Improving Management of Aquatic Invasions by Integrating Shipping Network, Ecological, and Environmental Data: Data Mining for Social Good

Jian Xu[*,1,2], Thanuka L. Wickramarathne[*,1,2,3], Nitesh V. Chawla[†,1,2,3], Erin K. Grey[3,4], Karsten Steinhaeuser[5], Reuben P. Keller[6], John M. Drake[7] and David M. Lodge[3,4]

[1]Dept. of Computer Science and Engineering, University of Notre Dame, USA
[2]Interdisciplinary Center for Network Science and Applications, University of Notre Dame, USA
[3]Environmental Change Initiative, University of Notre Dame, USA
[4]Dept. of Biological Sciences, University of Notre Dame, USA
[5]Dept. of Computer Science and Engineering, University of Minnesota, USA
[6]Institute of Environmental Sustainability, Loyola University Chicago, USA
[7]Odum School of Ecology, University of Georgia, USA

(jxu5, twickram, nchawla, egrey, dlodge)@nd.edu, ksteinha@umn.edu, rkeller1@luc.edu, jdrake@uga.edu

ABSTRACT

The unintentional transport of *invasive species* (i.e., non-native and harmful species that adversely affect habitats and native species) through the Global Shipping Network (GSN) causes substantial losses to social and economic welfare (e.g., annual losses due to ship-borne invasions in the *Laurentian Great Lakes* is estimated to be as high as USD 800 million). Despite the huge negative impacts, management of such invasions remains challenging because of the complex processes that lead to species transport and establishment. Numerous difficulties associated with quantitative risk assessments (e.g., inadequate characterizations of invasion processes, lack of crucial data, large uncertainties associated with available data, etc.) have hampered the usefulness of such estimates in the task of supporting the authorities who are battling to manage invasions with limited resources. We present here an approach for addressing the problem at hand via creative use of computational techniques and multiple data sources, thus illustrating how *data mining* can be used for solving crucial, yet very complex problems towards social good. By modeling implicit species exchanges as a *network* that we refer to as the *Species Flow Network (SFN),* large-scale species flow dynamics are studied via a *graph clustering approach* that decomposes the SFN into *clusters of ports* and *inter-cluster* connections. We then exploit this decomposition to discover crucial knowledge on how patterns in GSN affect aquatic invasions, and then illustrate how such knowledge can be used to devise effective and economical invasive species management strategies. By experimenting on actual GSN traffic data for years 1997–2006, we have discovered crucial knowledge that can significantly aid the management authorities.

Categories and Subject Descriptors

I.5 [**Pattern Recognition**]: Clustering; H.2.8 [**Database Management**]: Database Applications—*Data mining*

Keywords

Data mining, networks, clustering, risk assessment, invasive species, data mining for social good

1. INTRODUCTION

Background. Networks of human travel and trade transport different plants, animals and pathogens across the globe. Quoting from the National Invasive Species Council's 2008–2012 National Management Plan [6],

> *"Human activity such as trade, travel and tourism have all increased substantially, increasing the speed and volume of species movement to unprecedented levels. Invasive species are often unintended hitchhikers on cargo and other trade conveyances, (Page 4)." [6]*

Invasive species (i.e., non-native species that adversely affect habitats and bioregions) are among the top three drivers of global environmental change. Such invasive species include both plants and animals, and cause substantial economic and environmental harm by outcompeting or preying on native species. For instance, the impacts of *aquatic invasions* include increased diseases in humans (e.g., *cholera*) and aquaculture species (e.g., fish virus), losses of wild-caught fisheries (e.g, *comb jelly* invasion of the *Black Sea*), and losses of other ecosystem services. From an economic perspective, the estimated annual damage and control costs

*J. Xu and T.L. Wickramarathne have contributed equally.
†NV Chawla is the corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
KDD'14, August 24–27, 2014, New York, NY, USA.
Copyright 2014 ACM 978-1-4503-2956-9/14/08 ...$15.00.
http://dx.doi.org/10.1145/2623330.2623364.

of invasive species in the U.S. alone amount to more than USD 120 billion [19]. These species are introduced via the networks of human trade and travel, and analyzing these networks can illuminate potential management strategies, regulatory policies, incentive structures, and risks from changing climate.

Ship-borne invasive species problem. The Global Shipping Network (GSN) is the dominant global vector for unintentional translocation of non-native aquatic species [17]: species get translocated via *ballast water* (during ballast water uptake/discharge) and *biofouling* (i.e., the accumulation of microorganisms, plants, algae, or animals) on the surfaces of ships [8]. To reduce invasion risks, authorities (e.g., International Maritime Organization (IMO)) have proposed standards for the maximum density of organisms that can be discharged in ships' ballast water. These standards are based on the premise that reducing the concentration of live organisms in ballast tanks will reduce the number of invasions, but the extent to which this approach will actually reduce the invasion risk is unknown. In addition, this approach does not address invasions via biofouling, nor does it consider many poorly known, but likely significant, biological and ecological factors that influence invasion risk. Moreover, the problem cannot be understood at a regional level, because ship-borne species can arrive from anywhere via the GSN. *With all these uncertainties in place, decisions are still needed to be made about the most efficient ways to target limited management resources.*

Significance of the problem. In the few coastal areas with good invasive species monitoring, increased shipping connecting an expanding network of global ports is correlated with an accelerating accumulation curve of established species (e.g., San Francisco Bay), and is estimated to be responsible for 69% of known aquatic invasions [17]. Although only a portion of species transported via GSN become invasive (i.e., spread, become abundant, and cause harm), environmental and economic damages from these species are often large and increase over time [13, 15]. For instance, the annual loss to the US Great Lakes regional economy due to ship-borne aquatic invasive species may be as high as USD 800 million [22]. However, GSN undoubtedly provides enormous benefits to the US economy, and is also responsible for approximately 90% of global trade. Furthermore, global trade patterns are optimized based on economic and trade considerations, but not necessarily to safeguard against aquatic invasions. Therefore, imposing expensive and cumbersome regulations on the shipping industry could cause serious adverse effects to a country's economy and trade relationships.

Motivation and Goals. It is clear that a thorough understanding of ship-borne invasion risks in terms of overall data about trade patterns, ports, vessel types, etc. is necessary to devise practices and policies that are feasible, effective and capable of bringing to bear the net long-term benefits to human welfare. With this motivation, our goal is to develop computational and data-driven frameworks that can inform invasive species management policies and practices.

