Analysis of Port Activity Via Bipartite Graphs

David J. Marchette
dmarchette@gmail.com
Patricia H. Carter

Naval Surface Warfare Center
Code Q21

Hawaii, January 18, 2007

Outline

Introduction

Scan Statistics on Graphs

Bipartite Port Graphs

Cumulative Bipartite Graph

Discussion

Outline

Outline

NAVSEA WARFARE CENTERS DAHLGREN

Outline

DAHLGREN

Outline

Graph Definitions

- A graph is a pair (V, E) where V is a set (vertices) and E is a collection of unordered pairs of vertices (edges).
- We can consider directed graphs (V, A) where A (arcs or arrows) are ordered pairs.
- The order of the graph is $|V|$ and the size of the graph is $|E|$ (or $|A|$ in the case of directed graphs (digraphs)).
- A bipartite graph is a graph in which the vertex set admits a disjoint partition into two subsets V_1, V_2 such that all edges in E are between an element in V_1 and an element in V_2.

Graph Definitions Continued

- The (closed) neighborhood of a vertex v

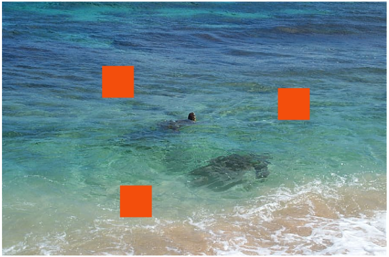
$$N(v) = N_1(v) = \{w \in V | vw \in E\}$$

is the set of vertices reachable in 1 step (we will assume the closed neighborhood — include the vertex v).

- The k-neighborhood $N_k(v)$ is the set of vertices reachable in at most k steps.

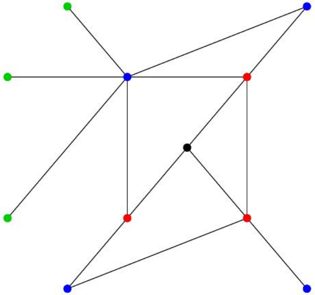
Scan Statistics on Graphs

Analogous to scan statistics on images:



- A scan region (of scale k) is a k-neighborhood of the graph.
- A locality statistic is a statistic on the scan region (eg: size of the induced subgraph).
- A scan statistic is the maximum locality statistic.

Example Scan Regions



- Scan 0: black, locality = 3 (degree).
- Scan 1: black + red, locality = 4.
- Scan 2: black + red + blue, locality = 11.
- Scan 3: black + red + blue + green, locality = 14.

Scan Statistic Details

- Let G be a graph.
- Scan region: induced subgraph of k-neighborhood: $\Omega(N_k(v;G))$.
- Locality statistic: $\Psi_k(v) = \text{size}(\Omega(N_k(v;G)))$.
- Scan statistic: $M_k = \max_v(\text{size}(\Omega(N_k(v;G))))$.

Incorporating Time

- Let $\{G_t\}$ $t = 1, \ldots$ be a time series of graphs.
- Scan region: induced subgraph of k-neighborhood: $\Omega(N_k(v; G_t))$.
- Locality statistic: $\Psi_{k,t}(v) = \text{size}(\Omega(N_k(v; G_t)))$.
- Scan statistic: $M_{k,t} = \max_v(\text{size}(\Omega(N_k(v; G_t))))$.
- Let τ be an integer (temporal window).

WARFARE CENTERS
DAHLGREN

Vertex Standardization

- We want to standardize the vertices ("loud" vertices don't drown out "quiet" ones).
- Vertex standardized locality statistic:

$$\widetilde{\Psi}_{k,t}(v) = \frac{\Psi_{k,t}(v) - \widehat{\mu}_{k,t,\tau}(v)}{\max(\widehat{\sigma}_{k,t,\tau}(v), 1)}$$

  where $\widehat{\mu}_{k,t,\tau}(v) = \frac{1}{\tau} \sum_{s=t-\tau}^{t=1} \Psi_{k,s}(v)$ and

  $\widehat{\sigma}^2_{k,t,\tau}(v) = \frac{1}{\tau-1} \sum_{s=t-\tau}^{t=1} (\Psi_{k,s}(v) - \widehat{\mu}_{k,t,\tau}(v))^2$.
- $\widetilde{M}_{k,t} = \max_v \widetilde{\Psi}_{k,t}(v)$.

NAVSEA WARFARE CENTERS DAHLGREN

Normalizing the Scan Statistic

- If we want to detect anomalies, we need to detrend.
- To do this we use the normalized version of $\widetilde{M}_{k,t}$:

$$S_{k,t} = \frac{\widetilde{M}_{k,t} - \widetilde{\mu}_{k,t,l}}{\max(\widetilde{\sigma}_{k,t,l}, 1)}$$

where $\widetilde{\mu}_{k,t,l}$ and $\widetilde{\sigma}_{k,t,l}$ are the running mean and standard deviation of $\widetilde{M}_{k,t}$ based on the most recent l time steps.

