IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 30, 2018, accepted October 3, 2018, date of publication October 23, 2018,
date of current version November 19, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2877659

Analyzing Congestion Interdependencies of Ports and Container Ship Routes in the Maritime Network Infrastructure

GEORGE STERGIOPOULOS[1], EVANGELOS VALVIS[1], DIMITRIS MITRODIMAS[2], DIMITRIOS LEKKAS[2], AND DIMITRIS GRITZALIS[1]

[1]Department of Informatics, Athens University of Economics and Business, GR-10434 Athens, Greece
[2]MarineTraffic Operations SA, 11525 Athens, Greece

Corresponding author: George Stergiopoulos (geostergiop@aueb.gr)

ABSTRACT Events, such as prolonged congestion in ports or unavailable ship routes in the maritime network, often initiate cascading congestions that block transportation and/or disrupt services over wide areas. Existing traffic flow analysis methods lack the ability to understand the cascading effects of delays in ship routes or how to reduce overall delays in greater maritime areas. Dependency risk graphs have been proposed as a tool for analyzing such cascading events using dependency chains. This paper proposes a risk-based interdependency analysis method capable to detect large-scale traffic congestions between interconnected ports and ship routes in the maritime network and provide solutions to improve flow. Presented dependency risk chains of ports along with graph theory help us analyze ship routes and detect ports that are affected most when other major ports are congested in the maritime network, detect the causes of bottlenecks, and provide valuable info in relieving delays across container ship routes. We apply the proposed method on historical container ship routing data provided by the MarineTraffic company that maintains a comprehensive maritime database worldwide for more than six million users monthly. This application-oriented, interdisciplinary effort culminated in a prototype tool is able to analyze the historical data for container ships in the entire global maritime network and detect congestion dependencies. The tool can be used to identify key shipping routes or ports that: 1) are prone to delays; 2) greatly affect the overall maritime network due to position, connections and risk of congestion; and/or 3) get affected the most by delays in previous route legs.

INDEX TERMS Maritime, sector, infrastructure, congestion, flow, container, port, dependency, graph, centrality, cascading, delay, risk, impact, delay, route.

I. INTRODUCTION

The U.S. department of Homeland security identifies the Maritime Transportation System as one of the seven key subsectors of the Transportation Systems Sector [1]. Efficient use of existing maritime network systems can result in reduced traffic congestion and cost while traffic flow prediction models can help both maritime users and authorities improve port traffic resilience, alleviate congestion or reduce traffic incidences in advance, by predicting the future states of traffic flow [2]. Dependency modeling, simulation and analysis of infrastructures have been studied extensively by researchers. Several methodologies and tools that focus on dependency analysis, estimate the impact [3], [4] or the risk derived from the dependencies within a critical infrastructure or among interdependent infrastructures [5]–[8]. The risk usually depends on two factors; the likelihood (or probability) of a negative event occurring and the impact (consequences) of that negative event. Such impact usually results in a complete non-operation or a partial malfunction in an infrastructure due to dependencies in infrastructure networks, usually called a failure.

A. CONTRIBUTION

We use a previous time-based dependency analysis methodology for Critical Infrastructure dependency modeling [6], [7], to analyze maritime port network traffic flows. We currently apply the proposed methodology in a dataset with ship

routes, journey time flows provided by MarineTraffic Corporation [9]. MarineTraffic records 800 million vessel positions, 18 million vessel and port related events on a monthly basis. The company provides details of over 650 thousand marine assets available (vessels, ports, lights) [9].

The dataset includes entry and exit calls for container ships in all ports that connect them, including journey times, anchor and port entry and exit calls along with traffic flows over a 3-year period. i.e. Entry calls denote the time when a ship is allowed to enter a port and dock in a specific assigned position, while exit calls denote the time when a ship leaves port and frees up a slot.

Our two major contributions are:

1. A methodology able to model maritime networks as dependency graphs and calculate the dependency risk between interconnected ports using assessment of traffic flow data. The methodology is able to assess the risk of congestion in ports and produce weighted risk dependency paths (i.e. how congestion in one port affects other connected ports through ship routes). Ports are represented as graph vertices, with ship routes linking one port to another as edges between the vertices. The methodology uses the min-max algorithm and statistical dynamic averages to calculate likelihood and impact of congestion.

2. An analysis of historical ship data for entry and exit calls in ports of the maritime network, detecting n-order port dependencies and automatically proposing ports for specific mitigation solutions that increase the overall resilience of the entire network. The overall risk of cascading congestions was computed, which indicates the probability of impact transmission in a path through the connected edges. Specific ports that initiate cascading traffic congestions were detected and simulation results indicated that applying traffic control on them may increase overall traffic flow resilience in wider areas while decreasing the risk of congestions up to 12%. We cross-referenced results with reports to check their validity.

II. RELATED WORK

During the past decade, modeling of infrastructures along with the flow of information between them has been a major topic of interest in research. This section summarizes models used in such research and focuses on similar work on the maritime and transportation infrastructures. We opt to present related work in these sectors due to similarities in methodologies used to model urban transportation highways between cities and shipping routes between ports. To this end, we should note here that the methodology presented in this paper has already been used successfully in urban transportations to predict high-risk road junctions and propose traffic congestion mitigation mechanisms [10].

A. INFRASTRUCTURE NETWORK MODELING

Many approaches exist in modeling infrastructure dependencies and information flow. Generally, infrastructure modeling appears to be associated with simulation techniques and mathematical models: (i) continuous time-step simulation; (ii) discrete time-step simulation; (iii) Monte Carlo simulation; (iv) decision trees; (v) geographical information systems; and (vi) risk management among the most famous [11].

