NAVSEA
WARFARE CENTERS
DAHLGREN

Ship Itinerary Predictability

Patricia H. Carter, Ph. D.
Computational Mathematics and Statistics Branch (Q21)
Naval Surface Warfare Center Dahlgren

QMDNS 2007
George Mason University

Outline

- Introduction and data description
- Measures for periodic (recurrent) behavior
- Maximum periodic gain predictor

Work supported by NSWC's ILIR and DTIP programs.

Outline

- Introduction and data description
- Measures for periodic (recurrent) behavior
- Maximum periodic gain predictor

Maritime Domain Awareness

Maritime Domain Awareness is "the effective knowledge of all activities associated with the global maritime environment that could impact the security, safety, economy, or environment of the United States".[1]

" There is also a need for more global tracking … to better **identify and analyze vessel behavior based on historical trends and characterization of normal shipping patterns/routes**. This track history will facilitate a more comprehensive risk evaluation of Vessels of Interest … that depart from known habits or expected behaviors …" [2]

1. Hearings of Committee on Transportation and Infrastructure, U. S House of Representatives, October 6, 2004 www.house.gov/transportation/cgmt/10-06-04/10-06-04memo.html
2. Mr. Jeffrey High, Director of the Coast Guard's Maritime Domain Awareness Program Integration Office

The Goal and its Mathematical Context

Goal: Predict the next port a ship will visit from the sequence of ports it has visited previously.

- Symbol statistics for time series analysis
- Sequence analysis in bioinformatics
- Symbolic signal processing

- Markov models
- Probability estimation of shipping routes

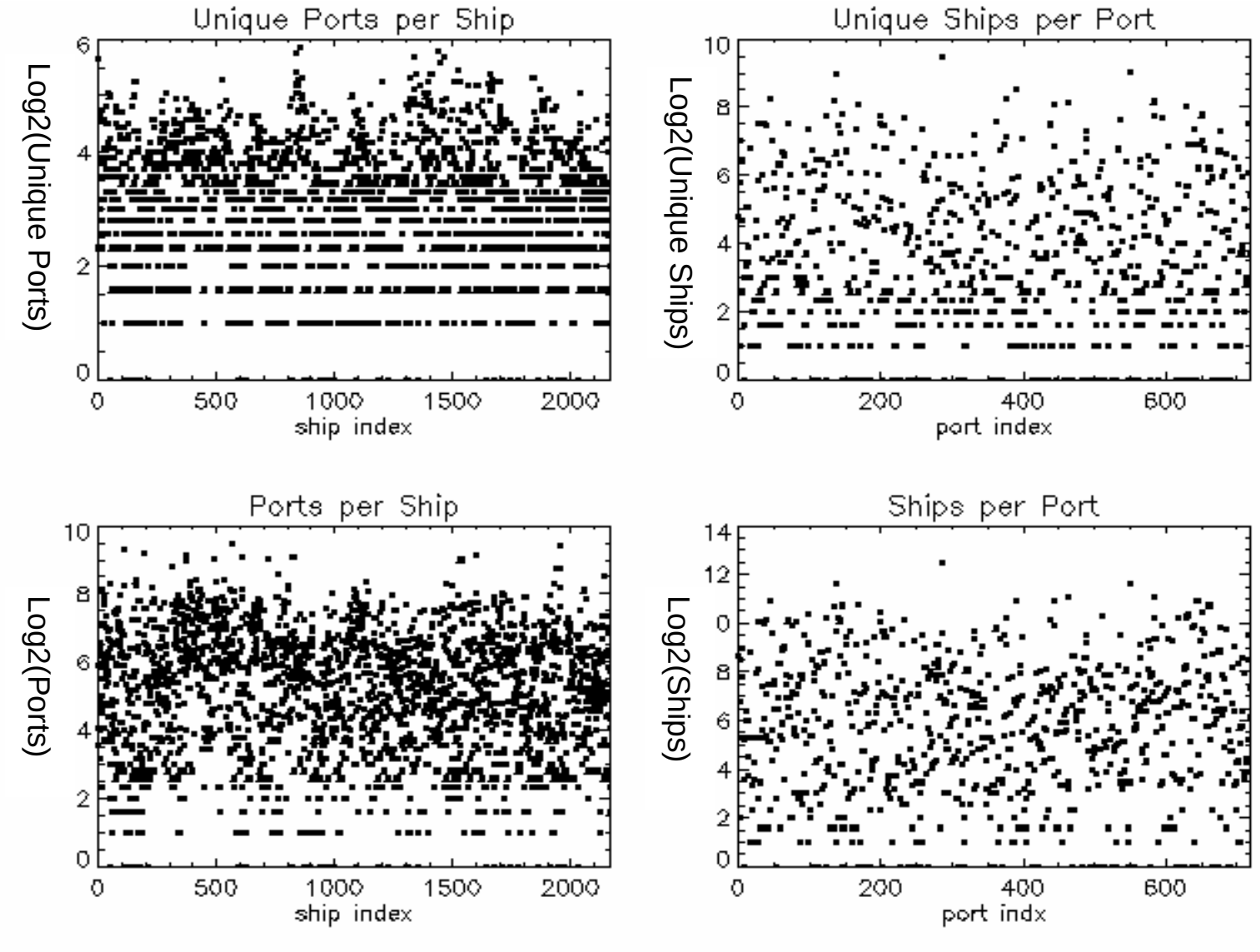
- Graph-based models (Dave Marchette, Q21)

