UNIVERSITY OF CALIFORNIA SAN DIEGO

Towards Anomalous Group Discovery in Maritime AIS Data

A Thesis submitted in partial satisfaction of the
requirements for the degree Master of Science

in

Data Science

by

Nigel F. Doering

Committee in charge:

Professor David Danks, Chair
Professor Biwei Huang
Professor Tauhidur Rahman

2023

The Thesis of Nigel F. Doering is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

To my fiancée Breona, thank you for your unwavering support, patience, and love over the course of my studies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my gratitude to Professor David Danks for his invaluable support and guidance throughout the course of my thesis. His expertise, insight, and encouragement have been pivotal in shaping both the direction and success of my research. I am also immensely grateful to Professor Tauhidur Rahman for his keen interest in my work and his instrumental role in helping me advance my research goals. His contributions have significantly enriched my academic journey. Furthermore, I extend my heartfelt thanks to Professors David Danks, Tauhidur Rahman, and Biwei Huang for their time and commitment in serving as members of my thesis committee. Their participation, feedback, and expertise have been crucial in bringing this research to fruition. This thesis would not have been possible without their collective wisdom, support, and dedication. I am deeply appreciative of their contributions to my academic and professional growth.

ABSTRACT OF THE THESIS

Towards Anomalous Group Discovery in Maritime AIS Data

by

Nigel F. Doering

Master of Science in Data Science

University of California San Diego, 2023

Professor David Danks, Chair

This research paper presents an innovative approach to anomaly detection in maritime data, focusing on the identification of anomalous vessel groups through Automatic Identification System (AIS) data. Traditional AIS anomaly detection systems are challenged by a narrow focus on individual vessel anomalies which leaves out critical collective action information. Our work seeks to overcome this limitation by introducing a graph modeling methodology that emphasizes the collective behavior of vessels, a critical aspect often overlooked in current frameworks. We work towards a system that detects unusual patterns of coordination among groups of vessels, which is particularly relevant for defense contexts where threats such as smuggling, piracy, and adversarial operations are typically carried out by networks of collaborating entities. The research

contributes to the field by offering a more comprehensive understanding of maritime traffic through the lens of collective dynamics, enhancing the detection capabilities of AIS monitoring systems. The implications of our methodology are far-reaching, providing a foundational strategy for future research in anomaly detection that could be applied to various domains where group coordination plays a pivotal role in defining anomalous behavior. This paper details our graph-based approach and suggests a trajectory for subsequent investigations into the broader applications of detecting coordinated anomalies.

Introduction

Anomaly detection in maritime data is a critical aspect of modern maritime surveillance and safety operations. With the increasing volume of maritime traffic, the Automatic Identification System (AIS) has become an indispensable tool for tracking and monitoring vessel movements across the globe. Anomaly detection algorithms applied to AIS data aim to identify patterns of behavior that deviate from established norms, which could indicate potential threats or unusual activities such as illegal fishing, piracy, or smuggling operations. These algorithms leverage a variety of statistical, machine learning, and rules-based techniques to process and analyze the vast amounts of data generated by AIS, providing insights that are essential for maritime authorities to ensure safe and secure navigation in international waters. The implementation of effective anomaly detection systems thus plays a pivotal role in enhancing maritime security, optimizing traffic management, and safeguarding marine ecosystems by enabling timely responses to irregular events.

Despite the potential of anomaly detection in AIS maritime data, the field faces significant limitations that challenge the robustness and reliability of these systems. One of the primary concerns is the quality and completeness of the data; AIS signals can be lost or go undetected, leading to gaps in tracking information that can obscure a comprehensive view of maritime traffic. Additionally, the AIS protocol lacks robust security measures, making it susceptible to spoofing and other forms of tampering, which can result in false data that misleads detection efforts. Furthermore, many anomaly detection techniques are tailored to specific types of anomalies, such as off-course movements or unexpected stops, and may not be as effective in recognizing other important but less obvious anomalous behaviors. This specialization can lead to a narrow

focus that overlooks a spectrum of irregularities, potentially leaving some threats undetected. The complexity of maritime environments and the dynamic nature of marine traffic add further layers of difficulty, necessitating more sophisticated and adaptive approaches to effectively identify and respond to anomalies in AIS data.

Building on the foundation of AIS anomaly detection, our research endeavors to advance the field by addressing some of its most pressing challenges. We propose a novel graph modeling approach that shifts the focus from individual vessels to the collective behavior of groups, recognizing that many maritime threats, such as smuggling, piracy, and adversarial military operations, are characterized by the coordinated actions of multiple vessels. By analyzing the intricate patterns of vessel coordination, our methodology aims to map out a baseline of normal group behavior within maritime traffic data. This enables identification of groups of vessels whose movements diverge from established patterns, potentially signaling malicious activities. The implications of this research extend beyond maritime surveillance, offering a versatile framework for understanding group dynamics in various domains where coordination is key to identifying collective anomalous behavior. Our comprehensive study not only proposes a methodological shift in anomaly detection but also lays the groundwork for future exploration in the broader context of collective action across different fields.

Chapter 1

The Automatic Identification System

The Automatic Identification System (AIS) began development in the early 1990s as a sophisticated tool to support greater ship-to-ship awareness in addition to radar and visual observation [8]. This development was spurred by the rapid acceleration of maritime activities, primarily owing to increased globalization and population growth [13]. These factors underscored the need for a robust system to preempt vessel collisions and streamline navigation in congested marine areas. Recognizing this, the International Maritime Organization (IMO) assumed authority to develop AIS, with the aim to enhance ship-to-ship and ship-to-shore communication. The system had several foundational objectives. Collision avoidance was the main purpose in the development of AIS by facilitating ship-to-ship communication, for all ships in a certain area [20]. More specifically, AIS information for each ship includes a vessel's identifier number, latitude, longitude, speed (Speed Over Ground), course (Course Over Ground), and other common trajectory information. Armed with this information, ships in proximity can better anticipate and thus avoid potential collisions.

The AIS system shines particularly in areas bustling with maritime activities, like major ports. Here, meticulous coordination is paramount to handle the sheer volume of incoming and outgoing traffic. By making the reporting of AIS data to coastal authorities mandatory, the system facilitates smoother navigation and docking processes, especially for large vessels [7]. This necessary orchestration is evident in Figure 1.1, which showcases a heat map of AIS traffic

surrounding the Port of Long Beach. The intense yellow and red hues, especially closer to the port, signify the elevated coordination level, which evidently dissipates further away from the port, where greater coordination is not as necessary.

AIS transponders, the equipment which sends AIS messages, have found applications beyond ships. They are frequently affixed to maritime equipment like buoys, life jackets, and survival crafts, becoming invaluable assets in search and rescue missions. Moreover, the system empowers coastal countries to monitor and track vessels navigating their territorial waters, ensuring maritime sovereignty. Reinforcing the importance of AIS, an amendment to the Safety of Life at Sea (SOLAS) agreement in 1974 decreed that vessels weighing over 300 gross tons must be equipped with AIS [2]. Over the years, several countries have further tightened these regulations, offering an even more granular view of maritime traffic patterns.

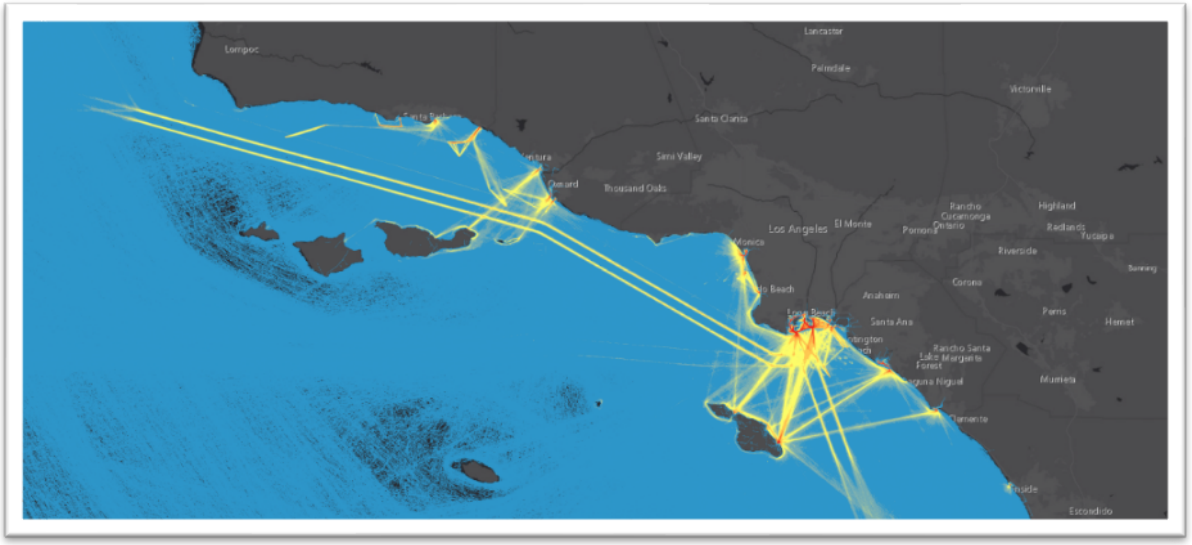

Figure 1.1. AIS Heat Map Traffic for Port of Long Beach. Each blue dot represents an AIS message recorded from a ship for January 1st-2nd 2020.

The Automatic Identification System (AIS) operates on a sophisticated messaging mechanism where transceivers, mounted on ships or positioned onshore, broadcast navigational and positional data via radio signals. This system crafts a localized maritime traffic overview, utilizing a Time-division multiple access (TDMA) protocol to allow multiple vessels to report

information concurrently [34, 38]. The horizontal reach of AIS signals is around 50km, limited by the Earth's curvature. This limitation was overcome with the implementation of satellite-based AIS (S-AIS), which employs satellite transceivers to propagate the transmission of AIS messages further [5, 31]. With S-AIS, the scope of maritime traffic understanding extends from coastal to global. Additionally, the frequency of AIS message reporting varies, with a 2-12 second interval for vessels underway and a 3-minute interval for anchored vessels, as stipulated by the IMO resolution MSC.74(69) [20]. These messages are dispatched automatically by the onboard AIS equipment, ensuring consistent and autonomous data flow. Additionally, there is a graphical interface for vessel crews to visualize nearby maritime traffic, enhancing situational awareness.

A pivotal concern, particularly relevant to the subsequent discussion on data quality, is the non-secure nature of the AIS messaging protocol. It lacks mechanisms for authentication, encryption, or integrity protection, rendering AIS data vulnerable to tampering and various cyber threats [18, 22]. These vulnerabilities may introduce a myriad of data quality issues, making modeling efforts more difficult. The system also does not account for the privacy of the crew, potentially linking AIS messages to specific individuals and exposing sensitive information.

AIS data is segmented into static, dynamic, voyage-related, and safety-related domains, as detailed in Table 1.1. The dynamic domain is particularly crucial for anomaly detection research, as it encompasses trajectory information that is essential for modeling and discerning standard vessel movement patterns. Due to the open architecture of the AIS protocol and its security vulnerabilities, virtually anyone with an AIS receiver can collect AIS messages. This has led to a proliferation of AIS data sources, which are varied in nature, including public, private, and synthetic datasets. One of the primary challenges in maritime anomaly detection research is the absence of a unified data source that consolidates all the information specified in Table 1.1. In practice, many of the features listed in Table 1.1 are frequently absent, resulting in an incomplete picture of a ship's trajectory.

The scarcity of verified anomaly labels poses perhaps the most significant obstacle to the development of effective anomaly detection methods for AIS data. Without a centralized

Table 1.1. Features of AIS Message Traffic. Content adapted from IMO Resolution 74.

Feature Type	**Message Data**
Static Features	• MMSI Identifier • Call Sign and Name • Length and Beam of Ship • Antenna Location
Dynamic Features	• Ship Position • Time in UTC • Course Over Ground (COG) • Speed Over Ground (SOG) • Heading • Rate of Turn • Navigational Status
Voyage Information	• Draught • Hazardous Cargo • Destination
Safety Information	• Text Message

source of anomalous labels, researchers are left to rely on proprietary anomalies that remain confidential, the generation of synthetic anomalies, or the arduous process of manual anomaly detection. These challenges will be addressed in greater detail in the following chapter. This concludes our examination of the AIS system's background. It is imperative to acknowledge that AIS was not initially conceived to identify anomalies or malicious activities, a fact that profoundly influences the limitations of current anomaly detection systems utilizing AIS data.

Chapter 2

Literature Survey

Our literature review commences with an exploration of anomaly detection within maritime data, recognizing the vast array of potential anomalous activities that defy categorization. Despite this unpredictability, domain experts have identified several well-established anomalous patterns, five of which are delineated in Table 2.1. We outline these five types below.

1. **Route Deviation:** Characterized by a vessel's AIS track diverging from established maritime lanes, which are often dictated by geographic constraints such as landmasses, water depths, and prevailing traffic conditions. Such deviations from heavily trafficked routes, particularly by cargo ships, are indicative of this type of anomaly.

2. **Unexpected Activity:** Frequently associated with illicit fishing practices, this type manifests as irregular gaps or sudden reappearances in a vessel's AIS track. This pattern may suggest intentional deactivation of AIS transceivers to conceal movement into restricted zones, such as for illegal fishing operations.

3. **Port Arrival:** These are identified when vessels make unscheduled stops at ports, which could be symptomatic of smuggling activities or the offloading of unauthorized catches in the context of illegal fishing.

4. **Close Approach:** Occurs when two or more vessels maintain a consistently close proximity over time, which is atypical for certain ship classes like cargo or tanker vessels and may

signal covert exchanges or transfers at sea.

5. **Zone Entry:** Involves a vessel entering a restricted area, such as a marine sanctuary or an exclusive economic zone, where general maritime traffic is not permitted.

