The complex network of ship movements in Europe

submitted in partial fulfillment for the degree of master of science

Nathalie van Veen
12413879

master information studies
data science
faculty of science
university of amsterdam

2020-07-02

	Internal Supervisor	External Supervisor
Title, Name	Dr Maarten Marx	Gerrit-Jan de Bruin, MSc.
Affiliation	UvA, FNWI, IvI	ILT
Email	maartenmarx@uva.nl	g.j.de.bruin@liacs.leidenuniv.nl

University of Amsterdam

Inspectie Leefomgeving en Transport
Ministerie van Infrastructuur en Waterstaat

ABSTRACT

Inspections of marine ships play a pivotal role in maintaining a clean and safe worldwide environment. In this work, we will contribute to insight regarding the inspection regime by examining ship movements in Europe using network analysis. This research will focus on the connection between ships, ports, and inter-ship timing. From a data set encompassing more than 3 million measurements of ships calling at more than 2000 ports in Europe between 2015-2019, we extract four different kinds of networks. First, in the ship-visiting network, nodes are ports and edges represent ships sailing between two ports. The obtained ship-visiting network has various properties common to the global ship network. Also, regional communities were found in the ship-visiting network, indicating ships have a preference for sailing in specific regions in Europe. Second, in the port-sharing network, nodes are ships and edges represent two ships visiting the same ports. Third, in the route-sharing network nodes are ships and edges represent two ships sailing the same routes. The port-sharing network and route-sharing network show ships tend to be selective in the ports they visit. Lastly, in the co-sailing network nodes are ships and edges represent pairs of ships that are systematically sailing together. There is some form of co-sailing behaviour present in Europe. This study improved understanding of ship traffic behaviour, which has implications for society and the environment. The results obtained can give suggestions on more effectively covering every ship during inspections. European collaboration between inspectorates is key to use limited inspection capacity to target the whole shipping industry.

KEYWORDS

shipping networks, infrastructure networks, co-sailing networks, network analysis

1 INTRODUCTION

Inspections of marine ships play a pivotal role in maintaining a clean and safe worldwide environment. This oversight comes with a challenge, because often the number of ships to inspect is larger than staff capacity allows. We propose a data-driven approach. In order to facilitate this, an understanding of the dynamics of the marine transport sector is required. This research will focus on the connection between ships, ports, and inter-ship timing. From a data set encompassing more than 3 million measurements of ships calling at more than 2000 ports in Europe between 2015-2019, we extract different kinds of networks. An overview of the networks that are investigated in this paper can be found in Table 1. We will explain the exact definitions and threshold to construct these networks in more detail in Section 3.3.

The global cargo ship network (GCSN) was studied by Kaluza et al. [17], defining a complex system as the network where nodes are ports and a link connects them if ship traffic passes between the ports. The data set used in this study is in principle a subset of the data set that was used to obtain the GCSN: Kaluza et al. used data regarding the world wide shipping network, while the data set used in this paper only includes information about ships calling at European ports. It is thus interesting to compare the characteristics of the GCSN with the network of only European ports. In the ship-visiting network $G_{visiting}$, nodes represent ports and the weight of edges will be the number of times ships sail between two ports. For this part, the main question is:

> 'Does the subnetwork of European ports have the same characteristics as the GCSN?'.

This will be answered by investigating if the network contains the small world property and the scale-free property. Also communities will be investigated to see whether ships have the tendency to sail in certain regions in Europe.

In addition to a network where nodes are ports, also two network where nodes are ships are made. Two different edge representations will be investigated. In the port-sharing network G_{port}, an edge is present between two ships if those ships visited the same ports in this data set. In the route-sharing network G_{route}, the edges between two ships will represent a route from one port to the other. For example, ship A and ship B will have a link if both ships travelled between Rotterdam and Antwerp. The port-sharing and route-sharing networks will allow us to better understand what characteristics ships have in a network. The main question for these networks is:

> 'What are the patterns of ship movements in Europe?'

This question will be answered by investigating if the network contains the small-world and scale-free property.

Lastly, we will investigate if there is a co-sailing network present. There is the expectation that there is a dependency in timing, by a process of smaller cargo ships feeding larger shiploads, mostly container ships. A study related to this was conducted by de Bruin et al. [11], in which the co-driving behaviour of truck drivers was examined using network analysis. In this co-sailing network $G_{co-sail}$, the nodes will represent ships. A co-occurrence of ships takes place when two ships are at the same location. Co-sailing is when this co-occurrence of two ships happens within a small time window. Those pairs of co-sailing ships that occur more than once, are defined as systematic co-sailing ships. The edges will represent this systematic co-sailing behavior in the co-sailing network. The main question for this part is:

> 'Is co-sailing behaviour present in Europe, and if it is, what factors can explain this?'.

In particular, attribute assortativity and communities will be investigated.

The remainder of this paper is organized as follows. After discussing related work in section 2, section 3 explains how the data was pre-processed and the network was constructed. Then, section 4 provides details on the results obtained. Conclusions and suggestions for future work are provided in section 5.

2 LITERATURE REVIEW

We start this section with describing general network science literature in Section 2.1. Then related work specifically to networks of

Network name	Symbolic name	Nodes	Weight of edges
Ship-visiting network	G_{ship}	Ports	Number of times ships sail between two ports
Port-sharing network	G_{port}	Ships	Number of different ports both ships visited
Route-sharing network	G_{route}	Ships	Number of different trajectories both the ships sailed
Co-sailing network	$G_{co-sail}$	Ships	Number of times two ships systematically sailed together

Table 1: Overview of network names and what nodes and weights of edges represent in each network

ships is described in Section 2.2. Lastly, network specific techniques that will be used in this paper are explained in Section 2.3.

2.1 Network science in general

Most real-world networks have common topological properties. These include a community structure, the scale-free property, and the small world property [12].

In random networks, the degree distribution is expected to be a binominal distribution where the majority of nodes are linked to a similar number of other nodes. In real world networks however, one common feature is the presence of hubs, which are nodes with a number of edges that greatly exceeds the average. The presence of hubs creates a long tail in the degree distribution. Such a degree distribution follows a power law distribution and is characteristic for scale-free networks [7].

In comparison, networks that have the small-world property have a higher density of edges as a result of their smaller diameter and higher clustering coefficient as defined by Watts and Strogatz [18]. In order to state that a network has the small-world property, the network has to have a higher average clustering coefficient than a random network with the same number of nodes and edges. A social network is defined to be a small-world network [14].

Ducruet and Zaidi [13] state that most real-world networks are both small-world and scale-free as a result of the combination of vertical (hierarchy) and horizontal (community) linkages. Also Barabasi et al. [8] state that the distances are smaller in a scale-free network than the distances in a similar random network, indicating that the presence of hubs increases the network's probability of having the small-world property. Broido and Clauset [10] however state that real-world networks are rarely scale-free networks and that log-normal distributions are a equally as good or a better fit than power laws. Also Barabasi et al. [8] state that in real systems a degree distribution that follows a pure power law is rarely observed. A truncated power law showing a low degree saturation and high degree cutoff is more frequently observed in degree distributions of real networks.