1.1 Data Mining for Social Good

Ship-borne invasions are a result of a complex interplay of ship traffic, ballast uptake/discharge dynamics, species survival during transport, various environmental/biological variables, etc [28]. Incorporating these complexities into a quantitative risk assessment framework is extremely difficult, since the majority of the governing relationships are poorly quantified. The few studies that have attempted to quantify invasion risks via probabilistic approaches [14, 24] have relied on multiple simplifying assumptions. Moreover, usefulness of these approaches towards development of efficient invasion management policies is further hampered by the inability to incorporate crucial invasion mechanisms (e.g., "stepping-stone" process) into risk assessment. However, numerous streams of data that capture vessel movement patterns, ballast uptake/discharge and other environmental/biological factors (that affect species transport and establishment) are increasingly being collected by several agencies for research/commercial purposes. Therefore, one can creatively combine domain expertise and computational data analysis to understand the underlying patterns of ship-borne invasions in order to develop a sufficient understanding towards the development of effective and economical management strategies. *Our work is in fact a multi-disciplinary attempt towards utilizing this data to create insights and knowledge that can eventually lead to decision-making tools for policy makers.*

Data. We now introduce the numerous data sources utilized for the research

(i) Vessel movement data: made available by Lloyd's Maritime Intelligence Unit (LMIU) contains travel information for vessels such as `portID`, `sail_date` and `arrival_date`, along with vessel metadata, such as `vessel_type` and `DWT` (Dead Weight Tonnage), etc. This information can be readily used to build a *network* to represent species flow paths and patterns among ports. Our experiments are based on LMIU data that spans four (4) two-years-long periods starting 1^{st} of May 1997, 1999, 2002 and 2005, totaling $6,889,748$ individual voyages corresponding to a total of $50,487$ vessels of various types that move among a total of $5,545$ ports and regions. However, none of the existing vessel movement datasets (including LMIU) provide explicit ballast water exchange amounts (or even whether a vessel discharged ballast water).

(ii) Ballast discharge data: made available by the National Ballast Information Clearinghouse (NBIC) contains the `date` and `discharge_volume` of all ships visiting U.S. ports from Jan. 2004 to present. As suggested in [24], NBIC data can be used to estimate an average ballast discharge based on `vessel_type` and `DWT` using a linear regression model.

(iii) Ecoregion data: are available via Marine Ecoregions of the World [26] and the Freshwater Ecoregions of the World [1], where *ecoregions* are defined by species composition and shared evolutionary history [26], and are thereby capable of providing an index of native ranges. Therefore, these definitions can be used for more realistic and qualitative invasion risk analysis, in comparison to, for example, geographic distance as used in [24].

(iv) Environmental data: on port temperature and salinity (i.e., the two most crucial variables for identifying survivability of species in non-native coastal environments) are available via Global Ports Database (GPD) [14]. These estimates can be used for calculating species establishment risk based on environmental similarity; the missing values in GPD can also be supplemented via estimates from the World Ocean Atlas 2009 [2, 16] when necessary.

Problem Statement. Given the complexity of the problem and lack of relationships that are required for robust risk assessment, we set forth to extract knowledge on large-scale patterns of GSN in order to obtain better insight towards ship-borne invasions of non-indigenous species. Furthermore, we will illustrate how such knowledge can be used to derive efficient invasion management strategies.

Framework. Our method is devised to tackle the limitations due to lack of data and governing relationships that are required for quantitative risk assessment. Towards this, we take the following approach: **(i)** a network that represents the general species flow tendency among ports is built; then, utilizing a graph clustering method [21] that operates on the basis of flow-dynamics, **(ii)** a *map* [12,27] of the species flow network, i.e., a cogent representation that extracts the main structure of flow while retaining information about relationships among modules (of main structure), is built; finally, using this map that summarizes the species flow dynamics in terms of *clusters* (or groups) of ports and highlights *inter-cluster* (i.e., between clusters) and *intra-cluster* (i.e., within cluster) relationships, **(iii)** the impact of GSN dynamics on aquatic invasions is studied in conjunction with ecological and environmental aspects that govern species establishment.

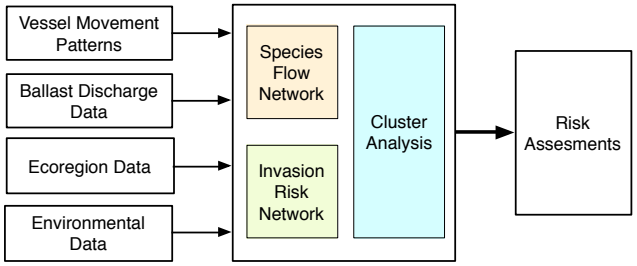


Figure 1: A **concept diagram** illustrating the integration of multiple data sources, modeling and data mining techniques for extracting useful knowledge.

1.2 Contributions and Broader Impact

We provide a data-driven foundation for more effective and efficient risk assessment and management by modeling the spread of aquatic non-indigenous species through the GSN, which is the most important vector of aquatic invasions. We have discovered vital information on patterns of GSN that can inform management strategies and regulatory policies. In a potential deployed configuration (see Fig. 2), the discovered knowledge can efficiently be used to analyze the invasion risks with respect to changing climate, policy and infrastructure. Understanding the structure of the component networks and the dynamic interactions between the different networks is crucial to the design of policies that could cost-effectively reduce invasions.

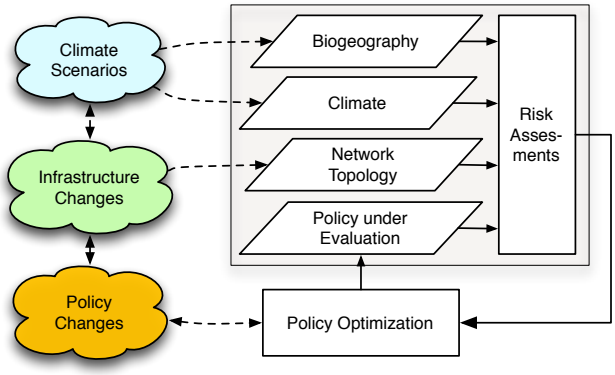


Figure 2: **Use of discovered knowledge in a potential deployed setting** for invasion risk assessment with respect to changing climate, policy and infrastructure.

This paper is organized as follows: Section 2 presents the formulation of species flow networks using LMIU and NBIC data, graph clustering approach for understanding the large-scale dynamics of GSN, and the main discoveries; Section 3 presents invasion risk assessment that incorporates ecoregion definitions and environmental conditions via a unique graphical approach; Section 4 presents how the emerging knowledge and methods can be potentially deployed towards development of species management strategies; and finally, Section 5 contains the concluding remarks.

2. SPECIES FLOW ANALYSIS

The basic idea behind our work is to find patterns of species flow in order to identify ports and shipping routes for which interventions would be the most effective in stopping invasions through the entire network. Such knowledge can then be further leveraged with auxiliary information (e.g., vessel types) in order to inform management strategies in a targeted manner. Since GSN naturally forms a graph, LMIU and NBIC data can be utilized to build a network to represent species flow among ports (see Fig. 3). This network can then be analyzed via graph mining or network science techniques to extract relevant insights.

2.1 Species Flow Network (SFN)

Let $\mathcal{G} \equiv (\mathcal{N}, \mathcal{E})$ be a directed weighted graph, where $\mathcal{N} \equiv \{n_1, \ldots, n_n\}$ and $\mathcal{E} \subset \mathcal{N} \times \mathcal{N}$ denote the set of nodes and edges of $\mathcal{G}$, respectively. Let the nodes in $\mathcal{N}$ correspond to ports visited by vessels in the GSN and the *weight* of the directed edge $e_{ij} \in \mathcal{E}$ given by $w_{ij} \in (0, 1]$ represents the *total probability of species flow* corresponding to all vessels traveling from port n_i to n_j (without intermediate stopovers), for all $n_i, n_j \in \mathcal{N}$.

