Time Series of Dot Product Graphs

David J. Marchette
dmarchette@gmail.com

Collaborators: Carey Priebe, Ed Scheinerman and others

Naval Surface Warfare Center
Code Q21

Compstat, Aug 2006

NAVSEA WARFARE CENTERS
DAHLGREN

Outline

Outline

Outline

NAVSEA WARFARE CENTERS DAHLGREN

Outline

Introduction

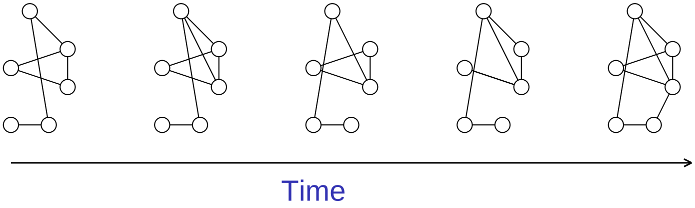
- Pattern recognition applications today involve very complex, high dimensional, and huge data sets.
- Traditional methods in $\mathbb{R}^d$ are no longer sufficient.
- What's needed are methods to operate on complex, changing data types.
- This talk will focus on one specific type of problem: recognition of changes (anomalies) in a time series of graphs.

NAVSEA WARFARE CENTERS DAHLGREN

What is a Time Series of Graphs?

A time series of graphs is:

- A graph valued random variable indexed by time.
- An ordered sequence of random graphs.



Why?

Time series of graphs are becoming more and more common:

- Communication graphs
- Social networks
- Biological networks

We need methods to:

- Model the time series
- Detect changes in the underlying model
- Interpret the model within the framework of the application

Analysis of Time Series of Graphs

The traditional way of analyzing time series of graphs is to convert to a time series of real numbers (or vectors) and use standard time series methods.

- Compute an invariant of each graph:
  - Number/density of edges
  - Maximal clique size
  - Scan statistic
- So $G_1, G_2, \ldots, G_t, \ldots \Rightarrow \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots$

Another Approach

Model the time series as a random variable indexed by time:

- Define a random graph model.
- Parameterize the model in some way.
- Let the parameter vary in time.

Erdös-Rényi Random Graphs

The simplest and most common type of random graph is the Erdös-Rényi random graph:

- Given a probability p
- Place an edge $v_i v_j$ between vertices v_i and v_j with probability p.
- This results in a “time series” of independent random graphs.
- We can change this to a more interesting time series by making p depend on time: p_t.
- How can we extend this to a more interesting/relevant model?

NAVSEA WARFARE CENTERS DAHLGREN

Social Network Motivation

- A social network is a set of vertices corresponding to "actors" (individual entities) and edges representing relationships.
- Our intuition is that actors with similar interests should be related: In some sense, the probability of an edge should be proportional to the amount of overlap of interests.

Intersection Graph

- Each vertex v_i has associated with it a (random) set S_i.
- Place an edge v_iv_j between vertices v_i and v_j if $S_i \cap S_j \neq \emptyset$ (or $|S_i \cap S_j| > \tau$).
- The randomness of the graph results from the random sets.
- Alternatively, one could place the edge with probability $p_{ij} = \frac{|S_i \cap S_j|}{|S_i \cup S_j|}$.
- In this case, the sets need not be random, but the graph sitll is, and the edges are no longer independent.

NAVSEA WARFARE CENTERS DAHLGREN

Random Dot Product Graphs (Scheinerman)

- Each vertex v_i has associated with it a vector x_i.
- Place an edge v_iv_j between vertices v_i and v_j with probability proportional to x_ix_j, the dot product of x_i and x_j.
- Thus $p_{ij} = f(x_ix_j)$. I will use the identity function in this talk.
- The edges in the random graph are no longer independent.
- Further, this can be interpreted in the manner of our social network motivation.
- Random dot product graphs are generalizations of intersection graphs, and a special case of the latent position model.
- We can now make a time series model by placing a time series model on the vectors x.

Social Networks Revisited

The attributes of an actor define the vector associated with the actor.

$$\left\{ \begin{array}{c} \mathit{Religion} \\ \mathit{Education} \\ \mathit{Sports} \\ \mathit{Food} \\ \mathit{TV} \end{array} \right\} \Rightarrow \left\{ \begin{array}{c} \mathit{categorical} \\ \mathit{ordinal} \\ \mathit{categorical} \\ \mathit{high-dimensional} \\ \mathit{categorical} \end{array} \right\} \overset{\not\Leftarrow}{\Rightarrow} \left\{ \begin{array}{c} 0.2 \\ 0.4 \\ 0.01 \\ 0.5 \end{array} \right\}$$

Note that the vector does not necessarily correspond to any simple (known or even knowable) mapping of the attributes.