According to Amaral et al. [6] there are three types of small-world networks: scale-free with power law, truncated scale-free, and single-scale networks. They explain that for airline networks, each airport will limit the number of landings/departures per hour because of space and time constraints. The number of possible edges attached to a given node is thus limited by a restricted capacity of a node and the physical costs of adding edges, reducing the number of hubs. Similar reasons apply to the maritime transport network. Therefore the chance that the maritime transportation network is found to be truncated scale-free is larger than for it to be scale-free with a pure power law.

2.2 Network science in shipping networks

Network science has been used in various papers in order to analyse the global cargo ship network (GCSN). Networks that are made in these papers have ports as nodes. For this reason, the papers discussed next are most relevant to the ship-visiting network that is constructed in this paper.

The GCSN was studied by Kaluza et al. [17], defining a complex system as the network where nodes are ports and a link connects them if ship traffic passes between the ports. The ships included in this study are of the types containers, bulk dry carriers, and oil tankers and have a minimum volume of 10,000 GT. The study shows the GCSN possesses the small-world property [18], which means the network has short path lengths despite a large clustering coefficient. This indicates ship traffic appears to be an ideal system of unidirectional, often circular, trajectories, rather than being composed by back-and-forth journeys. The GCSN's degree distribution does not follow a pure power law, and is therefore not exactly scale-free. The distribution of link weights does follow approximately a power law. This suggests the presence of hubs, which are a few substantial ports with a high clustering coefficient that the smaller ports use to transact their cargo. Characteristic movement patterns of different ship types were found: while container ships typically follow a rigid order, bulk dry carriers depend on the current supply and demand and thus often adjust their plan on short notice. Also oil tankers depend on momentary market trends. Several communities were detected, highlighting important canals and the division of groups of ports based on geographical location.

Hu et al. [16] also studied the GCSN, with a network where nodes are ports and container liners connecting the ports if they sailed between them. The study concludes the degree distribution follows a truncated power law distribution, which is in accordance with all kinds of other transportation networks. The GCSN was characterised as a small-world network, because of a small average shortest path and a high clustering coefficient. The small-world property was not defined or mentioned before it was stated the GCSN is a small-world network. Additionally, the hierarchy structure and rich-club phenomenon were revealed through analyzing weighted clustering coefficient and weighted average nearest neighbours degree. The hierarchy structure in this network is that ports with a low degree belong to interconnected communities and thus have a high clustering coefficient, while hubs connect many ports with a small clustering coefficient, so ports that are not directly connected. Since the weighted clustering coefficient is larger than

the unweighted clustering coefficient, it is stated that the rich-club phenomenon is present, indicating that ports with a high degree have the tendency to form links with other ports with a high degree.

Xu et al. [19] investigated the ship-transport network of China and included only the passenger liners, excluding the cargo transport that most other studies and the present study have focused on. They state the degree distribution follows a truncated power law and the network displays small world properties.

The co-sailing network made in this paper is based on the co-driving network made by de Bruin et al. [11]. In this network, nodes are trucks and edges represent pairs of trucks that are systematically driving together. Systematically driving together follows the definition of two trucks being at the same highway location within 8 seconds of each other more than once. The co-driving network structure has various properties common to real-world networks, such as a scale-free degree distribution. This means that the presence of hubs, indicating a few truck drivers drive with a large number of other trucks, whereas the majority only does so with a relatively small number of others. Distance metrics as well as community detection showed a highly modular structure. They found high assortativity metrics for the country attribute, meaning truck drivers from the same country are more likely to systematically drive together.

To the best of our knowledge, this paper is the first to investigate the phenomenon of ship co-sailing using network science methods and techniques.

2.3 Background

In this section the formal notation of the graph structure is introduced. This formalization was adopted from Barabasi et al [8].

A network can be formally mathematically represented as a graph. Therefore, the terms 'network' and 'graph' will be used interchangeably in the following section. In a graph, each object is called a node. If a connection is present between two distinct nodes, there will be an edge (or link) in the graph. Formally, we define a graph G as an ordered pair $G = (V, E)$ where V is a set of nodes, E is a set of edges, and each edge connects a pair of nodes [8]. In a directed graph, the edge from a node to another node is directed. Two nodes are called connected, or adjacent, if they share a common edge, in which case the common edge joins the two nodes. The *degree* of a node is the total number of nodes that share an edge with that node [8]. The nodes that share and edge with that node is also called the neighbourhood of a node. The average degree of a directed graph is calculated as $k = E/V$, which is the total number of edges divided by the total number of nodes [8]. The *density* of a graph $G = (V, E)$ measures how many edges are in set E compared to the maximum possible number of edges between nodes in set V [8]. The length of a path is the number of edges that it uses. For example, the length of the path between port A and port D, is three if port A is connected to port B, port B is connected to port C, and port C is connected to port D, and there are no other ports/edges connecting ports A and D.

A graph is *connected* if for every pair of nodes, there is a path between them [8]. A directed graph is strongly connected if every node is reachable from every other node following the directions of the edge. In other words, if for every pair of different nodes A and B there exists a directed path from A to B. A directed graph is weakly connected if the graph is connected when considering it as an undirected graph. In other words, for every pair of different nodes A and B there exists an undirected path (possibly running the opposite direction of the edge) from A to B. If a network is connected, the average shortest path length can be calculated and is defined as the average number of steps along the shortest paths for all possible pairs of network nodes.

The *clustering coefficient* is a measure of the degree to which nodes in a graph tend to cluster together [8]. It measures how complete the neighbourhood of a node is. The neighbourhood, or number of neighbors, of a node n, is the set of nodes that are connected to n. If every node in the neighbourhood of n is connected to every other node in the neighbourhood of n (closed triangles), then the neighbourhood of n is complete and will have clustering coefficient 1. If no nodes in the neighbourhood are connected, then the clustering coefficient will be 0. The *clustering coefficient* in a directed network can be calculated as: $c_i = E_i/(k_i(k_i - 1)$, where c_i is the clustering coefficient of node i, k_i is the number of neighbors of the node i, and E_i is the number of directed connections that exist between the k_i neighbors [8]. The clustering coefficient C for the whole network is obtained by averaging c_i over all nodes of the network. *Communities* are defined as groups of nodes that share many edges within the groups but few edges between different groups.

3 METHODOLOGY

Here we explain how the networks have been constructed. We start in Section 3.1 with the characteristics of the data. In Section 3.2 we describe how the data was prepared. Finally, in Section 3.3 we explain how the networks were constructed and how parameters were set.

3.1 Data description

In this paper we analyse a data set that was gathered by the European Maritime Safety Agency [1] over the period of five years, namely between 1-1-2015 and 30-10-2019, detailing the presence of more than 33,000 ships. The data set consist of more than 3,8 million measurements and contains all ships that visited one of 2030 ports in Europe. The data set consists of an unique ship ID, the type of ship, the port where it calls, the estimated time as well as the actual time of both the arrival and departure. It also contains information about the volume of the ships and its flag state.

3.2 Data preparation

The effects of the following selections on number of unique ships and number of port calls remaining in the data set can be found in Table 2. Other data cleaning steps that were taken can be found in Appendix A.