Species Flow Estimation. Estimation of exact amounts of species exchanged between ports is extremely difficult. However, as proposed in [24], vessel movement and ballast discharge data can be leveraged to estimate the likelihood of species exchange. We now briefly explain how LMIU and NBIC datasets are used to estimate species flow (i.e., the edge weights of $\mathcal{G}$) and refer the interested reader to [24] for a comprehensive discussion on probabilistic species flow

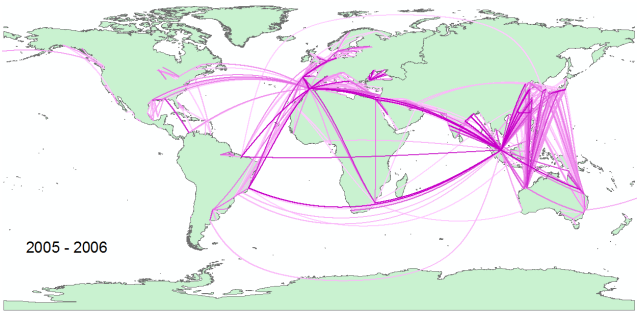


Figure 3: **Species flow between ports corresponding to vessel movements given in the LMIU 2005–2006 dataset.** The edges represent the aggregated species flow between ports, where the color intensity is proportional to the magnitude of flow. Approximately 2300 paths with the highest species flow are shown.

modeling including details on model assumptions, development and validation.

(a) Calculation of edge weights: Consider a vessel v traveling from port n_i to n_j in $\Delta t_{ij}^{(v)}$ time (without intermediate stopovers), during which the species in ballast water may die at a *mortality rate* of μ (is set to a constant average of $0.02/day$ for all routes r and vessel types). In addition, let $D_{ij}^{(v)}$, $\rho_{ij}^{(v)} \in [0,1]$ and λ denote the amount of ballast water discharged by vessel v at n_j, the efficacy of ballast water management for v for the route $n_i \to n_j$, and the characteristic constant of discharge, respectively. Then, the probability of vessel v introducing species from n_i to n_j (without intermediate stopovers) is given by:

$$p_{ij}^{(v)} = \rho_{ij}^{(v)} \left(1 - e^{-\lambda D_{ij}^{(v)}}\right) e^{-\mu \Delta t_{ij}^{(v)}}; \qquad (1)$$

the weight of the edge e_{ij} is taken to be the total probability of species introduction for all vessels traveling from n_i to n_j, and is given by:

$$w_{ij} = 1 - \prod_{\substack{r \in DB \\ r=v:n_i \to n_j}} (1 - p_{ij}^{(v)}), \qquad (2)$$

where the product is taken over all routes r in LMIU database DB s.t. a vessel v travels from port n_i to n_j.

(b) Estimation of ballast discharge: Information available on ballast discharge are incomplete, where estimation of exact quantities exchanged for each and every ship route r is impossible for most ports of the world: **(i)** ballast discharges in ports are not recorded globally, and are known to differ significantly by port and ship type; **(ii)** vessels may have intermediate stopovers, thus exchanging and mixing ballast water with existing water in ballast tanks; and **(iii)** data are largely unavailable for offshore discharges. Therefore, in order to mitigate the above difficulties, ballast discharge is estimated based on linear regression models on `DWT` per `vessel_type` as in [24]. Specifically, linear regression models on `DWT` for vessels of type `Bulk Dry`, `General Cargo`, `Ro-Ro Cargo`, `Chemical`, `Liquified Gas Tankers`, `Oil Tankers`, `Passenger Vessels`, `Refrigerated Cargo`, `Container Ships` and `Unknown/Other`). Furthermore, the relationship of ballast discharge amount to the likelihood of species introduction is not well defined. For estimation of (1), λ is chosen s.t. $p_{ij}^{(v)} = 0.80$ for a ballast discharge of $500,000\,m^3$, when $\rho_{ij}^{(v)} = 1$ and $\Delta t_{ij}^{(v)} = 0$, i.e., a discharge volume of $500,000\,m^3$ has a probability of 0.8 of introducing species if the vessel travels with zero mortalities and has no ballast management strategies in place.

Characteristics of the SFN. Summary of characteristics for SFNs generated for the four available years of data are shown in Table 1. The *path length* of a network identifies the number of stops required to reach a given port from another. An *average path length* of three (3) is observed in all four SFNs. This is perhaps mainly due to the presence of *hubs* (i.e., ports that are connected to many other ports) in GSN (e.g., Singapore). The *in-/out-degree* of a node is defined as the number of other nodes connected to/by it. Therefore, *average degree* in SFN describes the average number of direct pathways of species introduction. Furthermore, as empirical evaluations for *power law degree distribution* [5] suggest that SFNs fall under the category of *scale-free networks* [3], for degree $\geq$139.

Table 1: **Characteristics of Species Flow Networks**

Feature	97–98	99–00	02–03	05–06
Number of nodes	3971	4045	4264	4250
Number of edges	150479	150150	143560	145199
Average path length	2.987	2.998	3.018	3.041
Average in/out- degree	37.9	37.1	33.7	34.2
Diameter	8	7	7	9
Density	0.010	0.009	0.008	0.008

2.2 Clustering Analysis of SFN

Complex networks are efficient abstractions for highly complex systems that consist of numerous, often complex underlying patterns and relationships. However, these abstractions still remain too complex to derive useful inferences. Therefore, a decomposition that represents such complex networks via *modules* and their interactions [10, 18, 23] can be very useful in understanding the underlying patterns. We utilize a graph clustering approach in order to simplify the underlying flow dynamics of SFN. The clusters can capture the ship movement activity among ports leading to a better identification of risk corridors, which can then be used to estimate invasion risk based on ecological and environmental conditions.

Choice of Clustering Method. For the task at hand, we are interested in understanding how the structure of SFN relates to species flow across the network. Therefore, among many alternatives, *MapEquation* [21]—a graph clustering method that attempts to decompose the network with respect to *flow-dynamics* (in comparison to optimization of *modularity*)—is used. The basic principle of operation behind MapEquation-based clustering stems from the notions of information theory, which states the fact that a data stream can be compressed by a *code* that exploits regularities in the process that generates the stream [25]. Therefore, a group of nodes among which information flows quickly and easily can be aggregated and described as a single well connected module; the links between modules capture the avenues of information flow between those modules. MapE-

Table 2: **Ports that remained in the same cluster for the duration of 1997–2006**