Data

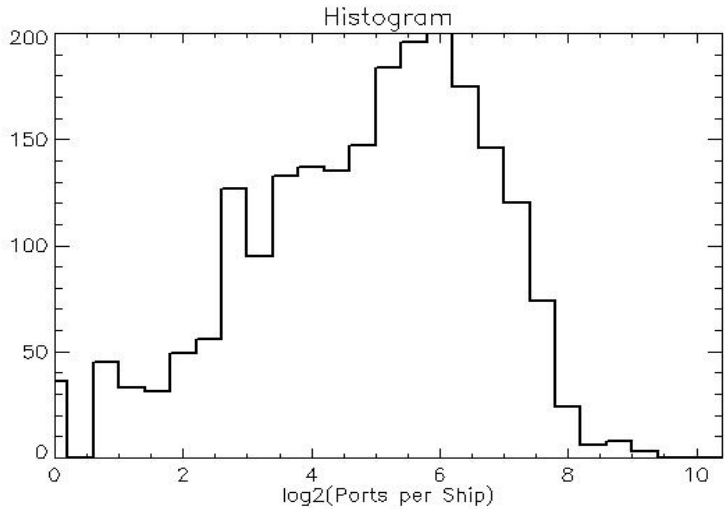
- Commercial ship schedules covering
  - 400 days (maximum)
  - 2000 commercial ships
  - 700 ports
  - total of 137,000 records
    "292305","TX174W","CMA CGM HUGO","2005-10-15 00:00:00.000",2,1,
    "Ship","OAKLAND","USOAK"
  - Abstracted to [ship, port] (in time order) for this study

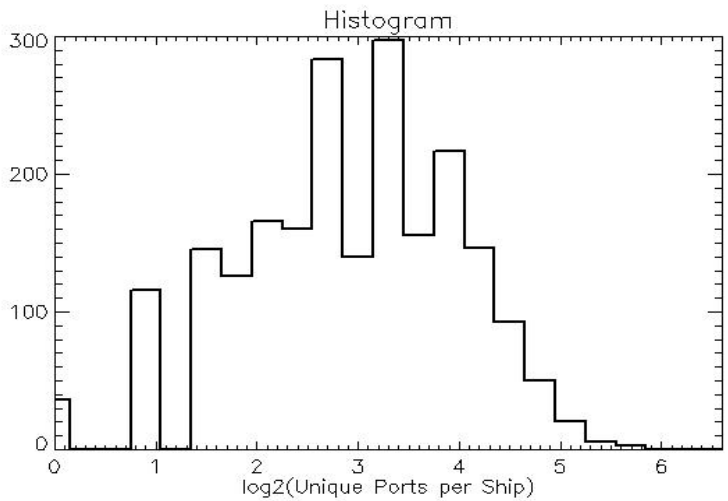
- A small % of global ship traffic
- Exhibits the problems of real world data (except scale)
- Purchased from Journal of Commerce
- Scheduled ports not sightings