One of the most cited publications [12] categorizes critical infrastructure protection methodologies and tools into (i) Empirical, (ii) System Dynamics, (iii)Agent-based and (iv) network-based. Empirical models are based on historical events, disaster data and expert knowledge to identify failure patterns. System Dynamics utilize top-down methods such as stock and flow to manage and analyze complex adaptive systems with interdependencies. Agent-based approaches model components of infrastructures as agents and analyze agent interaction based on sets of rules, while network-based approaches model infrastructures as network graphs whose nodes represent infrastructure components. In general, the most dominant approaches are agent and network-based approaches [11].

Our approach is purely network-based and creates a model of all ports in the global maritime network. Ports are modeled as nodes while ship routes between ports are portraited as graph links.

B. MODELING OF TRANSPORTATION INFRASTRUCTURES

Concerning traffic flow analysis, simulation models are widely used in research for modeling and understanding traffic flow and congestion in many critical infrastructure sectors; most noticeably in the Transportation Systems sector [2], [10], [13]. Many traffic flow models exist and are usually classified into one of the following categories as proposed by [2]:

1. Detail (submicro, microscopic, mesoscopic, macroscopic)
2. Independence scale (continuous, discrete, mixed)
3. Process representation (deterministic, stochastic)
4. Application area (stretches, links, junctions)
5. Type (traffic management, design, optimization)

According to research, short-term prediction models are often better at providing vicinity measurements of traffic flow in the subsequent instants of any occurred event of a traffic accident. Long-term predictions mostly provide generic, global measurements of traffic, which allows for trend analysis but fails to predict cases of atypical events (e.g. accidents in transportations).

Since most traffic flow analysis research is conducted for urban transportation systems, we opt to compare and classify the presented method with similar ones from this sector. Transportation system models range from Kalman filtering [14], exponential filtering [15], nonparametric statistical methods [16], [17], spectral and cross-spectral analyses [18], [19] and sequential learning [20] or predictions from pure time-series models [13].

Similarly with [10], the presented implementation is also considered to be a cyber-physical, deterministic, long-term optimization model that uses risk assessment, statistical analysis and graph theory to promote decision making for

container ships. The method cannot be considered neither stochastic nor short-term, since all data are gathered by AIS sensors onboard ships and refer to discrete measurements over extended periods of time.

C. MODELING OF MARITIME PORT SYSTEMS

Scientific literature on similar subjects is extremely limited. Physical embedding of geographical port locations can be used on spatial networks [22], but with limited benefits on shipping patterns due to the absence of track, data and port restrictions. Still physical locations mappings along with their spatial proximity along a given coastline [23] is implicit in studies for selecting ports based on their adjacency on various levels and are used for port analysis based on straits, basins and seas [21].

The frequency and density of trade linkages have also been used as indirect indicators of port influence based on the interdependencies that they create [24].

Other criteria that affect ship routes and port selection range from trade demand, quality to cost of service at and between ports in selected routes [25]. Some research specifically shows that especially for container ports, transport chain actors have considerable decision power over which ports will be selected on-route [26], which differentiates model graphs of the maritime infrastructure from similar social networks due to the existence of such external factors [21].

Research [21] and [27] also connects port size and location in a wider graph model with their influence on adjacent ports. Articles depict ports as graph nodes and ship routes between them as links. Spatial analysis of such models was conceptualized using graph centrality and intermediacy in [28], while [21] utilizes degree centrality metrics (i.e. number of links to other ports) as a key indicator of port influence in a maritime network.

Our work is similar to [21] since it also utilizes a graph model to depict global maritime networks and analyze port influence. Ducruet *et al.* [21] use graphs to reveal port relations among largest centrality ports and confirm the crucial importance of some ports, such as the Europe-Asia link through Singapore. The study also proposes the use of graph subgroups based on degree metrics to analyze port influence. Authors found out that while most subgroups of smaller ports are explained by spatial proximity, still exceptions exist revealing the permanency of specialized long-distance links between Western European ports and their former colonies.

Still, major differences exist between research in [21] and the work presented in this article. The biggest difference is that the work presented in this article does not simply use graph metrics to model port influence due to position, but can understand both the effect of congestion over all possible shipping routes for each ship and/or of the entire network in general. Contrary to [21], our work utilizes full historical data from all container ships routes from 2015 up until the second quarter of 2017 and does not rely on position of ports without any real-world information. Instead, the presented methodology uses formal likelihood metrics by taking into consideration traffic time of each ship route ever travelled and utilizes formal network metrics to calculate network congestion of all container ships and respective routes based on their own data.

Also, the dataset provided by MarineTraffic contains all Port Calls recorded during the last 2 years for Container Vessels plus some extra fields that were included to allow more flexible data filtering (beyond entry and exit call timestamps). By analyzing the complete set of potential dependency paths between ports, companies can project all the cascading congestion effects that may be realized and flag dependency risks of specific ship routes per season that are above a threshold for further attention. The presented work can also be used to run specific scenarios of interest to risk assessors. While the computation of the complete set of dependency risk paths may provide useful information and reveal ''hidden'' dependency risks between ports, assessors may also make use of this work to examine specific realistic scenarios. These include ''what-if'' scenarios that only consider initiating congestion events that affect one (or some) ports and allow mitigation through alternate shipping routes of avoiding ports on specific time periods.

III. DEPENDENCY ANALYSIS METHODOLOGY

The proposed methodology is based on a previous multi risk dependency analysis methodology by [6], [7], and [10]. The aforementioned methodology was also used to model the traffic flow of automobiles in the UK transportation system with significant success [10]. Similarly, the current implementation of this method also extends CIDA, a CI dependency analysis tool [29] and follows a similar approach, albeit for assisting MarineTraffic in modeling congestion between ports interconnected by ship routes. Here too, the tool utilizes the Neo4J graph database for handling very large graphs and relevant information [31]. Reviews and comparisons have showed that the Neo4J library performs better than other current solutions for large graph databases [32], [33].