Pacific %TP=28.33%, #P=818			**Mediterranean** %TP=15.61%, #P=513			**W. European** %TP=15.37%, #P=1117			**E. North_America** %TP=9.31%, #P=363			**Indian_Ocean** %TP=6.12%, #P=137			**South_America** %TP=3.41%, #P=80		
Port name	%TF	%CF	Port name	%TF	%CF	Port name	%TF	%CF	Port name	%TF	%CF	Port name	%TF	%CF	Port name	%TF	%CF
Singapore	2.82	9.96	Gibraltar	2.56	16.37	Rotterdam	0.87	5.68	Houston	0.52	5.57	Jebel Ali	0.25	4.07	Santos	0.42	12.37
Hong Kong	0.68	2.41	Tarifa	0.86	5.54	Skaw	0.60	3.93	New Orleans	0.37	3.94	Ras Tanura	0.22	3.67	Tubarao	0.33	9.70
Kaohsiung	0.58	2.05	Port Said	0.84	5.38	Antwerp	0.55	3.59	New York	0.35	3.80	Mumbai	0.20	3.29	San Lorenzo*	0.33	9.57
Port Hedland	0.52	1.83	Suez	0.48	3.09	Brunsbuttel	0.44	2.85	Baltimore	0.23	2.42	Juaymah Term.	0.19	3.12	Paranagua	0.21	6.11
Busan	0.50	1.76	Barcelona	0.29	1.83	Hamburg	0.42	2.76	Port Arthur	0.21	2.28	Kharg Is.	0.18	2.91	Rio de Janeiro	0.15	4.45
Hay point	0.49	1.72	Venice	0.24	1.52	Amsterdam	0.31	2.02	Santa Marta	0.20	2.17	Jubail	0.17	2.76	Bahia Blanca	0.15	4.31
Newcastle**	0.48	1.71	Genoa	0.23	1.47	Immingham	0.28	1.83	Tampa	0.20	2.16	New Mangalore	0.15	2.50	Rosario	0.14	4.07
Gladstone	0.47	1.67	Piraeus	0.22	1.39	St. Petersburg	0.27	1.73	Port Everglades	0.20	2.13	Mesaieed	0.13	2.08	Sepetiba	0.12	3.60
Nagoya	0.46	1.61	Leghorn	0.21	1.32	Tees	0.22	1.41	Mobile	0.19	2.04	Bandar Abbas	0.12	2.03	Rio Grande***	0.12	3.59
Incheon	0.45	1.60	Augusta	0.20	1.26	Zeebrugge	0.21	1.36	Savannah	0.18	1.95	Jebel Dhanna Term.	0.12	1.95	Praia Mole	0.12	3.50

Ports corresponding to highest %TF:=percentage flow w.r.t. total flow and %CF:=percentage flow w.r.t. flow within cluster are shown for six major clusters; for each cluster, the aggregated %TF:=percentage flow in the cluster w.r.t. total flow and number of ports in the cluster are given in the first row of table. Here, San Lorenzo*:=San Lorenzo, Argentina; Newcastle**:=Newcastle, Australia; Rio Grande***:=Rio Grande, Brazil.

quation identifies clusters by optimizing the entropy corresponding to intra- and inter-cluster in a *recursive* manner—the clusters identified cannot be further refined or partitioned.

Clusters of Ports based on Species Flow. Clustering analysis of SFN reveals several clusters of ports. These clusters represent groups of ports among which the species exchange is relatively higher; if inter-cluster pathways are controlled, species flow would be combinatorially reduced. Once such clusters are identified, species flow characteristics within clusters can be analyzed in conjunction with ecological and environmental data for invasion risk assessment. While clustering is derived based on species flow dynamics, geographical orientation of the major (in terms of aggregated flow) clusters is also intuitive (see Fig. 4). A few major clusters correspond to a significant proportion of total species flow among ports (see Table 2 for major ports that are in 6 of these major clusters). For instance, in 2005–2006, six (6) major clusters (out of 64 in total), viz., the clusters of `Pacific`, `Mediterranean`, `Western_Europe`, `Eastern_North_America`, `Indian_Ocean` and `South_America` contained 68.6% of total ports and corresponded to 76.3% of the total species flow.

Cluster Consistency. From a deployment perspective, perhaps the most crucial contribution of our analysis is that these major clusters continue to exist over the duration studied; for a given cluster, while some ports leave/join over time, the vast majority of the ports continue to remain in the same cluster (see Fig. 5). This provides a solid foundation for devising management strategies targeting clusters and inter-cluster connections to efficiently control species flow. Furthermore, evolution of clusters (or how the clustering patterns change over time) can reveal important information on how changes in vessel movement (and ballast discharge) patterns affect species flow dynamics. For instance, the exchange of the order of the two clusters `Mediterranean` and `Western_Europe` from 2002–2003 to 2005–2006 indicates a relative increase of species exchange among ports that belong to these clusters during 2005–2006, which can be attributed to the merger of a significant proportion of ports belonging to `Mediterranean` cluster with `South European Atlantic Shelf` cluster to form the `Tropical_East_Atlantic` cluster in 2005–2006. Another example is the formation of a new smaller cluster (the eighth in Fig. 5) in 1999–2000 by 21 ports in California and Hawaii, including ports such as San Francisco, Los Angeles and San Diego that previously belonged to the `Pacific` cluster in 1997–1998. Such changes can reveal large-scale trends that may be very useful in devising long term management strategies.

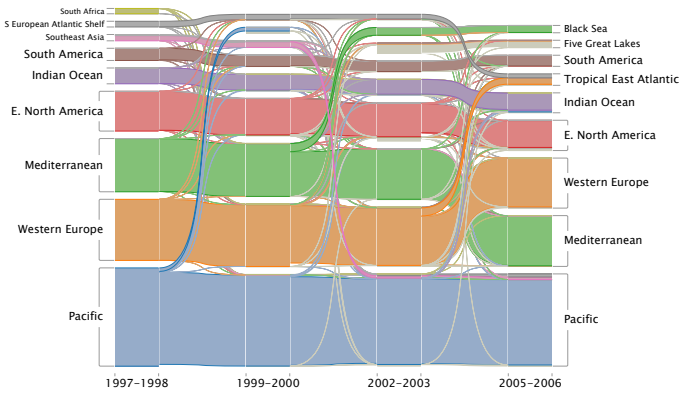


Figure 5: **Illustration of evolution of major clusters during the period of 1997–2006.** The clusters in *alluvial diagram* [21] are ranked by aggregated flow within the cluster. Here, the columns 1997, 1999, 2002 and 2005 represent the major clusters of SFN generated from LMIU datasets for 1997–1998, 1999–2000, 2002–2003 and 2005–2006, respectively.

3. INVASION RISK ANALYSIS

Quantification of invasion risk is a challenging problem because of the complex interactions between species and their abiotic and biotic environment [28]. Here, we shift our attention from inter-cluster species flow to *intra-cluster* (i.e., ports within a cluster) NIS invasion risk in order to gain insight into the plausibility of invasions in terms of environmental similarity. Previous studies have assumed that the invasion risk is proportional to *Euclidean* distance between annual averages of temperature and salinity [9, 14, 24]. However, this assumption may not be valid for invasive species that often exhibit broad environmental tolerances [7, 11]. We take an approach that is based on biogeographic patterns, and empirically observed temperature and salinity tolerances for ranking invasion risks.