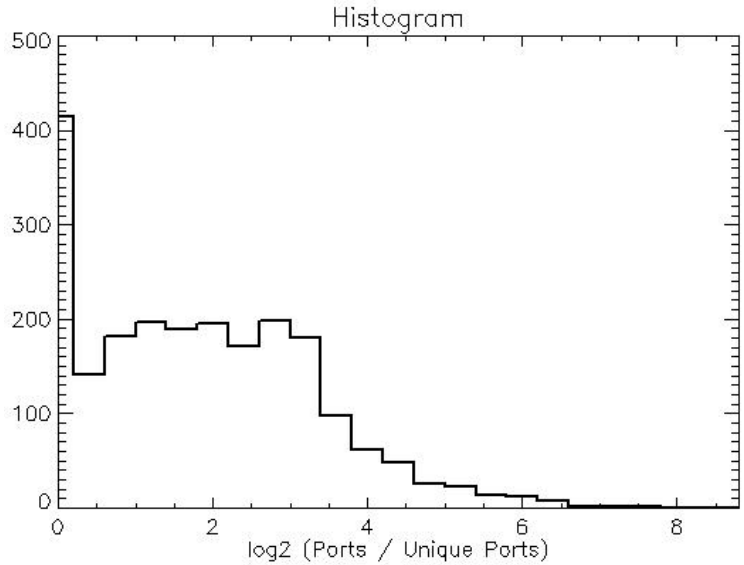
Ports per Ship and Ships Per Port



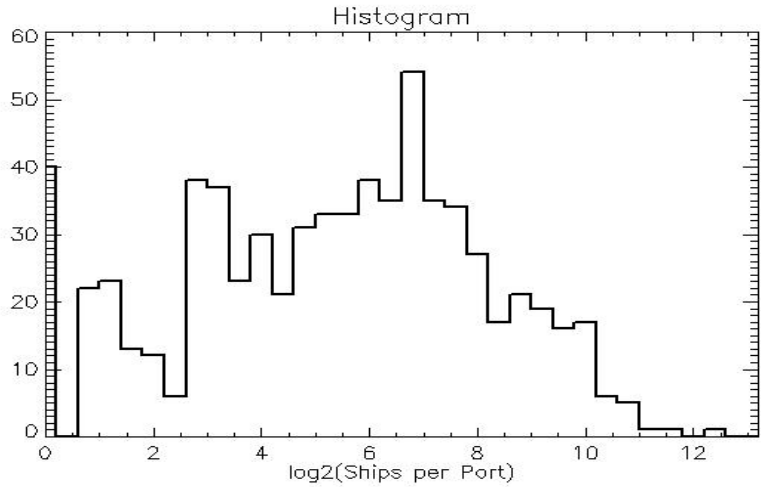
Statistics for Ports per Ship

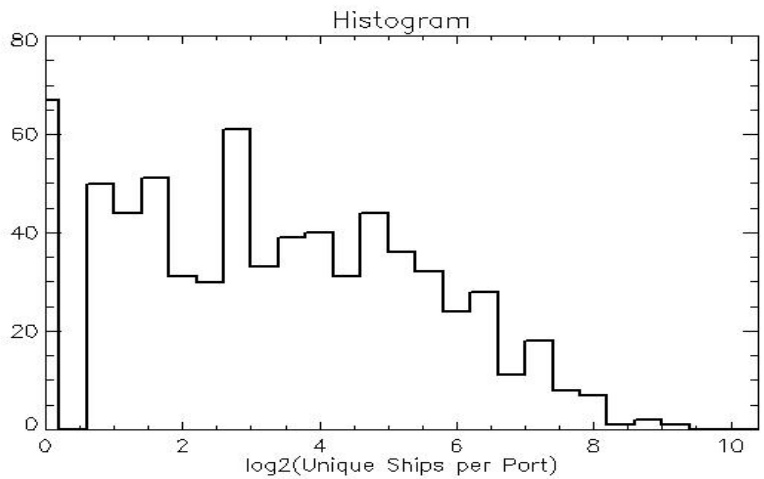


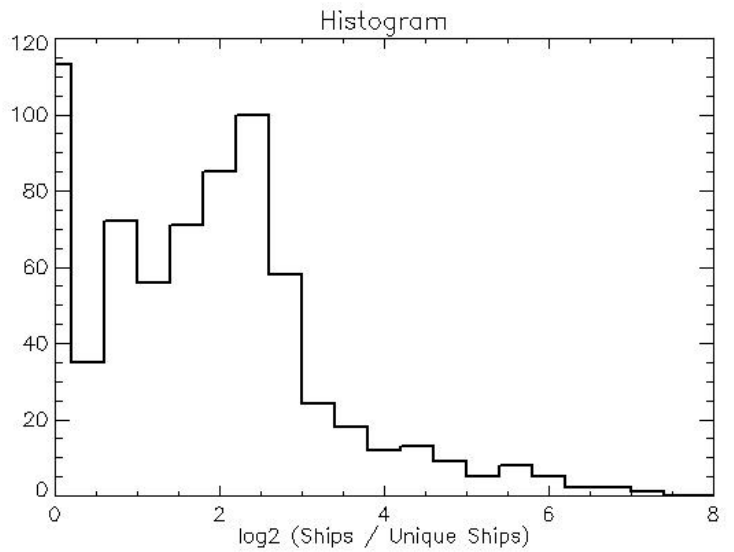




Statistics for Ships per Port







Container Ship *Scheduled* Itinerary (one cycle)
Port Sequence Represented by Integer Sequence

Ports	ID number
New York ,New York	16
Norfolk, Virginia	1
Savannah, Georgia	2
Valencia, Spain	3
Genoa, Italy	4
Gioia Tauro, Italy	5
Jeddah, Saudi Arabia	6
Khor Fakkan, UAE	7
Port Qasim, Pakistan	8
Mundra, India	9
Nava Sheva, India	10
Jebel Ali, UAE	11
Jeddah, Saudi Arabia	6
Alexandria, Egypt	12
Gioia Tauro, Italy	5
La Spezia, Italy	13
Fos sur Mer, France	14
Barcelona, Spain	15
Valencia, Spain	3
New York, New York	16

Niederelbe Schiffahrtsgesellschaft Buxtehude

- mv Ibn Sina and the mv London Senator
- container ships
- cycle 77days

Höegh Trotter

Integer Sequence Representing Port Itinerary

Highlighting: 8 10 9 7 11

6 5 8 10 9 7 11 2 6 5 8 10 9 7 11 2 5 6 8 10 9 7 11 2 6 5 4 8 1
10 9 7 11 2 6 5 8 10 9 7 11 3 2 6 6 5 5 8 8 10 10 9 9 7 11 3 2
5 6 8 10 9 7 11 3 2 6 5 8 5 8 10 9 7 11 3 2 6 5 8 10 9 7 11 3 2

Highlighting: 8 10 9 7 11 and Highlighting: 2 6 5

6 5 8 10 9 7 11 2 6 5 8 10 9 7 11 2 5 6 8 10 9 7 11 2 6 5 4 8 1
10 9 7 11 2 6 5 8 10 9 7 11 3 2 6 6 5 5 8 8 10 10 9 9 7 11 3 2
5 6 8 10 9 7 11 3 2 6 5 8 5 8 10 9 7 11 3 2 6 5 8 10 9 7 11 3 2

Highlighting: 8 10 9 7 11 2 6 5

6 5 8 10 9 7 11 2 6 5 8 10 9 7 11 2 5 6 8 10 9 7 11 2 6 5 4 8 1
10 9 7 11 2 6 5 8 10 9 7 11 3 2 6 6 5 5 8 8 10 10 9 9 7 11 3 2
5 6 8 10 9 7 11 3 2 6 5 8 5 8 10 9 7 11 3 2 6 5 8 10 9 7 11 3 2