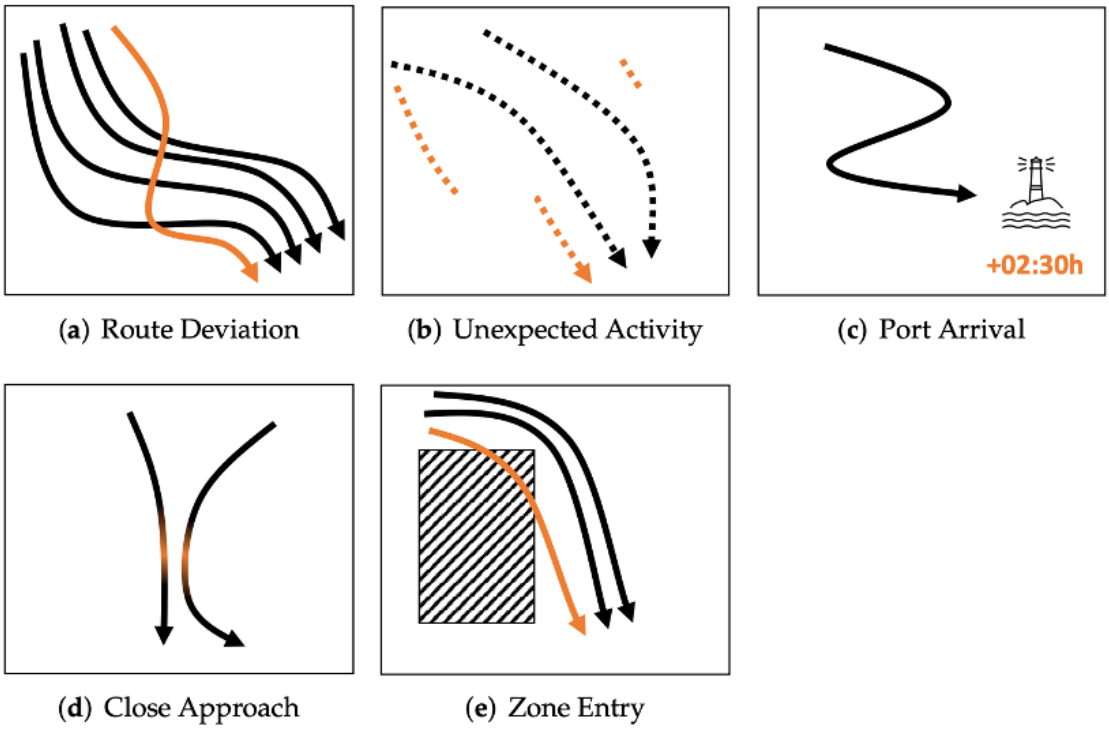

Figure 2.1. Different Types of Maritime Anomalies [43]

A significant challenge in maritime anomaly detection is the inherent instability of AIS data collection. Vessel tracks often exhibit large gaps, which can arise from technical issues with AIS equipment, radio interference, or deliberate manipulation, such as switching the equipment on and off for malicious purposes [29, 21]. Distinguishing between these causes is not straightforward, complicating the construction of accurate vessel tracks from raw AIS data. Furthermore, the vulnerability of AIS to spoofing or tampering can introduce anomalies into the data [3]. Another hurdle is the sheer volume of maritime traffic and the global dispersion of vessels, making manual inspection of all tracks impractical and underscoring the need for automated anomaly detection systems.

While AIS was initially designed for collision avoidance, it has increasingly been adopted for surveillance and security purposes [14, 8, 10]. As the primary source for global vessel tracks, AIS data is pivotal for identifying potentially dangerous or illicit activities. Military applications also benefit from detecting deviations in maritime traffic patterns, which can provide valuable intelligence.

Researchers often focus on identifying specific types of anomalies for particular applications. For example, some studies have concentrated on detecting smuggling operations through the close approach of vessels at sea [23], while others have targeted the on-off switching behavior of AIS equipment [40] or route deviations [37]. These targeted approaches can yield effective application-specific models but require a clear understanding of the expected anomaly patterns within the context of the application. However, given the unpredictable nature of anomalies, this strategy is not without limitations.

Generally, anomaly detection involves modeling 'normal' behavior and then measuring deviations from this model for new data points. This method is favored because it is more feasible to define normality than to anticipate every possible anomaly. Despite the challenges posed by the complexity and imperfections of AIS data, this is the strategy most maritime anomaly detection research adopts. Our survey has identified several methodologies for constructing these models of normal maritime behavior, which we outline in Figure 2.2 [43].

A diverse array of methodologies has been employed to develop automatic anomaly detection models using AIS data, with a significant emphasis on machine learning techniques such as neural networks and clustering, as illustrated in Figure 2.2. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) emerges as a favored method due to its proficiency in identifying clusters of arbitrary shapes—typical in AIS data—and its minimal requirement for input parameters [12]. DBSCAN's versatility also extends to incorporating multi-dimensional features, enabling the inclusion of Speed Over Ground (SOG) and Course Over Ground (COG) values in cluster formation.

The prevalent strategy involves using DBSCAN to delineate clusters representing high-

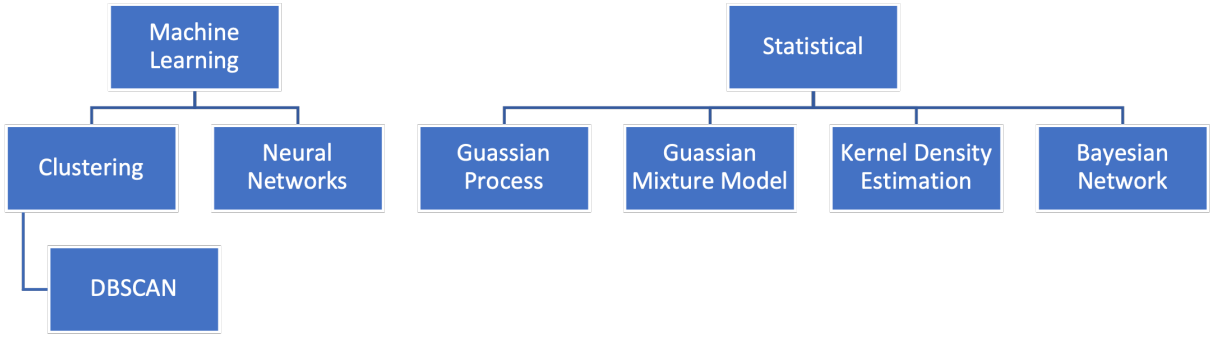

Figure 2.2. A Summary of Techniques for Maritime Anomaly Detection

traffic areas within a geographic region from raw AIS data. Subsequently, AIS tracks are transformed into reference tracks, with each point in the track assigned to the nearest cluster [15, 17, 26]. Normalcy is typically defined by the proximity of AIS points to the cluster centers, with points further away considered anomalous. Tracks with a significant number of points far from any cluster are flagged as anomalous. Variations in DBSCAN-based methodologies often pertain to the preprocessing of AIS data into coherent tracks. For example, [15] employs Piece-wise Linear Segmentation (PLS) to condense vessel tracks, reducing data sparsity. They also propose a technique to infer missing data in vessel tracks by referencing historical data from similar tracks [15]. However, caution is advised with such inferences, as gaps in data could signify illicit activities, and filling these gaps with 'normal' historical data might inadvertently obscure anomalies. A primary limitation of DBSCAN is the dependency on the availability of quality data. In regions with sparse data, DBSCAN may struggle to identify clusters that accurately reflect normal vessel movements. Additionally, the necessity to recalibrate clusters for different geographic regions introduces complexity and potential inconsistencies, which could

hinder the practical deployment of such models in real-world scenarios.

Within the domain of machine learning strategies for anomaly detection, neural network architectures have gained traction for modeling the normal patterns of vessel tracks. One of the pioneering works in this area is presented in [35], which critiques the limitations of the once-prevalent rule-based systems. These systems, reliant on predefined rules set by domain experts, often fail to accommodate the dynamic nature of maritime routes, which can be influenced by various factors such as mission objectives, weather conditions, and geopolitical developments.

In contrast, [35] proposes a learning-based framework where a neural network is trained to understand what constitutes normal maritime behavior. This model allows for domain expert intervention, where experts can adjust the classification of tracks from anomalous to normal during the training phase, thereby refining the learning process. However, the practicality of this method is constrained by the availability of accurately labeled data. AIS data typically lacks explicit normal versus anomalous labels, making it challenging to implement a system that relies on human-labeled data. Such a system would only be viable in a limited scope, with sufficient monitoring resources. Moreover, the assumption that operators can consistently identify anomalies is optimistic, especially when considering the potential for deliberate deceptive practices by vessels.

In the studies [32, 33], AIS messages are treated as imperfect observations of a vessel's underlying state. To capture this 'true state', the authors employ a recurrent neural network (RNN) architecture enhanced with latent variables. The RNN framework is adept at processing sequential data, such as vessel tracks, while the latent variables are introduced to manage the inherent noise and diversity within the AIS data. They adopt the variational RNN (VRNN) model, as formulated in [6], which integrates the latent variables into the learning process. The effectiveness of their model for anomaly detection is assessed by training the VRNN with data from the Gulf of Mexico and testing it with synthetically altered vessel tracks that simulate route deviation anomalies. This method demonstrates considerable success in handling the complexities of AIS data, making it a potentially robust solution for anomaly detection in

maritime surveillance.

Beyond machine learning, statistical methods have also been prominent in AIS anomaly detection research. An excellent example is the work by Laxhammar et al., which applies conformal prediction techniques to AIS data for independent and real-time anomaly detection [25]. Conformal prediction is a statistical framework that enhances machine learning predictions with quantifiable confidence measures [16]. It provides a prediction set that, with a specified level of confidence, say 95%, is likely to contain the true label. Laxhammar et al. created a conformal predictor tailored to AIS track data. Their algorithm forecasts the subsequent position in a vessel's track using historical AIS data. When the conformal predictor assigns a low probability to the actual subsequent position, it flags the behavior as anomalous. One of the key advantages of this method is its adaptability; the conformal predictor can be updated on-the-fly with incoming data, making it particularly suited to the dynamic maritime environment where AIS data is continuously streamed.

Building on the foundational work in AIS anomaly detection using conformal prediction, Smith et al. introduced innovative refinements to the methodology. They proposed employing kernel density estimation to calculate nonconformity scores, a departure from the k-nearest neighbor method previously applied by Laxhammar et al. [25]. This shift to kernel density estimation offers a different statistical perspective on measuring how a new data point compares to historical data. Furthermore, Smith et al. advanced the field by suggesting a p-value averaging approach. This technique aggregates the p-values associated with a set of predictions, aiming to bolster the reliability of the conformal predictor in unsupervised learning environments [41]. Their contributions represent a significant step in enhancing the precision and applicability of statistical methods for anomaly detection in maritime data.

A significant hurdle in advancing anomaly detection within AIS data is the absence of a universally accepted dataset for model training and benchmarking. Researchers often resort to utilizing datasets from a variety of sources, which introduces inconsistencies. Some draw from military or defense contractor databases [24, 23], while others depend on commercial

platforms [9, 19]. There are instances where researchers have procured AIS data independently using personal transceiver equipment, leading to a proliferation of private datasets [40, 36, 41, 4, 11, 27, 30, 44, 9]. The challenge is compounded by the scarcity of ground truth labels for anomalies. As a result, researchers often have to fabricate anomalous tracks to evaluate their models [25, 33, 28]. There are, however, instances where limited ground truth data is accessible, such as reported incidents in European waters [45], suspected illegal fishing activities [9], or through the expertise of domain specialists [42]. This diversity in data sources and the ad hoc nature of ground truth labeling severely hampers the ability to conduct comparative analyses of different methodologies. The field is in dire need of a comprehensive, standardized database of AIS messages coupled with verified anomalies to truly gauge and enhance the efficacy of emerging anomaly detection techniques.

Lastly, we revisit and introduce some of the major limitations to field of anomaly detection for maritime trajectories. First, as shown in Figure 2.1, there are several different types of known maritime anomalies. The bulk of research is, however, focused solely on detecting route deviations and lacks diversity of anomalies able to be detected. This is a considerable barrier to a robust system that can find different manners of suspicious behavior. Second, most research is focused on developing models of "normal" maritime behavior. However, there is severe limitations to this approach when there is a lack of data for specific maritime regions. Additionally, any errors in reporting AIS messages or anomalies themselves may be in the training data, introducing non-normal behavior during model development. Third, as discussed, the lack of an established database for AIS data and known anomalies is a serious limitation to the pursuit of anomaly detection systems for maritime traffic. As summarized in [43], without a common dataset, research is hindered by a lack of transparency, inability to replicate results, and failure to compare the effectiveness of different methodological approaches. Lastly, as recently documented in [39], only around 30% of total maritime vessels report AIS information. Out of the 70% that remain invisible, small recreational craft make up 53% of this portion [39]. Essentially, the lack of a full picture of maritime traffic seriously limits the performance of

models trained only on AIS data. A more robust and operational future approach will need to use multi-modal data, possibly combining AIS messages, satellite imagery, and radar, to build the next generation of models.

Chapter 3

Data

The data in our work is sourced from MarineCadastre.gov [1], a government database for a variety of ocean data. MarineCadastre.gov is a collaboration between the Bureau of Ocean Energy Management (BOEM) and the National Oceanic and Atmospheric Administration (NOAA). The database serves commercial and research efforts aimed at developing a variety of projects such as offshore wind energy, mining efforts, environmental studies, as well as marine traffic analysis. We take advantage of the AccessAIS tool from MarineCadastre.gov, which allows us to highlight a portion of U.S. coastal waters and download all the AIS shipping data from a specified period. We select a data sample from two days, January 1st-2nd, 2020, as our original sample. The methods we develop will be applicable to all data from AccessAIS, but for initial development we focus on a narrow time slot to aid visualizations. A sample of the data is shown in Figure 3.1 below.

Our dataset initially comprised 382,590 AIS messages, encompassing various vessel and trajectory features such as MMSI, time of message, latitude, longitude, Speed Over Ground (SOG), Course Over Ground (COG), heading, vessel name, and more. For our study, we focused on the MMSI, time, latitude, longitude, SOG, and COG features. The original dataset required cleaning to address several issues for more effective analysis. Firstly, we narrowed our scope to include only cargo and tanker vessels, as these types of ships typically provide a consistent and stable series of AIS messages that accurately reflect their routes. A significant challenge in the

	MMSI	BaseDateTime	LAT	LON	SOG	COG	Heading
0	309955000	2020-01-02 09:05:40	33.76864	-118.25179	0.5	0.769690	3.752458
1	309955000	2020-01-02 09:04:31	33.76850	-118.25200	0.8	0.884882	3.822271
2	309955000	2020-01-02 09:02:12	33.76816	-118.25237	0.5	0.689405	3.839724
3	309955000	2020-01-02 09:01:11	33.76802	-118.25249	0.7	0.530580	3.839724
4	309955000	2020-01-02 09:00:02	33.76781	-118.25264	0.8	0.616101	3.822271

Figure 3.1. A Sample of AIS Data from MarineCadastre.gov

data was the inclusion of non-moving or anchored ships. Our objective was to exclude vessels that remained stationary for extended periods. For example, if a ship consistently moved for two hours but then halted for one hour, we retained the data from the moving period and discarded the stationary segment. This was achieved by eliminating AIS messages where the SOG was below 0.5 knots. Although this approach introduced gaps in the ship tracks—where a vessel paused before resuming its journey—we later addressed these gaps with additional cleaning techniques. Next, we focused on eliminating ships with sparse data records. We established a threshold: if a ship had fewer than 100 records over the two-day data collection period, we deemed the data too limited for accurate trajectory interpolation and thus removed all corresponding AIS messages. This step ensured that our analysis was based on sufficiently detailed and representative data.

