Predicting Unobserved Links in Covert Networks

David J. Marchette
dmarchette@gmail.com

Naval Surface Warfare Center
Code Q21

Feb 8, 2007

Outline

Motivation

Definitions

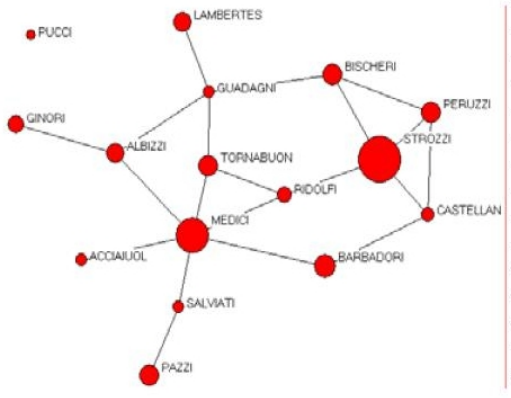
The Model

The Results

Social Networks

- A model of the relationships between entities.
- Also used to study insurgent groups, terrorist cells, etc.
- Relates actors (nodes in the network) through relationships (edges in the network).
- Typically used for small groups, with full knowledge of all links.

Marriage Network



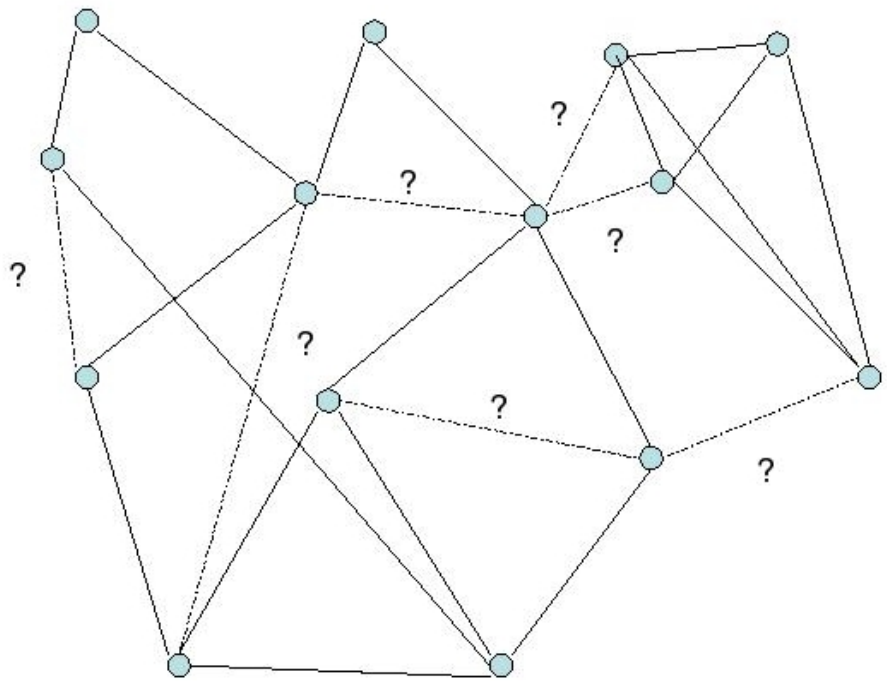
Family	Wealth	Betw.	Eigenv.	Degree
ACCIAIUOL	10	0.0	0.13	6.7
ALBIZZI	36	19.3	0.24	20.0
BARBADORI	55	8.5	0.21	13.3
BISCHERI	44	9.5	0.28	20.0
CASTELLAN	20	5.0	0.26	20.0
GINORI	32	0.0	0.07	6.7
GUADAGNI	8	23.2	0.29	26.7
LAMBERTES	42	0.0	0.09	6.7
MEDICI	103	47.5	0.43	40.0
PAZZI	48	0.0	0.04	6.7
PERUZZI	49	2.0	0.28	20.0
PUCCI	3	0.0	0.00	0.0
RIDOLFI	27	10.3	0.34	20.0
SALVIATI	10	13.0	0.15	13.3
STROZZI	146	9.3	0.36	26.7
TORNABUON	48	8.3	0.33	20.0

Korrelaatiot:

Wealth & Betweenness c.	0.3512
Wealth & Eigenvector c.	0.5366
Wealth & Degree c.	0.5590

Covert Networks

- Actors have a vested interest in not being observed.
- Networks may be very large.
- Some links are known to be there, some known to be missing, but others are unknown (then there's the "unknown unknowns"...).