This study focuses on ship types that are present in the global maritime transport economy and that the Inspectorate of Human Environment and Transport oversees. Therefore several ship types have been excluded from the data set. This excluded approximately 7000 unique ships. The remaining ship types are divided into six categories based on domain knowledge, which are described in Appendix A.

After this selection on ship types, port calls of ships where the current port and previous port are the same were excluded. This returning to the same port can have several reasons, for example the ship might have left Europe or has been at anchorage. Beside the fact that this behaviour is not informative for the network, excluding those port calls also has the advantage of preventing self-loops in the network. This step excluded nearly 2000 ships.

For the ship-visiting network, port-sharing network, and route-sharing network there is the expectation that ships with a volume less than 10,000 gross tonnage (GT) do not play a significant role in the global maritime transport economy based on domain knowledge and literature [17]. Also computationally it is beneficial to exclude ships with a volume of less than 10,000 GT. After excluding ships with a volume less than 10,000 GT the data set contains 16,524 unique ships with in total 1,411,993 port calls at 990 ports. The median number of port calls per unique ship is 15. In the co-sailing network there will be no limit of 10,000 GT, because the expectation is that smaller cargo ships feed larger shiploads, and those smaller ships would thus be excluded if the minimum of volume is applied. Figure 1 shows the median volumes per ship type of all ships present in the original data set. The red line indicates 10,000 GT and shows that almost all ship types that are being excluded based on domain knowledge in most cases don't have a median volume above 10,000 GT. This confirms that deleting those ship types does not have a large impact on the transportation network that is viewed as relevant. Figure 2 shows the frequency distribution of volumes of ships, after selecting on ship type and no return to same port, where the red line indicates 10,000 GT.

For the port-sharing network, route-sharing network and co-sailing network only the port calls from the year 2019 were used for computational reasons. This is not expected to have an effect on the network characteristics, since every year in the data set is similar. In 2019, the median number of unique ports a ship visits is 4, the mean is almost 6.

3.3 Network construction

Network metrics were computed using NetworkX [15]. For every network that is constructed the clustering coefficient and average shortest path were calculated in order to determine if the network has the small-world property. Also, plots were made of the degree distributions and the link weight distributions in order to investigate whether the distributions reveal a power law relationship. If the degree distribution follows a (truncated) power law distribution and the power law exponent is negative, it is stated the network possesses the scale-free property. To determine whether the best fit for the degree distribution is a power law or another distribution, the powerlaw package of Python was used [5]. For the ship-visiting network and co-sailing network, communities were investigated with the Louvain method [9]. Attribute assortativity was used for a network-driven understanding of the co-sailing network.

For the ship-visiting network, port-sharing network, and route-sharing network, an edge was only included if the weight was three or more. The aim of this study is to find systematic behaviour instead of random behaviour. Two ships sailing between two ports is considered as possible random behaviour, while three ships sailing between two ports is considered as relevant. The same reasoning

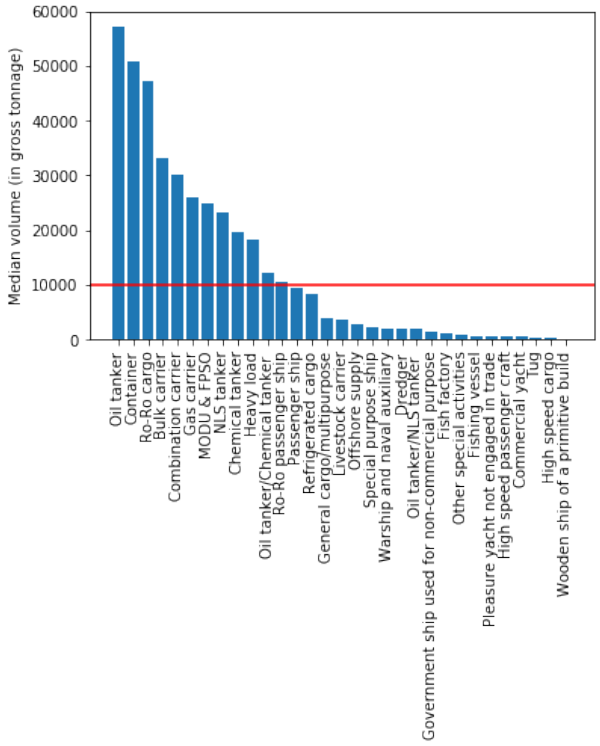

Figure 1: Median volume per ship type for unique ships, measured in gross tonnage, where the red line indicates 10,000 gross tonnage.

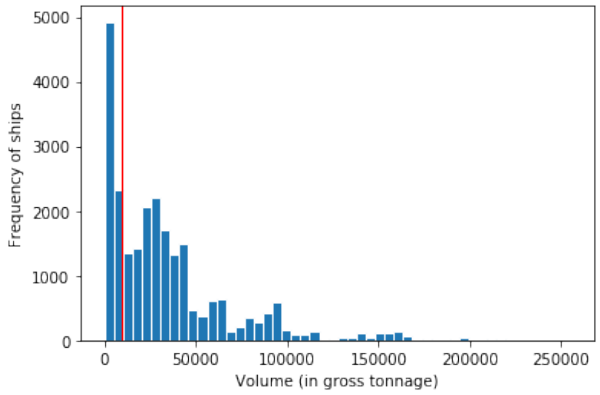

Figure 2: Frequency distribution of volumes of ships, measured in gross tonnage, where the red line indicates 10,000 gross tonnage.

applies to the port-sharing and route-sharing networks: two ships visiting one or two of the same ports or sailing one or two of the same trajectories is considered as possible random behaviour, while two ships visiting three of the same ports or sailing three of the same trajectories is considered as relevant.

Filter	Unique ships remaining	Port calls
Original	33,178	3,882,638
Ship type	25,351	3,619,804
No return to same port	23,507	2,983,111
Minimum 10,000GT	16,524	1,411,993
The calendar year 2019	10,214	264,862

Table 2: Effects of each data pre-processing step on the number of unique ships and port calls remaining in the data set

We will use assortativity for a network-driven understanding of co-sailing behaviour. Assortativity is a preference for a network's nodes to attach to other nodes that are similar in some way. This will allow us to investigate what type of common attributes explain the formation of links in the co-sailing network.

3.3.1 Co-sailing network parameter settings. In the co-sailing network, nodes are ships and edges represent systematically co-sailing ships. A co-occurrence of ships takes place when two ships are at the same location. Co-sailing is when this co-occurrence of two ships happens within a small time window. The time window is the maximum time interval between the arrival of one ship and the arrival of the other ship. The ships are not required to be at the same port at exactly the same time, because the cargo of ships is not being transferred directly from ship to ship, but via land. This is also why the retention time of ships in a port is high: most ships are for a long period of time in a port, because the discharging and loading of cargo takes time. It's relevant to see whether the cargo can be transferred, not whether the ships have to be at literally the same time at a port.

Those pairs of co-sailing ships that show co-sailing behaviour more than a certain value that has to be determined, are defined as systematic co-sailing ships. The edges represent this systematic co-sailing behavior in the co-sailing network. The network is undirected weighted, where the weight is the number of times the two ships systematically sailed together. The following definitions are employed to determine which pairs of ships are systematically sailing together.

Definition 1. A co-occurrence of ships takes place if ship A and ship B are at the same port.

