Extracting Useful Information from Complex Data

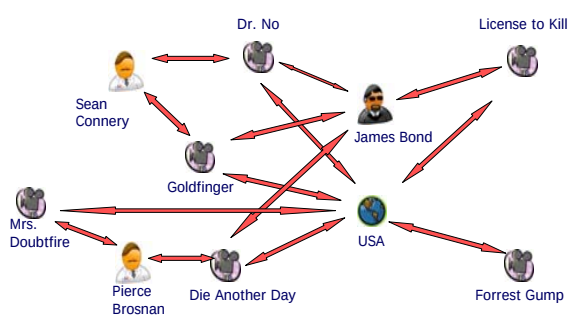
Brian Tran, Stanford University,
Information Operations & Analysis Center, Norman Franke,
Lawrence Livermore National Laboratory

Finding relationships in large, heterogeneous data can prove to be extremely vital yet perplexing. Graph clustering and hierarchical cluster trees help us easily visualize and interpret data. With it we can even discover new relationships and predict future outcomes.

Introduction

- A graph is a set of "nodes" or objects, and "edges" connecting related nodes together
- Graphs can be used to visualize almost any type of information
- Being able to fully interpret these graphs can be quite useful
- ❖ Can we find **hidden relationships** in our data that we were not able to previously recognize?
- ❖ Using our known data, can we **predict** future outcomes?
- With such large datasets, can it be done efficiently?

Movie	Character	Actor	Location
Die Another Day	James Bond	Pierce Brosnan	UK / USA
Goldfinger	James Bond	Sean Connery	UK / USA



A simplified James Bond movies graph

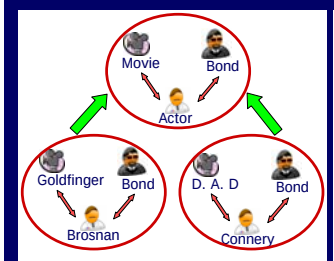
Methods

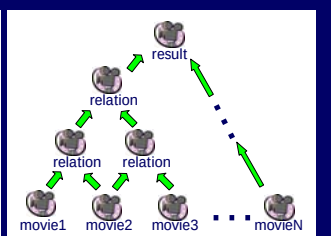
- Graphs representing large, heterogeneous data can themselves look too convoluted
- **Graph clustering** organizes the graph in visual space
  - The algorithm is done by minimizing the "total energy" (entropy) of the graph
  - Related nodes are grouped together
  - Unrelated nodes are pushed apart

$$Energy\ (p) = \underbrace{\sum_{\{u,v\} \in E} w_e \left\| p_u - p_v \right\|}_{\text{Node attraction}} - \underbrace{\sum_{\{u,v\} \in V^{(2)}} \ln \left\| p_u - p_v \right\|}_{\text{Node repulsion}}$$

w_e = weight of each edge E
p_u, p_v = positions of node u and node v in the set of node pairs $V^{(2)}$

- The user can then choose which cluster to interpret
  - Gives the user the power to choose with ease
  - Makes relational data mining more efficient
- A **hierarchical cluster tree** is built using selected nodes
  - The graphs are merged to find relations

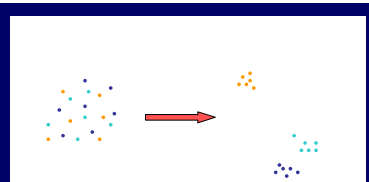




All possible movie combos are merged

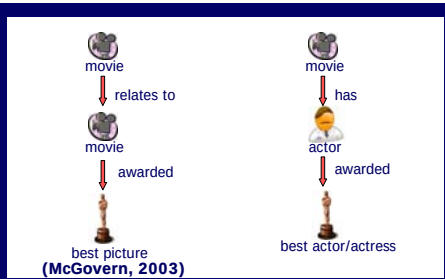
Results & Discussion

- Due to security reasons, the results on our data cannot be shown
- Instead, possible results are shown for predicting movie success



Sample clustering of movies

- Relational structures predicting movie success are then found



(McGovern, 2003)

- With the resulting relational structure, **predictions** can be easily made
- Some resulting relational structures may convey **hidden** relations or concepts never acknowledged before
- The method can be applied to any type of data (e.g. telecommunications, security, WWW, biological, social, etc.)

This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.