Outline

- Introduction and data description
- Measures for periodic (recurrent) behavior
- Maximum periodic gain predictor

Recurrent Behavior Definitions

- Sequence similarity
- Subsequence(s) Indicator
- Periodicity 1: fraction of sequence which is an exactly periodic subsequence
- Periodicity 2: fraction of sequence covered by repetitions of one subsequence
- Periodicity 3: fraction of sequence covered by repetitions of subsequences from a collection of subsequences

Sequence Similarity

The length of the sequence A is $L(A) = L(a_1, a_2, ..., a_n) = n$

For A and B sequences of same length their similarity is the fraction of places they agree:

$$\sigma(A, B) = \frac{\left|\{i \ : \ a_i = b_i\}\right|}{L(A)}$$

Example: $A = (1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5)$

$B = (1, 2, 4, 4, 5, 6, 1, 2, 3, 4, 7)$

$$\sigma(A, B) = \frac{9}{11}$$

Subsequence Indicator

Suppose B is a subsequence of A, the subsequence indicator is a sequence A_B

$$A_B(k) = \begin{cases} a_k & \text{if } \exists j : (a_{j,} a_{j+1}, \ldots, a_{j+q-1}) = B \text{ with } j \le k \le j+q-1 \\ 0 & \textit{otherwise} \end{cases}$$

If B is a subsequence ($B \propto A$) of A the number of times B occurs in A is

$$N_B = \frac{\sigma(A, A_B)}{L(B)}$$

Subsequence Indicator Examples

Example 1: $A = (1,2,3,4,1,2,3,6,5,1,2,3,7)$

$B = (1,2,3)$ $A_{(1,2,3)} = (1,2,3,0,1,2,3,0,0,1,2,3,0)$ $N_B = \frac{9}{3}$

$B = (6,5)$ $A_{(6)} = (0,0,0,0,0,0,0,6,5,0,0,0,0)$ $N_B = \frac{2}{2}$

Example 2: $A = (1,2,1,2,1,3,3,1,2,1,2,1)$

$B = (1,2,1)$ $A_{(1,2,1)} = (1,2,1,2,1,0,0,1,2,1,2,1)$ $N_B = \frac{10}{3}$

Subsequences Indicator

If **B** is a collection of subsequences of A

$$A_{\mathbf{B}}(k) = \begin{cases} a_k & \text{if } \exists B \in \mathbf{B} \text{ and } \exists j \ : (a_{j,} a_{j+1}, ..., a_{j+q-1}) = B \text{ with } j \le k \le j+q-1 \\ 0 & \textit{otherwise} \end{cases}$$

It follows that $A_{\mathbf{B}}(k) = \max\{A_B(k) : B \in \mathbf{B}\}$

Example: $A = (1,2,3,7,4,5,1,2,3,4,5,6)$

$\mathbf{B} = \{(1,2,3),(4,5)\}$

$A_{\mathbf{B}} = (1,2,3,0,4,5,1,2,3,4,5,0)$

Periodicity via Maximum Periodic Subsequence

If A is the concatenation of n copies of the length m sequence B, then A is ***exactly*** (n,m)-periodic.

Example: $A = (1,2,3,1,2,3,1,2,3,1,2,3)$

$= (1,2,3,1,2,3) + (1,2,3,1,2,3)$ (2,4)-periodic

$= (1,2,3) + (1,2,3) + (1,2,3) + (1,2,3)$ (4,3)-periodic

A measure of A's closeness to being exactly periodic is

$$\rho_e(A) = \frac{\max\{nm \ : \ C \propto A, \ C \ is \ (n,m) - periodic, n \geq 2, m \geq 2\}}{\mathrm{L}(A)}$$

Example: $A = (1,2,3,4,5,1,2,3,4,5,6,6)$

$A_C = (1,2,3,4,5,1,2,3,4,5,0,0)$ $\rho_e(A) = \frac{10}{12}$.

Periodicity Relative to a Subsequence

A measure of the periodic content of A due to the one subsequence B, where N_B is at least 2 is

$$\rho(B) = \frac{\sigma(A, A_B)}{L(B)}$$

A single subsequence measure of periodic content of A

$$\rho_s = \max\left\{\rho(B) : B \propto A, L(B) \geq 2, N_B \geq 2\right\})$$

Example: $A = (1,2,3,4,7,4,5,1,2,3,4,5,6)$

$B = (1,2,3)$ $\quad \rho(B) = \frac{6}{13}$

$B = (4,5)$ $\quad \rho(B) = \frac{4}{13}$

$B = (1,2,3,4)$ $\quad \rho_S = \rho(B) = \frac{8}{13}$

Periodicity Relative to a Set of Subsequences

Define the periodic content of A due to the collection of subsequences **B**

$$\rho_{\mathbf{B}} = \frac{\sigma(A, A_{\mathbf{B}})}{L(A)}$$

Example: $A = (1,2,3,7,4,5,1,2,3,4,5,6)$

$$\mathbf{B} = \{(1,2,3),(4,5)\}$$

$$A_{\mathbf{B}} = (1,2,3,0,4,5,1,2,3,4,5,0)$$

$$\rho_{\mathbf{B}} = \frac{10}{12}$$

Outline

- Introduction and data description
- Measures for periodic (recurrent) behavior
- Maximum periodic gain predictor

Insight for Prediction

Consider the sequences A of length n and A+(w). Find a predictor w for w by maximizing a measure of periodic (3) content of A+(w).