"It's only a model." M. Python.

NAVSEA WARFARE CENTERS DAHLGREN

Groups

- We will assume that there are a small number (K) of distinct vectors.
  - For regularization.
  - To detect underlying groupings of the data.
- These groups correspond to sets of vertices with the same within-group probability and inter-group probabilities.

Outline

NAVSEA WARFARE CENTERS DAHLGREN

Fitting the Model

Given a graph (or a collection of graphs) how do we determine the vectors?

- Important issues:
  - How many distinct x vectors are there? (What is K?)
  - What is the dimensionality d of x?
- Several methods have been suggested, given that we know d, the dimensionality of the vectors x. We will assume an arbitrary value for d ($d = 2$).

Maximum Likelihood Approach

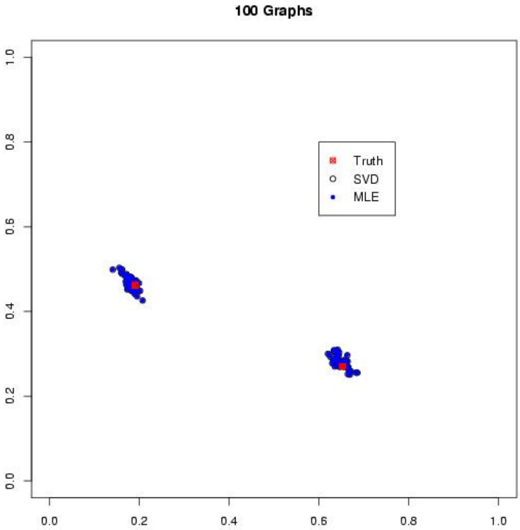
The edges are not independent, but they are conditionally independent (given the vectors x). Thus the likelihood is:

$$\begin{aligned} L(G) &= \left(\prod_{ij\in E}(x_i x_j)\right)\left(\prod_{ij\notin E}(1-x_i x_j)\right) \\ &= \prod_{ij}(x_i x_j)^{a_{ij}}(1-x_i x_j)^{1-a_{ij}} \\ l(G) &= \sum_{ij} a_{ij}\log(x_i x_j)+(1-a_{ij})\log(1-x_i x_j) \end{aligned}$$

Select X to maximize $l(G)$. Note that the extension to multiple graphs is easy, assuming independence.

NAVSEA WARFARE CENTERS DAHLGREN

Example



Comments

These aren't really what one gets when one runs these algorithms.

- Any rotation of a solution is a solution.
- Thus the solutions needed to be rotated into the first quadrant.
- I rotated them so that the center fell on the 45° line.
- This is critical when considering time series, and detecting change points.
- I used procrustes to match the vectors.

Outline

NAVSEA
WARFARE CENTERS
DAHLGREN

Model Selection: d

How does one choose the dimension d?

- Relying on a questionaire or one's intuition is a **bad** idea.
- Try $d = 1$. If the fit is good, stop. If not try $d = 2 \ldots$
- Pick d arbitrarily and proceed blindly.

Model Selection: K

Suppose there are only K distinct values (associated with K distinct populations) for the x vector; how to find K?

- Proposal 1:
  1. Derive a penalty for K.
  2. Use maximum penalized likelihood.
- Proposal 2:
  1. Fit x unconstrained (using whatever method works).
  2. Treat the n values of x as d-dimensional observations in $\mathbb{R}^d$ (or appropriate sub-manifold). Use a model selection algorithm in this space to select K.
  3. Re-fit x constrained by the selected model.
- Proposal 3:
  1. Cluster the vertices of the graph.
  2. Use this clustering to select K.
- I'll use proposal 2.

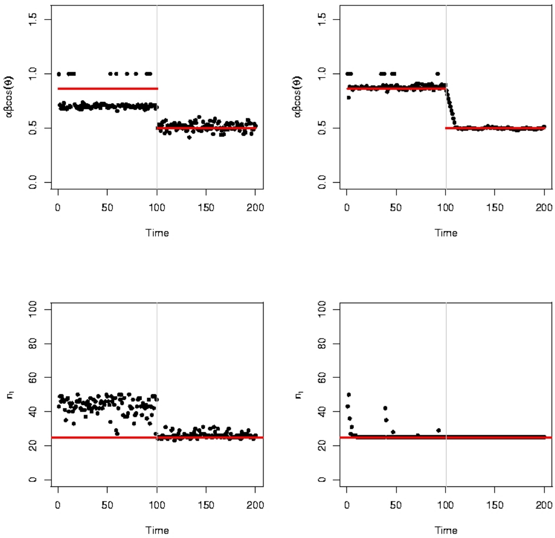
NAVSEA WARFARE CENTERS DAHLGREN

Time Series

- Rather than treat each graph independently, we'll average within a window W.
- The adjacency matrix now becomes the average of the adjacency matrices, an estimate of the edge probabilities.
- This assumes (approximate) independence in the window.
- Want W small to detect changes.
- Want W large to improve fit.

