Uncovering vessel movement patterns from AIS data with graph evolution analysis

Emanuele Carlini
Institute of Information Science and Technologies, National Research Council (CNR)
Pisa, Italy
emanuele.carlini@isti.cnr.it

Vinicius Monteiro de Lira
Institute of Information Science and Technologies, National Research Council (CNR)
Pisa, Italy
vinicius.monteirodelira@isti.cnr.it

Amilcar Soares
Institute for Big Data Analytics
Dalhousie University
Halifax, Canada
amilcar.soares@dal.ca

Mohammad Etemad
Institute for Big Data Analytics
Dalhousie University
Halifax, Canada
etemad@dal.ca

Bruno Brandoli Machado
Institute for Big Data Analytics
Dalhousie University
Halifax, Canada
brunobrandoli@dal.ca

Stan Matwin
Institute for Big Data Analytics
Dalhousie University
Halifax, Canada
stan@cs.dal.ca

ABSTRACT

The availability of the large amount of Automatic Identification System (AIS) data has fostered many studies on maritime vessel traffic during the recent years, often representing vessels and ports relationships as graphs. Although the continuous research effort, only a few works explicitly study the evolution of such graphs and often consider coarse-grained time intervals. In this context, our ultimate goal is to fill this gap by providing a systematic study in the graph evolution by considering voyages over time. a three years of AIS data from the coastal waters of United States. By mining the arrivals and departures of vessels from ports, we build a graph consisting of vessel voyages between ports. We then provide a study on topological features calculated from such graphs with a strong focus on their temporal evolution. Finally, we discuss the main limitations of our approach and the future perspectives that will spawn from this work.

1 INTRODUCTION

Maritime transportation represents 90% of international trade volume and plays a paramount role in today's economy, in terms of cargo shipping, passenger transportation, leisure navigation, and fishing operation[21]. Globalization and multiple mode transportation of goods in the shipping industry resulted in a large extension of maritime vessel route network. The study of vessel movements is a well-established source of information to understand the role of maritime routes and ports in economic, social, and environmental contexts. These studies include maritime traffic control and prediction [18], human migration flows [8], bioinvasion [9] and maritime piracy [20]. However, such a role cannot be properly unraveled by looking at ports and routes in isolation, but rather they must be put in relation to one another. This allows the study of the interplay of all the components in the complex maritime network, and it is even more important for understanding the evolution over time of that interplay.

A central concept for the analytical study of vessel routes is the Global Shipping Network (GSN), in which nodes are ports and edges are the routes between ports of cargo ships. Figure 1 illustrates the GSN resulting from the vessels' routes of 2017 in the American coast using the MarineCadastre.gov dataset [16].

Since the automatic identification system (AIS) for vessels was made mandatory in 2004 [1], there has been a surge of studies on the GSN and other maritime networks that use such data. Many works modeled the GSN based on graph theory [19, 22], but only a few of them analyzed the network in terms of its evolution over the years [14, 17]. Also, those works which studied the network evolution used private data, and performed exciting but high-level and coarse-grained analysis, such as in [6].

The main goal of our analysis is to provide a systematic study of the evolution aspect regarding maritime vessel routes, with the purpose of identifying recurrent patterns in their evolution. The analysis is based on publicly available data and with a defined and well documented data model. This aspect is fundamental for the reproducibility of our analysis and its expansion and updating of results when new data arrives. Also, it considers the two necessary dimensions of *time* and *layers* (i.e., the evolution of the network can be observed for multiple types of vessels, such as cargo and passengers).

Such an ambitious objective has some inherent challenges that must be tackled. First, the analysis of AIS datasets typically presents a Big Data challenge: the datasets are usually very large and data can arrive at a high rate. For example, ExactEarth alone claims to consistently track 165,000 vessels and compiling over

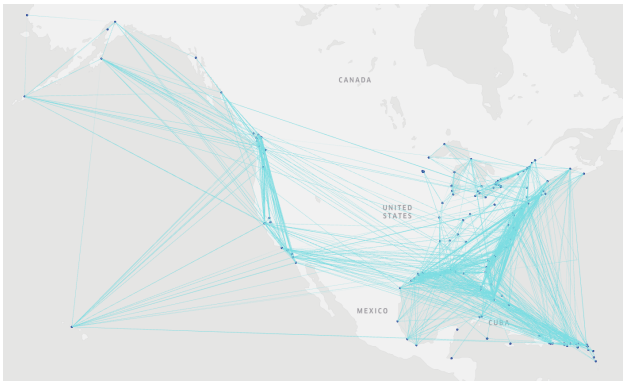


Figure 1: American coast vessels' routes in 2017. The nodes represent ports, and the edges are voyages between two ports

7,000,000 AIS messages daily[1], but data still need to be processed in a reasonable time.

Second, the purpose of any network analysis is to abstract the complexity of a system in order to extract meaningful information which are not directly available when the individual components are examined separately. Therefore, the definition of a network that encompasses time information is a complex task. Suitable approaches need to be carefully selected to study the evolving network.