Example: $A = (1,2,3,1,2,3,1,2,3,1)$

$$A = (1,2,3,1,2,3,1,2,3,1, w)$$

w should be one of 1,2 or 3

w = 2 seems to follow the pattern, given only this information

Maximum Periodic Gain (MPG) Predictor

Given the sequence A of length n and A+(w). Find a predictor $\underline{w}$ for w by maximizing the periodic content of A+(w) relative to sets of subsequences.

The only subsequences of A+(w) which can ***increase*** periodic content from that of A, are, with w a symbol in A, the suffixes of A+(w),

$$B_w = \{(a_n, w), (a_{n-1}, a_n, w), ..., (a_{n-q+1}, ..., a_n, w)\}$$

Subsequences must have length at least 2 and occur at least twice

$$\tilde{B}_\omega = \left\{B \in \mathbf{B}_\omega \ : \ N_B \geq 2, \ 2 \leq L(B)\right\}$$

The measure of the periodic content in A+(w) due to $\tilde{B}_w$ is

$$\rho\left(\tilde{B}_w\right) = \frac{\sigma\left(A+(w), (A+(w))_{\tilde{B}_w}\right)}{L(A+(w))}$$

The predictor $\underline{w}$ maximizes the periodic content of A+(w) is

$$\underline{w} = \arg\max_{w \in A}(\rho(\tilde{B}_w))$$

MPG Predictor Example Computation

Example:

$$A+(w) = (3,4,3,5,1,2,3,6,7,1,2,3,4,1,2,3, w)$$

Details of finding the w that maximizes the periodic content of A+(w) are in the next five views.

Maximum Periodic Gain Predictor Step 1
the Sets B_w

$$B_w = \{(a_n, w), (a_{n-1}, a_n, w), \ldots, (a_{n-q+1}, \ldots, a_n, w)\}$$

$$A+(w) = (3,4,3,5,1,2,3,6,7,1,2,3,4,1,2,3,w)$$

Each B_w consists of a set of suffixes of A+(w)

$$B_w = \left\{ \begin{array}{l} (3,w), \\ (2,3,w), \\ (1,2,3,w), \\ (4,1,2,3,w), \\ (3,4,1,2,3,w), \\ (2,3,4,1,2,3,w), \\ (1,2,3,4,1,2,3,w) \end{array} \right\}$$

$$w \in \{1,2,3,4,5,6,7\}$$

Observation about Subsequences

If $(b_0,b_1, b_2,\ldots,b_k)$ occurs at least two times in A then $(b_1, b_2,\ldots,b_k)$ occurs at least two times. The contrapositive: if $(b_1, b_2,\ldots,b_k)$ does not occur at least two times then $(b_0,b_1, b_2,\ldots,b_k)$ cannot occur two times.

Maximum Periodic Gain Predictor Step 2
Reduce Sets of Subsequences

$$B_w = \left\{ \begin{array}{l} (3, w), \\ (2,3, w), \\ (1,2,3, w), \\ (4,1,2,3, w), \\ (3,4,1,2,3, w), \\ (2,3,4,1,2,3, w), \\ (1,2,3,4,1,2,3, w) \end{array} \right\}$$

w=1 $A+(1) = (3,4,3,5,1,2,3,6,7,1,2,3,4,1,2,3,1)$

(3,1) only occurs once in A+(1) so it isn't in $\tilde{B}_1$

Since (3,1) occurs only once any subsequence containing (3,1) can occur at most once, hence $\tilde{B}_1 = \phi$

Similarly for w = 2,3 and 7

$A+(2) = (3,4,3,5,1,2,3,6,7,1,2,3,4,1,2,3,2)$

(3,2) only occurs once, so the others in B_2 occur at most once, hence $\tilde{B}_2 = \phi$

$A+(3) = (3,4,3,5,1,2,3,6,7,1,2,3,4,1,2,3,3)$

(3,3) only occurs once, so the others in B_3 occur at most once, hence $\tilde{B}_3 = \phi$

$A+(7) = (3,4,3,5,1,2,3,6,7,1,2,3,4,1,2,3,7)$

(3,7) only occurs once, so the others in B_7 occur at most once, hence $\tilde{B}_7 = \phi$

Maximum Periodic Gain Predictor Step 2
Reduce Sets of Subsequences

$$B_w = \left\{ \begin{array}{l} (3, w), \\ (2,3, w), \\ (1,2,3, w), \\ (4,1,2,3, w), \\ (3,4,1,2,3, w), \\ (2,3,4,1,2,3, w), \\ (1,2,3,4,1,2,3, w) \end{array} \right\}$$

$A+(4) =$

$(3,4,3,5,7,2,3,6,7,1,2,3,4,1,2,3,4)$

$A+(5) =$

$(3,4,3,5,7,2,3,6,7,1,2,3,4,1,2,3,5)$

$A+(6) =$

$(3,4,3,5,7,2,3,6,7,1,2,3,4,1,2,3,6)$

In A+(4) (3,4) occurs three times, (2,3,4) occurs twice, (1,2,3,4) occurs twice, but (4,1,2,3,4) occurs only once.