Definition 2. Co-sailing ships are those co-occurrences of ships that take place within a time window of at most t_{max}.

Definition 3. Systematically co-sailing ships are those co-sailing ships that occur at least θ times.

Accordingly, to derive the co-sailing network, parameters t_{max} and θ must be set.

When investigating the port call data to determine t_{max}, it appears there is a clear distinction between the ship belonging to the ferry category, refrigerated cargo ships, and high speed cargo ships on the one hand and all other categories on the other hand in terms of how many hours a ship stays at a port. The ships that have a low retention time at a port are excluded to prevent a over-occurrence of these ships in the network. Other studies investigating ship traffic networks have also focused on either cargo ships or passenger ship because of differences between these categories [13, 16, 17, 19].

Since the median retention time in a port is 26 hours, the expectation is that the right setting for t_{max} could range between 1 hour and 48 hours, depending on how close to each other ships enter the port. We derive the right parameter setting in a data-driven manner. Figure 3 shows the results of network metrics density, diameter, and average distance for increasing values of t_{max}, when using θ = 2. A high value of t_{max} will result in a high probability that a pair of co-occurring ships is seen as co-sailing by chance. We can conclude that the density is almost at its minimum at t_{max} = 24 hours, which means from 24 hours onwards the network has the least chance of being a random network. Also, the diameter increases strongly at 36 hours, which means from 36 hours onwards the network has a chance of being a random network. Combining these findings leads to setting the value for t_{max} to 24 hours, with the lowest density value combined with the lowest diameter value.

In Figure 4 network characteristics are shown for increasing values of θ, when using t_{max} = 26. We can conclude the network's density is at a minimum at θ = 2, indicating at this value the network is the most non-random. Also, since the diameter is lower at value 2 than 3, the network at θ = 2 can be regarded as a non-random network. We expect the probability that two ships randomly co-sail twice is sufficiently small. Therefore, we identify non-random and thus systematic co-sailing by setting θ = 2.

4 RESULTS

We start this section with an analysis of the ship-visiting network in Section 4.1. Then the characteristics of the port-sharing network can be found in Section 4.2. In Section 4.3 the analysis of the route-sharing network is described. Finally, Section 4.4 shows the results of the co-sailing network. In each section first general information about the network will be discussed, followed by the investigation of the small-world property and then the scale-free property. All results are described in Table 3.

4.1 Ship-visiting network

In this network the nodes represent ports. Two ports are connected, i.e. have an edge between them, if a ship has sailed from one port to the other. This edge is directed: it points from the port the ship departed from to the port the ship sailed to. The weight of edges is the number of times ships sail between two ports. An edge was only included if the weight was three or more. The number of nodes and thus ports in this network is 728. There are 18,142 edges present.

Metric	Ship-visiting network	Port-sharing network	Route-sharing network	Co-sailing network
Number of nodes	728	7358	4033	6614
Number of nodes (GC)	726	7341	3515	6541
Number of edges	18,142	1,224,972	74,804	112,979
Number of edges (GC)	18,140	1,224,962	74,178	112,942
Density (GC)	0.03	0.05	0.01	0.005
Diameter (GC)	7	6	16	9
Clustering coefficient	0.58	0.67	0.49	0.24
Average shortest path (GC)	2.49	2.25	3.77	3.23
Power law exponent	-0.24	0.06	-0.27	-0.62

Table 3: Statistics of the networks used in this study and their giant component (GC)

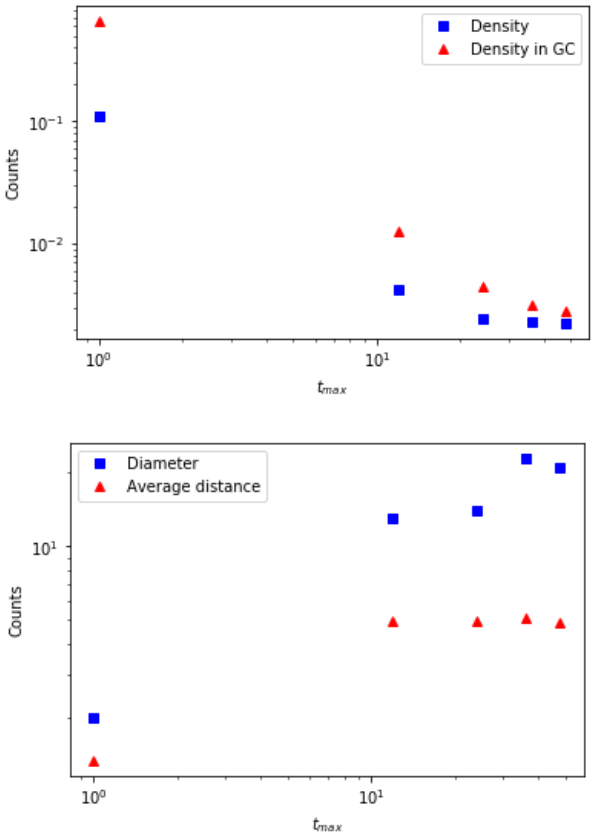

Figure 3: Co-sailing network statistics for increasing values of time window t_{max}.

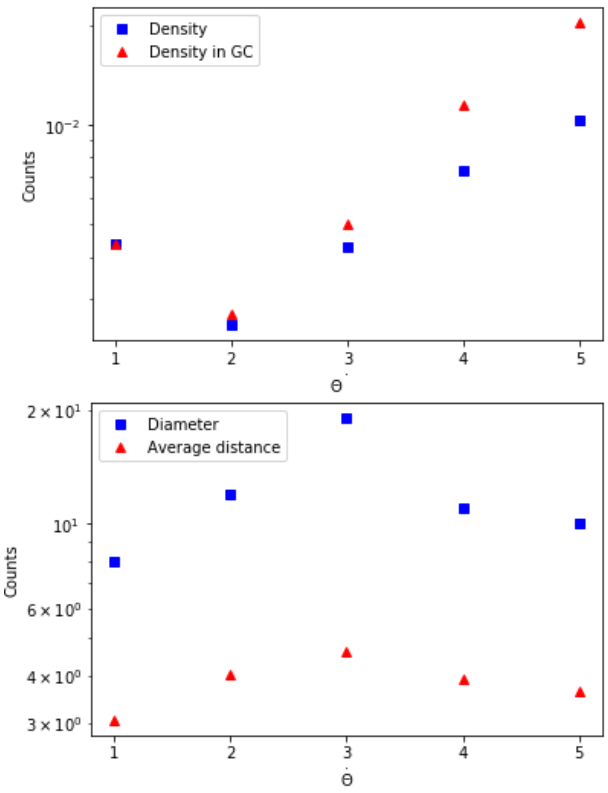

Figure 4: Co-sailing network statistics for increasing values of minimum occurrences θ.

4.1.1 Small-world property. The network is not connected. The largest weakly connected component, or giant component, however consists of 726 nodes, which means that only 2 nodes of the whole network are not connected to the giant component. The largest strongly connected component has 666 nodes. The average shortest path of the giant component is 2.49 which is considered small. This is very similar to the average shortest path in the GCSN research, which is 2.5.