Dependency analysis assesses the risk of nth-order dependencies by applying results of organization-level risk assessments performed by critical infrastructure owners and operators. A dependency can be defined as a ''one-directional reliance of an asset, system, network or collection thereof (within or across sectors) on an input, interaction or other requirement from other sources in order to function properly.'' Directional graphs $G = (N; E)$ are often used to model such dependencies [11], where N is a set of nodes and E is a set of edges. In this work, N is the set of ports of the maritime infrastructure and E is the set of the links among these components (i.e., the ship routes that connect the ports). The graph is directional to represent dependencies from one port to other ports within the maritime infrastructure. An edge from a port N_i to port N_j, i.e., $N_i \rightarrow N_j$, depicts the dependency relationship between the two nodes. Potential congestion disruption transferred through this dependency can be described by the values impact $I_{i;j}$ and likelihood $L_{i;j}$. The combination of these two values indicates the dependency

risk $R_{i;j}$of port N_j to port N_i due to its dependence, which is denoted by the edge $N_i \rightarrow N_j$. The dependency risk is quantified as an integer scaled [0...5], with 0 representing no risk of serious delays and 5 severe risk of serious delays. This value as associated with each edge, refers to the level of cascade derived risk for the receiver due to the dependency.

The results of ship route analysis are used as input to this method. If $CI_{Y0} \rightarrow CI_{Y1} \rightarrow \ldots \rightarrow CI_{Yn}$ is a chain of port dependencies based on a ship route, $L_{Y0...Yn}$ is the likelihood of the nth-order cascading congestion effect and $I_{Yn1,Yn}$ is the impact of the $CI_{Yn1} \rightarrow CI_{Yn}$ dependency, then the cascading risk exhibited by CI_{Yn} due to the nth-order dependency is computed as

$$R_{Yo,\ldots,Yn} = L_{Y0,\ldots\ldots,Yn} * I_{Yn-1,Yn} = \prod_{i=0}^{n-1} L_{Yi,Yi+1} * I_{Yn-1,Yn} \tag{1}$$

The cumulative dependency risk considers the overall risk exhibited by all the critical infrastructures in the sub-chains of the *n*th-order dependency. Let $CI_{Y0} \rightarrow CI_{Y1} \rightarrow \ldots \rightarrow CI_{Yn}$ be a chain of dependencies of length *n*. The cumulative dependency risk, denoted as $DR_{Y0,Y1,\ldots,Yn}$, is defined as the overall risk produced by an *n*th-order dependency:

$$\begin{aligned} DR_{Yo,\ldots,Yn} &= \sum_{i=1}^{n} R_{Y0,\ldots,Yi} \\ &= \sum_{i=1}^{n} \left(\prod_{j=1}^{i} L_{Yj-1,Yj} \right) * I_{Yi-1,Yi} \end{aligned} \tag{2}$$

Eq. (2) computes the overall dependency risk as the sum of the dependency risks of the affected nodes in the chain due to a failure realized in the source node of the dependency chain. The risk computation employs a risk matrix that combines the likelihood and incoming impact values of each vertex in the chain. Interested readers are referred to [6] and [7] for additional details about dependency risk estimation.

A. "LIKELIHOOD OF CONGESTION" FORMAL METRIC FOR ANTICIPATING PORT DELAYS

Each relationship is assigned with a likelihood value, which declares, how likely the port described by the current relationship is, to be congested. Intuitively, this value is a probability, based on which we can make predictions about each port's state, at different times. In order for this to be calculated, we firstly need to check whether each relationship is proportionally fair to the other relationships describing the same port.

1) MIN-MAX FAIRNESS AS A LIKELIHOOD METRIC

Generally, when talking about proportional fairness, we are referring to a system, in which two or more competitive entities are battling for resource control, and how we can maintain balance between them [30]. For example, in a computer network, our goal is to maximize the total throughput while at the same time allowing all the users to experience at least a minimal level of service. The above is calculated according to [30] as follows:

A vector or rates x_r is proportionally fair if it is feasible, that is $x_r \geq 0$, and if for any other feasible vector x_r^*, the aggregate of proportional changes is zero or negative:

$$\sum_{r \in R} \frac{x_r^* - x_r}{x_r} \leq 0 \tag{3}$$

In our case, x_r is the flow of the currently examined relationship, and x_r^* is the vector containing all the flow values for all other relationships referring to the same port as x_r. We mark each of our relationships as "good" iff they satisfy (3), and "bad" otherwise. Note that we do not have to check for feasibility, since both vectors contain traffic flow values, which are always positive.

After marking all the relationships as either "good" or "bad," the likelihood value for a relationship R is calculated as follows

$$L_R = \frac{Number\ of\ times\ R\ appears\ marked\ as\ "Bad"}{Total\ number\ of\ times\ R\ appears} \tag{4}$$

2) LINK FLOW CALCULATION

The aforementioned network metric needs flow values for all links in a network. In order to calculate average wait time flows for ship routes in ports, we calculate each container ship flow metric from any port A to any port B it has ever traveled. In a given ship route that involves multiple ports, a single connection between two ports (i.e. an edge between two nodes in the maritime route graph) is called a *leg*.

To calculate flow, we define flow of each leg as the number of ships passing from a reference point per unit of time (in our case, the amount of entry + exit port calls measured in ships per month). Link flow (leg flow) for a single port-to-port connection (i.e. for every edge in the graph ship route legs) is defined as *the amount of time passed from the time point when a container ship received an exit call to leave a port of origin, until the time when it received an entry call from a destination port*. Entry and Exit Call timestamps provide us with a number that shows how many milliseconds have passed between 01.01.1970 and the time the timestamp was taken. Thus, deducting the timestamp of an entry call given by a destination port from the timestamp of the exit call from the port of origin gives us the amount of time passed while the ship was on route (travel time) between the two ports. This time value also includes any potential congestion time that the ship had to endure by remaining anchored outside the port and waiting for a slot. We should note here that, contrary to other types of ships, container ships do not transfer cargo in and out of the ship while anchored in anchorages outside ports. Since this research aims to model only container ships, we can safely assume that *all time in anchorages for container ships can be considered as delay*.