A significant challenge in our data cleaning process involved handling ship trajectories characterized by dense activity periods, followed by extended stationary phases, and then resuming movement. Segmenting these tracks into distinct trajectories was complex due to the lack of a uniform pattern applicable to all ships. Our solution involved analyzing the time intervals between consecutive AIS messages for each ship, identified by its MMSI. Specifically, we calculated the time difference between each AIS message sent. If the average time difference for a ship exceeded two minutes, we excluded all its records from our dataset. While this approach resulted in the loss of potentially useful data, it was a necessary compromise given

our time constraints, and we still retained ample data for our analysis. Future efforts will aim to refine this data cleaning process, potentially incorporating techniques from sources like [33].

For the AIS messages that remained in our dataset, standardizing the index for each ship track was essential. Since each AIS message is recorded independently, the timestamps for different ships often do not align, complicating further analysis. To address this, we averaged the latitude, longitude, SOG, and COG values for each ship's AIS messages within three-minute intervals. Given that the interval between AIS messages for a single ship is typically 2-3 minutes, this method introduced only minor alterations to the original data. In cases where gaps still existed in a ship's trajectory (i.e., periods exceeding three minutes without an AIS record), we employed linear interpolation to estimate the missing values.

An important step in our data processing was the transformation of the Course Over Ground (COG) values from each AIS message. COG is originally recorded in degrees, which can be misleading as a COG of 359 degrees is almost identical in direction to a COG of 1 degree, yet numerically they appear vastly different. To address this, we applied a circular statistical transformation. Specifically, we converted the COG values to radians and then applied a modulo operation with 2π. This method effectively repositions higher values, such as 359 degrees, closer to lower values like 1 degree, accurately reflecting the small difference in actual direction.

After completing our data cleaning process, we compiled a dataset comprising 27 vessel tracks with a total of 5307 AIS messages. These tracks represent densely recorded ship movements around the Port of Long Beach, as depicted in Figure 3.2. Although the number of tracks may seem limited, they proved to be adequate for the development of our methodology. This approach is scalable and can be applied to much larger datasets in future research."

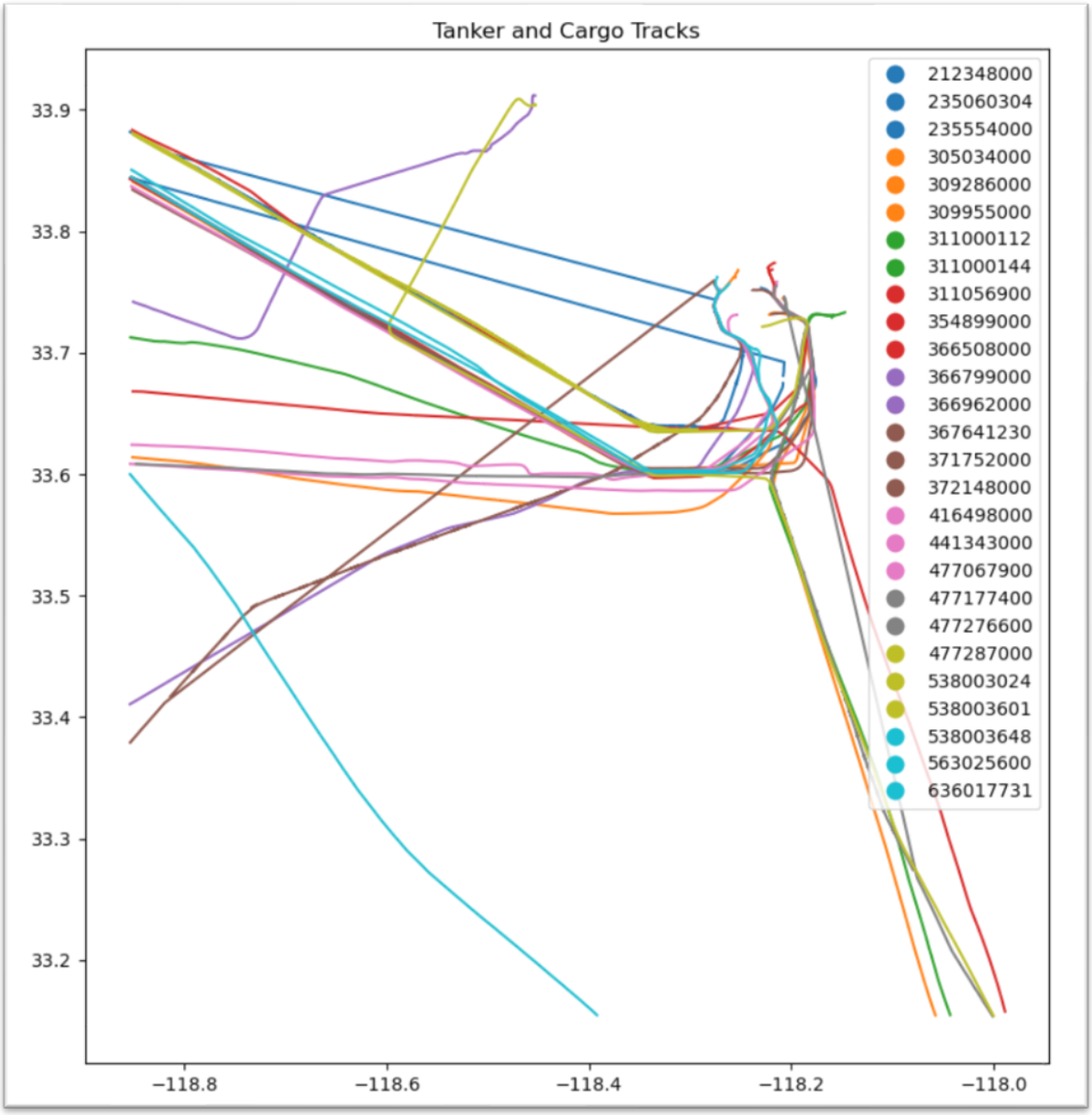

Figure 3.2. Tanker and Cargo Vessel Tracks around the Port of Long Beach.

Chapter 4

Methodology

In our review of existing methodologies, we observed a common approach in anomaly detection for maritime traffic: learning a model of 'normal' behavior based on individual vessel tracks. This model typically incorporates parameters like latitude, longitude, speed over ground (SOG), and course over ground (COG). Anomalies are then identified by measuring deviations from this 'normal' model, often focusing on specific events like route deviations (as illustrated in Figure 1.1). However, this approach primarily targets anomalies at the level of individual data points or events.

In this paper, we propose a shift in focus. Beyond the realm of maritime data, anomaly detection often centers on identifying singular instances, individuals, or events. While this is useful in many scenarios, it may overlook the complexity inherent in collective anomalous behaviors. Our research delves into the detection of groups or networks of anomalies, a concept that holds significant relevance across various domains. In maritime contexts, this approach could be pivotal in the early detection of military naval formations in strategic regions like the Black Sea, or in uncovering smuggling networks orchestrating complex operations to transport illicit goods. Extending beyond maritime scenarios, such as in environmental studies, the objective could be monitoring unusual movement patterns in groups of animals, which may indicate changes in health or environment. This concept of group anomalies is not confined to the maritime domain but is a prevalent phenomenon across numerous fields.

Our objective is to develop a methodology that facilitates the algorithmic discovery of group-level anomalies within maritime data. To achieve this, we introduce a graph construction methodology, enabling us to create a graphical model that captures the interplay and coordination among different vessel tracks. This approach is designed to reveal intricate patterns of coordination and interaction, which traditional single-instance anomaly detection methods might overlook. By doing so, we aim to provide a deeper and more nuanced understanding of group-level anomalous activities across various domains. Furthermore, this graph modeling technique lays the groundwork for future research endeavors, where more sophisticated models of normalcy can be developed based on the graph's structure and characteristics.

4.1 Graph Construction

In the context of identifying group anomalies, our focus extends beyond the individual behavior of each vessel to the collective interactions and coordination among multiple vessels. To capture this dynamic, we have devised a straightforward methodology for quantifying coordination using AIS data. This metric of coordination is then utilized to construct a graph that visualizes the interconnections between vessels within a defined timeframe. Our aim is to assess the potential of these graphs as tools for detecting group-level anomalies and to lay the groundwork for further research.

To clarify our approach, let's first establish a clearer definition of the terms we've been employing, particularly in relation to vessel coordination. In our framework, a ship or vessel is considered a unique entity that transmits AIS messages. An AIS message includes data such as the MMSI identifier of the ship, latitude, longitude, SOG, and COG values detailing the current movement of the ship. Finally, a track or trajectory then refers to the sequence of AIS messages associated with a single vessel, after the data cleansing process detailed in Chapter 3. These tracks correspond to the route the ship took and all the information associated with its movements.

Coordination between vessels can be quantified through various methods, including correlation metrics, statistical tests for independence, or distance measures. We utilize Pearson's correlation coefficient to assess the level of coordination between two vessels. Nonetheless, this computation is not straightforward due to the data's characteristics. Each vessel's trajectory covers different time periods, which may or may not have intersections with others. To ensure a meaningful comparison, we calculate the correlation only over the overlapping segments of the trajectories, determined by the timestamps of their AIS messages. Furthermore, to focus our analysis on significant interactions, we only consider pairs of vessel tracks whose intersection in time exceeds a predefined threshold, ε. For our test case, we have set ε at 90 minutes, ensuring that we only calculate the Pearson correlation coefficient for vessel pairs that have a substantial period of concurrent activity.

To integrate the various features of each vessel's trajectory into a single, comprehensive measure of coordination, we employ Principal Component Analysis (PCA). PCA is a statistical procedure that utilizes orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component, in turn, has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set.

For our application, PCA serves a dual purpose. First, it reduces the dimensionality of the data, simplifying the four-dimensional feature space (latitude, longitude, SOG, and COG) into a single dimension that captures the most significant variance in the data. This is particularly useful when dealing with multidimensional data, as it distills the information into a more manageable form without sacrificing the data's underlying structure and relationships. Second, by transforming the data into principal components, we mitigate the issue of multicollinearity among the features, which could otherwise skew the correlation analysis.

Algorithm 1. Graph Generation for Vessel Track Coordination

Let *tracks* be the set of all vessel tracks after data cleaning
Let ε be the time intersection threshold (e.g., 90 minutes)
Let δ be the correlation threshold (e.g., 0.5)
Initialize an empty graph G
for each *track* in *tracks* **do**
    Fit PCA to the track to reduce dimensions to 1
    Transform the track using the fitted PCA model
end for
for each pair of tracks $(track_i, track_j)$ in *tracks* **do**
    Find the intersection of timestamps between $track_i$ and $track_j$
    if the intersection duration $> \varepsilon$ **then**
        Calculate the Pearson correlation coefficient ρ of the intersecting segments
        if $\rho > \delta$ **then**
            Add an edge between nodes representing $track_i$ and $track_j$ in G
            Set the weight of the edge to ρ
        end if
    end if
end for
Output the graph G as the model of vessel coordination

In practice, for each vessel track, we fit a PCA model and then use this model to reduce the feature data to a single dimension. After transforming each vessel track from four dimensions to one, we calculate the Pearson correlation coefficient for the intersecting segments of each pair of vessel tracks. We then construct a graph where each vessel is represented as a node. An edge is created between two nodes if the correlation between their first principal components exceeds a predefined threshold, which in our case is set at 0.5. This threshold ensures that only significant correlations contribute to the graph structure, thereby focusing on the most relevant vessel interactions. We formally detail this process in Algorithm 1. The constructed graph thus offers a visualization of the coordination among all vessels for the period under review. The graph for our dataset, representing the two day sample, is depicted in Figure 4.1, illustrating the network of significant correlations that may indicate coordinated behavior among vessels.

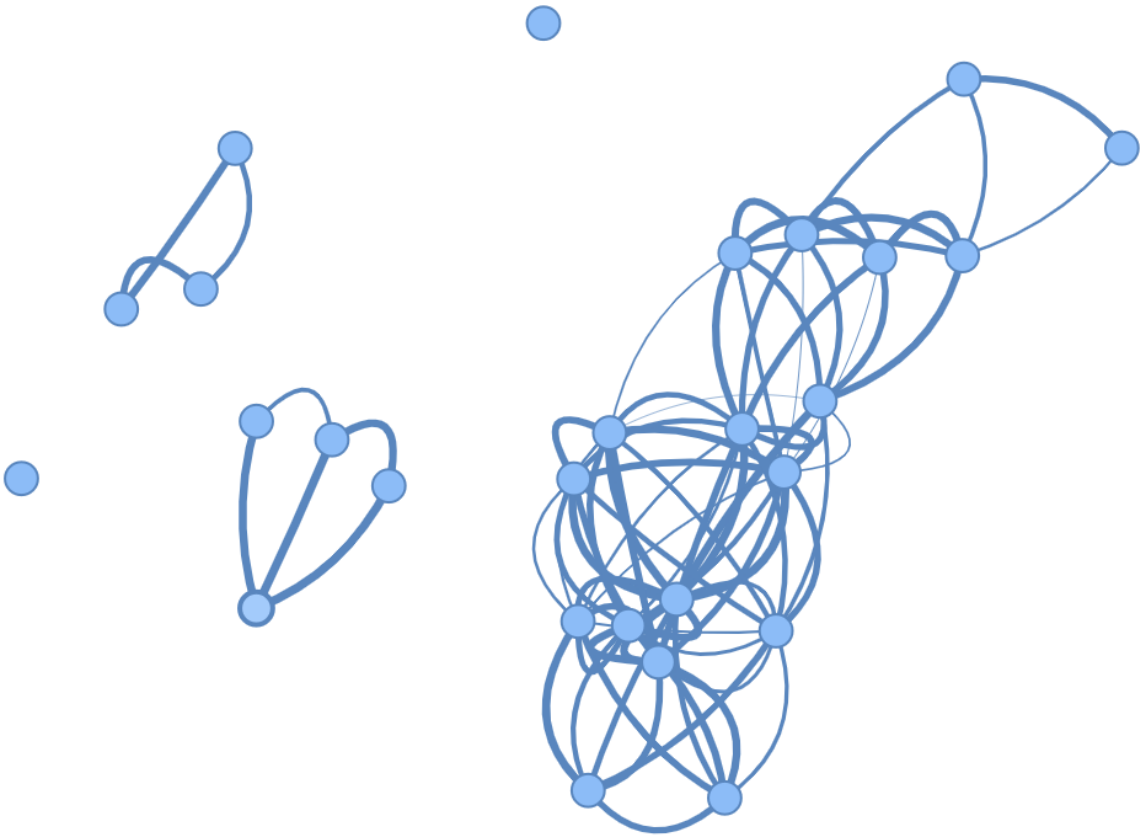

Figure 4.1. Graph of Vessel Trajectories. Each edge represents the coordination between two different vessels, where the thickness corresponds to the strength of coordination.