NAVSEA
WARFARE CENTERS
DAHLGREN

An Example: Changepoint in Vectors

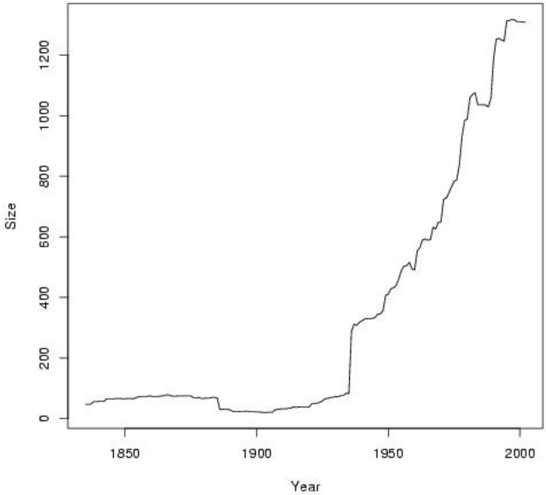


NAVSEA WARFARE CENTERS DAHLGREN

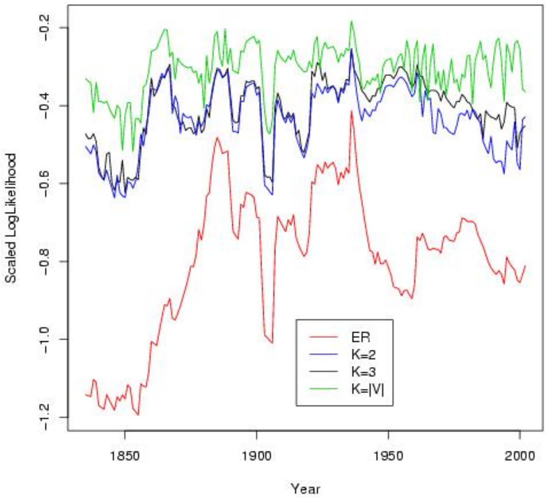
Interstate Alliances

- Data from 1816-2000.
- One graph per year:
  - Each country (nation state) is a vertex.
  - Each edge indicates an alliance between countries:
    - 0 or NA – no alliance.
    - 1 – defense pact.
    - 2 – neutrality pact.
    - 3 – nonaggression pact.
  - We'll binarize to alliance vs. no alliance.
- $W = 20$.

Sizes of the Graphs

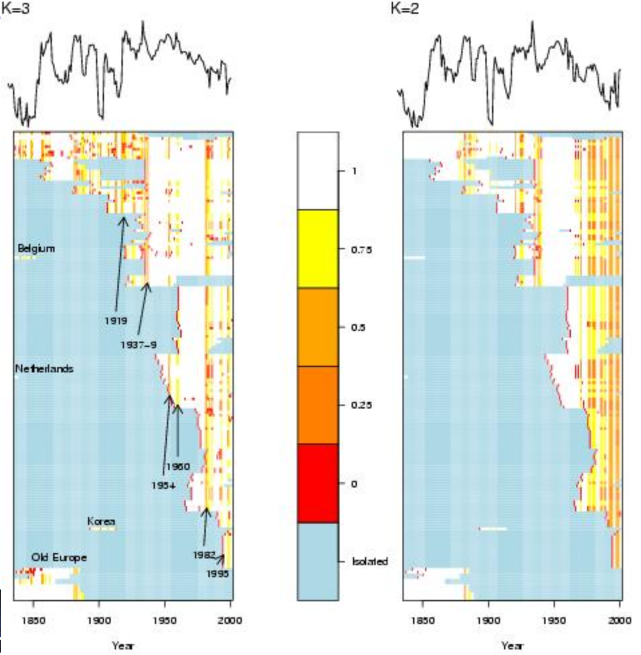


Scaled Log Likelihoods



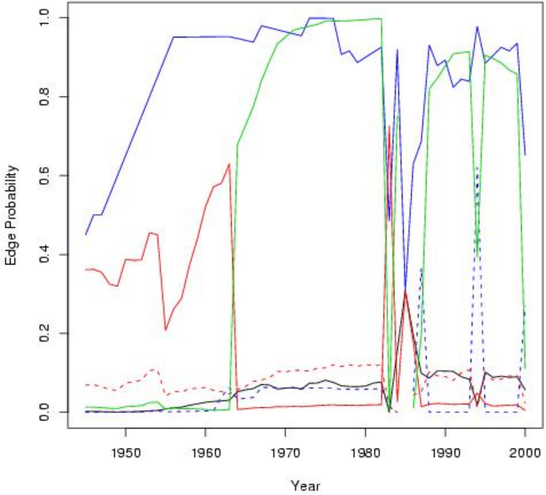
Scaled by $\frac{|V|(|V|-1)}{2}$.