The clustering coefficient in this network is 0.58, which is slightly higher than in the GCSN where it was 0.49. Random networks

with the same number of nodes and links only yield a clustering coefficient of 0.07 on average. The ship-visiting network - like the GCSN - can be regarded as a small-world network possessing short path lengths despite substantial clustering [18].

4.1.2 Scale-free property. The weighted degree distribution is, the same as in the GCSN, not exactly scale-free as it doesn't follow a pure power law, as can be seen in Figure 5. The degree distribution of the ship-visiting network follows a truncated power law distribution [5]. The power law exponent is -0.24. This indicates that the ship-visiting network has the scale-free property, and thus the network has hubs. The two ports that have the highest weighted degree, and thus function as hubs, are Rotterdam (the Netherlands), with a weighted degree of 121,138, and Antwerp (Belgium), with a weighted degree of 74,504. These ports are also the busiest cargo ports in Europe [4].

Figure 6 shows the link weight distribution of the ship-visiting network. A large majority of nodes have a low weight, but a small number have a higher weight. This means that a lot of ports are connected by in total a few ships sailing between them and a few ports have a lot of ships sailing between them. The median weight is 9. The highest weight is 22,073, which means at most 22,073 times ships sailed between two ports. In this network those ports are Vanasadam (Estonia) and Helsinki (Finland) which have a ferry route between them.

4.1.3 Community detection. Seventeen communities were detected in the ship-visiting network using the Louvain method [9], of which six communities consist of more than 20 ports. The modularity value is 0.64. Figure 7 shows these six communities, each in a different colour on a map of Europe. The largest community, shown by the red dots, contains 31% of all nodes in the network. The figure shows the communities have a strong geographical base, meaning ports are more densely connected to ports that are geographically close. These geographical communities can explain the high clustering coefficient: ports that are close have a relatively high density of links.

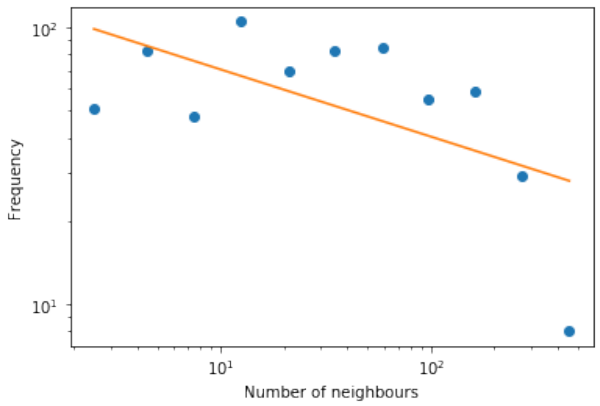

Figure 5: The weighted degree distribution reveals a truncated power law relationship for the ship-visiting network.

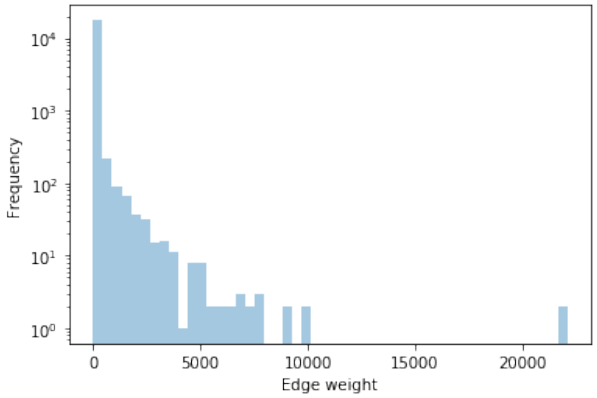

Figure 6: The link weight distribution shows most edges have a low weight, while only a few edges have a high weight for the ship-visiting network.

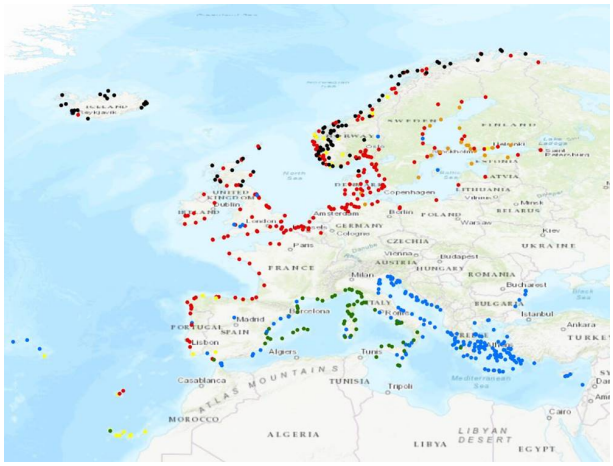

Figure 7: The network plotted in geographical coordinates, where the colours indicate communities. It is clearly visible the communities have a strong geographical base.

4.2 Port-sharing network

The port-sharing network is a network where nodes are unique ships and edges are present if the two ships visited the same ports. The weight of the edge is the number of the same ports the two ships visited. An edge was only included if the weight was three or more. The port-sharing network contained 7358 nodes or unique ships. There are 1,224,972 edges present.

4.2.1 Small-world property. The network is not connected, and has 9 connected components. The giant component contains 7341 nodes out of the 7358 in total. The average shortest path for the giant component is 2.25.

The average clustering coefficient is 0.67, which is considered high. This means that when ship A is linked to ships B and C, there is a very high probability that there is also a connection from B to C, and thus that all these three ships visited the same port(s).

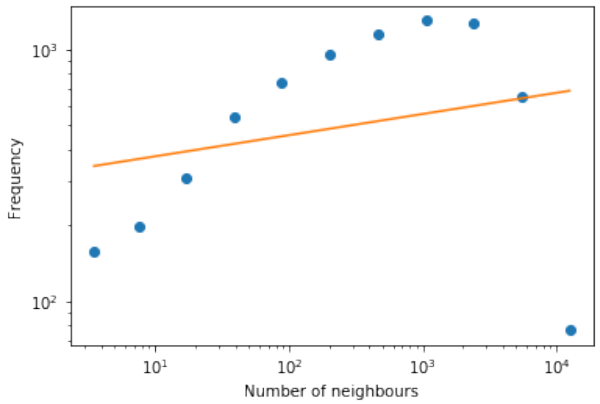

Figure 8: The weighted degree distribution shows no (truncated) power law relationship for the port-sharing network.

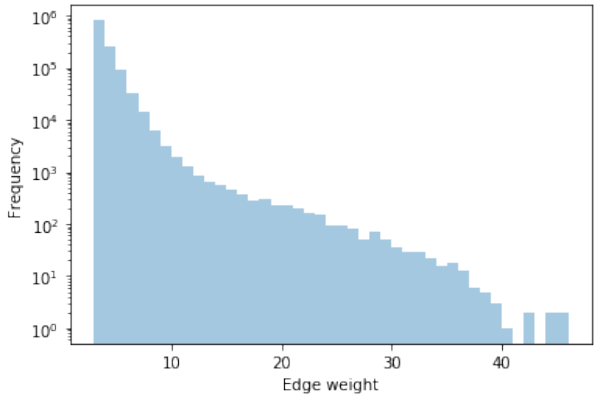

Figure 9: The link weight distribution shows most edges have a low weight, while only a few edges have a high weight in the port-sharing network

Random networks with the same number of nodes and links only yield a clustering coefficient of 0.04 on average.