4.2 Synthetic Coordinated Data Manipulation

In order to test our graphical model, we need to have data that resembles anomalous groups that we are searching for. Unfortunately, labelled datasets of such anomalies do not exist. As discussed, a general barrier to research in marine traffic anomaly detection is the lack of any public dataset that contains labelled anomalies [43]. To circumvent this, we devise a data manipulation strategy to create coordinated groups of vessel trajectories that we believe would mimic the kinds of anomalous groups we want to find. We can use these coordinated groups of vessels as a means to verify that our graphical model is able to identify varying levels of group coordination. This would serve as an initial validation of the graphical model as a useful tool in

group-level anomaly detection.

4.2.1 Coordinated Data Strategy I

Our first strategy for generating anomalous group data is simple in nature, but mimics the occurrence of a close approach anomaly involving two or more vessels. We start by selecting a single cleaned vessel track from our dataset. Given this track, we duplicate it to create two, or more, identical tracks. Each track in this set then shares the identical data for latitude, longitude, SOG, COG, and time stamps for each AIS message. However, we alter the MMSI identifier to make it appear that there is another ship following the exact same path as the original. Once we have our set of duplicate tracks, we time shift the time stamps for each individual track in the set of duplicates. We can more precisely define this process in the following way. Given a track τ_1, which is a 6xn matrix, we duplicate this track d times to create a set $S = \{\tau_i\}$ for i in 1:d. Taking each duplicated track, we adjust the time feature by a small delta, such as 6 minutes, to create the effect of a small group of ships travelling in a very similar direction near the same time. Algorithm 2 clearly defines this process.

Algorithm 2. Generate Coordinated Group Data

Let $originalTrack$ be the set of cleaned vessel tracks
Let $numDuplicates$ be the number of duplicate tracks to make
Let $timeShiftDelta$ be the time offset for the duplicate tracks
$S \leftarrow \{\}$ ▷ Initialize an empty set for tracks
APPEND(S, $originalTrack$) ▷ Add original track to set
for $i \leftarrow 1$ **to** $numDuplicates$ **do**
    $duplicatedTrack \leftarrow$ COPY($originalTrack$) ▷ Duplicate the track
    $duplicatedTrack.MMSI \leftarrow$ GENERATENEWMMSI ▷ Assign new MMSI
    for all $timestamp \in duplicatedTrack$ **do**
        $timestamp \leftarrow timestamp + (i \times timeShiftDelta)$ ▷ Time-shift
    end for
    APPEND(S, $duplicatedTrack$) ▷ Add to set
end for
return S

4.2.2 Coordinated Data Strategy II

Our second strategy for data generation involves simulating coordination between two vessel tracks that are traveling in a similar direction. Consider the two tracks illustrated in red in Figure 4.2. They follow a similar route, but are not necessarily synchronized since they may have been recorded at different times. To create an artificial sense of coordination, we first identify the track with a greater number of AIS data points and downsample it to match the data point count of the less populated track. This results in two tracks with an equal number of AIS messages, traveling in a similar direction.

Next, we synchronize the tracks by replacing the timestamps of one track with those of the other. Consequently, both tracks now appear to be moving in tandem, in the same direction, and at the same time, suggesting a high degree of coordination. To introduce variations in this coordination, we can apply a time offset to one of the tracks, for example, by 3, 6, 9 minutes, etc. This adjustment allows us to generate tracks with different levels of apparent coordination. Our graphical model should then be capable of detecting these varying degrees of coordination, providing a nuanced view of the data.

Algorithm 3. Simulate Artificial Coordination Between Tracks

Let $trackA$ and $trackB$ be the set of two similar tracks
Let $timeShiftDelta$ be the time offset for the given tracks
Let $numDuplicatesTrackA$ and $numDuplicatesTrackB$ be the number of duplicates for each track
if LENGTH($trackA$) $>$ LENGTH($trackB$) **then**
    $trackToDownsample \leftarrow trackA$
    $referenceTrack \leftarrow trackB$
else
    $trackToDownsample \leftarrow trackB$
    $referenceTrack \leftarrow trackA$
end if
$downsampledTrack \leftarrow$ DOWNSAMPLE($trackToDownsample$) ▷ Make track length equal
$downsampledTrack.timestamps \leftarrow referenceTrack.timestamps$ ▷ Align timestamps

---

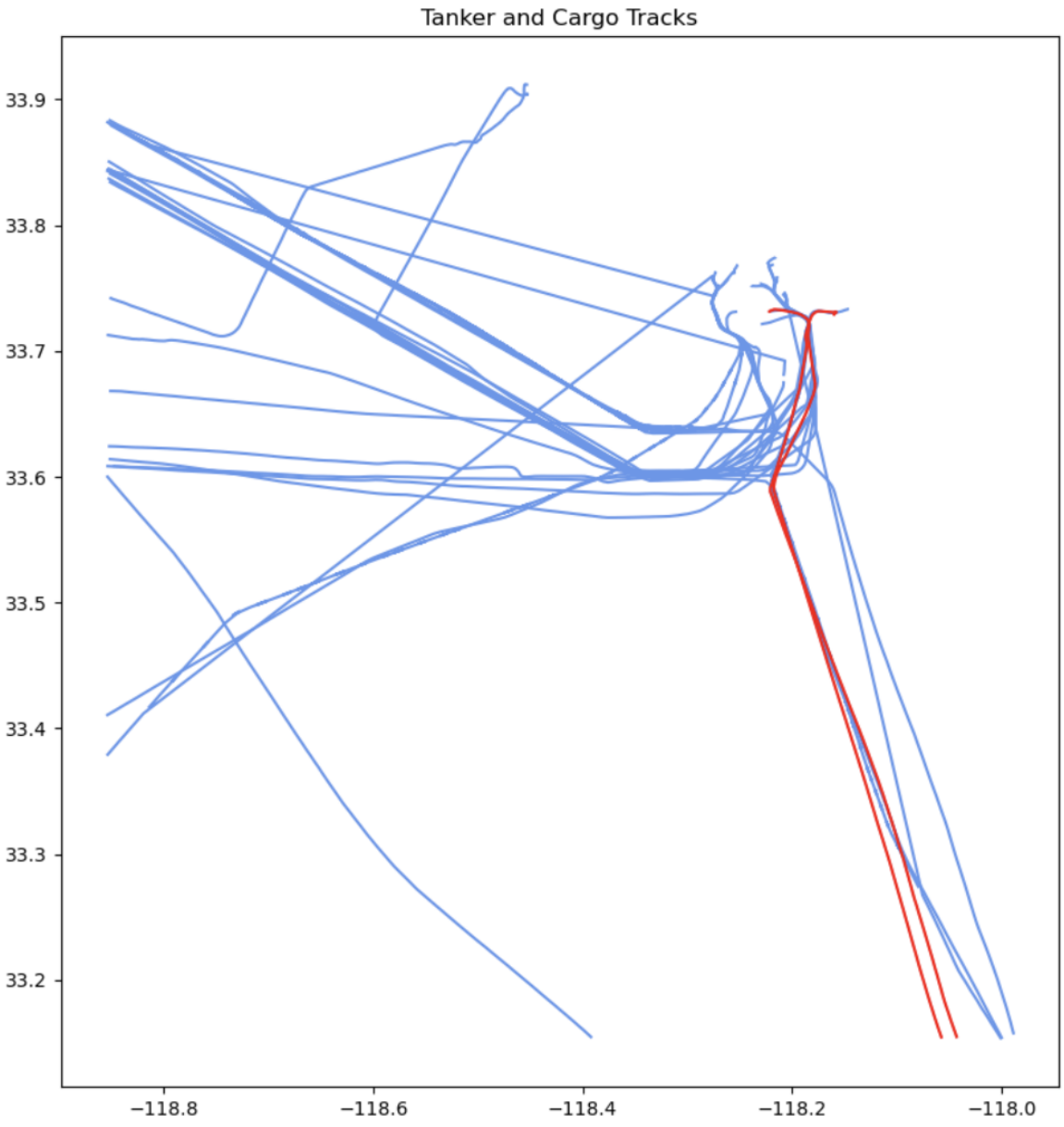

Figure 4.2. Data Generation Two Example Tracks. The two tracks in red share a similar route path. We will change the timestamps so they appear to occur at the same time.

Chapter 5

Results

Given the lack of a good ground truth dataset for our anomaly detection task, we created several ulterior methods for analyzing the effectiveness of our graph visualization strategy in identifying coordinated anomalous groups. First, we examine the constructed graphical model on our original data and examine if the cliques correspond to groups that would likely be coordinated in the real world using a contextual understanding of marine traffic. Second, we use data generation strategy I to generate artificial coordinated groups and examine if our graphical model can identify these groups as coordinated, and the degree to which they are coordinated. Third, we use data generation strategy II to examine slightly more complex groups of tracks to evaluate the effectiveness of the graphical model in a similar fashion. Lastly, to test the generalization of our model, we use data from different geographical ports and verify if the identified cliques correspond to coordinated tracks.

5.0.1 Results on Original Data

Using our original data consisting of 27 vessel tracks over 2 days, we can visualize the coordinated groups as seen in Figure 4.1. Looking at Figure 5.1, we can see five distinct connected groups of tracks. Our proposed graphical model has found that each of these groups share some level of coordination between their respective tracks. We will visually inspect and analyze some of these groups to elicit whether the model is finding coordination as expected.

Beginning with the two disconnected nodes in red and yellow, if our graphical model

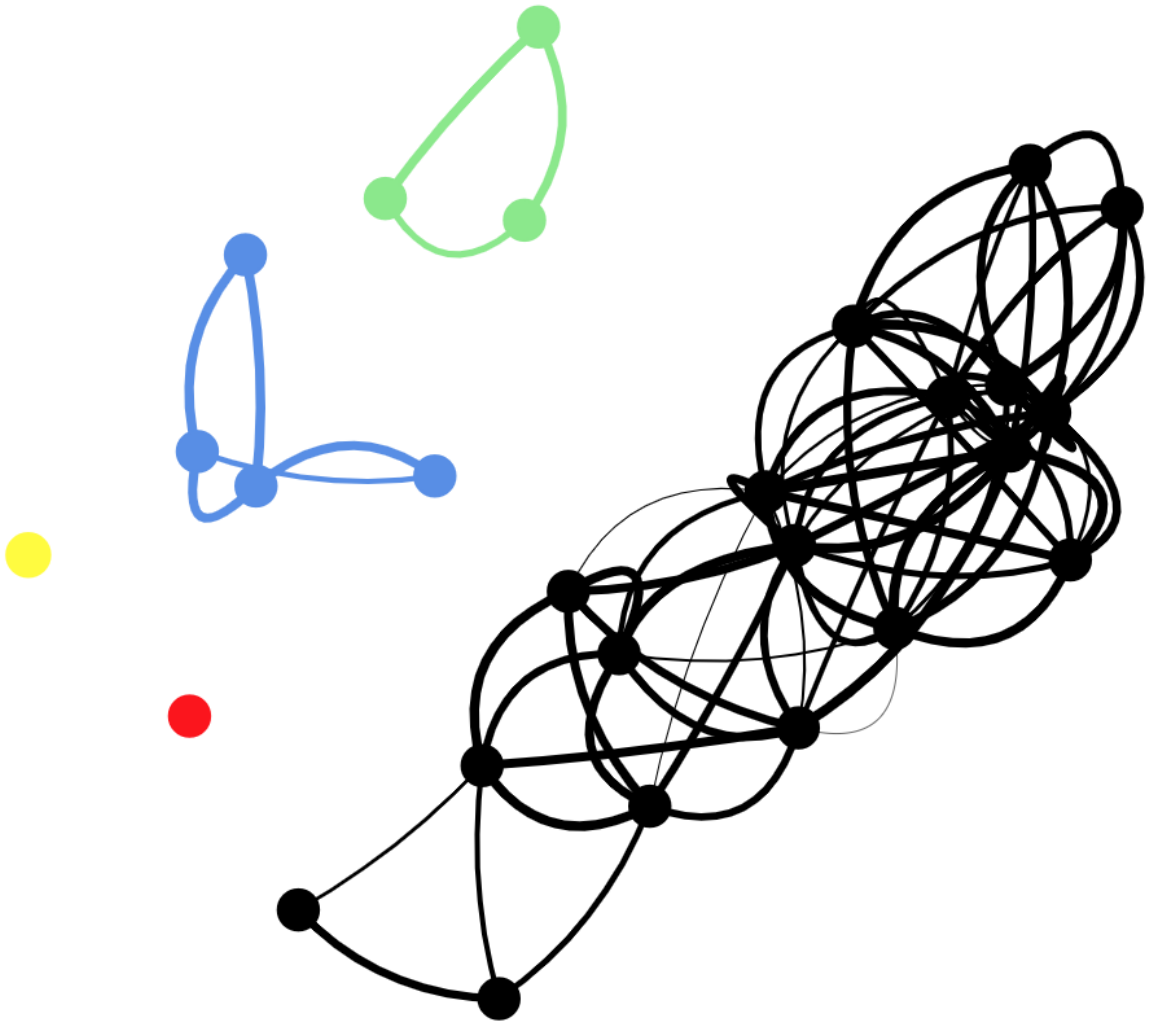

Figure 5.1. Different Coordinated Groups. Each color represents a different groups that has been found to share some level of coordination.

truly finds coordinated groups, then these two tracks should appear clearly not coordinated with the other tracks, either via space, time, speed, or some combination of the track's original data. Looking at the tracks against all other tracks in Figure 5.2, we do not see any obvious movements that look starkly non-coordinated with the other tracks. However, we created six different plots that visualize each track based upon their average timestamps. That is, for each track, we took the timestamps of all AIS data messages and found the average time. Given the