The contribution of this work can be outlined as the following:

- We propose an approach that uses AIS data to extract connections between ports derived from the vessels' movements. From these connections (or *voyages*) we build graphs in which the vertices correspond to the ports, whereas the edges or links correspond to the vessel voyage between two different ports. In addition, each edge has a semantic defined by the vessels' type.
- We study several topological properties of the temporal graphs generated from vessels' movements and how these features evolve over time. Specifically, we investigate features relative to graphs dimension, ability to form clusters, and geographical spatiality.
- We design and run an extensive analysis using a real dataset with AIS Data collected from vessels navigating in the U.S. coastal area. In our experiments, we investigate the aspect of stationarity of the time series of the topological properties of the graph and discuss the obtained insights.

The rest of the paper is organized as follows. Section 2 discusses related works. Section 3 introduces some concepts used through the paper. In Section 4 we describe our approach for deriving time-series of topological properties from graphs based on vessels' visits to ports. We perform some analysis in Section 5 and draw conclusions envisioning future works in Section 6.

2 RELATED WORK

The work done in [11] is one of the first to study the concept of GSN as a complex network. They use information about the itineraries of 16363 ships of three types (bulk dry carriers, container ships, and oil tankers) during the year 2007 to build a network of links between ports. The work of [11] shows that the three categories of ships differ in their mobility patterns and networks. Their results show that container ships follow regularly repeating paths, whereas bulk dry carriers and oil tankers move less predictably between ports. They also show that the network of all ship movements possesses a heavy-tailed distribution for the connectivity of ports and the loads transported on the links with systematic differences between ship types [11].

The work of [14] also uses a sample of the Lloyds database with the world container ship fleet movements from Chinese ports from the years of 2008 to 2010. The objective of their work is to look at changes in the maritime network prior to and after the financial crisis (2008-2010) and to analyze the extent to which large ports have seen their position within the network change. The authors show how the global and local importance of a port can be measured using graph theory concepts. They also show that the goods transportation network was contracted with respect to port throughput, but no contraction in the distribution capacity of the main hub ports was found [14]. Finally, the authors show that there are new port regions placed in the entrance and exit of the Panama Canal, and there are several significant business opportunities in that region.

A study of topological changes in the maritime trade network is shown in [13]. The authors propose two new measures of network navigability called *random walk discovery* and *escape difficulty*. Their results show that the maritime network evolves by increasing its navigability while doubling the number of active ports. The authors suggest that unlike in other real-world evolving networks studied in the literature up to date, the maritime network does not densify over time, and its effective diameter remains constant [13].

In [6], the author investigates the degree of overlap among the different layers of circulation composing global maritime flows. His work uses several methods from complex network analysis to understand the dynamics affecting the evolution of ports and shipping. The results show that there is a strong and path-dependent influence of multiplexity on traffic volume, range of interaction, and centrality from various perspectives (e.g., matrices correlations, homophily, assortativity, and single linkage analysis) [6]. When growing the network and concentrating the analysis around large hubs over time, results show that the traffic distribution is place-dependent due to the reinforced position of already established nodes [6].

The work of [22] builds a GSN using the 2015 AIS data of the world with multiple spatial levels. Their process mainly consists of five steps, where the first three generate the network nodes and the last two create the network links. The work of [22] applies the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to detect where ships stops and cross this information with terminal candidates of ports. A directed GSN is generated with the trip statistics between two nodes as the edges. Their work evaluates features such as average degree and betweenness centrality of each node, average shortest path length between any two nodes, and community clusters of the GSNs.

Following a similar idea of building GSNs, but with focus on anomaly detection, the work of [19] provides a mechanism that classifies vessel behavior in normal and abnormal, using historical information about similar vessels that operate in a particular area. In [19], the authors identify waypoints (i.e., a region of interest for a given application) that characterizes the operations and the sort of movement patterns that they follow (i.e., the nodes). As edges, the work of [19] uses the subtrajectories that links two waypoints, using also the extracted features of those subtrajectories for analysis. They identified each edge by the subtrajectory that links two ways points. Features of each edge is generated using a trajectory mining library introduced by [7]. Their analysis tries to detect outliers from the subtrajectory features (e.g., course over ground, speed over ground, etc) and using transition probabilities as the edges of the network.

Differently from [6, 11, 13, 14], our work use AIS data to determine vessel routes. Differently from [22], which uses stop points as nodes and evaluate centrality, shortest-path and communities, our work is focused on the evolution of the network instead of using waypoints as nodes and being focused on anomaly detection like [19], or using stop points as nodes to evaluate centrality, shortest-path, and communities like [22], we use the ports as nodes and we evaluate the evolution of the network as our primary task.

