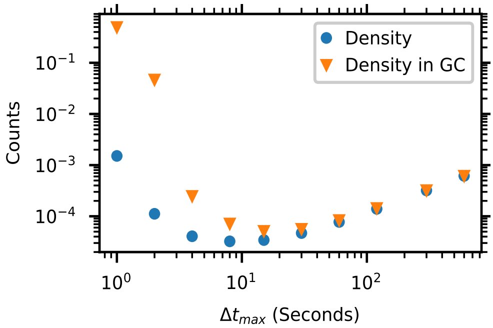
Figure 5.1: Summary statistics of the cargo truck data. (Note the logarithmic axes.)



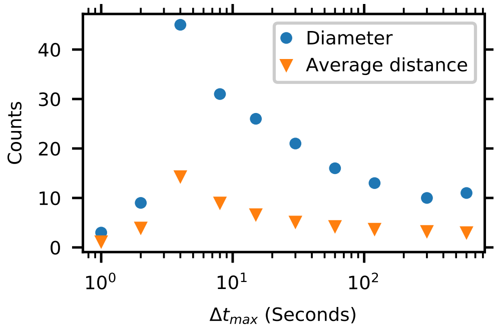

(a) Number of nodes in the network and its GC.




(b) Number of edges in the network and its GC.

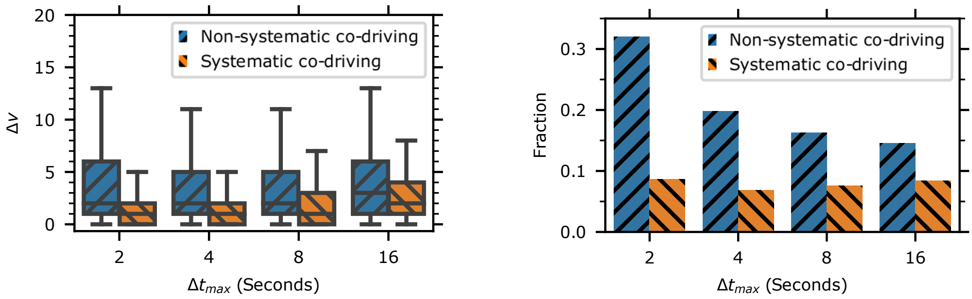



(c) Density of the network and its GC.




(d) Diameter and average shortest path length in the GC.

Figure 5.2: Statistics of the co-driving cargo truck network (for increasing $\Delta t_{\max}$).




(a) Distributions of the difference in speed $\Delta v$ (for increasing values of $\Delta t_{\max}$).

(b) The fraction of trucks driving on different lanes (for increasing values of $\Delta t_{\max}$).

Figure 5.3: Validation metrics for establishing systematic co-driving.

result is most evident for smaller values for $\Delta t_{\max}$, up to $8\,\mathrm{s}$. It hints that we selected the systematic co-driving events in a correct way.

The second validation metric is $h_a = h_b$, which means whether the considered pair of trucks are driving in the same lane. For a truck pair $(a, b)$ driving in the same lane it holds that $h_a = h_b$. In the case of systematic co-driving behavior, it is more likely that two trucks are in the same lane since they do not have to overtake each other to drive together. Figure 5.3b shows that, indeed, the fraction of trucks driving on a different lane ($h_a \neq h_b$) is more than two times higher for non-systematic co-driving than for systematic co-driving trucks. Thus, also this validation metric hints that we correctly selected the systematic co-driving events.

The two validation checks (see Figure 5.3) convince us that the derived co-driving network captures actual systematic co-driving behavior.

### 5.3.5 Regional co-driving network

Although trucks from various countries are observed in our data, we have additional information on Dutch trucks obtained from the Netherlands Vehicle Authority (RDW) (Dutch: RijksDienst voor het Wegverkeer). We use the additional information to construct a major contribution of our research, being a Dutch *regional co-driving network* which consists of trucks for which (1) the country was equal to the Netherlands (NL) (59% of the nodes) and (2) all systematic co-driving links between these trucks, having the following additional node attributes: (1) *city* where the truck is registered; (2) empty mass $m_{empty}$ of the truck; (3) maximum mass $m_{\max}$ of the truck; (4) *capacity* of the truck; (5) *company* that owns the truck; (6) registration date (*regdate*); and (7–10) $zip_{\{1,2,3,4\}}$ the zip code where the vehicle is registered with a higher number marking higher geographic precision. The regional co-driving network, together with the mentioned additional node attributes, are used in our research in (1) reducing traffic congestion and (2) optimizing fuel usage.