Groups



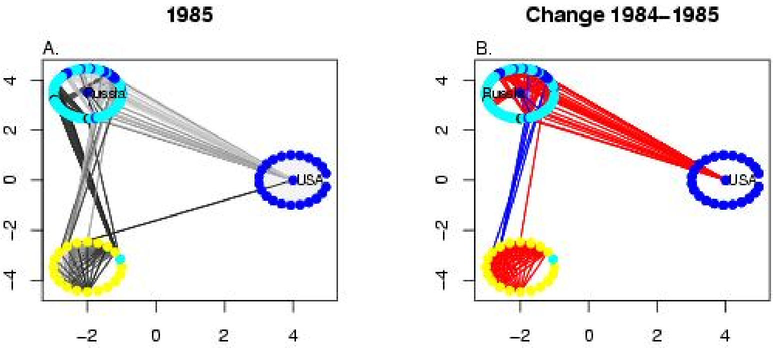
K=3
K=2
Belgium
1919
1937–9
Netherlands
1960
1954
Korea
Old Europe
1982
1995
1850
1900
1950
2000
Year
1
0.75
0.5
0.25
0
Isolated

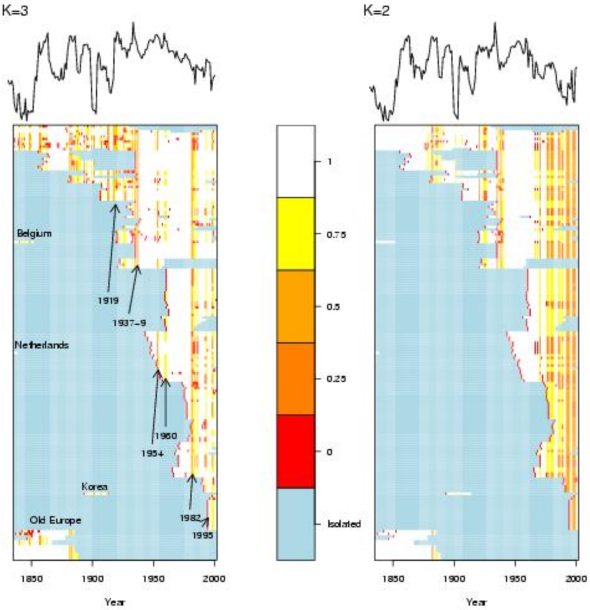
Group Probabilities



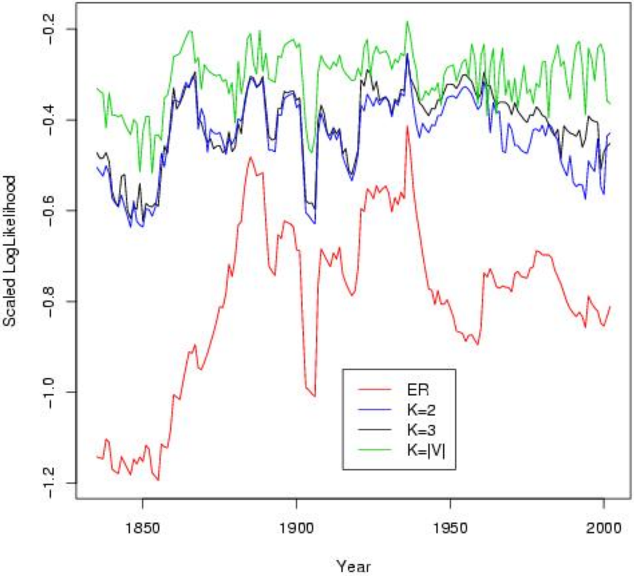
Russia=Red, USA=Blue, Others=Green

What Happened in 1985?




K=3
K=2
Belgium
Netherlands
Korea
Old Europe
1919
1937–9
1950
1954
1982
1995
1
0.75
0.5
0.25
0
Isolated
1850
1900
1950
2000
Year

What is the right model?



NAVSEA
WARFARE CENTERS
DAHLGREN

Discussion

- These models provide rich classes from which to do inference on time series of graphs.
- Time series of graphs appear naturally in several areas of science.
- Much work remains to be done, both theoretically and in applications.
- Collaborators are welcome.