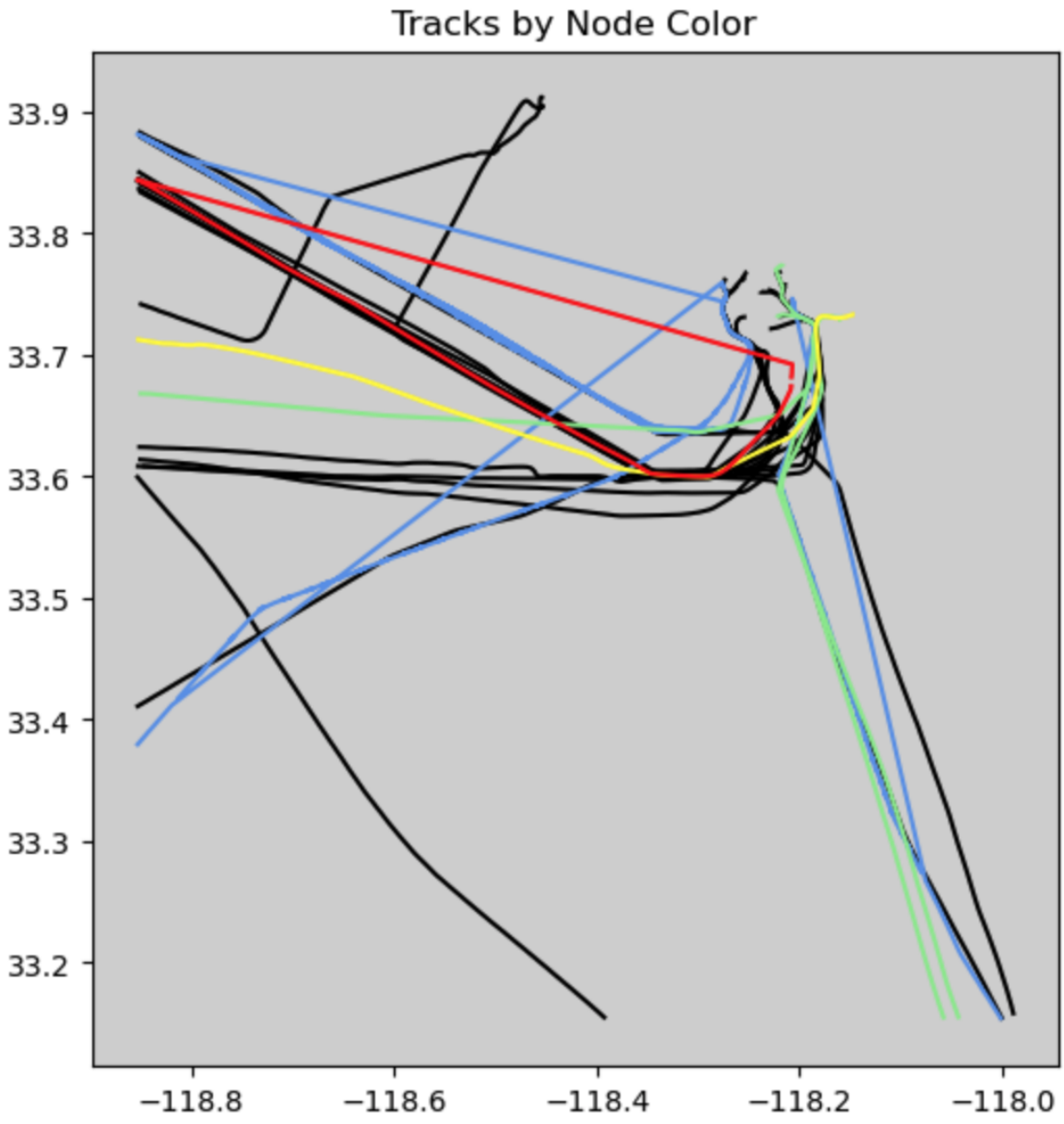

Figure 5.2. Tracks by Node Color. The tracks shown here are colored based on the node color in Figure 5.1.

average timestamp, we drew the tracks in separate plots according to which eight hour interval their average falls within. Using this visualization method, we can see clearly from Figure 5.3 that the two disconnected tracks, shown in their respective colors, red and yellow, appear to have taken place at unusual hours of the day when no other tracks were being recorded, or the other

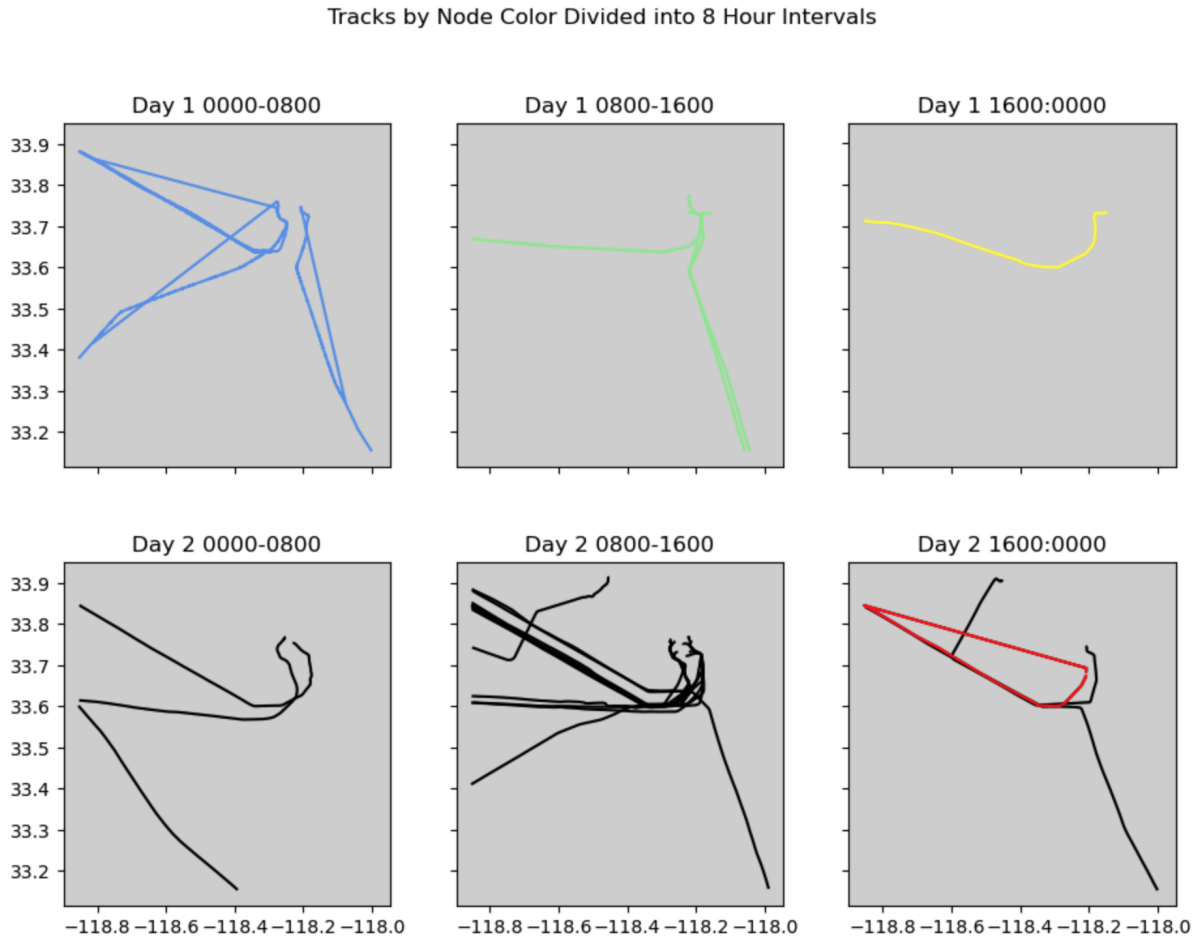

Figure 5.3. Tracks by Node Color Divided into 8 Hour Intervals. The tracks shown here are colored based on the node color in Figure 5.1. We then display each track within the 8 hour window that it's average timestamp value fell within.

tracks show very different routes. This would support the findings of the graphical model that these two tracks are not part of a coordinated group.

Next, we want to examine the groups shown in blue and green shown in Figure 5.1. We should see that each of these two groups shows some level of coordination different from the rest of the tracks. Again, it is easy to see based on Figure 5.3 that for both the blue and the green tracks, there appears some level of coordination as both groups are the only tracks recorded within their respective 8 hour interval. We do see some level of variation among the blue and green tracks in terms of the respective routes they took. However, we can assume that the time

dimension must have been of greater importance when calculating the correlation between each of the track pairs. It is also possible, and will be examined more closely in future work, that the blue and green tracks are travelling in the same direction at roughly the same time. This would indicate another level of coordination that is not easily apparent in our visualizations.

Lastly, and probably the most complex of the groups, is the black nodes in the graph in Figure 5.1. This group has by far the most tracks and, as seen in Figure 5.3, the tracks are not necessarily in the same time interval, nor share identical routes. The best explanation for the coordination of this group is most likely that each track appears to be heading into port, mostly following a very similar path as seen in the Day 2 0800-1600 time interval plot. Also, of note, is that not every track within this group shares the same degree of coordination as measured by the correlation coefficient. Looking at Figure 5.1, we can see the strength of coordination by the width of the edges between the black nodes. It is easy to see that some nodes appear more or less coordinated than others, possibly showing why there would be tracks in different time intervals found in this same group. Lastly, our visualization method, although helpful, is not perfect. It is possible that there is significant overlap of each track in more than one time interval, thus making it possible to have tracks from one coordinated group in more than one time interval.

Regardless of the limitations of the graphs, the visualizations of the different coordinated groups as found by the proposed graphical model in Figure 5.1, appear to work well in separating out tracks that follow somewhat similar routes during similar times of day. As an early experiment showing the efficacy of coordination measurements to find anomalous groups, the initial results appear promising.

5.0.2 Results using Data Generation Strategy I

Next, we examine some of the findings using our data generation strategy II. As a quick reminder, this strategy is focused on creating coordinated groups from a single track. That is, we take a single track, duplicate this track one or more times, and offset the duplicates by some time delta from the original track. This creates a group which appears to be travelling in

identical routes, but at varying times. Using this strategy, we can test several characteristics of our graphical model. First, we can discover whether the model weakens or strengthens edges appropriately as tracks become more or less coordinated. We can verify that groups are found appropriately when there is a tight coupling of duplicated tracks, and that as the time delta increases for a specific track, this track becomes dissociated from the rest of the group.

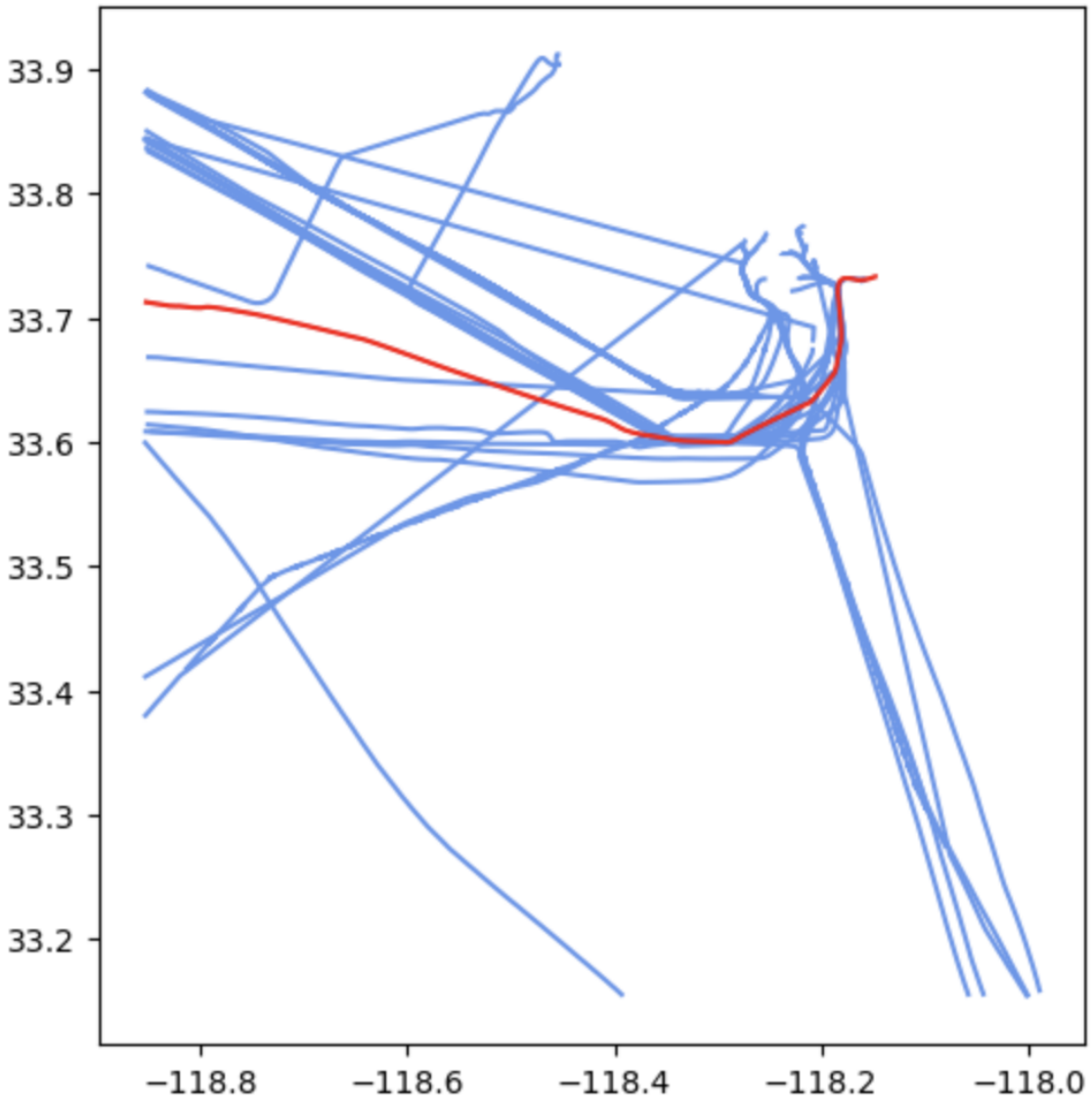

Figure 5.4. Single Disconnected Track Highlighted. All tracks here are shown, with the yellow node track from Figure 5.1, shown in red.