Since our experiments analyze route traffic per month to generate trends and allow seasonal congestion detection, all link flow values are aggregated per month (flow of ships per month).

B. THE "IMPACT" METRIC FOR CONGESTED SHIP ROUTES

Aside from the likelihood value, each relationship is assigned with an Impact value as well. As the name suggests, this metric declares how severe a possible congestion will be on that edge. Impact thresholds are calculated based on averages of best and worst-case data entries. The metric is described in levels, ranging from 1 to 5, with 5 being the highest. The impact level for each relationship of a port is calculated as follows:

1. We create a [1; 5] scale from our domain of flows (flow values for a port).
2. For each relationship, we rescale its flow value into the correct impact level in the range [1; 5].

We used the well-known linear scaling algorithm to achieve the above. Its result depicts the number of units of the original interval which are equal to 1 unit of the new interval.

The approach described above is a dynamic one. There is no common scale for all the ports. This happens, because in order to calculate the impact level for each port, we rescale based on its own maximum and minimum flow, so each produced impact level describes how impactful a flow value is for the currently examined port only. Thus, for each port, the scale describing its impact levels adapts to its own characteristics (its min and max flow values).

C. RISK CALCULATION PER SHIP ROUTE - DEPENDENCY PATH

Overall risk chains are calculated for each potential shipping route that involves more than two ports. The extension of the CIDA tool computes the security risk and/or impact evolution of ship route delays over all possible container shipping routes for each container ship. The tool represented thousands of dependent ship routes for all ships along with their ports as a weighted, directed graph. Each connection (edge) between two ports in a shipping route gets a likelihood and impact values that are derived from the impact and possibility of congestion between two ports in a ship route leg. Overall risk of each path is then calculated based on the equation presented above.

It should be noted that the current software and model calculates shipping route risk values by assuming that a single port acts as the initiator of congestion that cascades to a ship's overall routes. It does not cover common-cause failures that simultaneously affect several, independent ports. Also, modeling congestion is not affected by the type of event that cause the port congestion or the ship route delay. Accidents, weather and human-initiated strikes can all account for introducing delays in container ships. For example, a common-cause bad weather event may concurrently affect ports due to their physical proximity. Still, all time delay introduced is depicted in the differences in the amount of time for a ship to complete a leg of its journey. Since we monitor trends over multiple years, any event that is repeated periodically will be captured through the periodic time delay (congestion) whereas random congestions will only affect a port's / route's overall statistical probability of being congested.

Previous empirical research [35] proposes that cascading effects beyond the fifth-order rarely affect infrastructures. We were also able to confirm this in [10], since the likelihood of a cascading congestion was very small after the 5^{th} junction. Consequently, in this paper we accept the fifth-order as the upper limit on the number of port dependencies evaluated.

IV. MARITIME PORT CONGESTION ANALYSIS AND MITIGATION FOR CONTAINER SHIPS

First, we describe the data set used in our analysis and then we describe the proposed methodology in detail.

A. DATA SET PROVIDER

The dataset was provided by MarineTraffic [9]. It is a proprietary asset and was pseudonymized before delivery to protect relevant rights from ship owners. The data set contains all Port Calls (i.e. entry or exit alerts of a ship into/from a port) recorded by Container Vessels from the beginning of 2015 until the end of the first half of 2017 plus some extra fields (beyond the timestamp) were included to allow more flexibility. Specifically, data entries are container ship routes between two ports; an exit call marks the beginning of a ship route from port A and a relevant entry call marks the end of a ship's journey to the next port B. All entries contained the following information:

- VESSEL_IDENTIFIER: Generated ship identifier for tracking ship's route history.
- TEU_CAPACITY: The maximum number of containers that the ship can carry, indicative of the size of a Container Vessel. Used for filtering per size segments.
- PORT_NAME: The name of a port / berth.
- PORT_TYPE: P = Port, A = Anchorage, C = Canal / Strait, M = Marina, T = Offshore Terminal, Y = Shipyard, Demolition Yard, S = Shelter
- PORT_SIZE: Large, Medium, Small (L, M, S) based on the total traffic at each port.
- COUNTRY_CODE: The country to which a port belongs
- AREA_CODE: Larger geographical area to which a port belongs
- CALL_TIMESTAMP: Event log time
- MOVE_TYPE: 0 = Arrival, 1 = Departure
- INTRANSIT: 1 = Ship crossing from a defined geometry without stopping
- DRAUGHT: The ship's draft at Port Call's time.

B. DATA VALIDATION

To ensure the quality of the dataset for experiments, we opted to clean data entries from inconsistencies and useless information. Since our dataset only contained container ship calls over time, the first step was to transform the data into a balanced panel by removing or extrapolating tuples (i.e. couples of problematic exit-entry calls) to ensure that all ship routes have calls corresponding to the same intervals. In other words, we opted to remove records where a ship would leave a port (exit call) but would never reach a destination (missing

entry call) due to erroneous AIS data. The algorithm used was the following:

1. Remove entries to shipyards and demolition yards since it is out of scope. A total of 875 entries were removed.
2. Remove unanswered calls of ships that entered but never came out of ports (or vice versa) as follows:
    a. Create map based on vessel id and fill with entries.
    b. Sort entries based on (a) vessel id, and then based on (b) time stamp. After sorting was finished, we had all the entries for each vessel grouped together, in chronological order.
    c. For each vessel id, parse each entry.
    d. If current entry is entry call
    e. If next entry is an exit call and port name matches, then record both the entry and exit call on the output file.

We removed all entries regarding anchorages, since we are only interested in the total time of a vessel that traveled from a source port to its next destination port; and that includes the time it might have spent in an anchorage. To calculate the total time (duration) of an itinerary, we just need to subtract the source time stamp from the destination time stamp. We should note here that container ships do not load or unload cargo on anchorages or mid-sea. This allows us to ignore potential congestion on anchorages since they have no fixed size (i.e. ''areas'' to anchor always exist outside a port).