Goal

Given a network with imperfect information:

1. Predict which potential links are most likely to exist.
2. Prioritize further information collection aimed at improving the quality of the network.
3. Give a best guess as to the true structure of the network.

In this talk I will assume the vertices are known. Only the edges may be imperfectly observed.

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)).
- Vertices are sometimes called “nodes” or “actors”.
- Edges are sometimes called “links” or “relations”.
- The adjacency matrix $A = (a_{ij})$ is the $|V| \times |V|$ binary matrix with a 1 in those places where an edge occurs in the graph.

Probabilistic Framework

- We place a probability structure on the network.
- This means we fit a **generative** model to the graph.
- This allows us to estimate the probability of a missing (unknown) link.
- We can bring node attributes into the model.
- We are essentially choosing the “most likely” graph given the model assumption and the observed edges.

Random Dot Product Graphs

- Each vertex v_i has associated with it a vector x_i.
- Place an edge $v_i v_j$ between vertices v_i and v_j with probability proportional to $x_i \cdot x_j$, the dot product of x_i and x_j.
- Thus $p_{ij} = f(x_i \cdot x_j)$.
- If there are attributes z_i associated to the nodes, we can augment this as $p_{ij} = f(x_i \cdot x_j + y_i z_i \cdot z_j y_j)$
- The edges in the random graph are no longer independent.
- We need to estimate the x_i (and y_i) from the observed graph.
- We can extend the model to directed graphs by having in- and out-vectors x_i^I and x_i^O with p_{ij} proportional to $x_i^O \cdot x_j^I$.

Maximum Likelihood Approach

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

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

$A = (a_{ij})$ is the adjacency matrix. Select X to maximize $l(G)$. Given X we can predict the unknown edges: large values of p_{ij} mean high probability of an edge.

Models

- A further twist: we want to reduce the complexity of the model.
- Suppose we constrain the vectors to attain at most K distinct values.
- Thus our model is that there are K possible vectors, and each node selects one of these.
- This gives an automatic grouping of the nodes.
- The penalty is that we have to somehow select K.

Attributes

- We do not assume all attributes are observed.
- Unobserved attributes will be imputed via maximum likelihood.
- In some cases, for example in a time series of graphs, other approaches may be appropriate.
- We'll see that incorporating attributes can improve edge prediction.