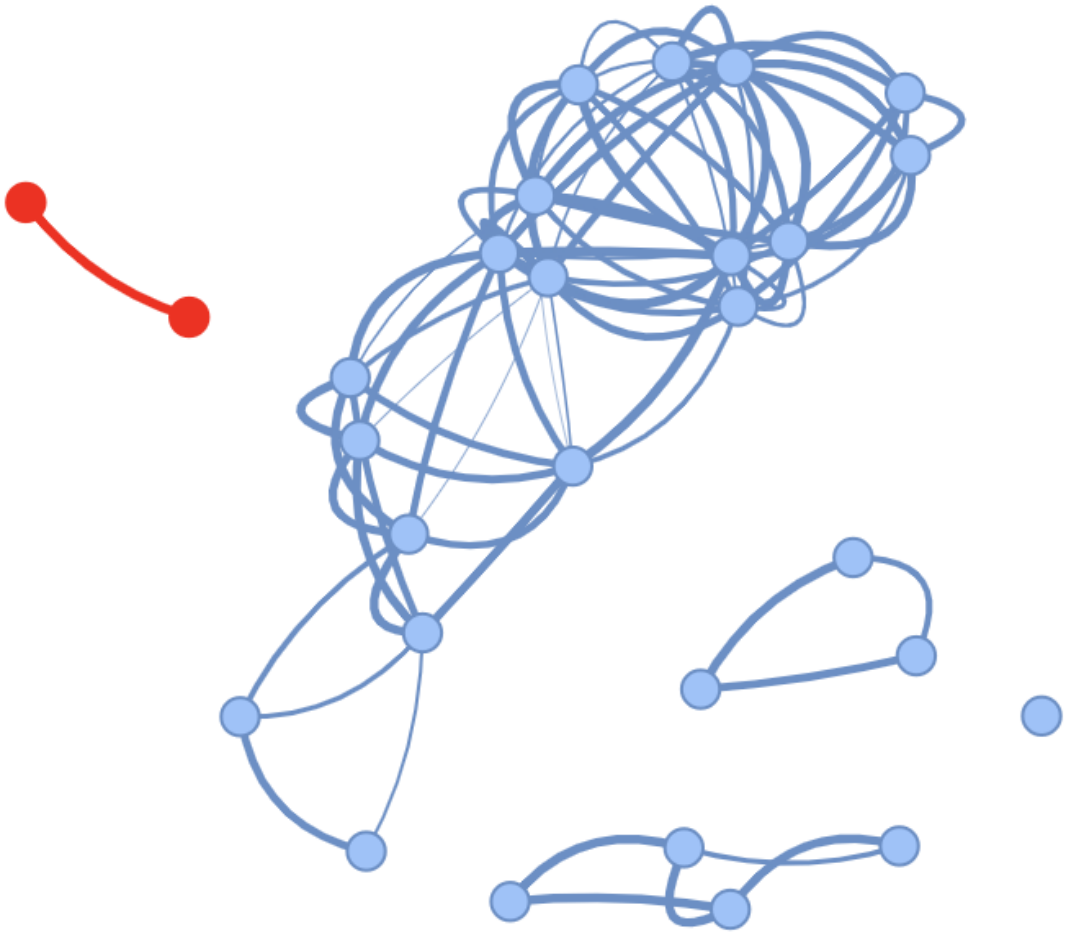

Figure 5.5. Graph with Duplicate Track Added. The graphical model is shown with the original track and it's duplicate shown in red.

First, we examine a single track, as highlighted in Figure 5.4. Recall, from Figure 5.1, that this track is originally disconnected from all other tracks in our graphical model. We start by creating a single duplicate of this track and examining if there exists an edge between the original track and the duplicated track at different time offsets. We confirm in Figure 5.5, that when a duplicate track is created, with a time offset of three minutes, that the duplicate track shares an edge with the original track. We continue with this line of examination, noting that as the time delta increases, the strength of the edge between the two nodes weakens. Finally, after a

time offset of sixty minutes, the edge is removed completely, indicating there is no coordinated group between the original and duplicate tracks. This follows what we would expect for the graphical model. As two very similar tracks are increasingly separated in time, the coordination between the two of them should decrease.

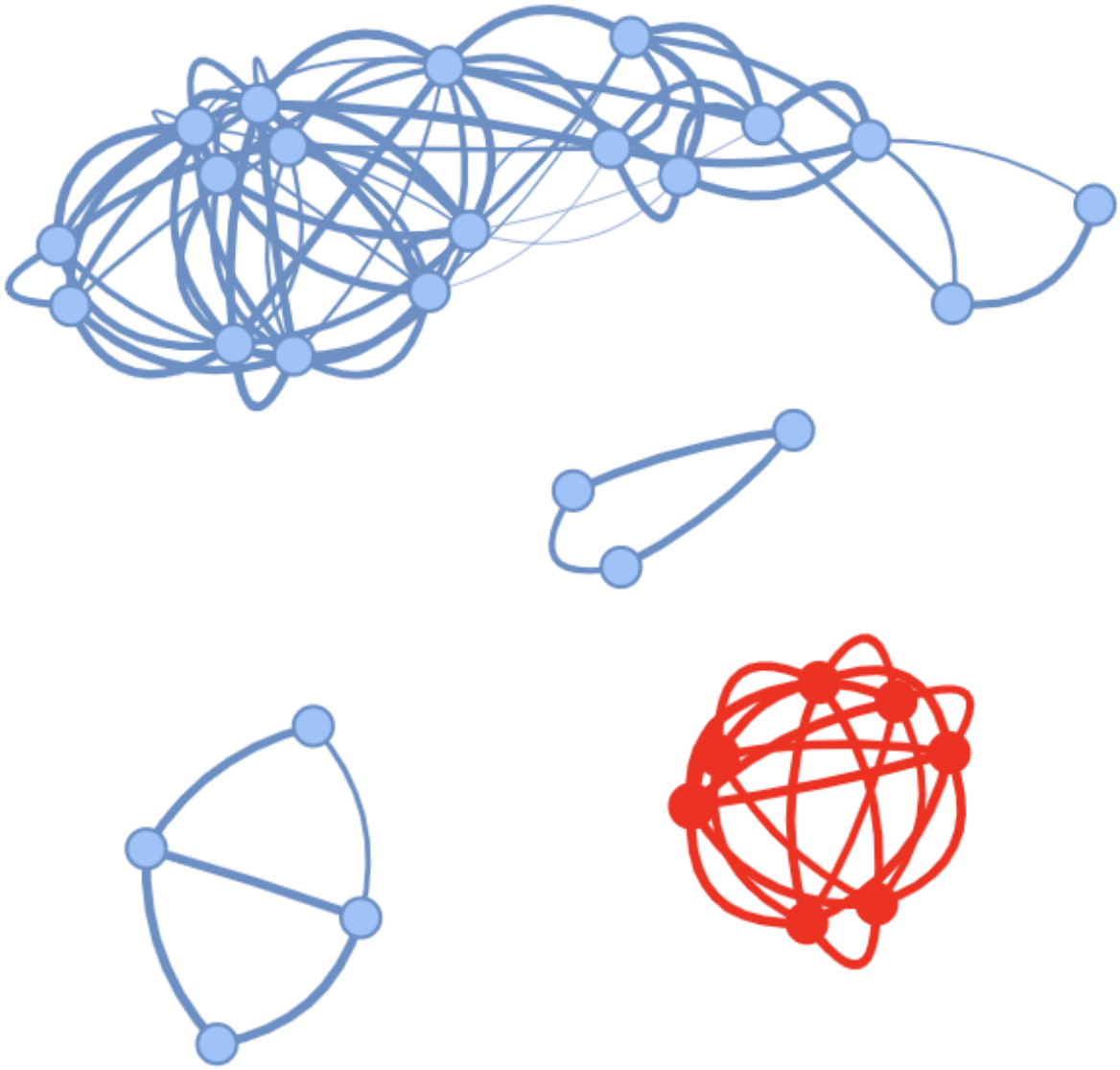

Figure 5.6. Graph with Duplicate Group of Tracks Added. The graphical model is shown with the original track and all of it's duplicate shown in red.

Next, we take a look at how the proposed graphical model handles *groups* of coordinated

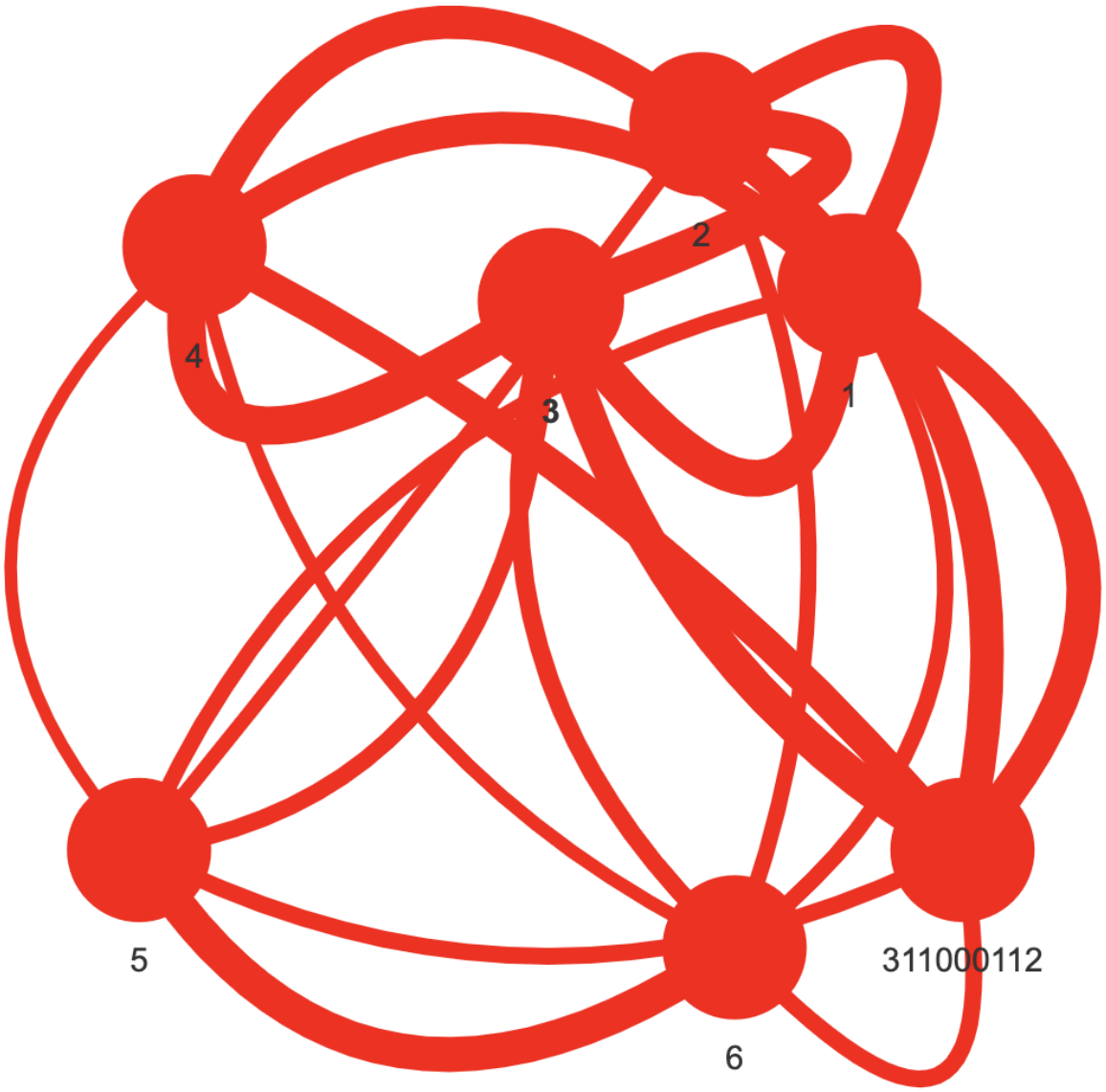

Figure 5.7. Graph Showing Duplicated and Original Tracks. Here, the original track is shown with the node 311000112. The tracks with small time offsets are shown in node 1, 2, 3, and 4. Lastly, tracks with large time offsets are shown in node 5 and 6.

tracks. We take the same original track as previously used, and we create five duplicates, with increasing time offsets from 3 minutes to 18 minutes. The results of the graphical model match our intuition about how this group should appear. We see in Figure 5.6, that all of the duplicated tracks form a sub graph with the original track. This is expected, but what would happen if we took 2 of the duplicated tracks and offset the time index by a very large number, such as 120

and 123 minutes? We should see that the edges disappear or, at the very least, are weakened. Indeed, examining Figure 5.7, we see that for the duplicated tracks that were offset by 120 and 123 minutes, node 5 and 6 respectively, there is a strong edge between the two, but weaker edges shared with the rest of the nodes. We would have expected to see the complete removal of edges, as tracks that are over two hours apart should not be considered coordinated. However, it took increasing the large offsets to well over 3 hours to fully remove the edges. This highlights a possible weakness with our graphical model. Likely, the original track takes place over a long period of time, such as several hours. If the track is similar throughout, then offsetting the original track by a number smaller than its full time length may create artificial coordination where no such coordination exists.

5.0.3 Results using Data Generation Strategy II

Further, we evaluate the graphical model according to data generation strategy II. This strategy differs slightly from strategy I, as it uses two completely different tracks, that appear to share a similar route, and artificially creates coordination between the two. This is done by essentially replacing the timestamps for one track's AIS messages with the timestamps of the other, to make it appear that the two ships were travelling very close together in time. Still need to complete this section.

As a case study, we start by examining the two tracks shown in red in Figure 5.8. Following strategy II, we create an artificially coordinated group between the two tracks in which they share identical timestamps for their AIS messages. The graphical model appears as we would expect in Figure 5.9, where there is a clear and strong edge between the two. However, what would happen if we increase the time offset between the two tracks? Similar to our earlier examination, we should see a weakening of the edge indicating the strength of coordination. Eventually, the edge should be removed. This is indeed what we observe. As the time offset is increased in increments of 15 minutes between the two tracks, there is a steady degradation of the edge between the two nodes. Eventually, after 1 hour, the two nodes no longer share an edge.