This approach also aids in avoiding cases of drifting, i.e. when a container ship ''drifts'' in and out of a dynamic anchorage area due to high winds, which results in having to consecutive exit and entry calls in 2 to 20 minute intervals.

This algorithm also covers some strange cases where the ship could enter a port X and then exit from a different port (i.e. two calls were missing). Then, the dataset was split based on vessel id. Fortunately, among millions of records, only 319.620 tuples were found to be erroneous.

It is important to note here that since the task at hand is not that of a mainstay recognition classifier but rather a reliability analysis of dependencies that arise in the maritime infrastructure, the notion of a biased dataset is not as relevant as in a, for example an image classifier. It is our belief that, since our dataset is made of real traffic data spanning several years, any potential imbalance of classes is a real-world representation and should not be tampered with. In addition, the amount of data was big enough to exclude unrealistic biases in data. Instead of rebalancing the classes we opted to report results based on the ratio of valid data entries per ship to the total amount of data entries per ship, as that is a much better indicator of the overall quality of data per container, and ensures that impact representation of congestion for ship routes per ship will lead to the correct identification of classes even in cases where the dataset is naturally unbalanced (i.e. some ports like Singapore are congested much more often than others and consequently have way more data to work with).

V. RESULTS

Our tool modeled the global maritime network of ports and container ship routes, along with respective anchorages, route flow (i.e. time passed between an exist call and following entry call for all containers ever to have sailed between 2015 and Q2 2017). Then, our tool generated all dependency chains for each ship and for each route travelled from the MarineTraffic dataset graph database. Path generation had significant requirements in time and memory. For example, on a cluster of two Core i7 PCs with a total of 32GB RAM and only for one container ship, the overall analysis took more than 20 minutes to complete whereas analyzing all ship routes for all ships and instances present in the dataset took two days only to create results for one ship against the entire global network. Similar to previous tests [10], the tool was again developed in the Java language, using the Neo4J graph database [31]. The tool accepted all ports (nodes) and ship route legs (edges) as input from converted CSV files. All csv dataset data were imported into a Neo4J graph database and then fed to CIDA for risk path calculation.

A file was created per ship which was then aggregated for data analysis to examine critical congestion dependencies. Thorough analysis of the entire dataset for years 2015 through Q2 2017 was performed on two different granularity levels for calculating dependency chains of ports in shipping routes: (i) average flow metrics per month, and (ii) average metrics per year. This allowed us to: (a) compare an output dataset in the order of many gigabytes with a more useful one and see if detailed results agree with monthly averages and (b) to have enough granularity to answer specific questions about specific ports during specific seasons and pinpoint potential factors that affect average monthly flows in the entire global container shipping network.

A. INDICATIVE RESULTS – CONTAINER SHIP 82274

As an example, Table 1 and 2 below depict results generated by our tool. Table 1 depicts port influence metrics calculated by analyzing the bulk of all dependency chains.

TABLE 1. Port influence metrics on paths (ship 82274).

MOST OCCURING PORT OF ORIGIN	Average Risk SEATTLE
SEATTLE	2.202
MOST OCCURING DESTINATION PORT	**Number of HONG KONG occurrences as DESTINATION**
HONG KONG	42
MOST OCCURING PORT (TOP 5 PATHS)	**Average Risk NORFOLK (All records)**
NORFOLK	2.533
COVARIANCE	
0.013	Ship route risk to end port covariance
-0.002	Ship route risk to starting port covariance

TABLE 2. Top 10 worst dependency ship routes (ship 82274).

RISK	FIRST PORT	1ST ROUTE RISK	SECOND PORT	2ND ROUTE RISK	THIRD PORT	3RD ROUTE RISK	4TH PORT	4TH ROUTE RISK	5TH PORT	5TH ROUTE RISK	6TH PORT
3,5	NOR	1	SAV	1	CHRL	0,5	PKLNG	0,5	SINGA	0,5	JAKA
3,43	NY	0,86	NOR	0,86	SAV	0,86	CHRL	0,43	PKLNG	0,43	SINGA
3,24	NY	0,86	NOR	0,86	SAV	0,86	CHRL	0,43	PKLNG	0,24	HK
3,21	NY	0,86	NOR	0,86	SAV	0,86	CHRL	0,43	PKLNG	0,21	CAIM
3,03	NOR	1	SAV	1	CHRL	0,5	PKLNG	0,29	HK	0,24	YANTIAN
3	NOR	1	SAV	0,5	PKLNG	0,5	SINGA	0,5	SINGA	0,5	JAKA
3	NOR	1	SAV	1	CHRL	0,5	PKLNG	0,5	SINGA		
3	NY	0,86	NOR	0,86	SAV	0,43	PKLNG	0,43	SINGA	0,43	JAKA
3	NY	0,86	NOR	0,86	SAV	0,86	CHRL	0,43	PKLNG	0	HLFX
3	PKLNG	1	SINGA	1	SINGA	1	JAKA				

The Risk column presents the overall congestion risk as the sum of the dependency risks of the affected port destinations in the ship route / chain of ports for the given ship. ROUTE RISK cells present the risk of congestion for the given shipping route leg that connects a port of origin with a destination port (e.g. "FIRST PORT" connects with "SECOND PORT" with a given congestion risk of "1ST ROUTE RISK" value).

Table II port abbreviations: NOR = NORFOLK, NY = NEW YORK, SAV = SAVANNAH, CHRL = CHARLESTON, PKLNG = PORT KLANG, SINGA = SINGAPORE, SHA = SHANGHAI, HK = HONG KONG, JAKA = JAKARTA, HLFX = HALIFAX, CAIM = CAI MEP

Table 2 depicts the top 10 worst dependency chains (container ship complete routes) detected for Container Ship 82274 from 2015 up until Q2 2017. Generally, our tool can detect:

- Ship routes (dependency paths) with highest overall risk to be congested and have major impact on ships in the entire network.
- Ship routes (dependency paths) with highest overall impact for specific ships per season.
- Alternate routes per ship or per destination. Our tool can propose alternate paths and avoid high-risk ports depending on historical data for specific months.
- Highest influence ports in terms of both impact (how much delay they introduce to ship journeys) and influence (how many routes they affect).