## 5.4 Chapter research methodology

Here we describe the techniques used to understand systematic co-driving behavior from a network perspective. We will start by outlining how assortativity can explain the driving forces in edge formation in Subsection 5.4.1, followed by the approach to detect communities within the co-driving network in Subsection 5.4.2.

### 5.4.1 Understanding co-driving behavior by assortativity

We will use assortativity to investigate what type of common node attributes explain the formation of links in the co-driving network. *Assortativity* is a measure of the preference of nodes in a network to connect with other nodes that are alike in some way [129], as explained in Subsection 1.2.1. The assortativity metric $r_a$ can be computed for each network's nominal and numerical node attribute $a$ using the definitions given in [127]. It should be noted that degree assortativity is the assortativity computed for the (numerical node attribute) degree, see Definition 8.

An assortativity value $r_a$ closer to $1$ indicates that nodes have more links to nodes with equal node attribute $a$. A value closer to $-1$ indicates disassortativity, meaning that nodes with different values for a node attribute $a$ are more likely to be connected. An assortativity of $0$ for an node attribute means no preferential attachment of edges between nodes based on the value of a node attribute $a$.

### 5.4.2 Understanding co-driving behavior by community structure

To better understand the co-driving network, we investigate the *community structure*, which can provide insights into the different groups of truck drivers. We use the well-known Louvain algorithm [19] to detect communities. It takes as input the structure of a weighted network and outputs an assignment of each node to a community. It furthermore has a *resolution parameter* $\gamma$ that predicts whether a more fine-grained or coarse-grained partitioning into communities should be found [105].

The Louvain algorithm uses heuristics to optimize the so-called *modularity* value $Q$, indicating the quality of the partitioning of the network into communities. A modularity value close to $1$ indicates that there are more edges *within* communities and fewer edges *between* communities. When adjusting the resolution parameter mentioned above, the value of modularity and the number of discovered communities $C$ change. At different resolutions, $\gamma$, similar values of $Q$ can be measured, each with a different number of communities $C$. This so-called modularity landscape *must* be explored to obtain the partitioning of the network into communities (and corresponding $\gamma$) that best explain the formation of groups in the underlying system [66].

We will propose to use the available node attribute information to explore these solutions automatically. Subsequently, we determine the assortativity for each node attribute and average that per community. After that, we take the partitioning of the node attribute with the highest assortativity for each community. We take the average over all communities, obtaining the proposed metric of average maximal community assortativity $R = \frac{1}{C}\sum_c \max_a r_a^{G(c)}$.

In this equation, $C$ is the number of communities, $c$ is one of the communities (defined as the subset of nodes in this community), $a$ is a node attribute, $G(c)$ is the subgraph induced on the nodes in the community $c$ and $r_a^{G(c)}$ is the assortativity $a$ in subgraph $G(c)$. Based on the value of $R$ for different network partitions into communities as a result of varying the resolution parameter $\gamma$, we select the partition into communities for which $R$ is highest because that partition allows for the best explanation of the communities observed.

## 5.5 Analysis of co-driving behavior

In Subsection 5.5.1, we start by providing statistics of the co-driving network. The results of applying the two approaches to understanding the formation of links outlined in Section 5.4 are discussed in Subsection 5.5.2 and Subsection 5.5.3.

### 5.5.1 Network statistics

*Network metrics*, of which definitions can be found in Section 1.2, were computed using NetworkX [72], whereas *distance metrics* were computed using teexGraph [185]. The python-louvain package was used for community detection [7].

In Table 5.1, we list (1) basic network statistics for the full network and (2) the regional co-driving network of measured Dutch trucks. We note that the majority of activity is captured in the GC. The degree distribution for both networks (all trucks vs Dutch trucks only) is given in Figure 5.4, showing a power-law distribution, suggesting that the co-driving network is scale-free. It means that a few truck drivers drive with many other trucks, whereas the majority only do so with a relatively small number of others trucks. The weight distribution in Figure 5.4 shows that some co-driving trucks frequently drive together. The diameter of the GC (which is affected by distant outliers) is relatively high, with a value of $31$ and $28$ for the full and regional network, respectively. In contrast, the average shortest path length is higher than $6$, which is common in many real-world networks. However, with a value of $9$, the average shortest path length is still substantially lower than average shortest path lengths encountered in random networks with similar sizes [10]. The power-law exponent of the degree distribution is $3.6$. Together, the three metrics (diameter, average shortest path length, and power-law exponent) indicate that although the network has a very skewed degree distribution, nodes are not as close to each other as in other real-world networks.