[1] https://www.exactearth.com/products/exactais

3 PROBLEM FORMULATION AND DEFINITIONS

Vessels report their location through AIS messages while navigating. A vessel may send AIS messages with a frequency that varies from a few seconds to a few minutes, depending on the type of message. When they are at the underway, they may send AIS messages every 2 to 10 seconds, while when they are at anchor, this time window can increase to every 3 minutes [22]. Therefore, positional information extracted from AIS messages can be seen as a representation of the spatial-temporal movement of a traveling vessel. We are interested in this spatial information with the intent of understating when a vessel is visiting a port. By merging subsequent visits to ports of vessels, it is then possible to build the sequence of vessels' voyages. Then the idea is to construct a *graph* (or network) representation out of these voyages in given interval of time, to be able to study the evolution of the graph with complex network mechanisms.

Graphs have some properties useful to unravel interesting information about the dynamism between two and more entities. In particular, in the context of a voyages graph, the topological properties of the graph can help us identify relevant characteristics within a network that would not have meaningful information if the individual entities were examined separately [6, 13]. Topological properties can be applied to the network as a whole or to individual nodes and edges. In particular, for our study, we are concerned with global network properties. Below, we provide a compact formalization of a AIS message as well as of some other important concepts that covers the scope of this paper.

Definition 3.1. (*AIS Message*): An AIS Message m_e is a tuple (x, y, t, c) that represents the GPS coordinate (x, y) at a time stamp t assigned to a vessel e of type c. We define M as the set of all AIS messages.

Definition 3.2. (*Port*): A Sea Port p is represented as a tuple (id, x, y), where x and y are the latitude and longitude coordinates of its geographical center, and id is the code that identifies the port. We also define the spatial function $\alpha(p, r)$ that defines a circular area of radius r centered on the coordinates of port p.

Definition 3.3. (*Visit*): Given a radius r, we formally define a visit $v = (p, m_e, t)$ of the vessel e to a port p when it exists at least one $m_e \in M$ at time t whose coordinates x and y are in the area defined by $\alpha(p, r)$.

Definition 3.4. (*Voyage*): A voyage $vj = (v_1, v_2)$ is a pair of visits, such that $v_1(p) \neq v_2(p)$ and $v_1(e) = v_2(e)$ and $v_1(t) < v_2(t)$ and they are consecutive (there is no visits to other ports between them for the same vessel). The ports of v_1 and v_2 are called respectively origin and destination ports. The duration of the voyage $vj(d)$ is the time of the last visit of e in the origin port and the time of first visit in the destination port.

Definition 3.5. (*Voyage Graph*): The Voyage Graph (VG) is a graph $G = (V, E)$ built according to a set of voyages VJ, in which V contains all the ports in VJ and E contains an edge for each unique pair of ports in VJ.

Definition 3.6. (*Voyage Graph Snapshot*): The Voyage Graph Snapshot (VGS) $G_w = (V, E)_w$ is an extension of the VG that includes a temporal time window w used to create the snapshot of the graph. This interval is used to select the AIS messages that will be used to build the snapshot G_w of VG G.

Algorithm 1: Trips Graph Snapshot Extraction

Input : M: AIS Messages divided into areas
w: Time window size
s: Time shifting
r: Radius
P: Set of ports
Output: G: set of Voyage Graph Snapshots
Init : $G \leftarrow \emptyset$; $V \leftarrow \emptyset$

1 $V \leftarrow$ `Visits`(A_i, P, r);
2 $R_1 \leftarrow$ `Voyages`(V);
3 $R_2 \leftarrow$ `Clean`(R_1);
4 **foreach** $b \in$ `Buckets`(w, s) **do**
5 $\quad G \leftarrow G +$ `Graph`(R_2, b);
6 **return** G

4 METHODOLOGY

This section describes the procedural methodology that we use to transform raw AIS data to a set of time series observations on the topological features of voyage graphs. We first describe the algorithm to build the voyage graph from the source data, then which features we extract from the graph, and finally how we create the time series.

4.1 Building Voyage Graphs

Building the set of voyage graph snapshot directly from the original data would be possible, but also very unpractical. The dataset has a lot of noise, entries are not ordered in time, and much information is redundant. Therefore, we applied an incremental approach to process the data. First, we exploit the fact that the original dataset is already divided into geographical areas (i.e. the zones of the MarineCadastre.gov dateset, see Section5). For each area, we remove redundant information (e.g., subsequent entries for the same vessel in the same port, with the same timestamp) and compute the set of *visits*. Second, we compute the set of *voyages* using the *visits*, removing incorrect and noisy *voyages*. Third, from the *voyages* we can determine the *graph* and the *snapshot* according to the time window considered. This incremental process has several advantages: (i) graphs building is very fast, as the set of *voyages* is practically an edge list; (ii) the costly cleaning process is done only once, and from the clean collection of *voyages* it is possible to build multiple graphs; and (iii) it is interesting to study *visits* and *voyages* without transforming them into a graph.

The pseudocode in Algorithm 1 illustrates the steps that extract all the *voyage graph snapshots* $(G_0, \ldots, G_n)$ for a given set of AIS Messages M. Besides M, the algorithm also receives as parameters a Radius r, a set of ports P, a time window size w, and a shifting s. The shifting parameter s indicates the amount of days to move forward the time window for each successive graph. For example, with $s = 10$, if the first time window starts at the 1st of January, the second starts at the 10th of January, and so on.