3.1 Invasion risk modeling

For an exchange of species to become an invasion, the introduced species must be: **(i)** non-indigenous, viz., movements between non-contiguous ecoregions; and **(ii)** able to survive and establish in its new environment. Invasion risk between port environments can then be ranked by considering a species *assemblage* (or a collection) that contains "generalist" and "specialist" species. We have taken this ap-

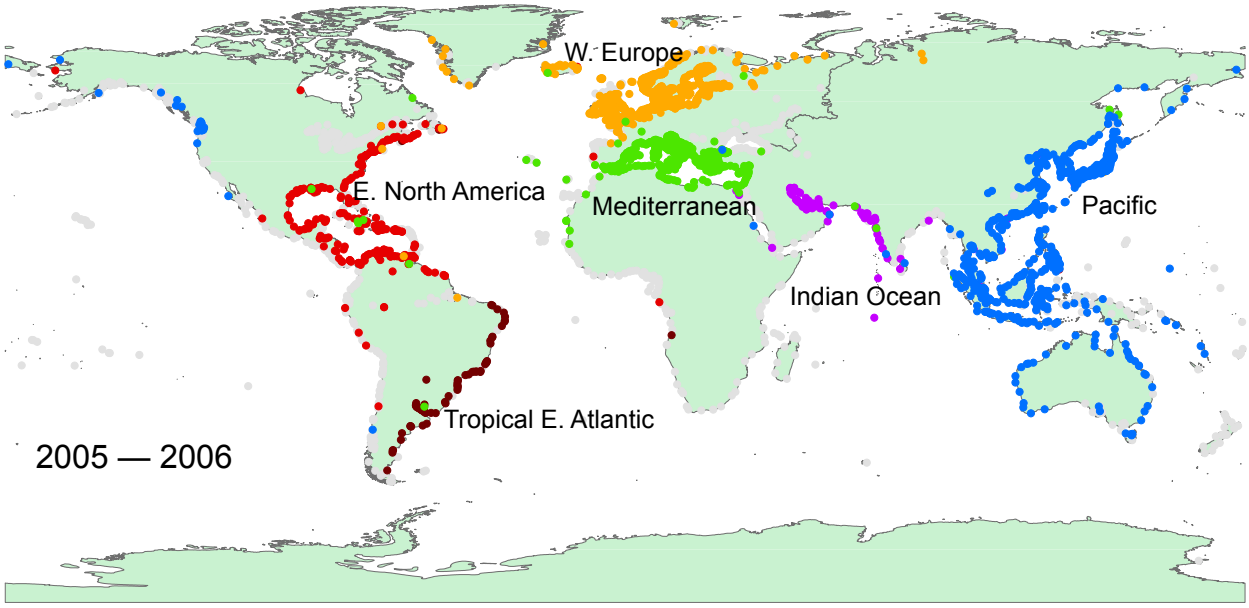


Figure 4: **The Six Major clusters of SFN during 2005–2006.** Color of dots correspond to that in Fig. 5, and white dots are not included in any of the six major clusters. Major clusters remain largely unchanged for the duration of 1997–2006, and contain a significant proportion of total species flow between ports.

proach to counteract the lack of relationships or data to calculate or estimate exact invasion risks. Based on temperature and salinity tolerance levels (empirically-estimated long term thermal tolerances of marine invertebrate taxa [20]), we define invasion risk in terms of number of species groups that can tolerate the given conditions. Specifically, six (6) different species tolerance groups based on two (2) temperature and three (3) salinity tolerance levels are considered (see Table 3). Here, salinity tolerance levels were set to capture species types that are completely intolerant to salinity (i.e., freshwater species), those that are restricted to marine waters (i.e., low tolerance), and species that can survive in a wide range of salinities (i.e., high tolerance). Risk between any two ports is then quantified as an index created by overlapping the species tolerance groups as shown in Fig. 6.

Table 3: **Grouping based on environmental tolerance**

Species Tolerance Group	Tolerance Levels	
	$\Delta T\,(^{\circ}C)$	$\Delta S\,(ppt)$
Tolerance Group 1	[0, 2.9]	[0, 0.2]
Tolerance Group 2	[0, 2.9]	[0, 2.0]
Tolerance Group 3	[0, 2.9]	[0, 12]
Tolerance Group 4	[0, 9.7]	[0, 0.2]
Tolerance Group 5	[0, 9.7]	[0, 2.0]
Tolerance Group 6	[0, 9.7]	[0, 12]

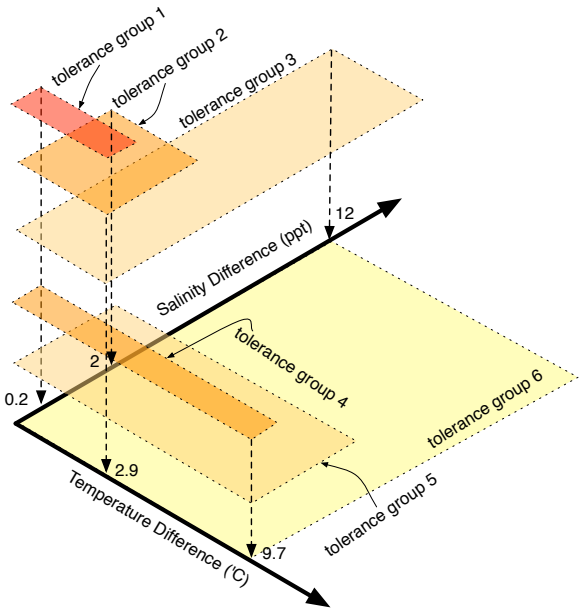
3.2 Environmental Similarity Network

Quantification of port-wise invasion risk is both difficult and not very useful in terms of species control and management. In order to gain insight into what ports are at risk based on species flow and environmental similarity, we utilize a graphical representation that is referred to as the *Invasion Risk Network (IRN)*. An IRN is built for every *major* cluster in SFN to intuitively represent the invasion risk based on how easily species can establish in the new environment (based on environmental tolerance, see Table 3). Note that IRN is an undirected weighted graph, since environmental match is symmetric, and the risk level (based on number of tolerance groups at risk) can vary between port pairs, respectively.

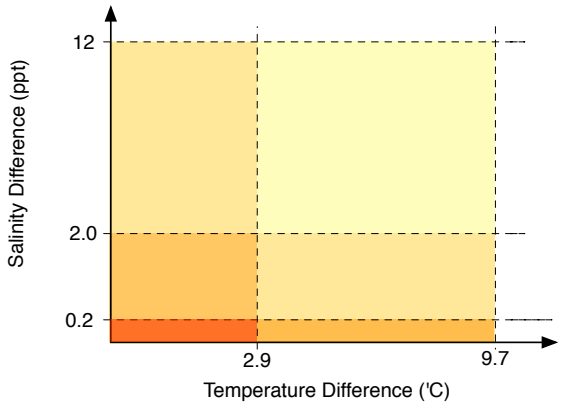
3.3 Clustering Analysis of IRN

With the edges representing the invasion risk between ports, clustering can help detect groups of ports that have similar environmental conditions (while belonging to different ecoregions). The basic idea here is to exploit the fact that the invasion risk between groups of ports that are very dissimilar (e.g., fresh-water ports and marine ports) is lower than ports within the same group (with relatively similar conditions). Clustering analysis (Section 2.2) can therefore again be utilized on IRN to identify groups of ports that are similar in terms of invasion risk. The clusters detected here are sub-clusters of SFN clusters (that are based on species flow dynamics); therefore, if two ports are in the same cluster of IRN, then it is very likely for an invasion to occur between these two ports. Furthermore, if adequate species flow control is not in place, given the frequent species exchanges and high chances of species establishment within ports in an IRN cluster, an invasion to one single port will immediately put all the other ports in the IRN sub-cluster at risk of an invasion. Therefore, IRN clusters identify groups of ports that would most benefit from some form of species flow control to avoid invasions.

