Exploring Graph Mining Approaches for Dynamic Heterogeneous Networks

Lisa Singh

Georgetown University

Department of Computer Science

singh@cs.georgetown.edu

Presentation Overview

- Graph mining problems
- Observational science data sets and issues
- Topological structures for heterogeneous graphs
- Visual mining
- Final thoughts

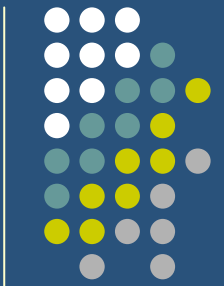
Graph, graphs everywhere

- Social networks
- Terrorist networks
- Corporate board networks
- Disease transmission networks
- Biological pathway networks
- Cellular networks
- Email networks
- Cognitive behavior networks
- Social media networks (blogs, MySpace, Flickr)

Graph Mining Problems

- Hidden community identification / group clustering
  - Kempe, Hopcroft, Girvan, Newman, Flake, Han, Gibson
- Information transmission / spread of influence
  - Sarkar, Moore, Gruhl, Liben-Nowell, Kempe, Kleinberg, Richardson, Domingo
- Group formation and evolution
  - Backstrom, Kleinberg, Wang
- Discovering meaningful graph approximations, metrics for prediction, visualization, etc.
  - Faloutsos, Borgatti, Getoor, Singh, Ahn, Lee

Observational scientific data



- Researchers monitor a subject for a specified period of time.
  - Subject – person, dolphin, celestial object
  - Observer – person, equipment
  - Event (Monitoring period) – behavior occurrence, set of observations
- Large number of events / observations for a small number of subjects.
- High dimensionality and complexity

Shark Bay Dolphin Research Project (SBDRP) Overview

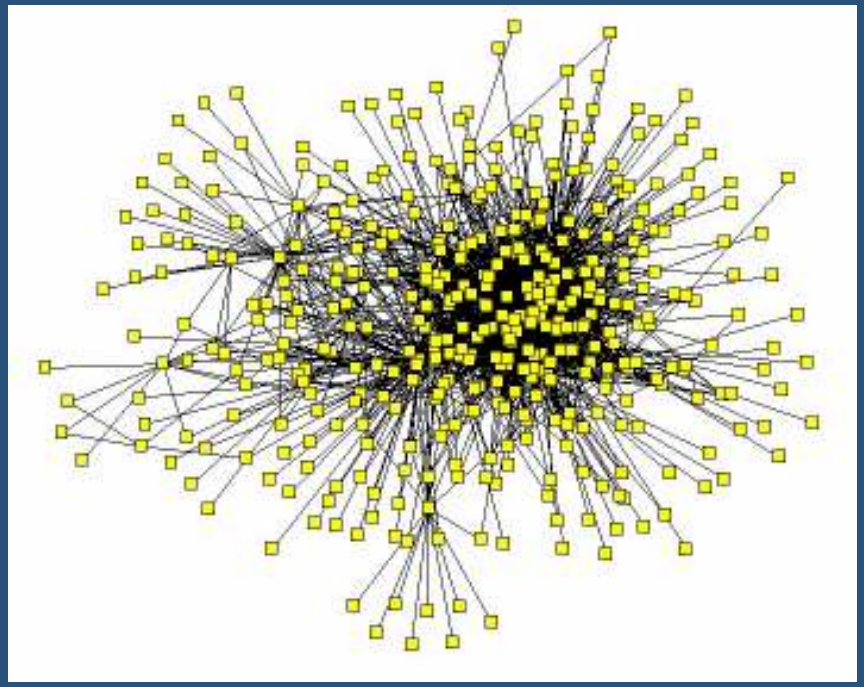
- Dolphins monitored by international team of scientists since 1984.
  - 13,400 surveys
  - Thousands of hours of focal follows
  - Thousands of pictures
  - GIS spatial data