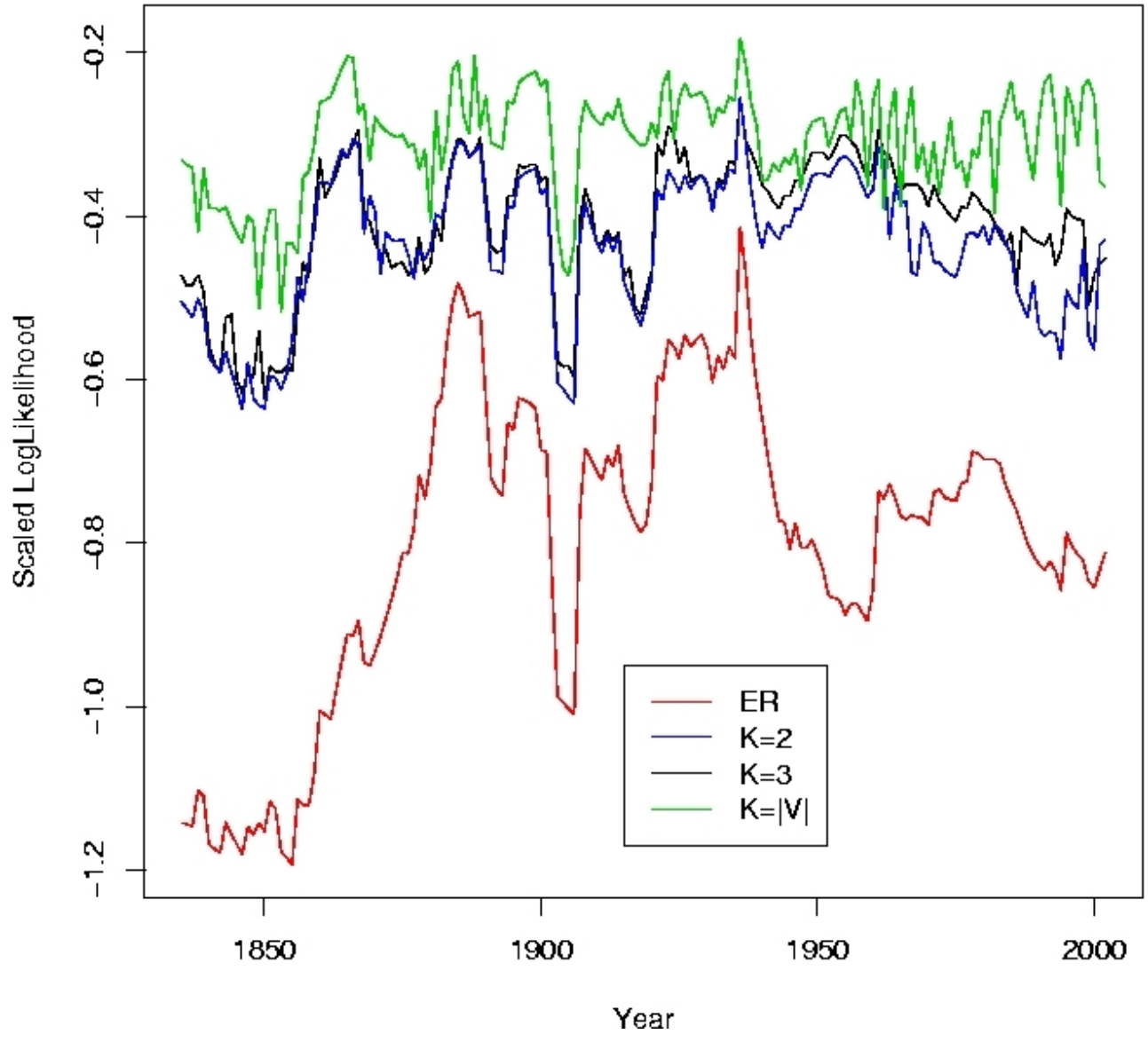
Evaluation

1. Split the set of possible edges into 4 distinct sets.
2. For each set:
   - 2.1 Set those edges to NA (unknown).
   - 2.2 Estimate the parameters of the graph.
   - 2.3 Assign a probability to each unknown edge.
3. Rank the edges by probability.
4. Plot the number of edges tested against the number of true edges found.
5. Repeat many times (100 times) to get an estimate of variability.