The port-sharing network can be regarded as a small-world network possessing short path lengths despite substantial clustering [18].

4.2.2 Scale-free property. Figure 8 shows the weighted degree distribution of the port-sharing network does not follow a pure power law or truncated power law distribution. The power law exponent is 0.06. This means that there are more nodes with a large degree and only a few nodes with a small degree. This is the opposite of a scale-free network, where a few hubs are present.

Figure 9 shows the link weight distribution of the port-sharing network. A large majority of nodes have a low weight but a small number, known as hubs, have a higher weight. The median weight is 3. The maximum weight of an edge present in the network is 46. Which means that the two ships that visited the most same ports visited 46 the same ports in 2019. For the port-sharing network this means that most combinations of two ships only visited one or a few of the same ports and only a few combinations of two ships visited more of the same ports.

4.3 Route-sharing network

The route-sharing network is a network where nodes are unique ships and edges are present if the two ships sailed the same routes. The weight of the edge is the number of the same trajectories the two ships travelled. An edge was only included if the weight was three or more. The route-sharing network contained 4033 nodes or unique ships. There are 74,804 edges present.

4.3.1 Small-world property. The network is not connected, as it has 191 connected components. The network does have one giant component of 3515 nodes out of the 4033 nodes in total. The average shortest path of the giant component is 3.77.

The average clustering coefficient is 0.53, which is considered high. This means that when ship A is linked to ships B and C, there is a very high probability that there is also a connection from B to C, and thus that all these three ships sailed the same trajectory or

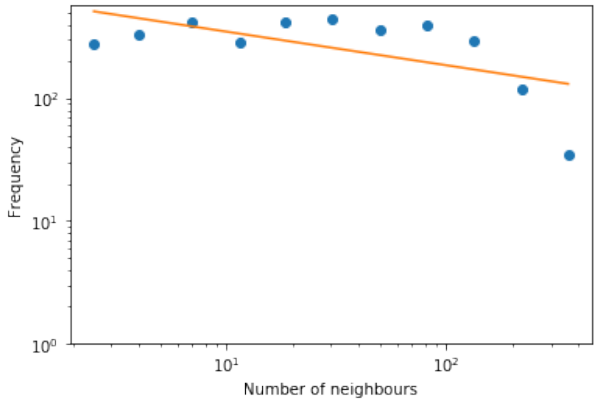

Figure 10: The weighted degree distribution reveals a truncated power law relationship for the route-sharing network.

trajectories. Random networks with the same number of nodes and links only yield a clustering coefficient of 0.009 on average.

The route-sharing network can be regarded as a small-world network possessing short path lengths despite substantial clustering [18].

4.3.2 Scale-free property. In Figure 10 the weighted degree distribution of the route-sharing network follows a truncated power law distribution [5]. The power law exponent is -0.27. This indicates that the route-sharing network has the scale-free property.

In Figure 11 the link weight distribution of the route-sharing network shows most ships have a very low weight and a few ships have a higher weight. For the route-sharing network this means that most combinations of two ships only sailed one or a few of the same trajectories and only a few combinations of two ships sailed more of the same trajectories. The maximum weight of an edge present in the network is 43. Which means that the two ships that sailed the most same trajectories had 43 of the same trajectories in 2019.

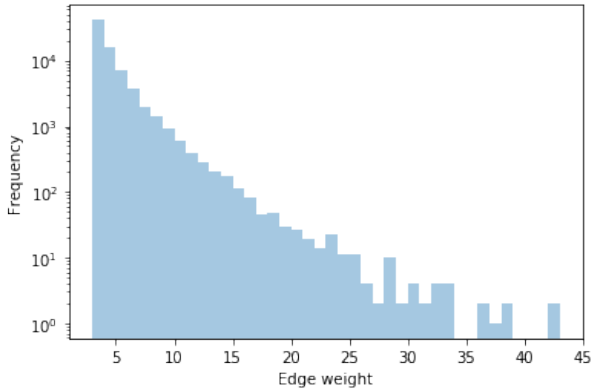

Figure 11: The link weight distribution shows most edges have a small weight and a few edges have a high weight in the route-sharing network.

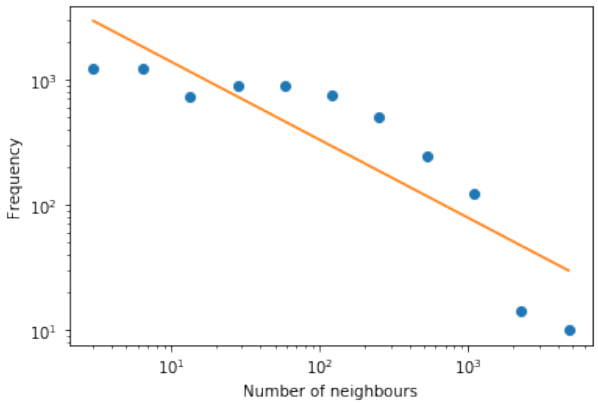

Figure 12: The weighted degree distribution reveals a truncated power law relationship for the co-sailing network.

4.4 Co-sailing network

A network where nodes are unique ships and edges are present if the two ships systematically co-sailed was made. The weight of the edge is the number of times the two ships co-sailed. Details of the network construction and parameters can be found in Section 3.3. The co-sailing network contained 6614 nodes or unique ships. There are 112,979 edges present.

4.4.1 Small-world property. The network is not connected, as it has 37 connected components. The network does have one giant component of 6541 nodes out of the 6614 nodes in total. The average shortest path of the giant component is 3.23.

The average clustering coefficient of the giant component is 0.24. Random networks with the same number of nodes and links only yield a clustering coefficient of 0.005 on average.

The route-sharing network can be regarded as a small-world network possessing short path lengths despite substantial clustering [18].

4.4.2 Scale-free property. In Figure 12 the weighted degree distribution of the co-sailing network follows a truncated power law distribution [5]. The power law exponent is -0.62. This indicates that the route-sharing network has the scale-free property.

In Figure 13 the link weight distribution of the co-sailing network shows most ships have a very low weight and a few ships have a higher weight. For the co-sailing network this means that most combinations of two ships only co-sailed a few times and only a few combinations of two ships co-sailed more often. The maximum weight of an edge present in the network is 274. Which means that the two ships that have co-sailed the most together did this 274 times in 2019.

4.4.3 Attribute assortativity. The chosen attributes to inspect the assortativity of, are the *ship type* and the *flage state* of the ship. The nodes in the co-sailing network only show a small positive assortativity for *ship type* and *flage state*, being 0.08 and 0.02 respectively. This means we cannot conclude that certain ship types or ships with the same flag state are more likely to systematically sail together.

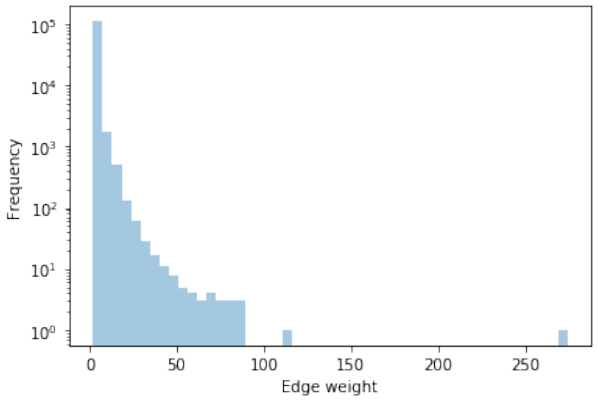

Figure 13: The link weight distribution shows most edges have a small weight and a few edges have a high weight in the co-sailing network.