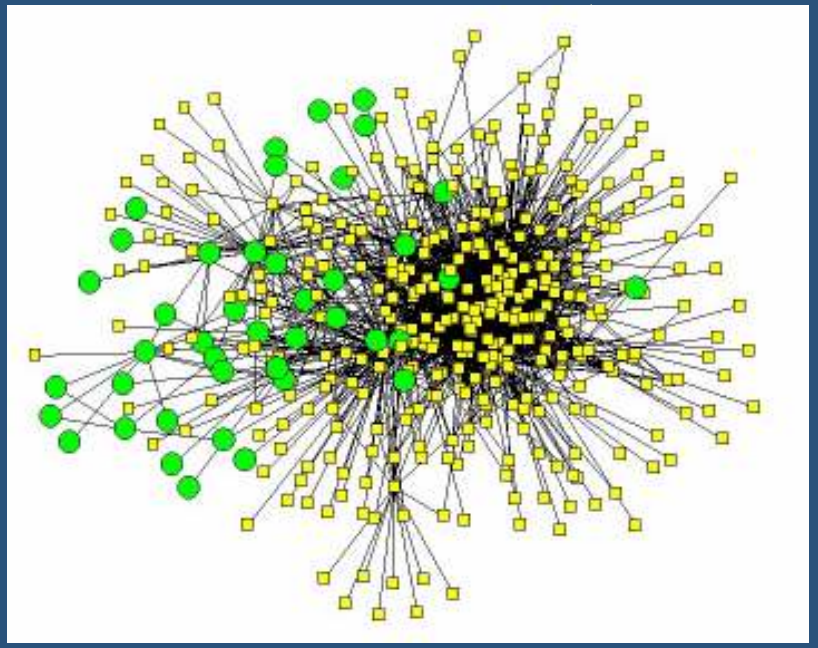
Data Set Details

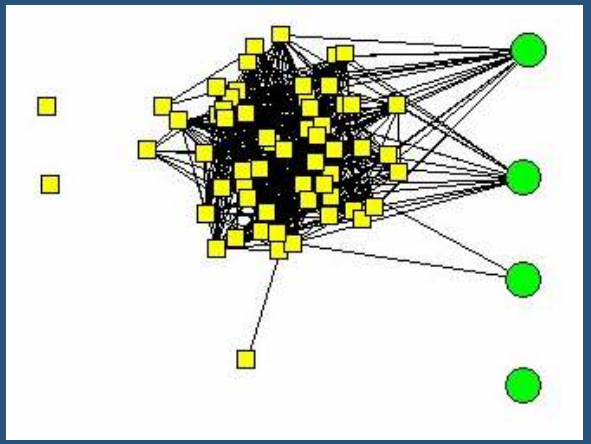
- Paper forms are used to collect survey and focal data prior to electronic entry.
- Survey data is entered into an Access database and converted to spreadsheets.
- Focal data is added to Excel spreadsheets.
- Special data sources have been created from the base data sets.
- The complete data set contains 17 repositories, 100s of attributes, missing data, inconsistent attribute values, and redundancy.
- Ad hoc query not possible; comprehensive analysis across data sources done using a ‘manual merge’ procedure.

HELP!!

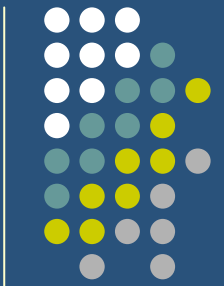
Uni-modal graph models







Complex Questions Observational Science Data



- Are observations conducted by different researchers on the team consistent or are there biases?
- Are the observations reliable or were there field conditions that impacted the quality of the observation?
- How does the community structure of prominent subjects having particular features changing over time?

Specific Research Questions Related to Dolphins

Interested in the relationship between protracted development and social-ecological complexity.

- Must calves meet specific ecological or social challenges before weaning?
- How does information move through networks in a fission-fusion social system?
- Are relationships and social bonds a predictor of female calving success?

M*3 Network

A topological representation for multi-modal, multi-relational, multi-featured network. (Singh et al, IV 2007)

$N = (\mathbf{A}, \mathbf{E}, \mathbf{R})$

$\mathbf{A} = \{A_1, A_2, \ldots A_{nAS}\}$

$\mathbf{E} = \{E_1, E_2, \ldots E_{nES}\}$

$\mathbf{R} = \{R_1, R_2, \ldots R_{nRS}\}$

$A_x\ (ID_{Ax}, B_{x1}, \ldots B_{xr})$ where $x = 1$ to n_{AS}

$E_y\ (ID_{Ey}, C_{y1}, \ldots C_{ys})$ where $x = 1$ to n_{ES}

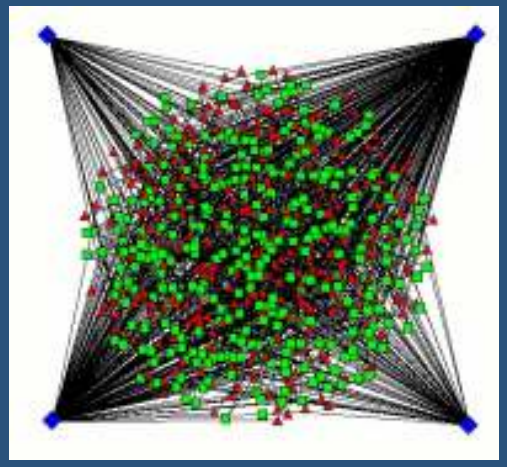
$R_z\ (ID_{Ax}, ID_{Ey}, D_1, \ldots D_t)$