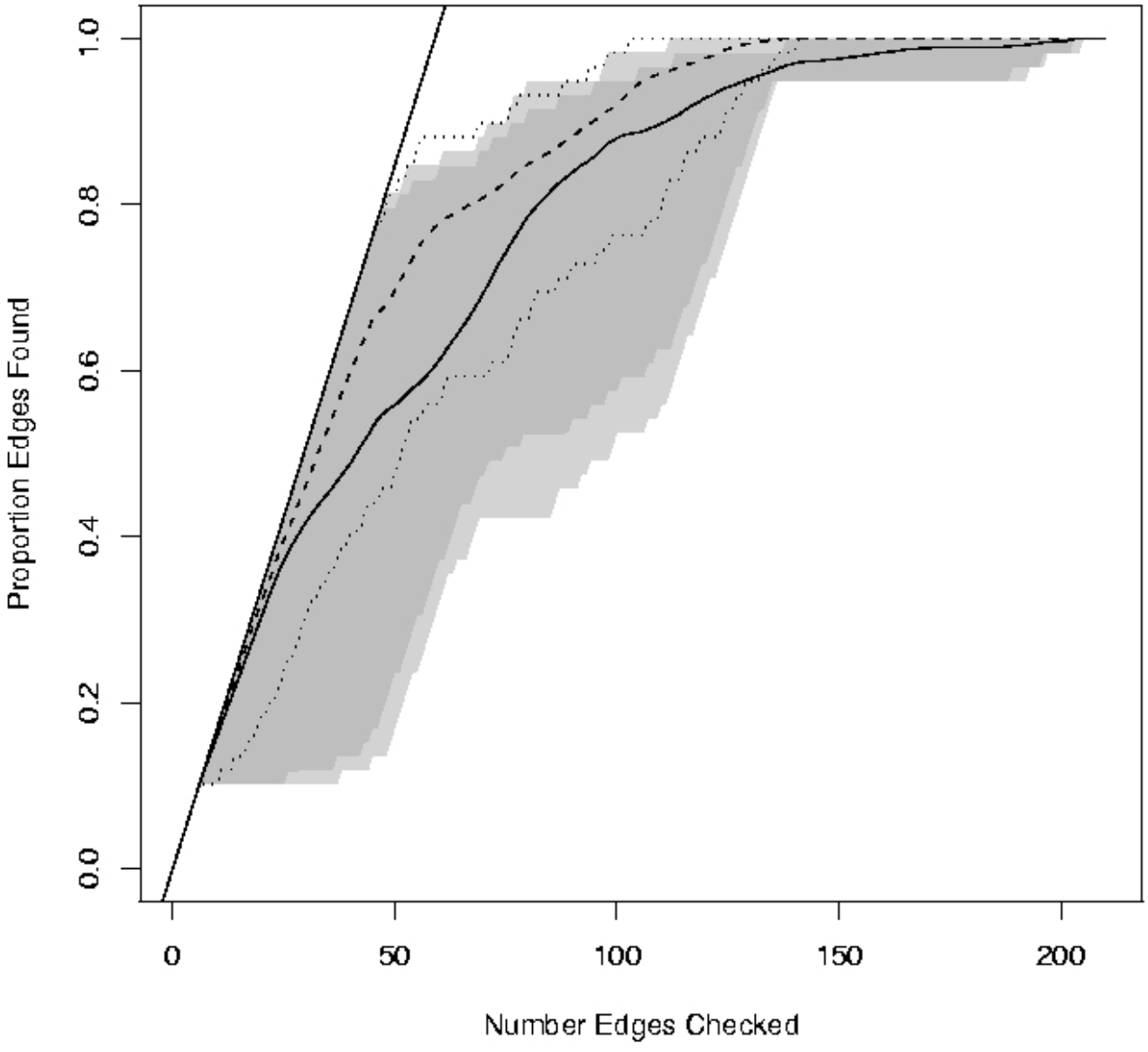
The Data

- Graphs of interstate alliances between countries between 1816 and 2000.
- Isolated countries removed from the graphs.
- Considered a subset of the years.
- Each vertex in the graph has three attributes associated with it:
  - Military expenditures.
  - Energy expenditures.
  - Democracy score.