In our example, by parsing the entire dataset concerning the top ten paths from each day, statistical analysis showed that some ports clearly appear more than others in congested dependencies and seem to greatly affect the network concerning ship 82274 (or parts of its route) in terms of delays.

Besides detecting the worst paths in terms of congestion and pinpointing ports and their influence on specific ships or the overall network per season, we were able to generate further interesting results.

Although Los Angeles appears to affect ship 82274 the most in terms of impact, since it appears as a key port in all high-risk congestion dependency paths, still the Seattle port seems to actually have a bigger indirect effect on the container's delays. By analyzing all of ship's 82274 routes (high and low risk alike), Seattle appears dozen times more often than Lost Angeles. In terms of weighted statistical analysis, our tool predicts that long-term delays at the Seattle port will eventually affect ship 82274 more, while few scarce high-impact congestions at the Los Angeles port with have greater effect.

The highest congestion that led to the worst overall delays was detected in Norfolk, while Port Klang appears to be one of the main destinations for the specific ship. Through dependency analysis, we were able to detect alternate, low-risk paths to that port although some of them were longer by significant miles. Still, in times of high congestion, results propose that following alternate, longer routes actually lowers delay instead of following the typical, highly-congested route.

Other uses of the tool's results are similar to research in [21], albeit with more granularity on the results, since our metrics not only consider degree centrality, but also weigh the influence of each port to other ports. In our example, Table 2 below depicts how often specific ports are detected as starting points in ship's 82274 routes in times where its journey was highly congested.

We calculated the Covariance of all likelihood and impact values produced by the presented methodology to search for unintended likelihood-impact biases; i.e. how much the two variables vary together. A positive covariance would indicate a positive linear relationship between the variables, and a negative covariance would indicate the opposite. A perfect methodology should provide a covariance rating close to zero, since likelihood and impact metrics have distinct purposes in our research; namely how often a congestion occurs and how big the impact (i.e. reduction of average speed) is. Fortunately, the covariance of the two variables was measured to be indeed low; around −0,2% to 1%. We consider this to be an excellent result, since both variables utilize traffic flows for their calculation but, on the other hand,

calculate different aspects of ship route congestion through ports.

After considering unintended likelihood-impact biases, we should also state that no overfitting issues exist concerning results. The algorithm does not predict future port congestion based on machine learning, but rather extracts congestion patterns and trends through existing, real-world sensor data. We do not opt for a generalization of training data to unseen data but rather to understand long-term congestion patterns in container shipping routes per ship and/or per season. Also, bias and variance that are forms of such overfitting errors provided no alarming values.

Multiple Regression (Multiple R) metrics measure the correlation between observed risk values of dependency paths and ports presented in them. A value of 1 (100%) means a perfect positive relationship and a value of zero means no relationship at all. In fields that attempt to predict human behavior (such as psychology), it is entirely expected that R-squared values will be low, typically lower than 50%.

R-squared values can be used on ship results to draw important conclusions about how changes on likelihood and impact of congestions in ports included in some ship dependency paths affect the overall risk of a ship route.

VI. CONCLUSION AND EVALUATION OF RESULTS

Overall data calculated were too big to depict in a single publication. To this end, only indicative results for one ship (ID 82274) were given.

We were able to detect the worst ports in terms of affecting the overall delay of a ship's routes, along with its highest-risk paths. Namely, for ship 82274, the worst congestion bottlenecks were found to occur due to the Seattle and Los Angeles ports. Besides being able to analyze port dependency paths and provide mitigation solution for delays per ship or per port, our tool also detected other ports affected by previous port congestions both for specific ships or as a trend, from the overall measurements of a given port and routes that pass through it. This enables us to create a global maritime dependency path table for the worst ports and how delay cascades into adjacent ports. We simulated applying mitigation solutions on routes (such as increased switching to different, lower-risk routes to alleviate overall delay). Results on some top worst initiating ports of origin showed a reduction in overall delay in ship routes in terms of 21%. Applying mitigation mechanisms on ports (i.e. increasing capacity thus lowering likelihood of congestion) could lead up to 15% reduction in delays over wider areas of the maritime network.

Modeling and analysis support the proactive study of large-scale congestion scenarios in ship routes of all container ships monitored by MarineTraffic. Results aid ship owners and decision makers to assess congestion risks of routes and ports before any delays are realized. The model can also be used to run specific scenarios of interest to ship owners concerning ship routes and reveal ''hidden'' risks under specific realistic scenarios (e.g. strike trends in specific ports or bigger overall delays in relatively shorter routes due to random

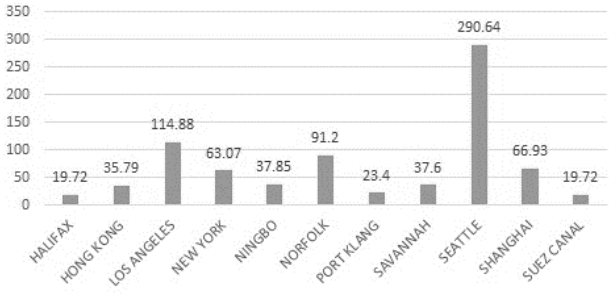

FIGURE 1. Risk sum for ports of origin (ship 82274) –Influence of ports of origin.