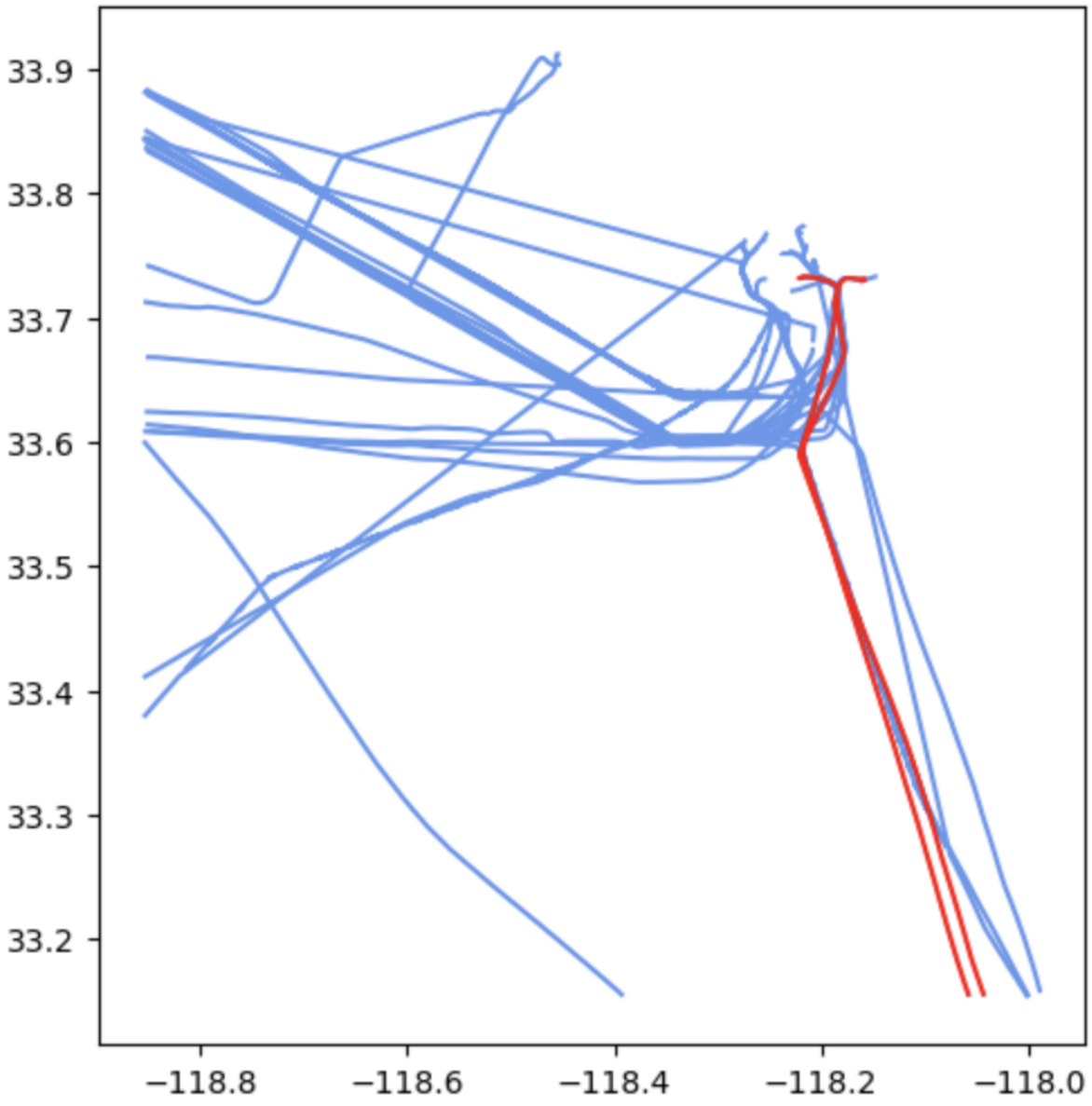

Figure 5.8. Two Similar Tracks. Two tracks, with very similar routes, are shown in red.

Again, this validates our graphical model, whereby coordinated groups should be found when they appear to be following a very similar path close together

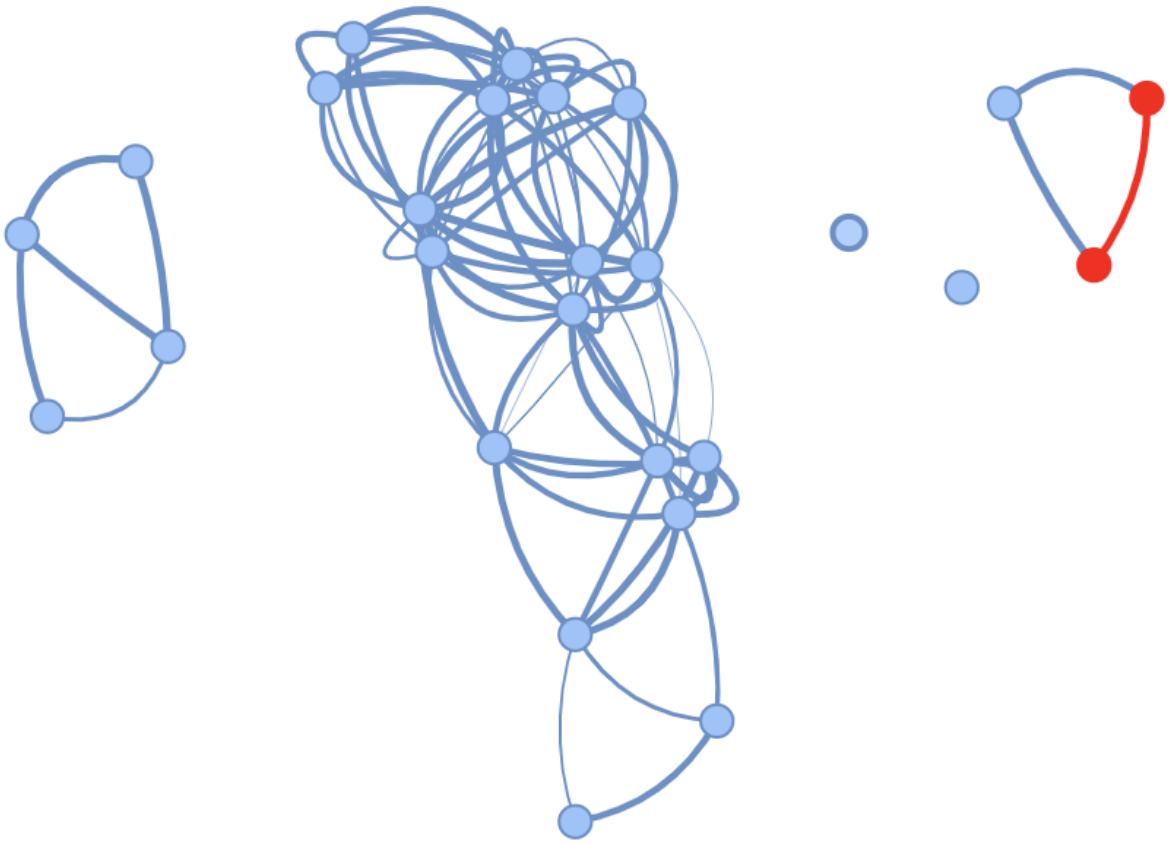

Figure 5.9. Graphical Model of Two Tracks with Same Time Index. The two tracks picture in 5.8 are shown in the graphical model after applying data generation strategy II.

5.0.4 Generalization Results

In our study, we sought to validate the generalizability of our graphical model beyond the initial dataset. To this end, we focused on a different geographical area, specifically around the Port of Oakland, and collected AIS data from the same dates, January 1st and 2nd, 2020. After applying our established data cleaning methods, we obtained 15 vessel tracks comprising 1344 AIS messages, focusing again on cargo and tanker tracks.

Upon applying our graphical modeling technique to this new dataset, we observed the formation of two primary cliques, depicted in blue and red, in the coordination model (Figure

5.10). Notably, several nodes remained isolated from these cliques. To further analyze these patterns, we visualized two dimensions of the data—latitude and longitude—over time, as shown in Figure 5.11. This visualization revealed distinct patterns corresponding to the separate nodes identified in the graphical model. For example, the green and purple tracks exhibited noticeable differences in their longitudinal approaches.

Interestingly, the blue clique comprised tracks spanning multiple time windows, suggesting that our model does not solely rely on temporal proximity for determining coordination. This observation implies that other features, beyond just latitude, longitude, and time, significantly influenced the coordination among these tracks. The presence of similarities among the blue tracks, despite their temporal dispersion, underscores the model's ability to discern coordination based on a broader set of features.

In summary, the application of our graphical model to the Port of Oakland data reinforces its potential for generalization across different geographical contexts. The model's ability to identify distinct cliques and isolated nodes in both datasets—Port of Long Beach and Port of Oakland—demonstrates its robustness in capturing complex patterns of vessel coordination. These findings not only validate the model's effectiveness in diverse settings but also highlight its versatility in analyzing maritime traffic behavior using various features beyond mere spatial and temporal data. The success in generalizing our approach paves the way for future research to apply this model in even broader contexts, potentially offering valuable insights into maritime traffic patterns and anomaly detection on a global scale.

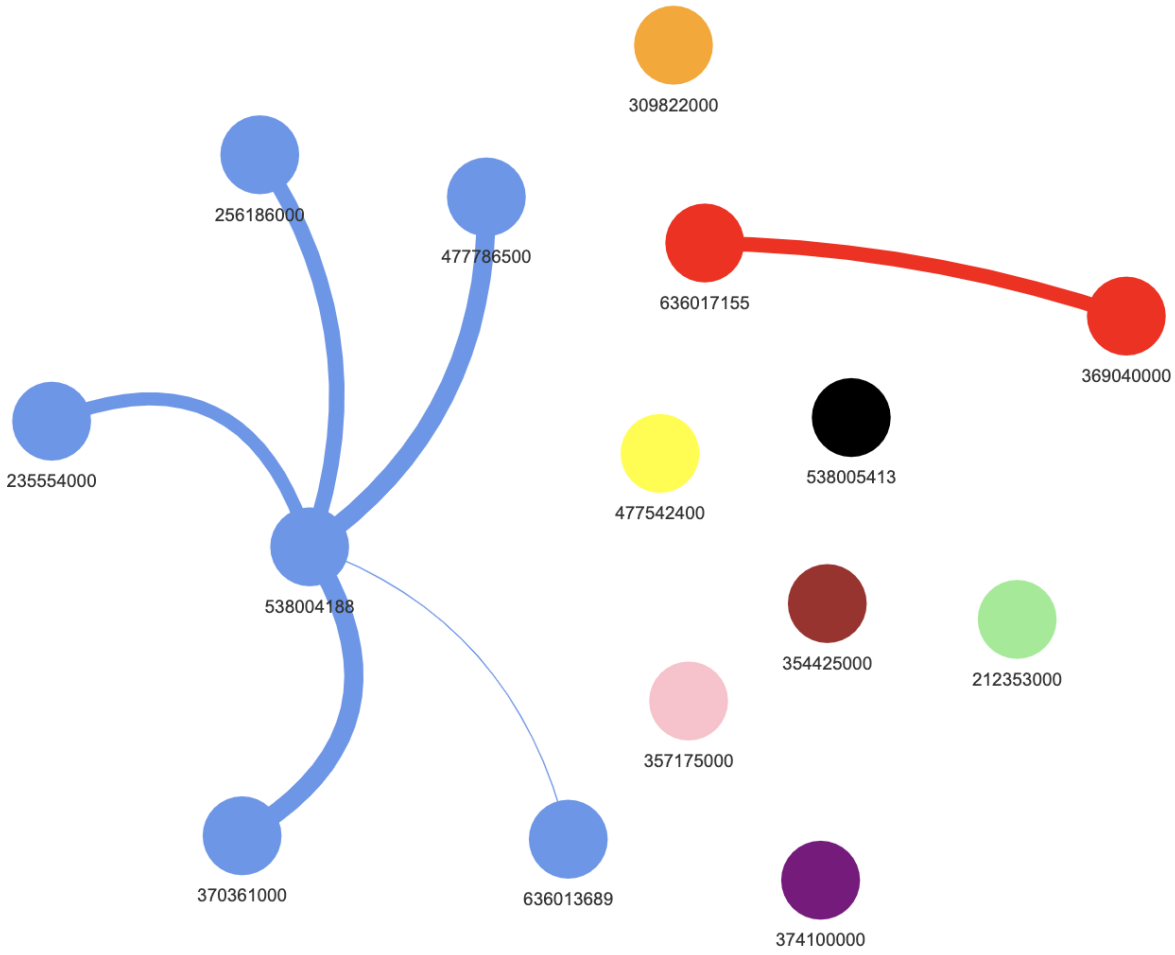

Figure 5.10. Graphical Model of Tracks Around the Port of Oakland. Each color represent a different coordinated group.

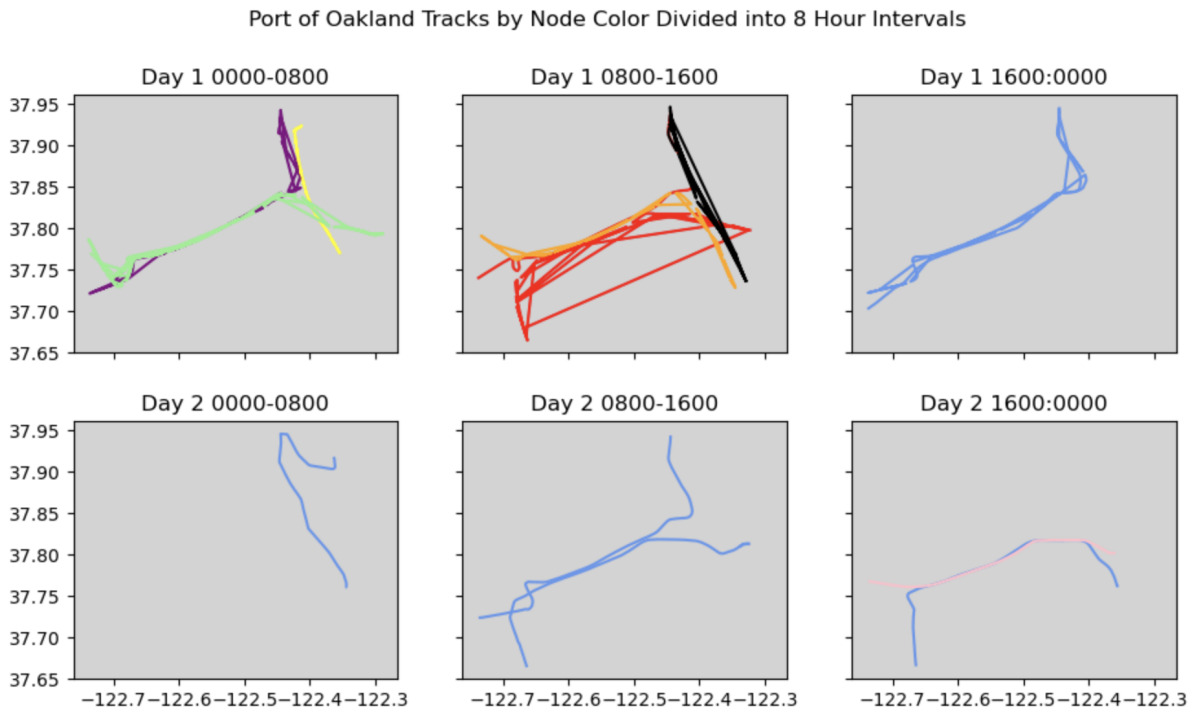

Figure 5.11. Port of Oakland Tracks by Node Color Divided into 8 Hour Intervals. The tracks shown here are colored based on the node color in Figure 5.10. We then display each track within the 8 hour window that it's average timestamp value fell within.

Chapter 6

Discussion

In this study, we introduced a novel approach to group anomaly detection, emphasizing the importance of entity coordination in the detection process. Our methodology involved the development of a graphical model tailored to maritime data, where vessels are represented as nodes and their pairwise coordination forms the edges. We devised two strategies to generate coordinated groups of trajectories, which served as a benchmark for our model. Upon analyzing both the original graph and the one constructed from our synthetically coordinated groups, we observed promising indications that our graphical model can effectively identify coordinated groups. The primary achievement of this research lies in the conceptualization and development of this graphical model, paving the way for further advancements in this area. However, this paper also touches upon a broad spectrum of topics, each presenting opportunities for more in-depth exploration and refinement in future research endeavors.