Note: Clustering based on flow-dynamics (simulating random walks) is used here for identifying ports with similar environmental conditions, since an approach only considering pair-wise distance is not capable of capturing the stepping-stone effect.



(a) Species tolerance groups



(b) Risk level definition

Figure 6: **Illustration of *risk level* definition based on species tolerance groups and between-port environmental differences.** Sub-figure (a): identifies six (6) different species groups that categorizes the risk of survival relative to given difference in temperature and salinity based on two (2) temperate tolerance levels (high = can survive up to $9.7°C$ and low = can survive up to $2.9°C$ temperature difference) and three (3) salinity tolerance levels (zero = $0.2ppt$, low = $2.0ppt$ and high = $12.0ppt$ tolerance). Sub-figure (b): definition of risk level, defined based on number of species groups as identified in (a); the colors are generated by overlapping the layers and later enhanced for clarity and ease of distinction. In this setting, risk level ranges from 0 to 6.

4. EMERGENT SPECIES FLOW CONTROL STRATEGIES

Clustering analysis on SFN has discovered a consistent pattern of port groupings in terms of potential species exchanges. One can easily identify such regions at risk, by overlaying the ecoregions to the IRN clusters above (see Fig. 9(a)). This allows one to consider the four factors of vessel movement, ballast discharge, environmental conditions, and ecoregion in a unique but an intuitive fashion. With this knowledge in place, species exchange among ports can be efficiently controlled by management strategies that target high species exchange pathways in order to isolate ports and clusters of ports.

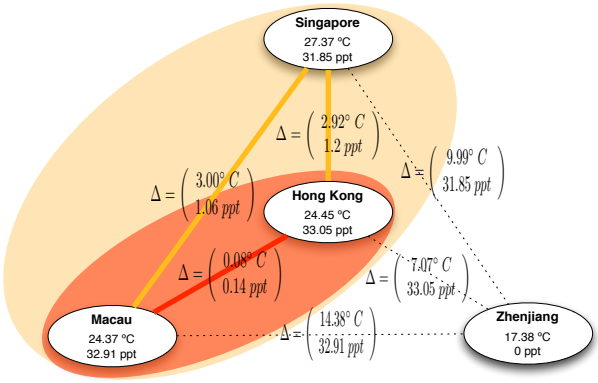


Figure 7: **Illustrating the generation of Invasion Risk Network (IRN).** The IRN is an undirected graph where nodes and edges are given by the ports visited in the GSN and invasion risk level, respectively. Shown here as examples are four ports along with annual average temperature and salinity, and pairwise salinity and temperature differences. Edges drawn in solid lines represent the risk level between ports as defined in Fig. 6; dotted-lines show zero (0) risk edges; colored-patches are used to show the overlap of species tolerance groups shared by a port-pair.

4.1 Managing inter-cluster exchanges

Consider Fig. 8 which illustrates the inter- and intra-cluster species flow among major clusters. While we observe some changes in inter-cluster connections, major clusters and species exchange pathways are virtually consistent over time. Therefore, by limiting species flow on inter-cluster connections, species exchanges among ports could be restricted to ports within clusters. This will combinatorially reduce the species introduction pathways.

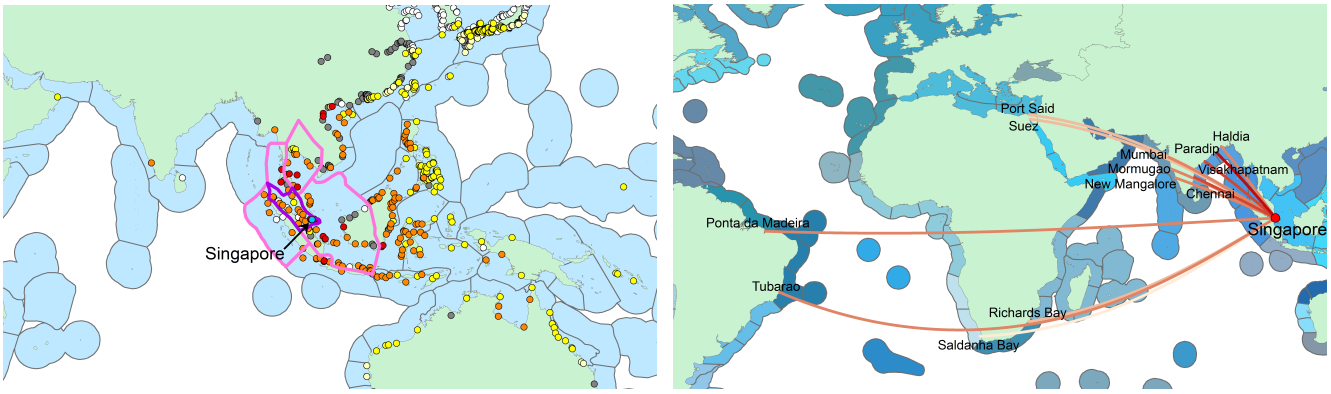
For instance, consider the `Pacific` cluster in year 2005–2006. There are 37,596 inter-cluster connections, where Table 4 tabulates the strongest connections. Singapore alone contributes to approximately 26% of total inter-cluster flow from/to `Pacific` cluster that contains 818 ports (see Fig. 9 for an illustration of invasion risk with respect to Singapore). Here, via targeted ballast management on inter-cluster con-

Table 4: **Highest inter-cluster flow for `Pacific` cluster in 2005–2006**

From Port	To Port
Singapore	Port Said
Singapore	Richards Bay
Mormugao	Singapore
Suez	Singapore
Paradip	Singapore
Visakhapatnam	Singapore
Tubarao	Singapore
Chennai	Singapore
Ponta da Madeira	Singapore

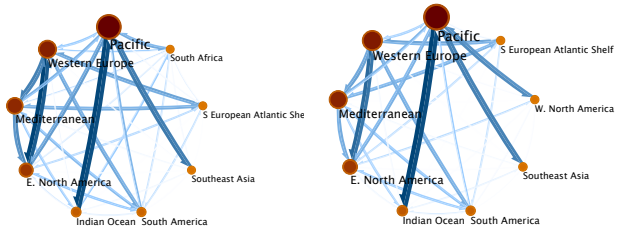
Table 5: **Major inter-cluster contributors in 2005–2006**

Cluster	Port
`Pacific`	Singapore
`Mediterranean`	Gibraltar
`W. Euro`	Rotterdam
`E. N. America`	New York
`Indian Ocean`	Mormugao
`S. America`	Tubarao
`Great Lakes`	Seven Islands
`Black Sea`	Istanbul
`W. N. America`	Long Beach



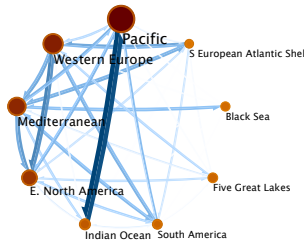
(a) Inner-Cluster Invasion Risk w.r.t. Singapore (b) Inter-Cluster Invasion Risk w.r.t. Singapore

Figure 9: **NIS invasion risk with respect to Singapore,** where the colors correspond to risk level definitions in Fig. 6.