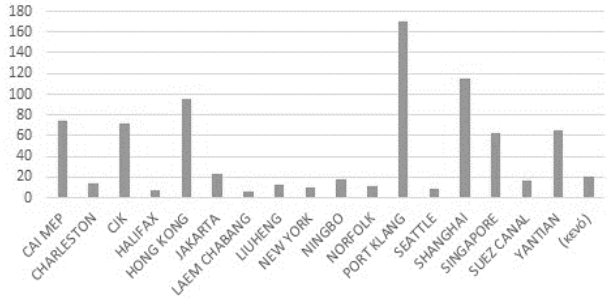

FIGURE 2. Risk sum for destination ports (ship 82274) – Depicts which ports are influenced the most from cascading delays from ports detected in ship's 82274 routes.

weather events). These include ''what-if'' scenarios that only consider delays that affect one (or some) ports/nodes.

According to MarineTraffic, the assessment of shipping routes and port influence will help increase scheduling and business decisions. Based on real input data, the tool can examine hundreds of scenarios, including previous incidents. Ship owners may model slight variations of shipping routes with different weights and even different dependencies to simulate the implementation of alternative container ship routes. For example, the tool can be used to project the effect of choosing different route legs or different anchoring times in ports to reduce the likelihood of congestion in specific ports.

The tool can also be used to identify key ports prone to delays, ports with great influence on the overall network due to (i) position, (ii) connections and (iii) risk of congestion, and/or ports that get affected the most by delays in previous ports. In this way, it is possible to evaluate the benefits of various alternative and/ or complementary routes, and provide arguments about their expected benefits.

ACKNOWLEDGMENT

We would like to thank the MarineTraffic corp. both for supplying us with the necessary data and also for their support during modeling, implementation and analysis of results.

REFERENCES

[1] U.S. Department of Homeland Security. *Critical Infrastructure Sectors*. Accessed: May 30, 2017. [Online]. Available: https://www.dhs.gov/transportation-systems-sector

[2] S. P. Hoogendoorn and P. H. L. Bovy, "State-of-the-art of vehicular traffic flow modelling," *Proc. Inst. Mech. Eng., I, J. Syst. Control Eng.*, vol. 215, no. 4, pp. 283–303, 2001.
[3] L. Franchina, M. Carbonelli, L. Gratta, M. Crisci, and D. Perucchini, "An impact-based approach for the analysis of cascading effects in critical infrastructures," *Int. J. Critical Infrastruct.*, vol. 7, no. 1, pp. 73–90, 2011.
[4] B. Robert, "A method for the study of cascading effects within lifeline networks," *Int. J. Crit. Infrastruct.*, vol. 1, no. 1, pp. 86–99, 2004.
[5] G. H. Kjølle, I. B. Utne, and O. Gjerde, "Risk analysis of critical infrastructures emphasizing electricity supply and interdependencies," *Rel. Eng. Syst. Saf.*, vol. 105, pp. 80–89, Sep. 2012.
[6] P. Kotzanikolaou, M. Theoharidou, and D. Gritzalis, "Cascading effects of common-cause failures in critical infrastructures," in *Proc. Int. Conf. Critical Infrastruct. Protection*. Berlin, Germany: Springer, 2013, pp. 171–182.
[7] P. Kotzanikolaou, M. Theoharidou, and D. Gritzalis, "Assessing n-order dependencies between critical infrastructures," *Int. J. Critical Infrastruct.*, vol. 6, no. 9, pp. 93–110, 2013.
[8] I. B. Utne, P. Hokstad, and J. Vatn, "A method for risk modeling of interdependencies in critical infrastructures," *Reliab. Eng. Syst. Safety*, vol. 96, no. 6, pp. 671–678, Jun. 2011.
[9] (2018). *MarineTraffic: Global Ship Tracking Intelligence | AIS Marine Traffic*. Accessed: Aug. 19, 2018. [Online]. Available: https://Marinetraffic.com
[10] G. Stergiopoulos, E. Valvis, F. Anagnou-Misyris, N. Bozovic, and D. Gritzalis, "Interdependency analysis of junctions for congestion mitigation in transportation infrastructures," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 45, no. 2, pp. 119–124, 2017.
[11] G. Stergiopoulos, E. Vasilellis, G. Lykou, P. Kotzanikolaou, and D. Gritzalis, "Classification and comparison of critical infrastructure protection tools," in *Proc. Int. Conf. Critical Infrastruct. Protection*, vol. 485. Arlington, VA, USA: Springer, Nov. 2016, pp. 239–255
[12] M. Ouyang, "Review on modeling and simulation of interdependent critical infrastructure systems," *Rel. Eng. Syst. Safety*, vol. 121, pp. 43–60, Jan. 2014.
[13] B. M. Williams and L. A. Hoel, "Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results," *J. Transp. Eng.*, vol. 129, no. 6, pp. 664–672, Nov. 2003.
[14] I. Okutani and Y. J. Stephanedes, "Dynamic prediction of traffic volume through Kalman filtering theory," *Transp. Res. B, Methodol.*, vol. 18, no. 1, pp. 1–11, 1984.
[15] P. Ross, *Exponential Filtering of Traffic Data*. 1982.
[16] G. A. Davis and N. L. Nihan, "Nonparametric regression and short-term freeway traffic forecasting," *J. Transp. Eng.*, vol. 117, no. 2, pp. 178–188, 1991.
[17] B. L. Smith, B. M. Williams, and R. K. Oswald, "Parametric and nonparametric traffic volume forecasting," in *Proc. Transp. Res. Board 79th Annu. Meeting*, 2000.
[18] Stathopoulos, Anthony, and M. Karlaftis, "Temporal and spatial variations of real-time traffic data in urban areas," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1768, pp. 135–140, Jan. 2001.
[19] A. Stathopoulos and M. G. Karlaftis, "Spectral and cross-spectral analysis of urban traffic flows," in *Proc. IEEE Intell. Transp. Syst.*, Aug. 2001, pp. 820–825.
[20] H. Chen and S. Grant-Muller, "Use of sequential learning for short-term traffic flow forecasting," *Transp. Res. C, Emerg. Technol.*, vol. 9, pp. 319–336, Sep. 2001.
[21] C. Ducruet and F. Zaidi, "Maritime constellations: A complex network approach to shipping and ports," *Maritime Policy Manage.*, vol. 39, no. 2, pp. 151–168, 2012.
[22] M. T. Gastner and M. E. Newman, "The spatial structure of networks," *Eur. Phys. J. B-Condens. Matter Complex Syst.*, vol. 49, no. 2, pp. 247–252, 2006.
[23] C. Ducruet, T. E. Notteboom, and P. W. De Langen, "Revisiting inter-port relationships under the new economic geography research framework," in *Ports in Proximity: Competition and Cooperation Among Adjacent Seaports*, T. E. Notteboom, C. Ducruet, and P. W. De Langen, Eds. Aldershot, U.K.: Ashgate, 2009, pp. 11–28.
[24] T. Notteboom, C. Ducruet, and P. de Langen, "Ports in proximity: Competition and coordination among adjacent seaports: Introduction," Tech. Rep., 2009, pp. 1–10.
[25] T. Notteboom, "Complementarity and substitutability among adjacent gateway ports," *Environ. Planning A, Economy Space*, vol. 41, no. 3, pp. 743–762, 2009.
[26] R. Robinson, "Ports as elements in value-driven chain systems: the new paradigm," *Maritime Policy Manage.*, vol. 29, no. 3, pp. 241–255, 2002.
[27] J.-P. Rodrigue and T. Notteboom, "Foreland-based regionalization: Integrating intermediate hubs with port hinterlands," *Res. Transp. Econ.*, vol. 27, no. 1, pp. 19–29, 2010.
[28] D. K. Fleming and Y. Hayuth, "Spatial characteristics of transportation hubs: centrality and intermediacy," *J. Transport Geography*, vol. 2, no. 1, pp. 3–18, 1994.
[29] G. Stergiopoulos, P. Kotzanikolaou, M. Theocharidou, G. Lykou, and D. Gritzalis, "Time-based critical infrastructure dependency analysis for large-scale and cross-sectoral failures," *Int. J. Crit. Infrastruct. Protection*, vol. 12, pp. 46–60, Mar. 2016.
[30] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan, "Rate control for communication networks: Shadow prices, proportional fairness and stability," *J. Oper. Res. Soc.*, vol. 49, no. 3, pp. 237–252, Mar. 1998.
[31] J. Webber, "A programmatic introduction to Neo4j," in *Proc. 3rd Annu. Conf. Syst., Program., Appl., Softw. Humanity*, 2012, pp. 217–218.
[32] S. Batra and C. Tyagi, "Comparative analysis of relational and graph databases," *Int. J. Soft Comput. Eng.*, vol. 2, no. 2, pp. 509–512, 2012.
[33] S. Jouili and V. Vansteenberghe, "An empirical comparison of graph databases," in *Proc. IEEE Int. Conf. Social Comput. (SocialCom)*, 2013, pp. 708–715, doi: 10.1109/SocialCom.2013.106.
[34] M. Van Eeten, A. Nieuwenhuijs, E. Luiijf, M. Klaver, and E. Cruz, "The state and the threat of cascading failure across critical infrastructures: The implications of empirical evidence from media incident reports," *Public Admin.*, vol. 89, no. 2, pp. 381–400, 2011.