The algorithm starts by extracting the set of visits V (see definition 3.3) from M in each geographical area A_i in the original dataset with the function $Visits()$ (line 1). This function returns $V \subset A_i$, such that V contains the AIS records transmitted inside the area $\alpha(p, r)$ (see definition 3.2) for each port $p \in P$. Depending on r, there could be overlapping port areas such that the same

AIS record results transmitted inside multiple ports. In this cases, we discriminate by associating the record to the closest port.

In turn, V feeds the function $Voyages()$ that extracts the voyages of the vessels (line 2). The underlying assumption is that if a vessel e is seen at the port $p_o \in P$ at time t_0 and e is also seen at the port p_f at time t_1 and $t_1 > t_0$, then we record a voyage (definition 3.4).

The set of voyages is then cleaned by the function $Clean()$ (line 3), to retain only the valid voyages. AIS data inevitably contains noise due to many reasons, including malfunctions, errors in transmission, and malicious use. In our context, such noise and mistakes translate into incorrect voyages. To remove them, we performed two cleaning actions. First, we removed those entries having null or invalid data in relevant fields (typically position or vessel type). For example, several incorrect entries had a vessel identifier, called Maritime Mobile Service Identity (MMSI), whose value is composed only by zeroes, which may indicate a placeholder for missing MMSI. We then removed the voyages that, depending on their duration (vj_d) and length, imply an impossible speed for a vessel. We set the length of a voyage by computing the geodesic distance between starting and arrival ports. The geodesic distance is the minimum distance between two points on earth, which can be used as the shortest possible maritime voyage. By using the length and the duration, we compute the estimated average speed based and removed those voyages whose speed exceeds 60 knots (which is still very high speed, but we left some margin to cope with a possible degree of approximation in the data). We did not remove the slow speed voyages as we cannot estimate how long is the actual maritime route between two ports with respect to the geodesic distance.

Finally, the algorithm builds the set of Voyage Graphs (definition 3.5) thought the function $Graph()$ using the clean set of voyages R_2 (line 5). This function uses the ports as vertices and the voyages as edges to build the graph. It considers the function $Bucket()$ that creates all the possible time windows for the parameters w and s.

4.2 Topological Voyage Graph Features

In order to study the evolution of the voyages graph, we employ a set of graph metrics that are related to the different aspects we wish to evaluate, called Topological Voyage Graph Features (TVGs). In this paper, we studied only global network metrics (i.e., related to the whole graphs). We used the *networkx* library [10] to compute such metrics.

An important metric in studying and comparing graphs is the dimensions of the graph. To this end, we have considered the *order* (i.e., amount of nodes/ports in the graph) (f_n) and the *size* (i.e., amount of edges) (f_e). Semantically, these two measures are connected. A higher *order* and *size* can indicate an increased overall vessel traffic, and/or the tendency to perform less conservative routes, as more ports are involved. Inversely, a lower order and size can indicate a decreased overall vessel traffic, and/or the tendency to perform more conservative routes, as less ports are involved.

A relevant aspect is the identification of cohesive subgroups of ports in the graph, as a way to identify those ports that share a strong tie in the traffic for a particular vessel type. The number of Connected Components (CC, f_c) is the number of subgraphs in which any node is connected to each other by edges, and which is not connected to another subgraph. The number of Connected Components (f_c) indicates how much the graph represents a global scale activity (low CC number), rather than composed by a set of not connected and local activities (high CC number). The Average Clustering coefficient (f_a), is the average of local clustering of nodes. The local clustering of each node in the graph is the fraction of triangles (set of 3 vertices such that any two of them are connected by an edge) that exist over all possible triangles in its neighborhood. In other words, this coefficient represents the probability that two neighbors of a node are neighbors themselves, and can serve to evaluate how many voyages happen around the same set of ports.

Another essential metric for the maritime networks is the extension of their geographical spatiality, as the distance between ports can be directed linked with various cost aspects as fuel consumption, maintenance rates, and insurance costs [6]. In addition, such a metric can give insights on whether certain vessel types are more oriented toward short or long routes. In this paper we have considered the average (f_d) and weighted average (f_w) link distance. The former represents the geodesic distances between connected ports in the graph, expressed in kilometers. The weighted version multiplies each distance by the number of voyages made between the two ports.