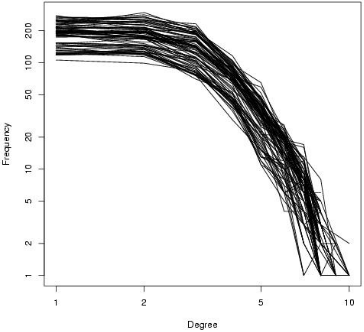
Bipartite Port Graphs

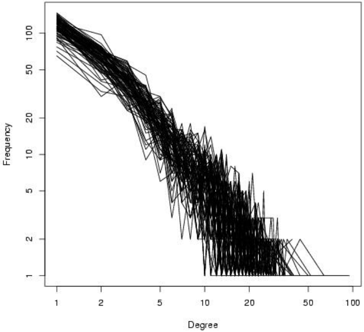
- The data consist of Vessel/Port interactions in Time; each record consists of:
  - Date Port Vessel
- These are converted to a series of bipartite graphs where:
  - Each graph corresponds to one week (non-overlapping).
  - V_1 is the set of vessels and V_2 is the set of ports.
  - There is an edge between vessel v and port p if the vessel was in the port during the corresponding week.

NAVSEA WARFARE CENTERS DAHLGREN

Degree Distributions

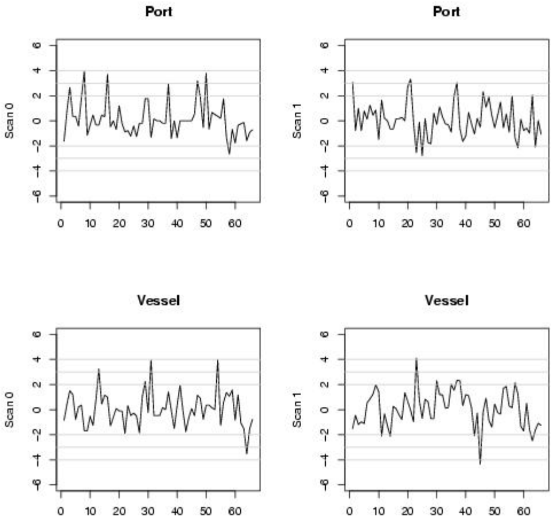
Vessels | Ports





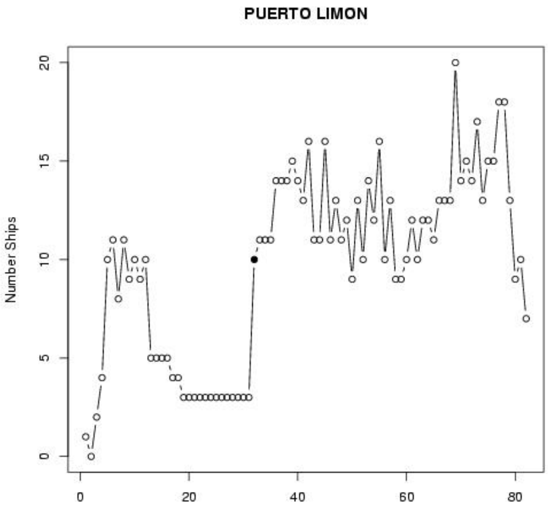
Each curve corresponds to one week.

Scan Stati



NAVSEA WARFARE CENTERS DAHLGREN

Scan Statistic Detection

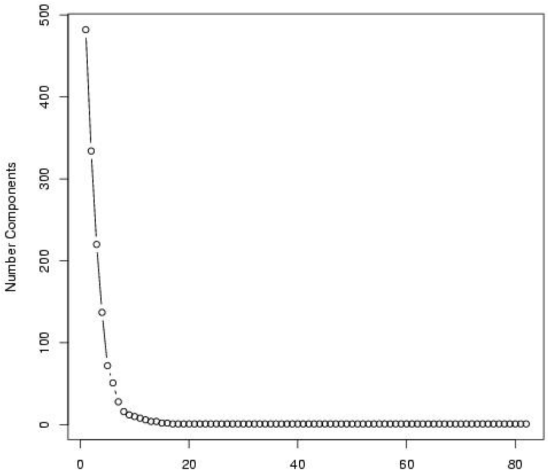


WARFARE CENTERS
DAHLGREN

Cumulative Bipartite Graph

- We can look at the cumulative bipartite graph: each week the edges are added to the preceding graph.
- Calculations on this graph can provide insight into the port activity.
- For instance, the number of components of the graph gives a measure of how “connected” the product flows are:
  - Many connected components means there are many “communities” of trade.
  - A single connected component means that (in principle) an item can go between any two ports in the time corresponding to the graph.

Connected Components of Cumulative Graph



After 17 weeks the graph is connected.

Discussion

- Graphs provide a convenient framework for analyzing the relationships between entities.
- We have considered bipartite graphs defined by Vessel/Port pairings.
- Scan statistics were introduced to illustrate change-point detection.
- Considering various types of time series of graphs (or dynamic graphs) illustrates several calculations that can be easily performed to learn about the activities of the entities of interest.
- A final note: the data we have are very buggy, many errors. This has required a lot of cleaning.
  Moral: **Always** perform exploratory data analysis.