where

$x = 1$ to n_{AS}

$y = 1$ to n_{ES}

$z = 1$ to n_{RS}



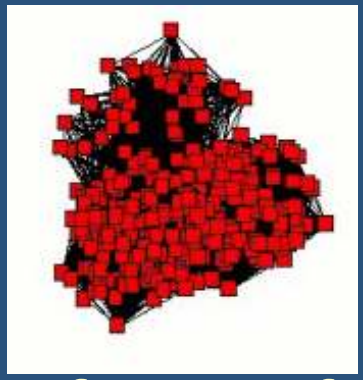
Network Topological Models

Uni-modal network consists of a set of actors A linked via a set of relationships R.

$N = (A, R)$ or generally, $G=(V,E)$

$A = \{a_1, a_2, \ldots a_{nA}\}$

$R = \{ (a_i, a_j) \mid a_i, a_i \in A \}$



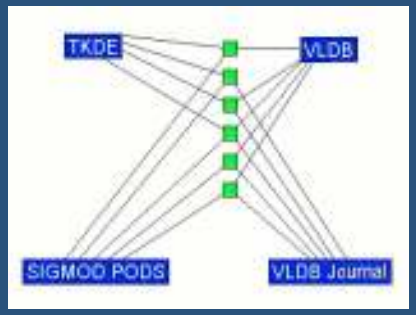
An affiliation network (bi-mode) consists of a set of actors A, linked via a set of relationships R to a set of events E.