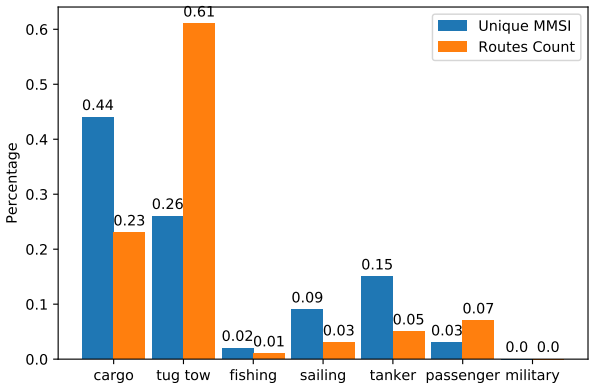


Figure 2: Total percentage of unique MMSI and voyages by vessel type

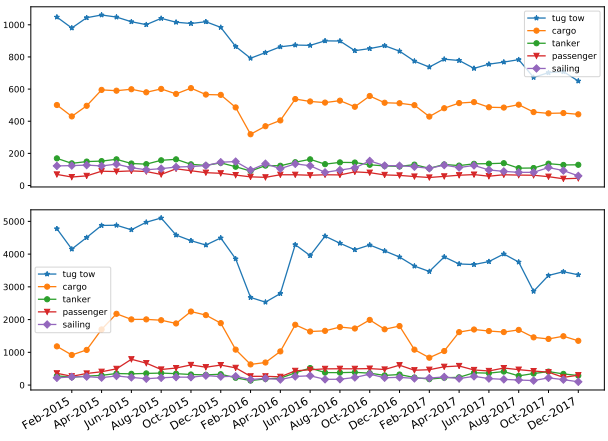
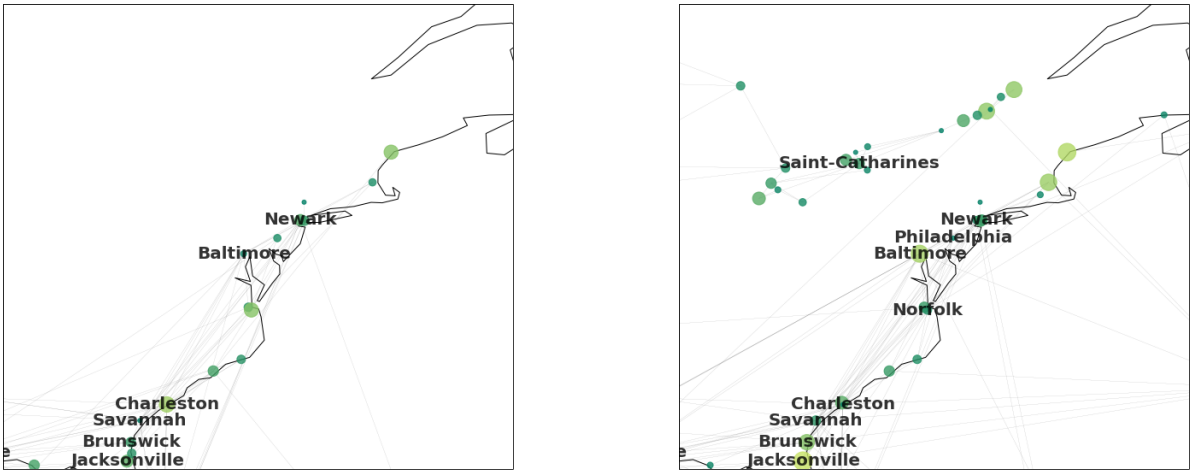


Figure 3: Unique MMSI and voyages count per vessel type over time



(a) February 2016

(b) August 2016

Figure 4: Graphs of North East US in winter (a) and summer (b). Larger and lighter nodes have an higher degree.

4.3 TVGs Time-Series

A time-series is a collection of observations made sequentially over the time [3]. Examples include (i) sales of a particular product in successive months, (ii) electricity consumption in a particular area for successive one-hour periods.

In particular, we build time-series of the TVG features. For this purpose, we use Algorithm 1 to return the set of Voyages Graph Snapshot G, compute the TVGs for each one of the graph $g_i \in G$ and store them as separated time-series T_j where j indicate the TVG feature (i.e. number of nodes (f_n), number of edges (f_e), average clustering (f_a), number of strongly connected components (f_c), the average (f_d) and weighted average (f_w) link distance.

Let $function_j$ be the function that compute the topological metric of a given graph g indicated by the index j, for instance if j is equal to f_n then $function_j$ computes the number of edges on the graph g. Also, let I be the set of existing temporal buckets in G and let J be the set of TVG features.
We then formalize the TVG time-series (T^f) with the following equations:

$$T_j \coloneqq \{\, function_j(g_i) \mid g_i \in G, i \in I, j \in J \,\}$$

where:

$$J \coloneqq \{f_n, f_e, f_a, f_c, f_d, f_w\}$$
$$I \coloneqq \{w_1, w_2, ..., w_n\}$$

5 EXPERIMENTS

In this section, we present the experiments conducted to assess the stationary behavior of the TVG features. We perform such analysis for the different spectrum of types of vessels. In the experiments, to build the Voyage Graph Snapshot using Algorithm 1, we assign respectively the values $r = 5km$, $s = 10days$, and $w = 30days$.

Our source of AIS data is the MarineCadastre.gov dataset [16], which contains filtered AIS records for the US coastal waters for the years 2015-2017, for a total of about 934 GB. Records are sampled to one minute and organized in the comma-separated value (CSV) format, for a total of about 8 billions of records.

Additionally, we use the Sea-Ports dataset [15] that contains spatial information, such as latitude and longitude, of all known seaports in the world. From the original datasets, we have retained only the ports of North America.