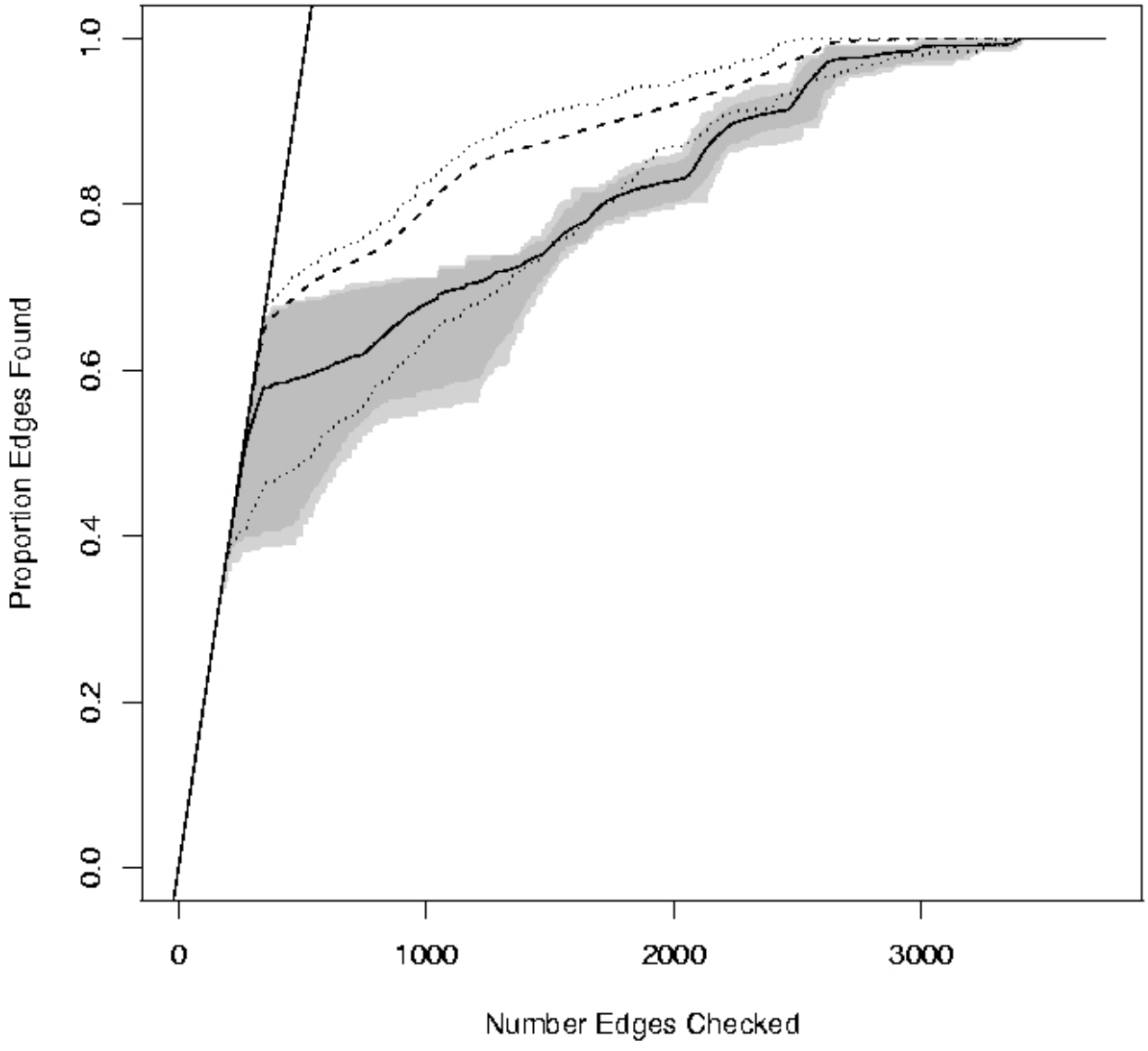
Model Selection



Typical Result: 1866



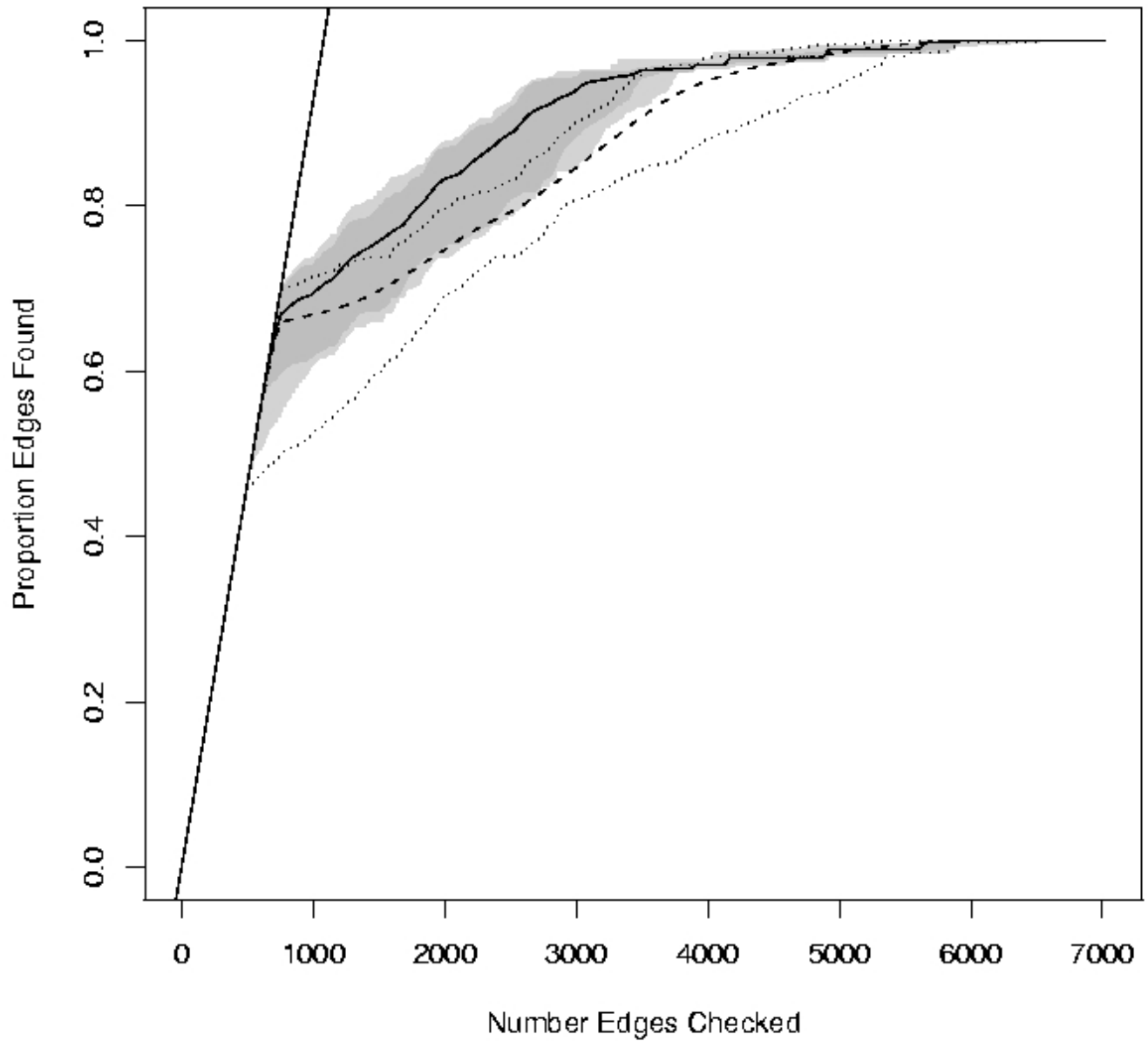
1970


Proportion Edges Found
0.0
0.2
0.4
0.6
0.8
1.0
0
1000
2000
3000
Number Edges Checked

Missing Attributes

- In this dataset, some of the attributes are missing from some nodes at certain times.
- We can impute these values various ways:
  1. Use past information: fit a model to past values and predict present.
  2. Use maximum likelihood to impute the values.
  3. Use Bayesian methods if prior information is available (or even if not).