$N = (A, E, R)$

$A = \{a_1, a_2, \ldots a_{nA}\}$

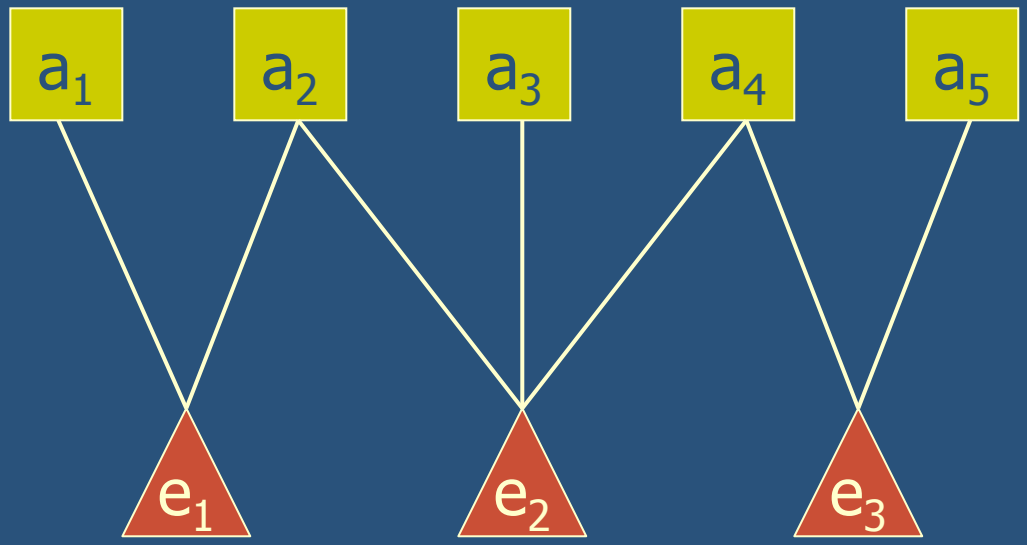
$E = \{e_1, e_2, \ldots e_{nE}\}$

$R = \{ (a_i, e_j) \mid a_i \in A \ \& \ e_i \in E \}$



Affiliation Graph

Two-mode Actor Event Node Graph

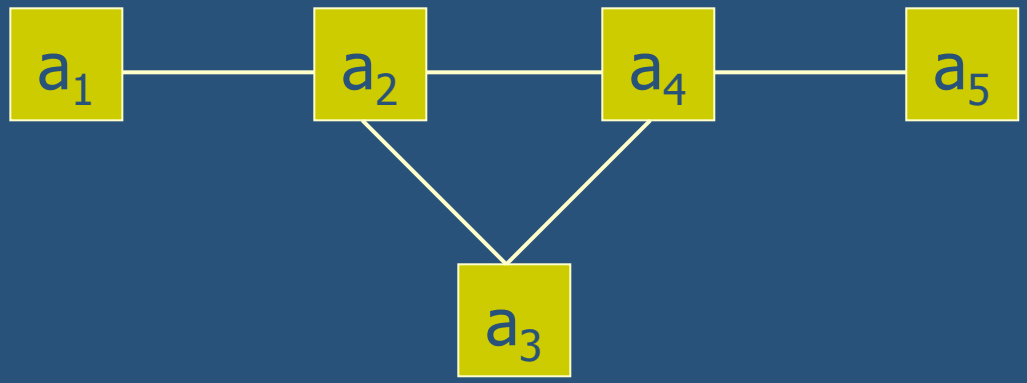


$$Nodes = A, E$$

$$Edges = R(Id_A, Id_E)$$

Co-membership Graph

Uni-mode Actor Node Graph



$$Nodes = A$$

$$Edges = \pi_{S.Id_A,\, R.Id_A}(\sigma_{S.Id_A < R.Id_A}(\rho_S(R) \underset{S.Id_E = R.Id_E}{\triangleright\!\!\triangleleft} R))$$

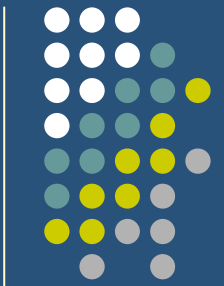
Event Overlap Graph

Uni-mode Event Node Graph

$$Nodes = E$$

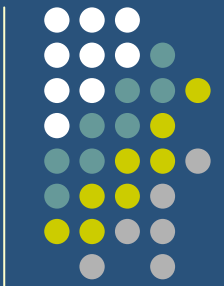
$$Edges = \pi_{S.Id_E,\, R.Id_E}(\sigma_{S.Id_E < R.Id_E}(\rho_S(R) \underset{S.Id_A = R.Id_A}{\triangleright\!\!\triangleleft} R))$$

Why prune or sample a social network?



- Social networks are large. Inefficient to search entire graph for some problems.
- Some nodes in a large network introduce noise.
- Efficiently search subgraphs that maintain relevant relationships in terms of predictive accuracy.

Approaches to Pruning



Structural properties

- Prune nodes that do not have a high degree, betweenness, eigenvector

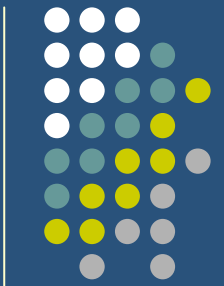
Descriptive attribute

- Prune edges by selecting on attributes of one or more relationships
- Prune nodes by selecting on attributes of a particular node type.

Random sampling

- Maintain only a random sample of the node/edge population for analysis.

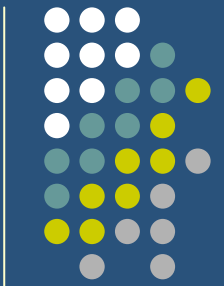
Pruned Classification Algorithm



- Given relations A, E, and R, identify the attributes that will be part of the analysis, where N = {AA, EE, RR}.
- Remove a subset of actors, events, and/or relationship attributes to create a pruned network, N' = {A', E', R'}.
- Determine prediction feature(s), P.f
- Create necessary aggregate values based on P.f
- Run classification algorithm, e.g. Bayes, C4.2, etc.

Singh, Getoor & Licamele, ICDM 2005

What must our models consider?



- Multiple node types and multiple edge types.
- Features associated with nodes and edges.
- Dynamic, time varying data
- Uncertainty in the observed graph
- Incomplete data
- Potentially massive size, streaming data sets

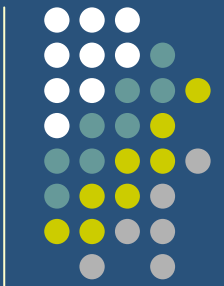
Extending graph topology approaches & pruning algorithms

- Develop a dynamic approach for finding ‘good’ pruning attribute based statistical distribution and network structural properties.
- Employ more robust classifiers that consider node dependencies – collective classification.
- Develop a generalized model for relational projections of social networks that also considers:
  - Multiple node types and multiple edge types.
  - Features associated with nodes and edges.
  - Dynamic, time varying data
  - Uncertainty in the observed graph
  - Incomplete data
  - Potentially massive size, streaming data sets