The experiments conducted aim to answer the following research questions comprehensively. Our research question can be summarized as the following: ***Are the TVG time series stationary?*** Stationarity is an essential characteristic of a time series. A time series is said to be stationary if its statistical properties do not change over time. In other words, it has constant mean and variance, and covariance is independent of time. Section 5.2 address this research question. To address this research question, we will use a statistical test designed to comment on whether a time series is stationary explicitly.

5.1 Data Overview

Figure 2 shows, in percentage, the amount of unique MMSIs and route counts by vessel type and for the whole dataset. In the original dataset, there are several vessel categories: cargo, passenger, sailing, military, fishing, tanker, and tug tow. Interestingly, for cargo and tanker, a relevant percentage of unique MMSI correspond to a much lesser portion of total voyages. This is expected as these types of vessels perform less but longer voyages concerning other vessel types. By comparison, tug boats with 26% unique vessels have the 61% of total routes. Unlike cargo and tankers, these routes are very short and relative to the vessels moving between nearby ports.

Among all categories, military and fishing account only for 1% of the total voyages, meaning that there is too few data to build meaningful graphs for these categories. We also did not consider tug tows as their routes mostly regard only two ports resulting in not interesting graphs. Therefore, from now on, we consider only the following categories: cargo (which includes also tanker) passenger, and sailing.

Figure 3 top and bottom shows respectively the amount of unique MMSIs and voyages count for each vessel type and for each month in the period considered. Looking at cargo and tug tows, we can notice an evident pattern of less naval traffic in the winter periods (for example from February to March 2016), in which both the number of unique MMSI and voyages has a clear drop. The drops is also visible (especially for cargo) in the same period of 2015 and 2017. Also, it interesting to notice the

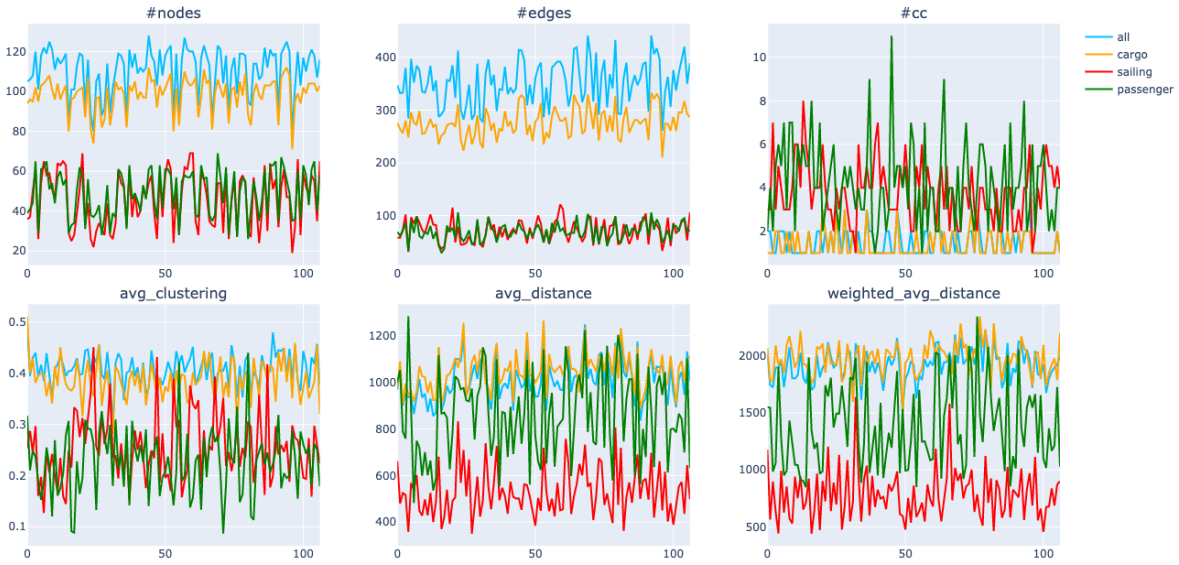


Figure 5: TVG Time-Series

similarity in the trend of the cargo and tug tow, which can be explained by the fact the tug boats are mostly used to help large vessels, such as cargoes, into ports.

Figure 4 shows the graph resulting from cargo vessels in the northeast US in February-2016 (Figure 4 a) and August-2016 (Figure 4 b). It is clear from the graph that the traffic is more abundant during the summer because of the better climatic conditions. The graphs also show the complete halt of voyages in the great lakes during winter due to the formation of ice in the lakes.

Figure 5 shows the extracted time-series using the MarineCadastre.gov dataset for the features: number of nodes, number of edges, number of connected components, and average clustering, average distances and weighted average distance. For each TGV feature, we consider four different series: *passenger*, *sailing*, *cargo*, and *all*.

5.2 Stationary behavior analysis

In this section, we address our research question of investigating the (non) stationary behavior of the TVGs. A time series is stationary if they do not have a trend or seasonal effects. This means that the statistics calculated on the time series such as the mean, variance, and auto-correlation of the observations are consistent over time [2]. Most statistical forecasting methods are based on the assumption that the time series can be modeled approximately stationary through the use of mathematical transformations [12]. Thus, stationary time series are easier to model. Indeed, statistical modeling methods assume or require the time series to be stationary to be effective.