4.4.4 Communities. 51 communities were detected in the co-sailing network using the Louvain method [9]. The five largest communities make up 95% of the network. The modularity value is 0.32.

Visual inspection of the frequency distributions of ship types in the five largest communities shows that in some communities certain ship types are more represented than in the distribution of the whole network. These frequency distributions can be found in Appendix B.

It is noted that ships with a small volume tend to have a higher degree in the co-sailing network, which can be seen in Figure 14. Investigating this, leads to the conclusion that the communities are not driven by differences in ship volumes, which can be seen in Figure 15, since there is no considerable difference in the distribution between communities.

5 CONCLUSION & DISCUSSION

In this paper we described how four networks can be extracted from European port call data. First, we will discuss the conclusions of

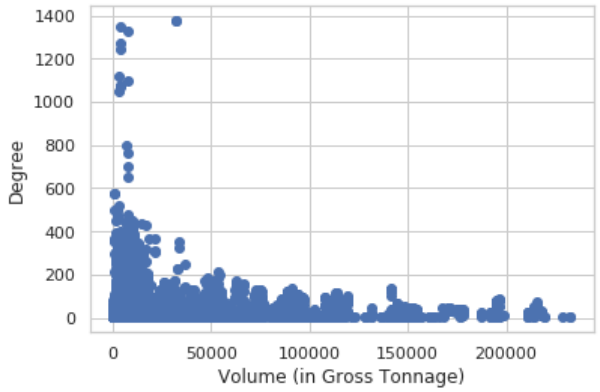

Figure 14: The degree-volume scatter plot for the co-sailing network.

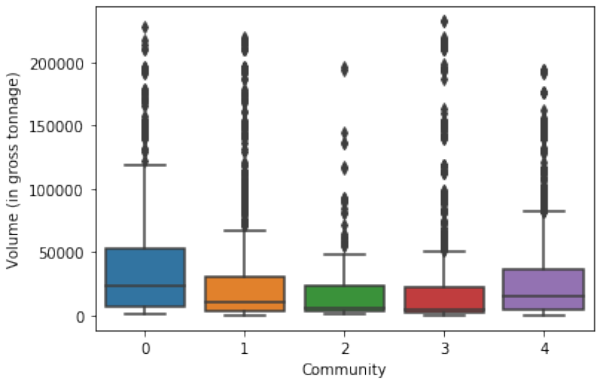

Figure 15: The boxplot plot for the co-sailing network shows there is no large difference in the distribution of volumes of ships between the communities.

the ship-visiting network, port-sharing network, route-sharing network, and co-sailing network which follow from results described in Section 4. Then limitations of the present study and future work are presented.

5.0.1 Ship-visiting network. The ship-visiting network shows comparable results as a global study on port networks. Both can be regarded as a social network possessing the small-world property and scale-free property. The small-world property means every port in the network is connected to every other port in just a few steps. This is also what was expected. The network shows the presence of hubs and the two ports that are the biggest hubs are also in the real-world the most busy ports. This indicates this data shows what is generally considered as true, and allows us to use it for more complex questions.

Geographical communities were found. The European inspection regime does at this moment not take into account these regional communities of ports. This can aid further development of this system. The inspection regime would want to make sure the whole shipping industry is targeted and to prevent being fixated on only certain groups of ships.

5.0.2 Port-sharing network. Since the port-sharing network is not scale-free, there are no hubs present. However, the network is rather well connected, so most ships have a lot of links to other ships. It appears that most ships only visit a small number of the same ports, with a median of three. This means that ships have the tendency to visit at least one port hub, and are therefore well connected to all other ships. So, ships do not have the tendency to visit every port, but ships tend to visit at least one port that has a lot of ships coming to this port.

5.0.3 Route-sharing network. The network shows that most ships only share routes with a few other ships. This means that while ships ensure a highly connected network of ports through their diverse sailing behaviour, as seen in Section 5.0.1, there is little sign of shared routes.

From an environmental perspective it could thus be argued that increasing efficiency through larger scale of ships and synchronized routes is not a solution [3]: because it appears the shipping system is already efficient, since this data does not show a lot of ships sailing the same routes.

5.0.4 Co-sailing network. The co-sailing network contains typical real-world network properties such as a scale-free degree distribution and the small-world property.

The co-sailing behaviour could not be explained by attribute assortativity. This means we cannot conclude from the presented analysis that certain ship types or ships with the same flag state are more likely to systematically sail together.

Investigating network communities leads to the conclusion that there is some form of co-sailing within communities. In contrast to the low attribute assortativity, it is found that certain ship types are more represented in some communities than in the whole network. It is thus possible that ships of the same ship type do have a preference of being together in communities.

5.0.5 Main conclusion. The research presented showed that network analysis facilitates a better understanding of the shipping industry. This can aid the inspectorate to target those ships that show specific behavior. We suggest that European collaboration between inspectorates is key to use limited inspection capacity to target the whole shipping industry.

5.0.6 Discussion & future work. It is possible that there are port calls present from ships that have been outside of Europe. These port calls are included as successive trajectory of the ships. This has been chosen because the only way to find out whether a ship has been outside of Europe is to look at the travel time, meaning how long the ship travelled from one port to another. If this travel time is longer than anticipated, it could be the case that the ship has been outside of Europe, but might be the case that the ship has been at anchorage for a while, or waiting outside a port for the most optimal price (oil tankers), or other reasons. Because there is no way to know from this data, it was chosen to keep the port calls with a longer travel time than expected in the data.

A possible limitation of the ship-visiting, port-sharing, and route-sharing networks is that the threshold of adding an edge was set

to three, because it was considered relevant. Especially in the port-sharing network you can see that this leads to a very densely connected network without the scale-free property. The threshold could haven been chosen in a data-driven manner.

For the co-sailing network the ferry category was excluded, while the other networks do contain this category. Future work regarding ship traffic behaviour should consider focusing on either ferry or cargo categories, because of differences between the two.

Future work regarding co-sailing behaviour should consider investigating whether this behaviour can be explained by the geographical preference of ships which was found in the ship-visiting network. It is also interesting to investigate the properties of the ships in the regional communities in the ship-visiting network further, for example to examine if ships in a certain group are non-compliant more often.