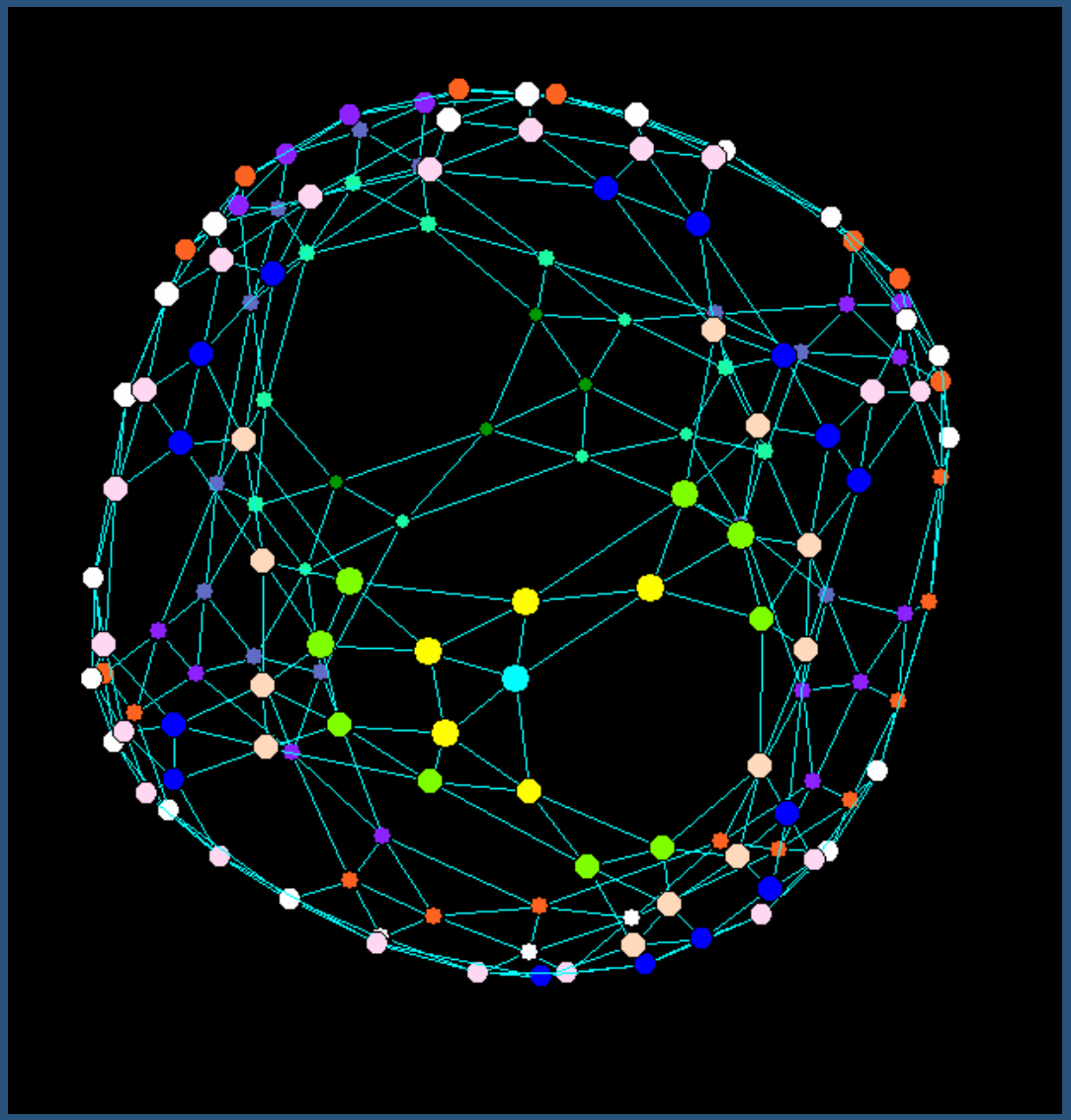
Value of generalize graph model

- Use richer set of features and network structural properties for descriptive and predictive tasks.
- Focus analysis using topological transforms.