There are different methods to verify whether a time series is stationary or not. Statistical tests are widely used to analyze if the requirements of stationary are met or have been violated. Here, we adopted the Augmented Dickey-Fuller Test [5] (ADF) that uses an auto-regressive model and optimizes an information criterion across multiple different lag values [4].

The null hypothesis of the test is that the time series is not stationary. The alternative hypothesis (rejecting the null hypothesis) is that the time series is stationary. When interpreting the p-value from the this test, values below a threshold (such as 5% or 1%) suggests to reject the null hypothesis, i.e., the time-series is stationary. While, p-values above the threshold suggests to do not to reject the null hypothesis, meaning that the time-series is non-stationary.

We performed an ADF test on the TVG Time-series extracted from our dataset. The idea is that the more negative (lower) this ADF statistic, the more likely we have a stationary time-series or does not have time-dependent structure. We report the ADF test results on Table 1. The table reports for each TVG feature, and for each serie, the ADF-Statistic, the p-value and the critical values (1%, 5%, and 10%).

By looking at the results of each TGV feature, it is possible to see that most of the series have statistic values lower than the critical value of 1%, except the features *avg_clustering*, for the Serie *sailing* (ADF statistic equals to -2.851), and *#cc*, for the Serie *passenger* (ADF statistic equals to -3.431). This indicates weak evidence against the null hypothesis, so for these 2 cases, considering critical values at 1% level, these two TVG time-series are non-stationary.

Therefore, our analysis using the MarineCadastre.gov dataset suggests that there is a stationary characteristic over the time of the vessel's voyages in the USA coast for the different considered vessels' type (i.e., *all*, *sailing*, *cargo* and *passenger*). This characteristic has been presented in most of the investigated TGVs time-series, and it implies that the voyage graphs at different ranges of time points keep constant on average, without showing significant trends or seasonality during the years of 2015 and 2017.

Serie	ADF	p-Value	Crit 1%	Crit 5%	Crit 10%
# nodes					
all	-9.355	8.104785e-16	-3.494	-2.889	-2.582
cargo	-9.831	5.036417e-17	-3.494	-2.889	-2.582
passenger	-10.303	3.341735e-18	-3.494	-2.889	-2.582
sailing	-8.957	8.412623e-15	-3.494	-2.889	-2.582
# edges					
all	-9.588	2.077671e-16	-3.494	-2.889	-2.582
cargo	-8.962	8.154913e-15	-3.494	-2.889	-2.582
passenger	-10.476	1.249137e-18	-3.494	-2.889	-2.582
sailing	-9.668	1.296871e-16	-3.494	-2.889	-2.582
avg_clustering					
all	-3.633	5.152565e-03	-3.499	-2.892	-2.583
cargo	-8.795	2.182865e-14	-3.494	-2.889	-2.582
passenger	-9.243	1.561285e-15	-3.494	-2.889	-2.582
sailing	-2.851	5.137153e-02	-3.497	-2.891	-2.582
#cc					
all	-8.284	4.431598e-13	-3.494	-2.889	-2.582
cargo	-11.461	5.540629e-21	-3.494	-2.889	-2.582
passenger	-3.431	9.952558e-03	-3.498	-2.891	-2.582
sailing	-8.160	9.189961e-13	-3.494	-2.889	-2.582
avg_distance					
all	-9.729	9.121595e-17	-3.494	-2.889	-2.582
cargo	-9.946	2.595040e-17	-3.494	-2.889	-2.582
passenger	-10.358	2.434362e-18	-3.494	-2.889	-2.582
sailing	-11.361	9.454762e-21	-3.494	-2.889	-2.582
weighted_average_distance					
all	-4.527	1.758368e-04	-3.495	-2.890	-2.582
cargo	-10.204	5.860322e-18	-3.494	-2.889	-2.582
passenger	-9.783	6.668246e-17	-3.494	-2.889	-2.582
sailing	-12.326	6.577707e-23	-3.494	-2.889	-2.582

Table 1: Augmented Dickey-Fuller (ADF) Test

6 CONCLUSION

This paper presented an analysis on the evolution of networks made by voyages of vessels between ports, based on several topological features of the network so called Topological Voyage Graph Features (TVGs). The networks were built in a bottom-up and data-driven fashion, considering 3 years of AIS data of the US coastal waters. The analysis unravelled insights about the stationarity of several features over time, including the number of nodes and edges, clustering coefficient and geodesic distance between the ports (nodes) of the network. The analysis were performed on the MarineCadastre.gov dataset containing vessels' voyages in the USA coast. In particular, for this dataset, we found that for most of the TVGs, the time-series do not present any trend or seasonality behaviors.