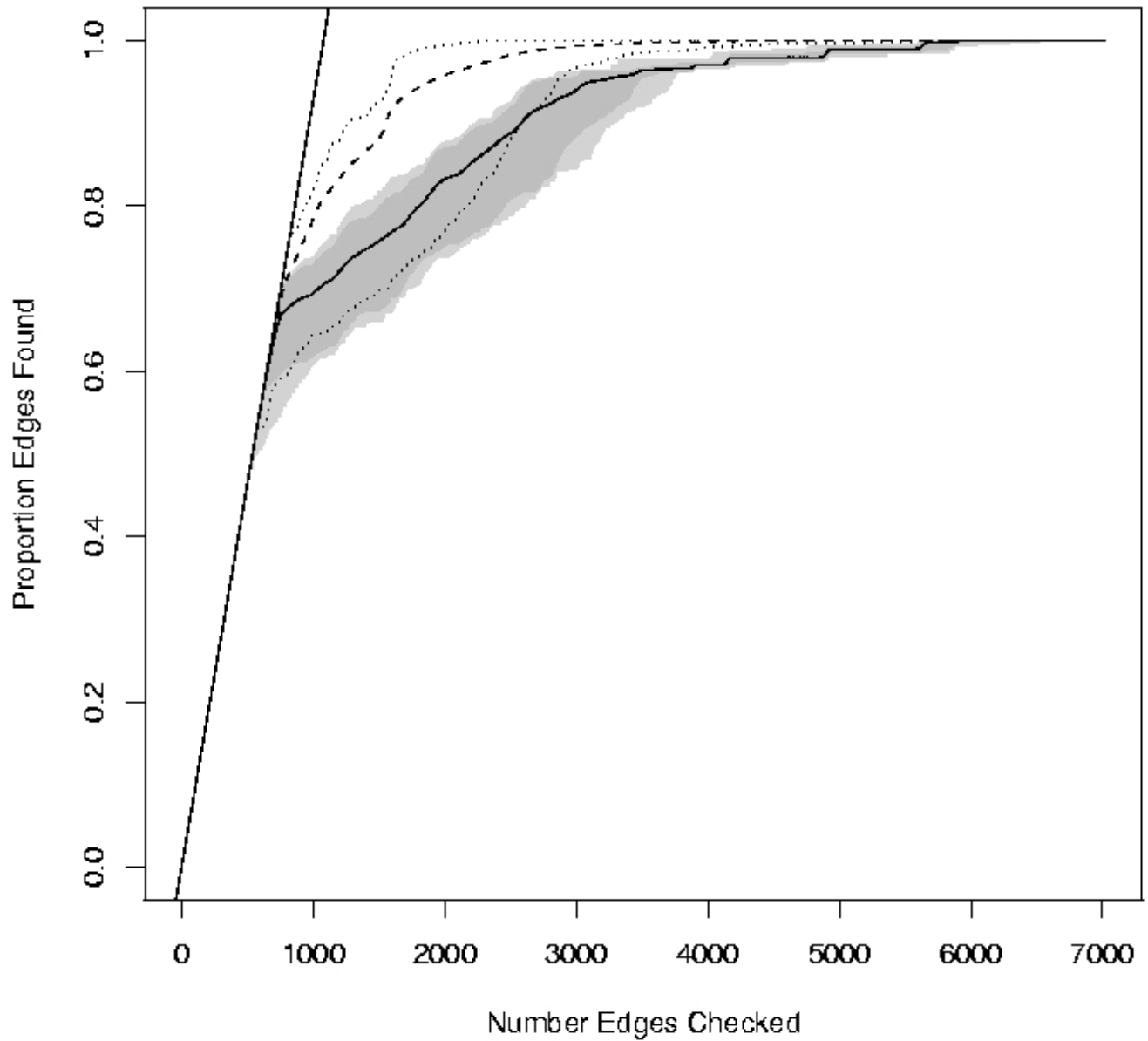
1995: Before Imputation



Quantitative Methods in Defense & National Security
QMDNS 2007

NAVSEA WARFARE CENTERS
DAHLGREN

1995: After Imputation



Future Work

- Can we detect edges that are wrong?
- Can we use other prior information to improve the detection:
  - Knowledge about degrees: most everyone makes at least N phone calls every day.
  - Knowledge about local structure: node x_i is “central” (in some precisely defined way).
  - Knowledge about global structure: there is a tendency to form tight cliques, there is a rough hierarchical command structure, etc.
- What about missing nodes?