REFERENCES

[1] European Maritime Safety Agency Information System THETIS. http://www.emsa.europa.eu/psc-main/thetis.html. Accessed: 2020-06-22.
[2] Google Maps. https://www.google.nl/maps. Accessed: 2020-06-01.
[3] Green Deal on Maritime and Inland Shipping and Ports. https://www.greendeals.nl/sites/default/files/2019-11/GD230%20Green%20Deal%20on%20Maritime%20and%20Inland%20shipping%20and%20Ports.pdf. Accessed: 2020-06-12.
[4] World Atlas The Busiest Cargo Ports in Europe. https://www.worldatlas.com/articles/the-busiest-cargo-ports-in-europe.html. Accessed: 2020-06-12.
[5] J. Alstott and D. P. Bullmore. powerlaw: a python package for analysis of heavy-tailed distributions. *PloS one*, 9(1), 2014.
[6] L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley. Classes of small-world networks. *Proceedings of the national academy of sciences*, 97(21):11149–11152, 2000.
[7] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.
[8] A.-L. Barabási et al. *Network science*. Cambridge University Press, 2016.
[9] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.
[10] A. D. Broido and A. Clauset. Scale-free networks are rare. *Nature communications*, 10(1):1–10, 2019.
[11] 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 *International Conference on Complex Networks and their Applications*, pages 223–235. Springer, 2018.
[12] L. Despalatović, T. Vojković, and D. Vukicevic. Community structure in networks: Girvan-newman algorithm improvement. In *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 997–1002. IEEE, 2014.
[13] C. Ducruet and F. Zaidi. Maritime constellations: a complex network approach to shipping and ports. *Maritime Policy & Management*, 39(2):151–168, 2012.
[14] D. Easley, J. Kleinberg, et al. *Networks, crowds, and markets*, volume 8. Cambridge University Press Cambridge, 2010.
[15] A. Hagberg, P. Swart, and D. S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.
[16] Y. Hu and D. Zhu. Empirical analysis of the worldwide maritime transportation network. *Physica A: Statistical Mechanics and its Applications*, 388(10):2061–2071, 2009.
[17] P. Kaluza, A. Kölzsch, M. T. Gastner, and B. Blasius. The complex network of global cargo ship movements. *Journal of the Royal Society Interface*, 7(48):1093–1103, 2010.
[18] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world'networks. *Nature*, 393(6684):440, 1998.
[19] X. Xu, J. Hu, and F. Liu. Empirical analysis of the ship-transport network of china. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 17(2):023129, 2007.

APPENDIX

A DATA DESCRIPTION

In this appendix, additional data cleaning steps are described. The data set contains 8 columns considered relevant to this study. These columns are *ship id*, *ship type*, *flag state*, *gross tonnage*, *time of arrival*, *time of departure*, *current port*, *previous port*.

The variable *previous port* was not present in the data set by EMSA [1], but was added based on the sequence of port calls. Port calls where *time of arrival* was later than *time of departure* were excluded. Port calls where *time of departure* of the ship was after *time of arrival* of the ship in another port were excluded.

If an observation had a missing value in any of the relevant columns, it was excluded. Also, a unique ship was excluded if the information in the columns *ship type*, *flag state* and/or *gross tonnage* was not identical in all observations of that ship.

Coordinates of ports were found using Google Maps [2] and connected to the port names present in the data set.

The categories that were mentioned in 3.2 are the following. The bulk category includes bulk carrier ships. The tanker category includes the ship types: chemical tanker, combination carrier, NLS tanker, oil tanker, oil tanker/chemical tanker, oil tanker/NLS tanker. The container category includes container ships. The category gas carrier includes the gas carriers. The cargo category includes the ship types: general cargo/multipurpose, high speed cargo, livestock carrier, refrigerated cargo, Ro-Ro cargo. And lastly the ferry category included the ship types: high speed passenger craft, passenger ship, Ro-Ro passenger ship.

B RESULTS CO-SAILING NETWORK

Figures 16, 17, and 18 show frequency distributions of ship types of the whole co-sailing network and two communities. It can be concluded that Figure 17, and 18 show different frequency distributions than 16.

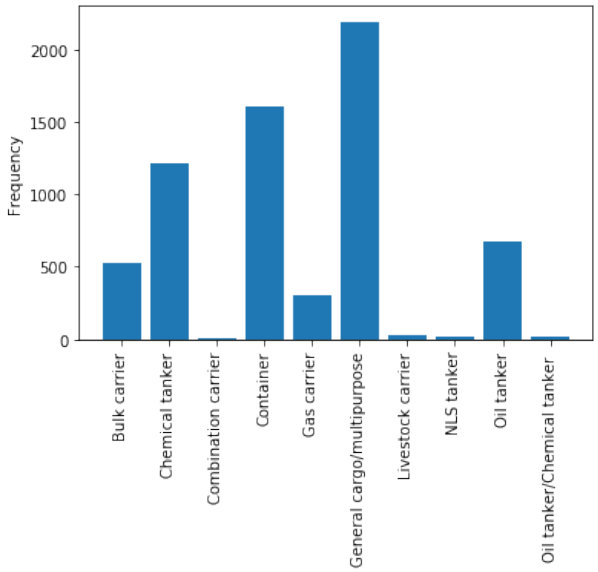

Figure 16: Frequency dist for the whole network for the co-sailing network.

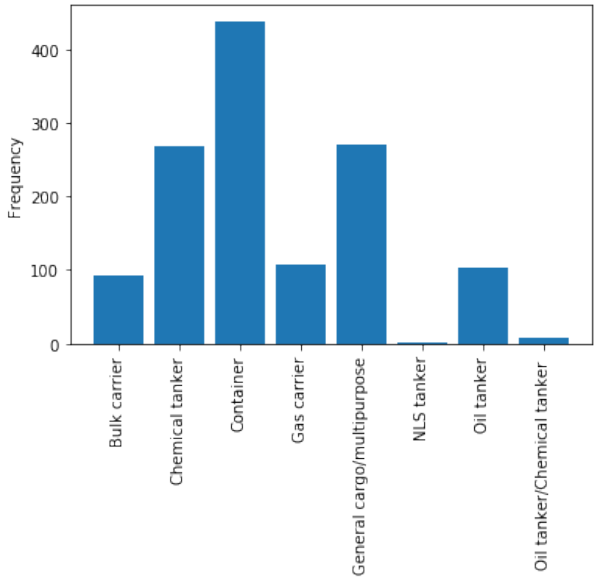

Figure 17: Frequency dist for community '0' for the co-sailing network.

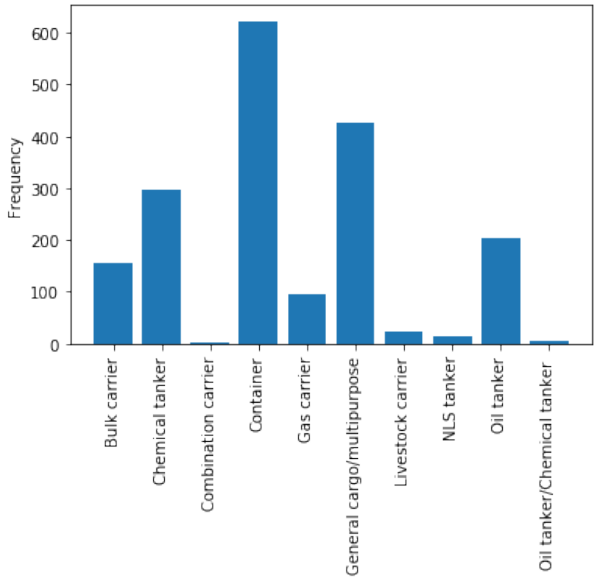

Figure 18: Frequency dist for community '4' for the co-sailing network.