GEORGE STERGIOPOULOS received the B.Sc. degree in informatics from the University of Piraeus, Greece, and the M.Sc. degree in information systems and the Ph.D. degree in critical infrastructure protection at software and information interdependency levels from the Athens University of Economics and Business, Greece. He is currently an Adjunct Lecturer and a Post-Doctoral Researcher with the Department of Informatics, Athens University of Economics and Business. He has published over 20 in peer-reviewed journals and international conferences. He was a Principal Investigator in multiple funded research projects in the areas of critical infrastructure protection, computer security, and network security. He is an expert in ISO 27001 and EU General Data Protection Regulation consulting.

EVANGELOS VALVIS is currently a Research Assistant with the Department of Informatics, Athens University of Economics and Business, Greece. He has worked in a number of research projects. His current research interests include software testing, computer security, and software engineering. He was a recipient of the Best Paper Award at the 16th International Working Conference on Source Code Analysis and Manipulation (SCAM 2016).

DIMITRIS MITRODIMAS received the Dipl.Eng. degree in mechanical engineering from the National Technical University of Athens, Greece, and the M.B.A. degree and the M.Sc. in business analytics from the Athens University of Economics and Business, Greece. He was a marine frontliner, involved in daily ship management operations and in projects aiming at improving operational performance. He is currently a Data Scientist with MarineTraffic, a global online ship-tracking service. His role revolves around the creation of value-adding data-driven products for the commercial maritime sector.

DIMITRIOS LEKKAS received the B.Sc. degree in mathematics from the University of Athens, Greece, the M.Sc. degree in information technology from Glasgow University, U.K., and the Ph.D. degree in information security from the University of the Aegean, Greece. He is the Founder and the Chief Product Officer of MarineTraffic, a global online ship-tracking service. His published scientific work includes over 40 journal and conference papers. His professional experience includes senior consulting, technical, and research positions at several private and public institutions.

DIMITRIS GRITZALIS received the B.Sc. degree in mathematics from the University of Patras, Greece, the M.Sc. degree in computer science from The City University of New York, USA, and the Ph.D. degree in information systems security from the University of the Aegean, Greece. He is currently the Associate Rector for Research and a Professor of ICT security with the Department of Informatics, Athens University of Economics and Business, Greece. He is also the Director of the Information Security and Critical Infrastructure Protection Research Laboratory and the Director of the M.Sc. Programme in Information Systems. His current research interests focus on cybersecurity, critical infrastructure protection, social media intelligence, data protection, risk assessment, and smartphone security. He has published over 150 papers in peer-reviewed journal and international conferences. He has served as an Associate Commissioner of the Greek Data Protection Commission and the President of the Greek Computer Society. He is the Academic Editor of the *Computers & Security* journal (Elsevier) and the Scientific Editor of the *International Journal of Critical Infrastructure Protection* (Elsevier).