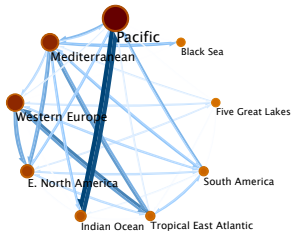


(a) 1997-1998

(b) 1999-2000



(c) 2002-2003



(d) 2005-2006

Figure 8: **Illustration of inter-cluster and intra-cluster flow.** Here, ratio of darker/lighter region explains the ratio of intra-cluster flow (i.e., flow between ports within a cluster) to inter-cluster flow (i.e., flow between ports belonging to different clusters). Therefore, in major clusters, species exchange among ports within clusters appears to be much higher compared to that of between clusters.

nections to/from Singapore and a few other "influential" ports, inter-cluster flow from/to `Pacific` cluster can be significantly reduced (see Fig. 9(b)).

Table 5 lists ports corresponding to the highest inter-cluster flow in major clusters for 2005–2006. Any practices that reduced species movements through these ports would potentially reduce a large proportion of inter-cluster species flow. Increases in species surveillance in these ports would strengthen the foundation for geographic allocation of risk management efforts. Finally, increased surveillance of ports would provide a baseline against which to measure the effectiveness of future risk reduction efforts–a baseline that is now largely absent globally (Costello et al. 2007)

4.2 Targeting hubs for species flow control

Average path length of three (3) that is observed on SFN indicates that species could be translocated between any two given ports within two (2) stopovers on average. This indicates a generally high risk of invasions in the absence of risk reduction practices. In order to understand the impact of targeted ballast management on average path length, a test scenario based on a hypothetical SFN—$\widehat{\text{SFN}}$ is derived as follows: **(i)** choose an SFN (SFN corresponding to 2005–2006 LMIU dataset was chosen for our experiment); then, **(ii)** identify 20% of all ports with the highest degree (see Table 6); and, finally **(iii)** generate $\widehat{\text{SFN}}$ by removing all edges to/from the above ports; this corresponds to ballast management with 100% efficiency, i.e., zero (0) species flow from/to these ports. *Then, the average path length increases to 6.4 indicating that it will be at least twice as difficult for species to be translocated from one port to another.* Furthermore, higher average path length also implies, **(i)** longer travel times (hence, very lower chance of survival for species during the voyage) and **(ii)** increased number of intermediate stop-overs (which is likely to dilute ballast water and expose organisms to multiple shocks).

Table 6: **Ports* with degree** > 1000 **in 2005-2006 that act as "hubs" in SFN**

Port name	Degree	Important pathways (connected ports)
Gibraltar	1882	Cape Finisterre, Tubarao
Dover Strait*	1747	Cape Finisterre, Rotterdam, Tubarao
Singapore	1569	Mormugao, Tubarao
Cape Finisterre	1387	Gibraltar, Rotterdam, Tubarao
Panama Canal*	1275	New Orleans
Tarifa	1224	Gibraltar, Cape Finisterre
Rotterdam	1126	Cape Finisterre, Dover Strait

* indicates locations in LMIU database, but do not correspond to actual ports; connected ports are listed in decreasing order of degree.

4.3 Vessel type based management strategies

The exact amount of species relocated by a vessel depends on many factors: ballast size, average duration per trip, frequently visited ports, etc. Furthermore, vessel types we observe in GSN are often chosen for specific tasks (e.g., oil transportation, vehicle transportation, etc.) and these vessels often have their respective frequent ports/routes. There-

fore, we investigate the relationship of vessel types to inter- and intra-cluster species flow in order to understand existing patterns that can be helpful in devising species management strategies (based on the 2005-2006 LMIU dataset).

(i) Frequent inter-cluster travelers: While not being the most active vessel in the GSN, `container carriers` correspond to 57,909, or equivalently 24% of all inter-cluster trips in 2005-2006. Among the most frequently seen vessel types, `bulkers`, `crude oil tankers`, `refrigerated general cargo ships` and `combined bulk and oil carriers` tend to travel inter-cluster for over 25% of the time. Furthermore, among the vessel types that do not travel frequently, some vessel types tend to travel inter-cluster in a majority of their trips (e.g., `wood-chip carriers`: 40.4%, `livestock carriers`: 34.3%, `semi-sub HL vessels`: 37.4% and `barge container carriers`: 55.7%).

(ii) Frequent intra-cluster travelers: Among the most frequently travelled vessel types, `passenger carriers` tend to stay within clusters for 97.6% of their trips, thus imposing only a very minimal risk in terms of inter-cluster species translocation. Similarly, `barge ships` also stay within the cluster for 98.1% of total trips.

4.4 Impact of environmental conditions

With proper species control on inter-cluster connections, species flow can be confined to clusters. Even though ports within a cluster have higher species exchanges among them, if the environmental conditions are significantly different, invasions are less likely to occur. On the other hand, for ports in the same IRN cluster (hence, environmental conditions are very similar), if proper species flow control is not in place, invasions will be nearly unavoidable. For instance, in the `Pacific` cluster, we observe that *Hong Kong, Qingdao* and *Kaohsiung* have higher species exchanges among them; and, following clustering analysis of IRN, we also notice that Hong Kong and Kaohsiung are in the same IRN cluster. Therefore, invasions are very likely to occur in between these two ports. On the other hand, Qingdao is in a different sub-cluster to Hong Kong, indicating these two ports have significantly different environmental conditions—invasions are less likely to happen between these two ports, even with high species exchanges.

5. CONCLUDING REMARKS

Aquatic invasions via the GSN are a result of a complex interplay of ship traffic, ballast uptake/discharge dynamics, species survival during transport and numerous environmental/biological variables. The inherent complexity of the invasive species problem has made risk assessment very difficult, and thereby has severely hampered the effectiveness of species management efforts. To that end, we have developed an approach for more effective and efficient risk assessment and management by modeling the spread of aquatic non-indigenous species through the GSN, which is the most important vector of aquatic invasions. Knowledge about the patterns of GSN, within the context of species flow and invasion risk, appropriate risk assessments can be generated to help inform management strategies and regulatory policies.