Despite being an initial analysis, this work opens new and exiting research perspectives. In the future we plan to extends the geographical range of our studies and consider a complete dataset to produce a worldwide network. In addition, the definition of the spatial area of ports can be improved to increase the precision in voyages detection, in a way similar to the one performed in [22]. Finally, in terms of network analysis, the metrics we considered in this paper are global, but an analysis of local node metrics (such as node centrality) is an interesting future improvement.

ACKNOWLEDGMENT

The authors acknowledge the support of the H2020 EU Project MASTER (Multiple ASpects TrajEctoRy management and analysis) funded under the Marie Skłodowska-Curie grant agreement No 777695.

REFERENCES

[1] 2004. Safety of Life at Sea (SOLAS), Consolidated Edition. (2004).
[2] Peter J Brockwell and Richard A Davis. 2016. *Introduction to time series and forecasting*. springer.
[3] Chris Chatfield. 2000. *Time-series forecasting*. Chapman and Hall/CRC.
[4] Yin-Wong Cheung and Kon S Lai. 1995. Lag order and critical values of the augmented Dickey–Fuller test. *Journal of Business & Economic Statistics* 13, 3 (1995), 277–280.
[5] David A Dickey and Wayne A Fuller. 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association* 74, 366a (1979), 427–431.
[6] César Ducruet. 2017. Multilayer dynamics of complex spatial networks: The case of global maritime flows (1977–2008). *Journal of Transport Geography* 60 (2017), 47–58.
[7] Mohammad Etemad, Amílcar Soares Júnior, and Stan Matwin. 2018. Predicting transportation modes of GPS trajectories using feature engineering and noise removal. In *Canadian Conference on Artificial Intelligence*. Springer, 259–264.
[8] Giorgio Fagiolo and Marina Mastrorillo. 2013. International migration network: Topology and modeling. *Physical Review E* 88, 1 (2013), 012812.
[9] Stephan Gollasch, Chad L. Hewitt, Sarah Bailey, and Matej David. 2019. Introductions and transfers of species by ballast water in the Adriatic Sea. *Marine Pollution Bulletin* 147 (2019), 8 – 15. https://doi.org/10.1016/j.marpolbul.2018.08.054
[10] Aric Hagberg, Pieter Swart, and Daniel S Chult. 2008. *Exploring network structure, dynamics, and function using NetworkX*. Technical Report. Los Alamos National Lab.(LANL), Los Alamos, NM (United States).
[11] Pablo Kaluza, Andrea Kölzsch, Michael T Gastner, and Bernd Blasius. 2010. The complex network of global cargo ship movements. *Journal of the Royal Society Interface* 7, 48 (2010), 1093–1103.
[12] Genshiro Kitagawa and Hirotugu Akaike. 1978. A procedure for the modeling of non-stationary time series. *Annals of the Institute of Statistical Mathematics* 30, 2 (1978), 351–363.
[13] Zuzanna Kosowska-Stamirowska, César Ducruet, and Nishant Rai. 2016. Evolving structure of the maritime trade network: evidence from the Lloyd's Shipping Index (1890–2000). *Journal of Shipping and Trade* 1, 1 (2016), 10.
[14] Fernando González Laxe, Maria Jesus Freire Seoane, and Carlos Pais Montes. 2012. Maritime degree, centrality and vulnerability: port hierarchies and emerging areas in containerized transport (2008–2010). *Journal of Transport Geography* 24 (2012), 33–44.
[15] marchah. (accessed November 2019). *Sea Ports Data*. https://github.com/marchah/sea-ports
[16] MarineCadastre.gov. (accessed November 2019). *Vessel Traffic Data*. https://marinecadastre.gov/ais/
[17] Carlos Pais Montes, Maria Jesus Freire Seoane, and Fernando González Laxe. 2012. General cargo and containership emergent routes: A complex networks description. *Transport Policy* 24 (2012), 126–140.
[18] Lokukaluge P Perera, Paulo Oliveira, and C Guedes Soares. 2012. Maritime traffic monitoring based on vessel detection, tracking, state estimation, and trajectory prediction. *IEEE Transactions on Intelligent Transportation Systems* 13, 3 (2012), 1188–1200.
[19] Iraklis Varlamis, Konstantinos Tserpes, Mohammad Etemad, Amílcar Soares Júnior, and Stan Matwin. 2019. A Network Abstraction of Multi-vessel Trajectory Data for Detecting Anomalies.. In *Proceedings of the Workshops of the EDBT/ICDT 2019 Joint Conference*.
[20] Michele Vespe, Harm Greidanus, and Marlene Alvarez Alvarez. 2015. The declining impact of piracy on maritime transport in the Indian Ocean: Statistical analysis of 5-year vessel tracking data. *Marine Policy* 59 (2015), 9–15.
[21] Michele Vespe, Ingrid Visentini, Karna Bryan, and Paolo Braca. 2012. Unsupervised learning of maritime traffic patterns for anomaly detection. (2012).
[22] Zhihuan Wang, Christophe Claramunt, and Yinhai Wang. 2019. Extracting global shipping networks from massive historical automatic identification system sensor data: a bottom-up approach. *Sensors* 19, 15 (2019), 3363.