Table 5.1: Statistics of the full and regional co-driving cargo truck networks (and their GC).

| Metric | Full network | Regional network |
|---|---|---|
| Number of nodes | 65, 290 | 35, 706 |
| Number of nodes (GC) | 37, 858 | 22, 511 |
| Number of edges | 68, 958 | 36, 885 |
| Number of edges (GC) | 51, 730 | 30, 851 |
| Density | $3.2 \times 10^{-5}$ | $5.8 \times 10^{-5}$ |
| Density (GC) | $7.2 \times 10^{-5}$ | $1.2 \times 10^{-4}$ |
| Diameter (GC) | 31 | 28 |
| Average shortest path length (GC) | 9 | 9 |
| Clustering coefficient | 0.06 | 0.07 |
| Power law exponent | 3.58 | 3.61 |

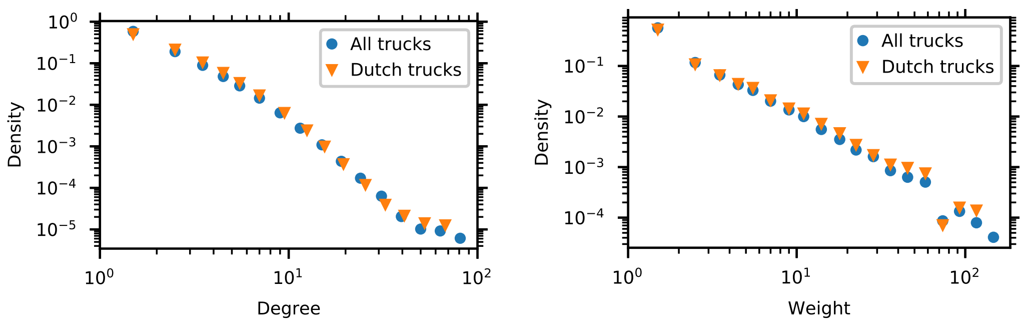



Figure 5.4: Degree (left) and weight (right) distribution of the (full and regional) co-driving cargo truck network.

### 5.5.2 Assortativity

The values reported in Table 5.2 were obtained by using the metric of *assortativity*, which was discussed in Subsection 5.4.1. The results indicate that actively co-driving trucks tend to be connected to other active co-driving trucks, as evidenced by the positive value for *degree assortativity*. The geographical information available about the trucks was found to be the most effective in explaining systematic co-driving behavior. Specifically, the *zip code* node attribute in the regional network showed substantially high assortativity metrics, and the *country attribute* in the full network had a value of $0.56$. These findings suggest that truck drivers from the same city or country are more likely to engage in systematic co-driving.

### 5.5.3 Average maximal community assortativity

The results of applying community detection to the GC of the entire network are shown in Figure 5.5. The number of communities and the modularity value are shown for increasing resolutions. A maximum value of $Q = 0.86$ is found for resolution $\gamma = 1$. This high value is the second evidence that our co-driving network is highly modular. We observe how there are several solutions with a similar modularity value but a very different number of communities.

To better understand these findings, we look at the average maximal community assortativity $R$ (see Subsection 5.4.2) shown in the bottom right of Figure 5.5. Although at $\gamma = 1$ the highest modularity is found, we see that for $\gamma = 2$ (as opposed to lower values of $\gamma$), the best community partitioning is obtained in terms of explainability using assortativity. For this value of the resolution, we find that 52 of the total 120 communities are best described using the *country* attribute. In contrast, the remaining attributes $\tilde{v}$, $n_\ell$, and $\ell_{\text{max}}$ explain 30, 29 and 9 communities respectively.