In a management context, the discovered knowledge could efficiently be used to analyze the invasion risks with respect to changing climate, policy and infrastructure. Understanding the structure of the component networks and the dynamic interactions between the different networks is crucial to the design of policies that could cost-effectively reduce invasions. The analyses outlined and performed in this paper could also be used to geographically prioritize species surveillance efforts using traditional organism sampling methods (e.g., water samples, nets) and/or newer genetic approaches [4]. Furthermore, our work illustrates the value of creative use of data mining for social good via the application to a significant societal problem.

6. ACKNOWLEDGMENTS

This work is based on research supported by the Notre Dame Office of Research via the Environmental Change Initiative (ECI), the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053, the U.S. Air Force Office of Scientific Research (AFOSR) and the Defense Advanced Research Projects Agency (DARPA) grant #FA9550-12-1-0405, NOAA CSCOR and EPA GLRI grants. We are grateful to the anonymous reviewers for their constructive comments.

7. REFERENCES

[1] R. Abell, M. L. Thieme, C. Revenga, M. Bryer, M. Kottelat, N. Bogutskaya, B. Coad, N. Mandrak, S. C. Balderas, W. Bussing, M. L. J. Stiassny, P. Skelton, G. R. Allen, P. Unmack, A. Naseka, R. Ng, N. Sindorf, J. Robertson, E. Armijo, J. V. Higgins, T. J. Heibel, E. Wikramanayake, D. Olson, H. L. Lopez, R. E. Reis, J. G. Lundberg, M. H. Sabaj Perez, and P. Petry. Freshwater ecoregions of the world: A new map of biogeographic units for freshwater biodiversity conservation. *BioScience*, 58(5):403–414, May 2008.

[2] J. I. Antonov, D. Seidov, T. P. Boyer, R. A. Locarnini, A. V. Mishonov, H. E. Garcia, O. K. Baranova, M. M. Zweng, and D. R. Johnson. World Ocean Atlas 2009, Volume S: Salinity. In S. Levitus, editor, *NOAA Atlas NESDIS*, volume 69, page 184. U.S. Government Printing Office, Washington, D.C., 2010.

[3] A.-L. Barabási, R. Albert, and H. Jeong. Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A: Statistical Mechanics and its Applications*, 281(1-4):69–77, June 2000.

[4] K. Bohmann, A. Evans, M. T. P. Gilbert, G. R. Carvalho, S. Creer, M. Knapp, D. W. Yu, and M. de Bruyn. Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology & Evolution*, 29:358–367, May 2014.

[5] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51:661–703, Apr 2009.

[6] N. N. I. S. Council. 2008-2012 national invasive species management plan, 2008.

[7] S. Devin and J.-N. Beisel. Biological and ecological characteristics of invasive species: a gammarid study. *Biological Invasions*, 9(1):13–24, 2007.

[8] J. M. Drake and D. M. Lodge. Global hot spots of biological invasions: Evaluating options for ballast-water management. *Proceedings: Biological Sciences*, 271(1539):575–580, Mar. 2004.

[9] O. Floerl, G. Rickard, G. Inglis, and H. Roulston. Predicted effects of climate change on potential sources of non-indigenous marine species. *Diversity and Distributions*, 19(3):257–267, 2013.

[10] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[11] B. Goodwin, A. McAllister, and L. Fahrig. Predicting invasiveness of plant species based on biological information. *Conservation Biology*, 13:422–426, 1999.

[12] R. Guimera and L. A. Amaral. Functional cartography of complex metabolic networks. *Nature*, 433:895–900, Feb. 2005.

[13] B. S. Halpern, S. Walbridge, K. A. Selkoe, C. V. Kappel, F. Micheli, C. D'Agrosa, J. F. Bruno, K. S. Casey, C. Ebert, H. E. Fox, R. Fujita, D. Heinemann, H. S. Lenihan, E. M. P. Madin, M. T. Perry, E. R. Selig, M. Spalding, R. Steneck, and R. Watson. A global map of human impact on marine ecosystems. *Science*, 319(5865):948–952, Feb. 2008.

[14] R. P. Keller, J. M. Drake, M. B. Drew, and D. M. Lodge. Linking environmental conditions and ship movements to estimate invasive species transport across the global shipping network. *Diversity and Distributions*, 17(1):93–102, 2011.

[15] R. P. Keller, D. M. Lodge, M. A. Lewis, and J. F. Shogren. *Bioeconomics of Invasive Species : Integrating Ecology, Economics, Policy, and Management: Integrating Ecology, Economics, Policy, and Management*. Oxford University Press, Apr. 2009.

[16] R. A. Locarnini, A. V. Mishonov, J. I. Antonov, T. P. Boyer, H. E. Garcia, O. K. Baranova, M. M. Zweng, and D. R. Johnson. World ocean atlas 2009, volume 1: Temperature. In S. Levitus, editor, *NOAA Atlas NESDIS*, volume 68, page 184. U.S. Government Printing Office, Washington, D.C., 2010.

[17] J. L. Molnar, R. L. Gamboa, C. Revenga, and M. D. Spalding. Assessing the global threat of invasive species to marine biodiversity. *Frontiers in Ecology and the Environment*, 6(9):485–492, Feb. 2008.

[18] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, June 2005.

[19] D. Pimentel, R. Zuniga, and D. Morrison. Update on the environmental and economic costs associated with alien-invasive species in the United States. *Ecological Economics*, 52(3):273–288, Feb 2005.

[20] J. Richard, S. A. Morley, M. A. S. Thorne, and L. S. Peck. Estimating long-term survival temperatures at the assemblage level in the marine environment: Towards macrophysiology. *PLoS ONE*, 7(4):e34655, Apr. 2012.

[21] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

[22] J. Rothlisberger, D. Finnoff, R. Cooke, and D. Lodge. Ship-borne nonindigenous species diminish great lakes ecosystem services. *Ecosystems*, 15(3):1–15, 2012.

[23] M. Sales-Pardo, R. Guimera, A. Moreira, and L. Amaral. Extracting the hierarchical organization of complex systems. *Proc. National Academy of Sciences of the United States of America*, 104:15224–15229, Sept. 2007.

[24] H. Seebens, M. T. Gastner, and B. Blasius. The risk of marine bioinvasion caused by global shipping. *Ecology Letters*, Apr. 2013.

[25] C. E. Shannon and W. Weaver. *A Mathematical Theory of Communication*. University of Illinois Press, Champaign, IL, USA, 1963.

[26] M. D. Spalding, H. E. Fox, G. R. Allen, N. Davidson, Z. A. F. na, M. Finlayson, B. S. Halpern, K. D. Martin, E. Mcmanus, J. Molnar, C. A. Recchia, and J. Robertson. Marine ecoregions of the world: A bioregionalization of coastal and shelf areas. *BioScience*, 57(7):573–583, July 2007.

[27] E. Tufte. *Beautiful Evidence*. Graphics Press, 2006.

[28] M. Wonham, J. Byers, E. D. Grosholz, and B. Leung. Modeling the relationship between propagule pressure and invasion risk to inform policy and management. *Ecological Applications*, Mar. 2013.