In A+(5) (3,5) occurs twice but (2,3,5) occurs only once

In A+(6) (3,6) occurs twice, (2,3,6) occurs twice, but (1,2,3,6) occurs only once.

$$\tilde{B}_4 = \left\{ \begin{array}{l} (3,4), \\ (2,3,4), \\ (1,2,3,4) \end{array} \right\}$$

$$\tilde{B}_5 = \{(3,5)\}$$

$$\tilde{B}_6 = \left\{ \begin{array}{l} (3,6), \\ (2,3,6) \end{array} \right\}$$

Maximum Periodic Gain Predictor Step 3
Calculate the Content Measures

$$\rho\left(\tilde{B}_w\right) = \frac{\sigma\left(A+(w),(A+(w))_{\tilde{B}_w}\right)}{L(A+(w))} \qquad \text{and} \qquad \underline{w} = \arg\max_{w\in A}(\rho(\tilde{B}_w))$$

$$A+(w) = (3,4,3,5,1,2,3,6,7,1,2,3,4,1,2,3,w)$$

$\tilde{B}_1 = \phi$, $\tilde{B}_2 = \phi$, $\tilde{B}_3 = \phi$, $\tilde{B}_7 = \phi$ imply $\rho(\tilde{B}_1) = 0$, $\rho(\tilde{B}_2) = 0$, $\rho(\tilde{B}_3) = 0$, $\rho(\tilde{B}_7) = 0$

$$\tilde{B}_4 = \begin{Bmatrix}(3,4),\\(2,3,4),\\(1,2,3,4)\end{Bmatrix} \quad \left(A+(4)\right)_{\tilde{B}_4} = (3,4,0,0,0,0,0,0,0,1,2,3,4,1,2,3,4) \quad \rho(\tilde{B}_4) = \frac{10}{17}$$

$$\tilde{B}_5 = \{(3,5)\} \quad \left(A+(5)\right)_{\tilde{B}_5} = (0,0,3,5,0,0,0,0,0,0,0,0,0,0,0,3,5) \quad \rho(\tilde{B}_5) = \frac{4}{17}$$

$$\tilde{B}_6 = \begin{Bmatrix}(3,6),\\(2,3,6)\end{Bmatrix} \quad \left(A+(6)\right)_{\tilde{B}_6} = (0,0,0,0,0,2,3,6,0,0,0,0,0,0,2,3,6) \quad \rho(\tilde{B}_6) = \frac{6}{17}$$

$$\underline{w} = \arg\max_{w\in A}(\rho(\tilde{B}_w)) = 4$$

Periodicity Study

How much periodicity did A have relative to the set of subsequences ending in the predicted w, i.e.,

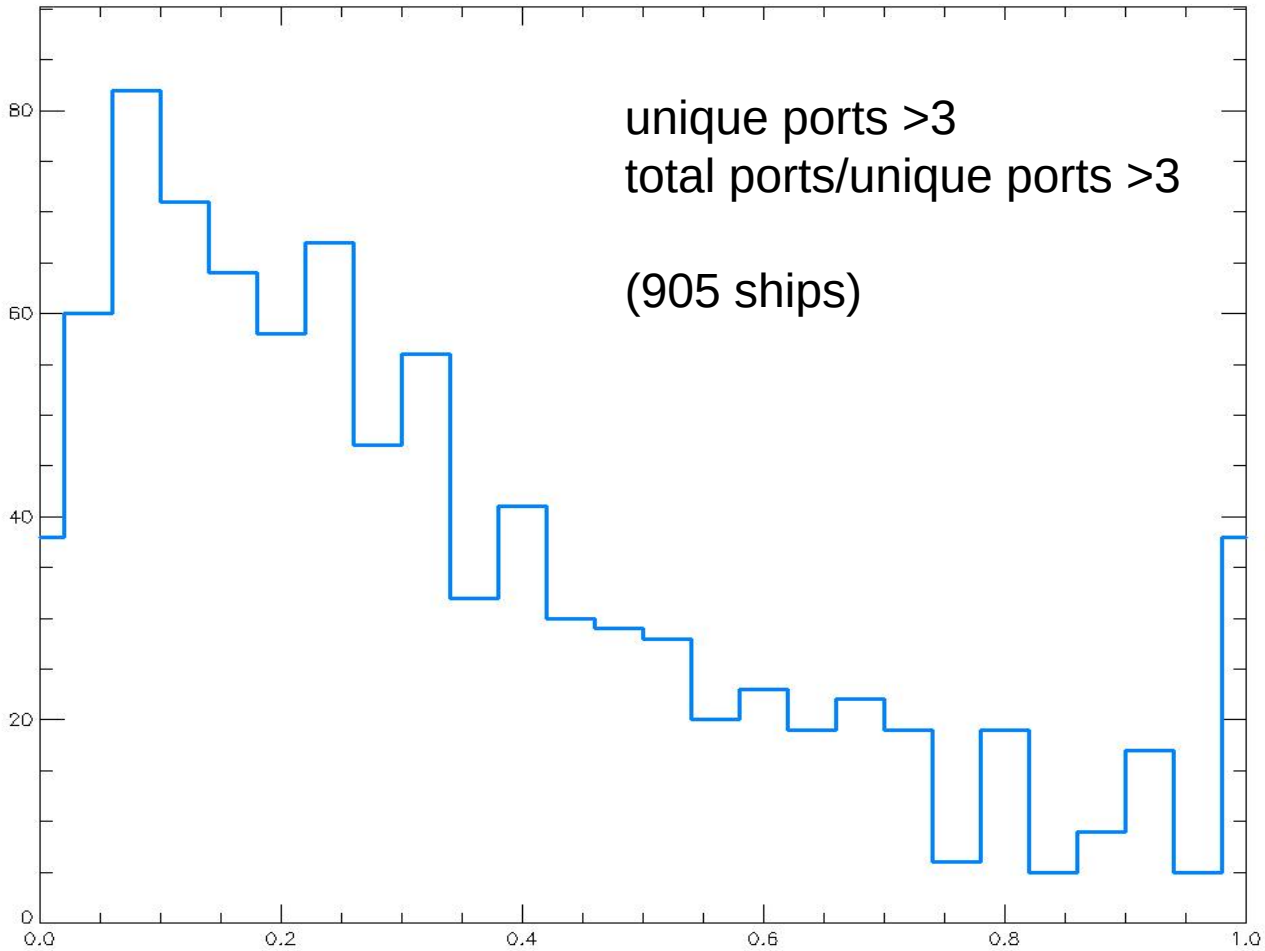
What was the distribution of the numbers