Table 5.2: Calculated assortativities of the full and regional truck co-driving network.

| Node attribute | Type | Full network | Regional network |
|---:|---:|---:|---:|
| degree | numeric | 0.12 | 0.12 |
| *country* | 17 categories | **0.56** | – |
| $\tilde{v}$ | numeric | 0.55 | 0.34 |
| $n_\ell$ | numeric | 0.45 | 0.40 |
| $\ell_{\max}$ | 17 categories | 0.25 | 0.21 |
| *city* | 1,319 categories | – | 0.33 |
| $m_{empty}$ | numeric | – | 0.30 |
| $m_{\max}$ | numeric | – | 0.35 |
| *capacity* | numeric | – | 0.32 |
| *company* | numeric | – | 0.29 |
| *regdate* | numeric | – | 0.13 |
| $zip_4$ | 1,975 categories | – | 0.32 |
| $zip_3$ | 718 categories | – | 0.33 |
| $zip_2$ | 90 categories | – | 0.35 |
| $zip_1$ | 9 categories | – | 0.41 |

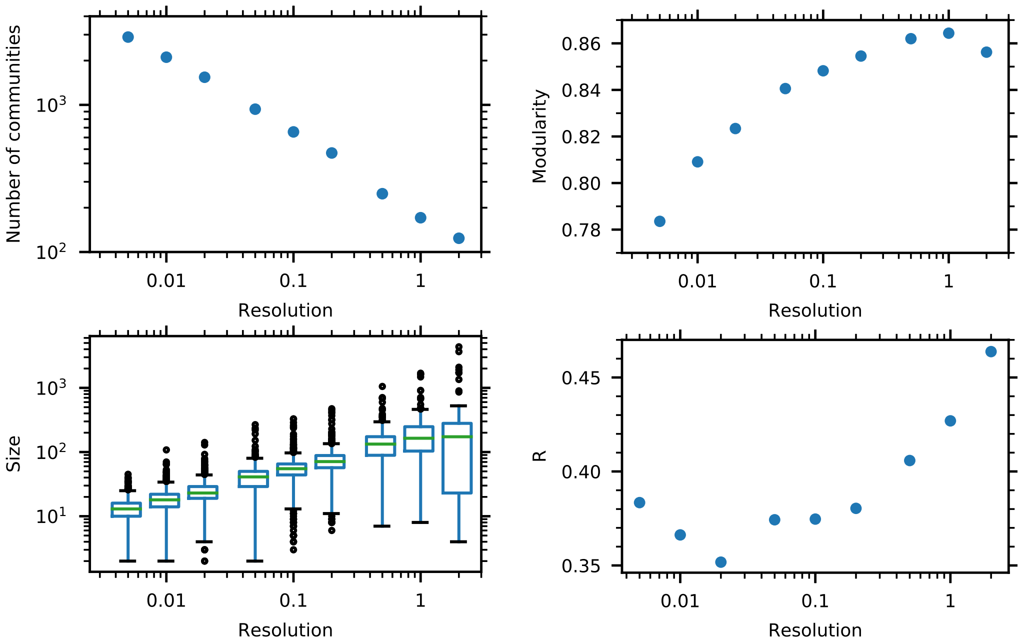



Figure 5.5: Properties of the communities (for various values of the resolution parameter). (Top left: Number of Communities; Top right: Modularity value $Q$; Bottom left: Average community size; Bottom right: Average maximal community assortativity $R$. Note the various logarithmic axes.)

## 5.6 Chapter conclusion

This chapter provides a detailed report on the extraction of truck measurement data for revealing its real-world properties and characteristics. Technically, we focus on a Giant Component, scale-free degree distribution, positive degree assortativity, and a highly modular community structure. The newly developed average maximal community assortativity metric is used to optimize the node attribute information to obtain a good partitioning into communities. Thereby we address Research question 4: “How can node attribute information be exploited to automatically create a good partitioning of a co-driving network into communities?” Our answer is that in the truck co-driving network as designed by us (see Section 5.3), we were able to establish that the highly modular community structure can be explained using different attributes’ assortativity in each community, dominated by geographical features.

### Chapter outlook

Additional investigation into the relationship between the observed network characteristics and the domain is on our list of further research. Timestamps will be incorporated to investigate the co-driving network’s dynamics, identifying which truck drivers initiate co-driving behavior and the conditions under which the behavior diffuses to other nodes.

Understanding the community structure of the truck co-driving network can lead to interventions to educate drivers on best practices. Moreover, truck drivers can save fuel and reduce traffic congestion by reduced aerodynamic drag when co-driving.