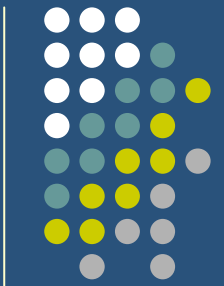
Metrics for understanding complex graph structures



- Uni-modal Metrics
  - Degree, betweenness, closeness, clustering coefficient, network density, etc.
- Multi-modal Metrics
  - Modal density
  - Multi-edge path length
  - Transmission rate
  - Network turnover



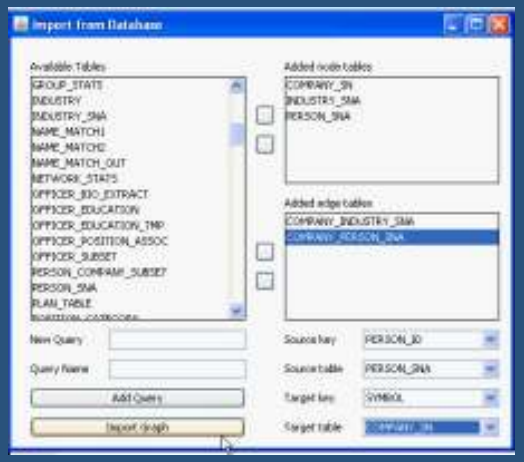
Graph Abstractions

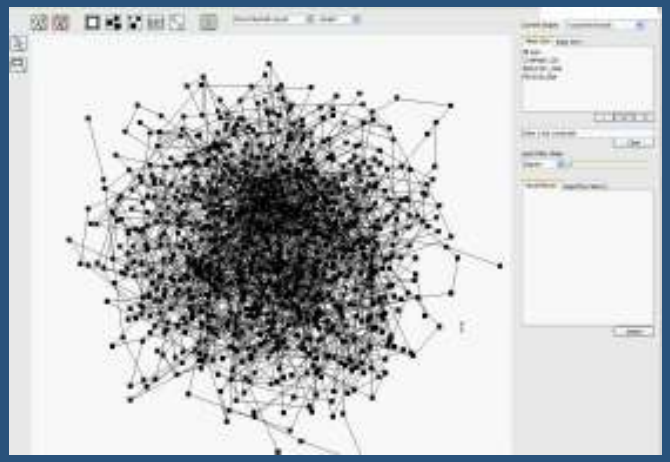


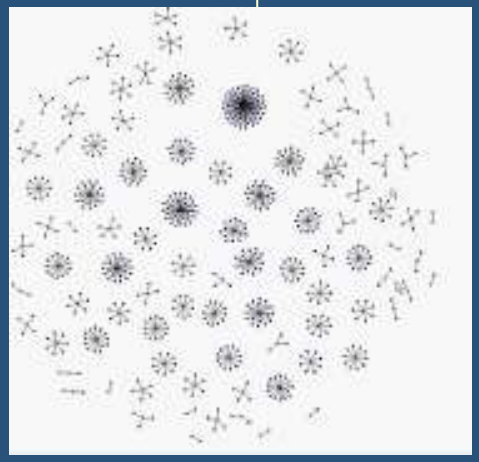
- Network element structural property
  - Degree, betweenness, etc.
- Network topological models
  - Uni-modal (co-membership), bi-modal (affiliation), etc.
- Mining algorithm
  - Clustering result, path result

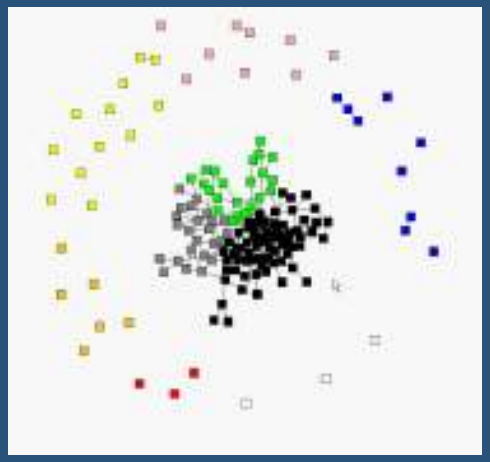
INTERACTIVE VISUAL MINING

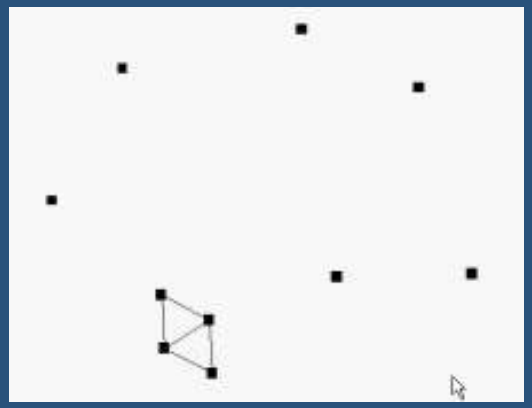
Invenio


Import from Database
Available Tables
GROUP_STATS
INDUSTRY
INDUSTRY_SNA
NAME_MATCH1
NAME_MATCH2
NAME_MATCH_OUT
NETWORK_STATS
OFFICER_BIO_EXTRACT
OFFICER_EDUCATION
OFFICER_EDUCATION_TMP
OFFICER_POSITION_ASSOC
OFFICER_SUBSET
PERSON_COMPANY_SUBSET
PERSON_SNA
PLAN_TABLE
Added node tables
COMPANY_SN
INDUSTRY_SNA
PERSON_SNA
Added edge tables
COMPANY_INDUSTRY_SNA
COMPANY_PERSON_SNA
New Query
Query Name
Add Query
Import Graph
Source key
PERSON_ID
Source table
PERSON_SNA
Target key
SYMBOL
Target table