The immediate progression for utilizing our graphical model involves integrating a learning component. This would enable us to use the graph's insights to train a model that understands typical patterns of group coordination. By learning what constitutes normal group behavior, we can then develop a model capable of assessing the abnormality of a group. A critical challenge in this process, as is common in AIS data research, is the establishment of reliable ground truth data for group anomalies. Collaborating with domain experts or devising sophisticated group anomaly generation strategies will be essential to create this ground truth.

The success of these efforts is pivotal, as it directly impacts our ability to accurately gauge the effectiveness of our proposed approach.

A critical area for future research is the improvement of AIS data cleaning techniques. As previously discussed, AIS data is fraught with challenges, including data gaps, potential manipulation, and incomplete coverage of maritime traffic. To address these issues, future studies may need to adopt multi-modal data collection methods. Incorporating satellite imagery and radar reconnaissance, where feasible, could significantly enrich the data, particularly in specialized research contexts like government or defense labs. Additionally, developing more sophisticated methods for transforming raw AIS messages into coherent vessel tracks is essential. In our current work, time constraints limited our ability to fully utilize the available data, leading to the exclusion of a substantial portion that could not be reliably processed into accurate tracks. While there are existing methodologies, such as those mentioned in [33, 25], the specifics of these approaches are not extensively detailed. Therefore, a more thorough exploration and development of robust data cleaning methods remain a priority for enhancing the quality and usability of AIS data in future research.

The initial definition of our graphical model construction methodology opens up numerous avenues for further research and refinement. The core concept is to leverage the coordination of trajectories as a key input in developing a model that defines normal maritime traffic behavior. Our current graph construction approach represents an initial step in this direction. However, there are several aspects that could be optimized to more effectively capture coordination information from the data. One area for exploration is the use of alternative metrics to the Pearson correlation coefficient for measuring coordination. Different metrics might provide new insights or more accurately reflect the nuances of vessel interactions. Additionally, reevaluating the methodology for edge creation in the graph could be beneficial. By exploring methods that retain more of the original data without resorting to dimensionality reduction through PCA, we might capture a richer set of information. Furthermore, considering alternatives to PCA for data transformation is another promising direction. Techniques like DBSCAN, commonly used in

related research, could offer a different perspective on trajectory analysis. Each of these potential modifications represents an opportunity to refine and enhance our approach, contributing to a more robust and insightful anomaly detection model in maritime data analysis.

Future research could greatly benefit from the exploration of advanced feature engineering techniques. Our initial model primarily utilizes COG, SOG, latitude, and longitude to construct the graph, but incorporating additional features could enhance our ability to differentiate between normal and abnormal vessel tracks. One such feature could be the quantification of stops within a track. For example, analyzing the frequency and duration of a vessel's pauses, particularly when these coincide with other vessels, might reveal patterns indicative of abnormal behavior. Another area for potential development is the categorization of individual track segments. Different phases of a vessel's journey, such as departing, en route, and arriving at ports, exhibit distinct characteristics compared to segments where the vessel is anchored or stationary. By classifying these segments and incorporating them into our model, we could achieve a more granular understanding of vessel behavior. This approach would be particularly insightful for identifying coordinated activities during specific phases of a vessel's journey, thereby enhancing the sophistication and accuracy of our coordination metric. We could also fairly easily incorporate more dynamical information such as the rate of change for the SOG values. It may add additional information to the coordination metric by incorporating categorical features such as whether the tracks is slowing down, speeding up, or at constant velocity. We can imagine in cases such as a naval formation, different vessels are going to either speed up or slow down all at a similar time and in a similar place so that they can for a tight group.

A significant enhancement for future research would be the development of more sophisticated visualization tools for maritime tracks. This task, primarily rooted in software development, would greatly benefit the analysis process. An interactive visualization tool would not only facilitate a more intuitive evaluation of whether tracks are coordinated or exhibit abnormal behavior, but it would also offer several other advantages. For example, imagine a tool that allows users to select a specific group of vessels and visually track their movements in real time, considering

latitude, longitude, and SOG. Such a tool could provide a clearer understanding of the dynamics between these vessels, making it easier to discern patterns of coordination or anomalous behavior. Additionally, interactive visualizations could enable researchers to manipulate time scales, toggle between different data features, and overlay additional contextual information, such as weather patterns or maritime traffic density. This level of interactivity and detail would not only aid in identifying coordinated activities among ships but also in understanding the broader context of their movements. Furthermore, it could facilitate the exploration of hypothetical scenarios, allowing researchers to simulate potential changes in vessel behavior and observe the resultant effects on the overall traffic pattern. The development of this kind of tool would be a valuable asset in maritime research, enhancing our ability to analyze complex data sets and extract meaningful insights.

In conclusion, this paper presents a novel approach to group anomaly detection, emphasizing the significance of entity coordination in maritime data analysis. By developing a unique graphical model that captures the intricate interactions between vessels, we have laid the groundwork for a new paradigm in anomaly detection. This methodology not only enhances our understanding of maritime traffic patterns but also holds promise for broader applications. Its potential extends to various domains where understanding complex group dynamics is crucial, such as environmental monitoring, defense strategy, and transportation logistics. As we continue to refine and adapt this model, it stands to offer valuable insights and tools for researchers and practitioners across a multitude of fields, paving the way for more sophisticated and comprehensive analyses of group behaviors and interactions.

Bibliography

[1] Marinecadastre.gov. https://marinecadastre.gov/.

[2] *Solas Chapter V Safety of Navigation*. IMO London, UK, 2002.

[3] Andrej Androjna, Marko Perkovič, Ivica Pavic, and Jakša Mišković. Ais data vulnerability indicated by a spoofing case-study. *Applied Sciences*, 11(11):5015, 2021.

[4] Mathias Anneken, Yvonne Fischer, and Jürgen Beyerer. Evaluation and comparison of anomaly detection algorithms in annotated datasets from the maritime domain. In *2015 SAI Intelligent Systems Conference (IntelliSys)*, pages 169–178. IEEE, 2015.

[5] Miguel A Cervera and Alberto Ginesi. On the performance analysis of a satellite-based ais system. In *2008 10th International Workshop on Signal Processing for Space Communications*, pages 1–8. IEEE, 2008.

[6] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. *Advances in neural information processing systems*, 28, 2015.

[7] National Research Council. Vessel navigation and traffic services for safe and efficient ports and waterways: Interim report. 1996.

[8] Kimbra Cutlip. Ais for safety and tracking: A brief history. *Global Fishing Watch*, 31, 2017.

[9] Enrica d'Afflisio, Paolo Braca, Leonardo M Millefiori, and Peter Willett. Maritime anomaly detection based on mean-reverting stochastic processes applied to a real-world scenario. In *2018 21st International Conference on Information Fusion (FUSION)*, pages 1171–1177. IEEE, 2018.

[10] DEPARTMENT OF HOMELAND SECURITY WASHINGTON DC. Small vessel security strategy. 2008.

[11] Gerben Klaas Dirk De Vries and Maarten Van Someren. Machine learning for vessel trajectories using compression, alignments and domain knowledge. *Expert Systems with Applications*, 39(18):13426–13439, 2012.

[12] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[13] INTERNATIONAL TRANSPORT FORUM. *ITF transport outlook 2021*. OECD, 2021.

[14] Mélanie Fournier, R Casey Hilliard, Sara Rezaee, and Ronald Pelot. Past, present, and future of the satellite-based automatic identification system: Areas of applications (2004–2016). *WMU journal of maritime affairs*, 17:311–345, 2018.

[15] Peiguo Fu, Haozhou Wang, Kuien Liu, Xiaohui Hu, and Hui Zhang. Finding abnormal vessel trajectories using feature learning. *IEEE Access*, 5:7898–7909, 2017.

[16] Alexander Gammerman and Vladimir Vovk. Hedging predictions in machine learning. *The Computer Journal*, 50(2):151–163, 2007.

[17] Mojtaba Goodarzi and Mahdi Shaabani. Maritime traffic anomaly detection from spatio-temporal ais data. In *The Second of International Conference on Management and Fuzzy Systems (ICMFS Series)*, 2019.

[18] Athanassios Goudossis and Sokratis K Katsikas. Towards a secure automatic identification system (ais). *Journal of Marine Science and Technology*, 24:410–423, 2019.

[19] Dini Oktarina Dwi Handayani, Wahju Sediono, and Asadullah Shah. Anomaly detection in vessel tracking using support vector machines (svms). In *2013 International Conference on Advanced Computer Science Applications and Technologies*, pages 213–217. IEEE, 2013.

[20] IMO IMO. Resolution msc. 74 (69): adoption of new and amended performance standards. *MSC 69/22/Add. 1*, 74(May):20, 1998.

[21] Clément Iphar, Cyril Ray, and Aldo Napoli. Uses and misuses of the automatic identification system. In *OCeAnS 2019-Marseille*, pages 1–10. IEEE, 2019.

[22] GC Kessler. Protected ais: A demonstration of capability scheme to provide authentication and message integrity. *TransNav: International Journal on Marine Navigation and Safety of Sea Transportation*, 14(2):279–286, 2020.

[23] Kira Kowalska and Leto Peel. Maritime anomaly detection using gaussian process active learning. In *2012 15th International Conference on Information Fusion*, pages 1164–1171. IEEE, 2012.

[24] Rikard Laxhammar. Anomaly detection for sea surveillance. In *2008 11th international conference on information fusion*, pages 1–8. IEEE, 2008.

[25] Rikard Laxhammar and Göran Falkman. Conformal prediction for distribution-independent anomaly detection in streaming vessel data. In *Proceedings of the first international workshop on novel data stream pattern mining techniques*, pages 47–55, 2010.

[26] Nicolas Le Guillarme and Xavier Lerouvreur. Unsupervised extraction of knowledge from s-ais data for maritime situational awareness. In *Proceedings of the 16th International Conference on Information Fusion*, pages 2025–2032. IEEE, 2013.

[27] Po-Ruey Lei. A framework for anomaly detection in maritime trajectory behavior. *Knowledge and Information Systems*, 47(1):189–214, 2016.

[28] Steven Mascaro, Ann E Nicholso, and Kevin B Korb. Anomaly detection in vessel tracks using bayesian networks. *International Journal of Approximate Reasoning*, 55(1):84–98, 2014.

[29] Fabio Mazzarella, Michele Vespe, Alfredo Alessandrini, Dario Tarchi, Giuseppe Aulicino, and Antonio Vollero. A novel anomaly detection approach to identify intentional ais on-off switching. *Expert Systems with Applications*, 78:110–123, 2017.

[30] Ashley McAbee, James Scrofani, Murali Tummala, David Garren, and John McEachen. Traffic pattern detection using the hough transformation for anomaly detection to improve maritime domain awareness. In *17th International Conference on Information Fusion (FUSION)*, pages 1–6. IEEE, 2014.

[31] Kristian Metcalfe, Nathalie Bréheret, Eva Chauvet, Tim Collins, Bryan K Curran, Richard J Parnell, Rachel A Turner, Matthew J Witt, and Brendan J Godley. Using satellite ais to improve our understanding of shipping and fill gaps in ocean observation data to support marine spatial planning. *Journal of Applied Ecology*, 55(4):1834–1845, 2018.

[32] Duong Nguyen, Rodolphe Vadaine, Guillaume Hajduch, René Garello, and Ronan Fablet. A multi-task deep learning architecture for maritime surveillance using ais data streams. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 331–340. IEEE, 2018.

[33] Duong Nguyen, Rodolphe Vadaine, Guillaume Hajduch, René Garello, and Ronan Fablet. Geotracknet—a maritime anomaly detector using probabilistic neural network representation of ais tracks and a contrario detection. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):5655–5667, 2021.

[34] International Maritime Organization. Revised guidelines for the onboard operational use of shipborne automatic identification systems (ais), 2015.

[35] Bradley J Rhodes, Denis Garagic, James R Dankert, Lauren H Stolzar, Majid Zandipour, Michael Seibert, and Neil A Bomberger. *Anomaly Detection & Behavior Prediction: Higher-Level Fusion Based on Computational Neuroscientific Principles*. INTECH Open Access Publisher, 2009.

[36] Branko Ristic, Barbara La Scala, Mark Morelande, and Neil Gordon. Statistical analysis of motion patterns in ais data: Anomaly detection and motion prediction. In *2008 11th International Conference on Information Fusion*, pages 1–7. IEEE, 2008.

[37] H Rong, AP Teixeira, and C Guedes Soares. Data mining approach to shipping route characterization and anomaly detection based on ais data. *Ocean Engineering*, 198:106936, 2020.

[38] M Series. Technical characteristics for an automatic identification system using time-division multiple access in the vhf maritime mobile band. *Recommendation ITU: Geneva, Switzerland*, pages 1371–1375, 2014.

[39] Norma Serra-Sogas, Patrick D O'Hara, Kim Pearce, Leh Smallshaw, and Rosaline Canessa. Using aerial surveys to fill gaps in ais vessel traffic data to inform threat assessments, vessel management and planning. *Marine Policy*, 133:104765, 2021.

[40] Sandeep Kumar Singh and Frank Heymann. Machine learning-assisted anomaly detection in maritime navigation using ais data. In *2020 IEEE/ION Position, Location and Navigation Symposium (PLANS)*, pages 832–838. IEEE, 2020.

[41] James Smith, Ilia Nouretdinov, Rachel Craddock, Charles Offer, and Alexander Gammerman. Anomaly detection of trajectories with kernel density estimation by conformal prediction. In *Artificial Intelligence Applications and Innovations: AIAI 2014 Workshops: CoPA, MHDW, IIVC, and MT4BD, Rhodes, Greece, September 19-21, 2014. Proceedings 10*, pages 271–280. Springer, 2014.

[42] Behrouz Haji Soleimani, Erico N De Souza, Casey Hilliard, and Stan Matwin. Anomaly detection in maritime data based on geometrical analysis of trajectories. In *2015 18th International Conference on Information Fusion (Fusion)*, pages 1100–1105. IEEE, 2015.

[43] Konrad Wolsing, Linus Roepert, Jan Bauer, and Klaus Wehrle. Anomaly detection in maritime ais tracks: A review of recent approaches. *Journal of Marine Science and Engineering*, 10(1):112, 2022.

[44] Ying Wu, Anthony Patterson, Rafael DC Santos, and Nandamudi L Vijaykumar. Topology preserving mapping for maritime anomaly detection. In *Computational Science and Its Applications–ICCSA 2014: 14th International Conference, Guimarães, Portugal, June 30–July 3, 2014, Proceedings, Part VI 14*, pages 313–326. Springer, 2014.

[45] Dimitris Zissis, Konstantinos Chatzikokolakis, Giannis Spiliopoulos, and Marios Vodas. A distributed spatial method for modeling maritime routes. *IEEE Access*, 8:47556–47568, 2020.