$$\max(\rho(\tilde{B}_w))$$

over the sequences A ?

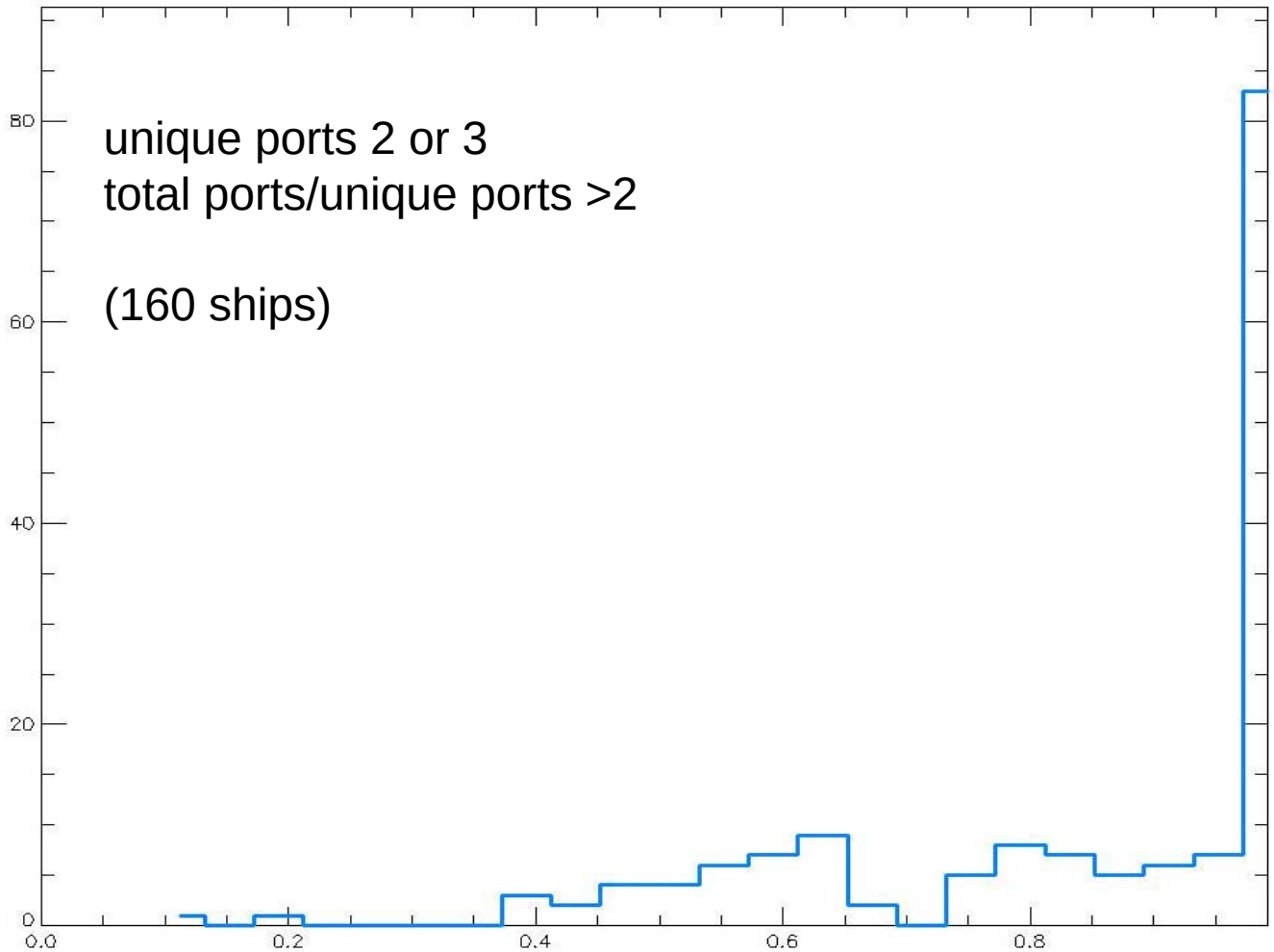
Periodicity Study 1

Histogram of $\max_{w \in A}(\rho(\tilde{B}_w))$



Periodicity Study 2

Histogram of $\max_{w \in A}(\rho(\tilde{B}_w))$



Performance of MPG Predictor

Compare the performance of

- maximum periodic gain predictor (MPG)
- periodic gain predictors based on a single subsequence (MPG-k)
- predictor based on Markov process models (Markov-k)

The predictors MPG-k and Markov-k are defined in the subsequent two views.