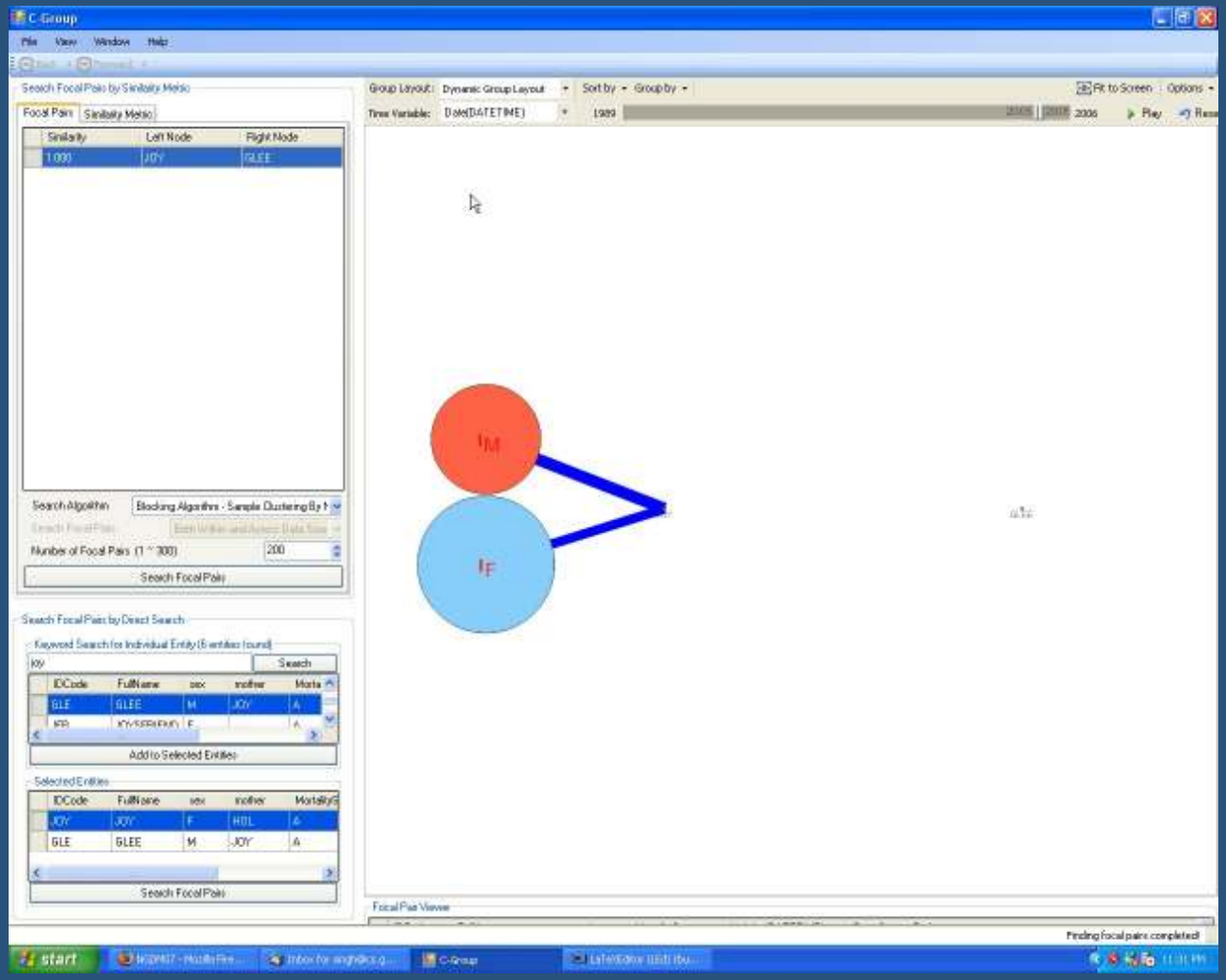






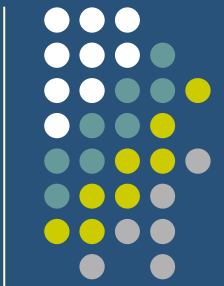


C-Group



Kang, Getoor, Singh, VAST 2007

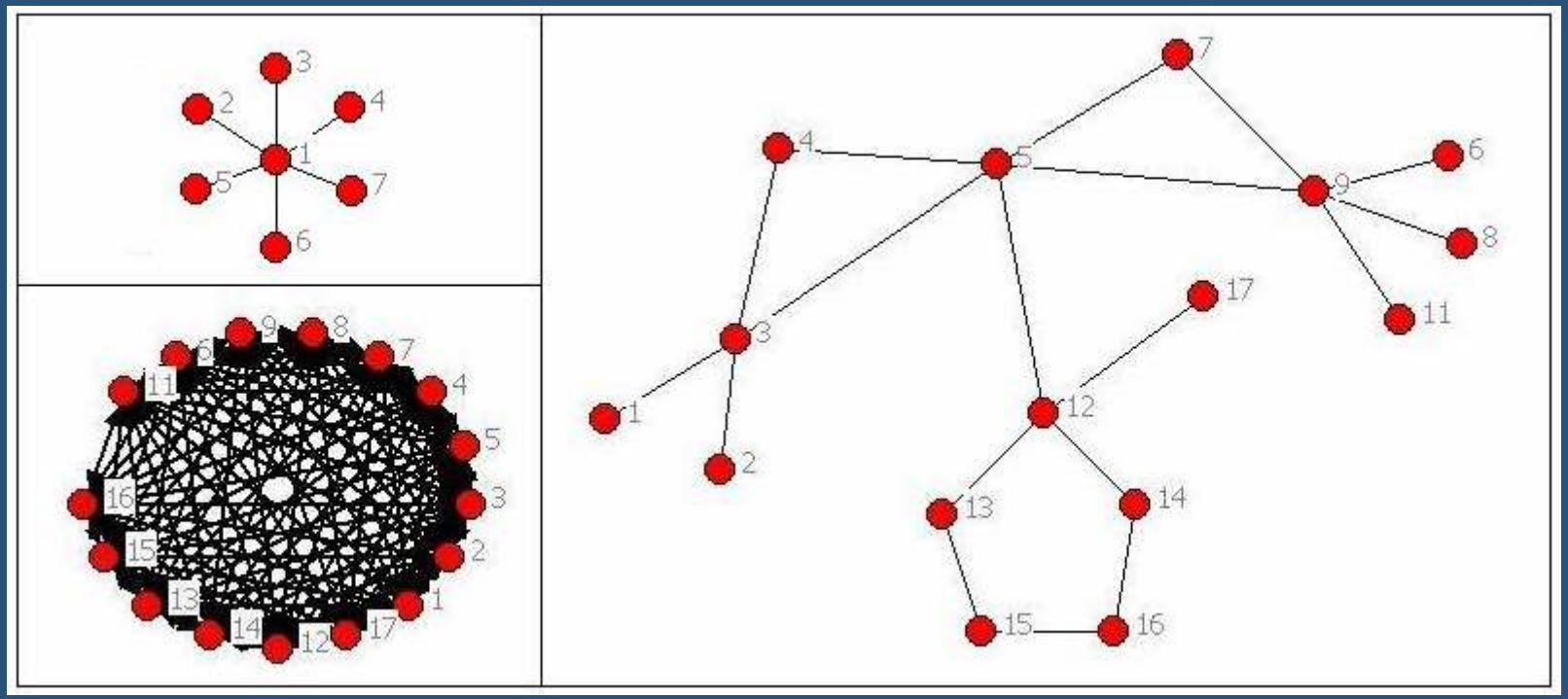
Complex social networks and privacy



Corporate email networks, customer referral networks, disease population networks.

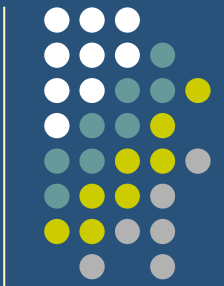
- What constitutes a privacy breech?
- How can we use the topological structure of complex networks to measure the level of anonymity in the network.
- To what degree is network topology a factor?
- How accurate does the representation need to be for reasonable graph mining results?

Understanding Topological Anonymity



Singh & Zhan, GrC 2007

Final thoughts



To date most graph mining algorithms have focused on simple graphs. The complexity of today's scientific networks forces us to consider ways to handle this complexity and integrate it into our descriptive and predictive mining algorithms.

Thank you

Questions?