MPG-k

With the same nomenclature as MPG, restrict the set of suffixes to the one with length k.

$$\mathrm{B}_w = \left\{(a_{n-k+2},\ldots,a_n,w)\right\}$$

If the one subsequence does not occur at least twice, then $\tilde{B}_w = \phi$ otherwise $\tilde{B}_w$ has just one element.

$$\tilde{\mathrm{B}}_w = \left\{B \in \mathbf{B}_w \ : \ N_B \geq 2, \ L(B) = k\right\}$$

$$\rho\left(\tilde{B}_w\right) = \frac{\sigma\left(A+(w),(A+(w))_{\tilde{B}_w}\right)}{L(A+(w))}$$

$$\underline{w} = \arg\max_{w \in A} \left(\rho(\tilde{B}_w)\right)$$

Markov-k

Markov-k refers to modeling the sequence A+(w) as a Markov process with the current value dependent on (k-1)-time steps, and choosing the predictor by maximizing the transition probability.

Approximate probabilities are computed from A+(w). Write

$$A = (a_1, ..., a_n) \quad \text{and} \quad B = (a_{n-k+2}, ..., a_n, w)$$

$$\Pr(w \mid B) = \frac{\left| j : B + (w) = (a_j, ..., a_{j+k-1}) \right|}{\sum_z \left| j : B + (z) = (a_j, ..., a_{j+k-1}) \right|}$$

$$\underline{w} = \arg\max_{w \in A} (\Pr(w \mid B)$$

Note that Markov-k and MPG-k are very closely related but not identical.

Achievement of Maxima

Of 1085 ships with unique ports > 3 and (total ports/unique ports) > 3 the number of values for w that achieved the maximum defining the predictor

	Single w gives maximum – only one potential solution	Multiple w achieve maximum -- several potential solutions	No non-zero values -- no potential solutions
MPG	1018	34	33
MPG-2	956	96	33
MPG-3	847	85	153
MPG-4	716	70	299
MPG-5	625	36	424
Markov 2	959	97	29
Markov 3	846	87	152
Markov 4	720	70	295
Markov 5	637	37	411

Prediction Performance

Of 1085 ships with unique ports > 3 and (total ports/unique ports) > 3

	Number Correct	Fraction Correct	Number with Maximum of Test	Fraction with Maximum of Test
MPG	761	.701	772	.712
MPG-2	684	.630	718	.662
MPG-3	679	.626	710	.654
MPG-4	601	.554	627	.578
MPG-5	525	.484	541	.499
Markov 2	682	.629	717	.661
Markov 3	678	.625	710	.654
Markov 4	603	.556	629	.580
Markov 5	536	.494	554	.511

Significance

Use the McNemar Test to compare the performance of the MPG predictor and the best of the Markov predictors:

	MPG C	MPG W
Mar C	656	26
Mar W	105	298

MPG correct and Markov correct - 656
MGP correct and Markov wrong - 105
MGP wrong and Markov correct - 26
MGP wrong and Markov wrong - 298

McNemar's chi-squared = 47.6412, df = 1, p-value = 5.118e-12 (computed in R)

Hence we conclude that the MPG predictor was significantly better than the Markov predictor.

Summary

- Scheduled ship itineraries exhibit different amounts of periodicity / recurrence

- Three measures of periodic / recurrent behavior of a sequence
  - maximum fraction that is exactly periodic
  - maximum fraction that is replications of one subsequence
  - maximum fraction that is replications of sequences from a set of subsequences

- The maximum periodic gain (MPG) predictor is a predictor based on maximizing the periodic / recurrent content

- The performance of the MPG predictor was significantly better than that of a Markov predictor

All computations except for the McNemar test were done in Research Systems' Interactive Data Language (IDL).

Some Backup Views

q-gram Profiles

$G_q(A, \bullet)$ is a map from the set of subsequences of length q to their number of occurrences

$$A = (1, 2, 3, 4, 1, 2, 3, 6, 6, 1, 2, 3, 7).$$

$G_1(A, \cdot)$

Sequence	Count
1	3
2	3
3	3
4	1
6	2
7	1

$G_2(A, \cdot)$

Sequence	Count
1,2	3
2,3	3
3,4	1
4,1	1
3,6	1
6,6	1
6,1	1
3,7	1

$G_3(A, \cdot)$

Sequence	Count
1,2,3	3
2,3,4	1
3,4,1	1
4,1,2	1
2,3,6	1
3,6,6	1
6,6,1	1
6,1,2	1
1,2,3	1
2,3,7	1

$G_4(A, \cdot)$

Sequence	Count
1,2,3,4	1
2,3,4,1	1
3,4,1,2	1
4,1,2,3	1
3,6,6,1	1
6,6,1,2	1
6,1,2,3	1
1,2,3,7	1

The all-gram profile G(A, .) is a map from the set of all subsequences to their number of occurrences.

The q-gram similarity

This is a measure of similarity which does not account for multiplicities. For the sets $\mathbf{S}_q(A)$ and $\mathcal{S}_q(B)$ of q-grams of A and of B

$$f_q(A,B) = \frac{|\mathcal{S}_q(A) \cap \mathcal{S}_q(B)|}{|\mathcal{S}_q(A) \cup \mathcal{S}_q(B)|}.$$