

Social Networks: A Statistical view

Mark S. Handcock

**Departments of Statistics and Sociology
and
Center for Statistics and the Social Sciences**

University of Washington

email: handcock@stat.washington.edu

James Moody

**Department of Sociology
The Ohio State University**

Carter T. Butts

**Department of Sociology
University of California - Irvine**

Workshop on Statistical Inference, Computing and Visualization for Graphs, August 1 - 2, 2003

What is meant by Networks?

A *network* is a set of entities and a set of relations between them. Conceptually, the entities are represented by *nodes* and the relationships by *edges*.

The study of networks is multi-disciplinary:

- plethora of terminology
- varied objectives, multitude of frameworks

Examples:

Network	Nodes	Edges
WWW	Webpages	a link between them
Aviation System	Airports	a non-stop flight
River system	intersections	river segments
Probabilistic Models	Variables	dependency

What is meant by Social Networks?

The conceptualization and analysis of a network with the objective of understanding social structure.

social structure: a system of social relations tying distinct social entities to one another

- Attempt to represent the structure in social relations via networks.
- Theory relating to types of observable social spaces and their relation to individual and group behavior.
- The data are of two forms:
 - individual level information on the social entities
 - relational data on pairs of entities

- Primary interest in the nature of relationships:
 - How the behavior of individuals depends on their location in the social network
 - How the qualities of the individuals influence the social structure

- Secondary interest is in how network structure influences processes that develop over a network
 - spread of HIV and other STDs
 - diffusion of technical innovations
 - spread of computer viruses

- Tertiary interest in the effect of interventions on network structure and processes that develop over a network

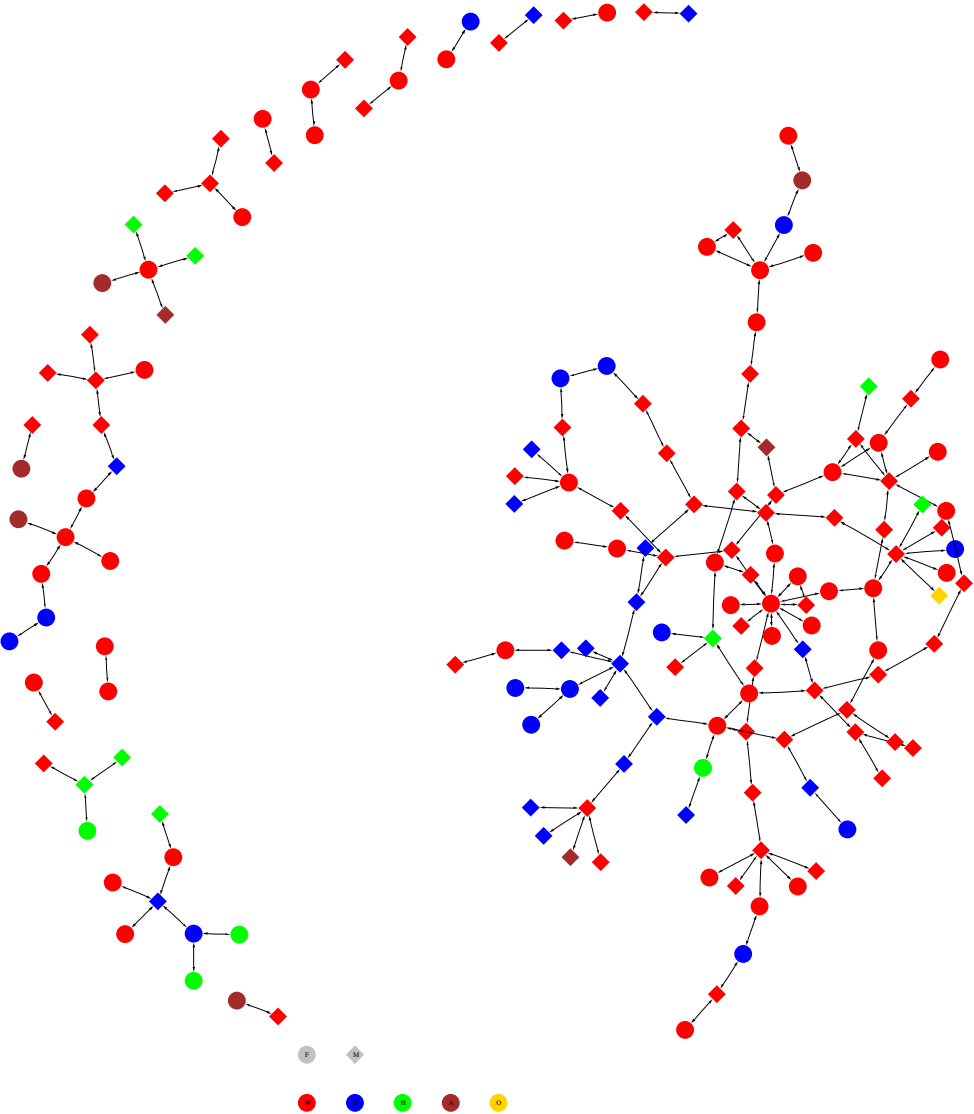
Network elements and terms

- Multiple types of both vertices and edges
- *nodes*:
 - e.g., sex, race, age
- *edges*:
 - *directed*: relationship has a direction associated with it
 - *weighted*: a discrete or continuous variable associated with it
- *networks*:
 - *directed*: all relationships have direction
 - *undirected*: all relationships do not have direction
 - *acyclic*: graph does not contain closed “loops” of edges

- *bipartite*: nodes of two types and edges only between different types
- *affiliation*: bipartite with nodes either a *group* or a person, and edges an affiliation of a person with the group
- *networks as elements*:
 - *time*: continuous and durational
 - *space*: localized networks with a spatial dimension
 - *embedding*: Nested networks (Hierarchy)
 - latent trait (clustering)
 - latent class (cohesion)

Application: Social networks of IV drug use

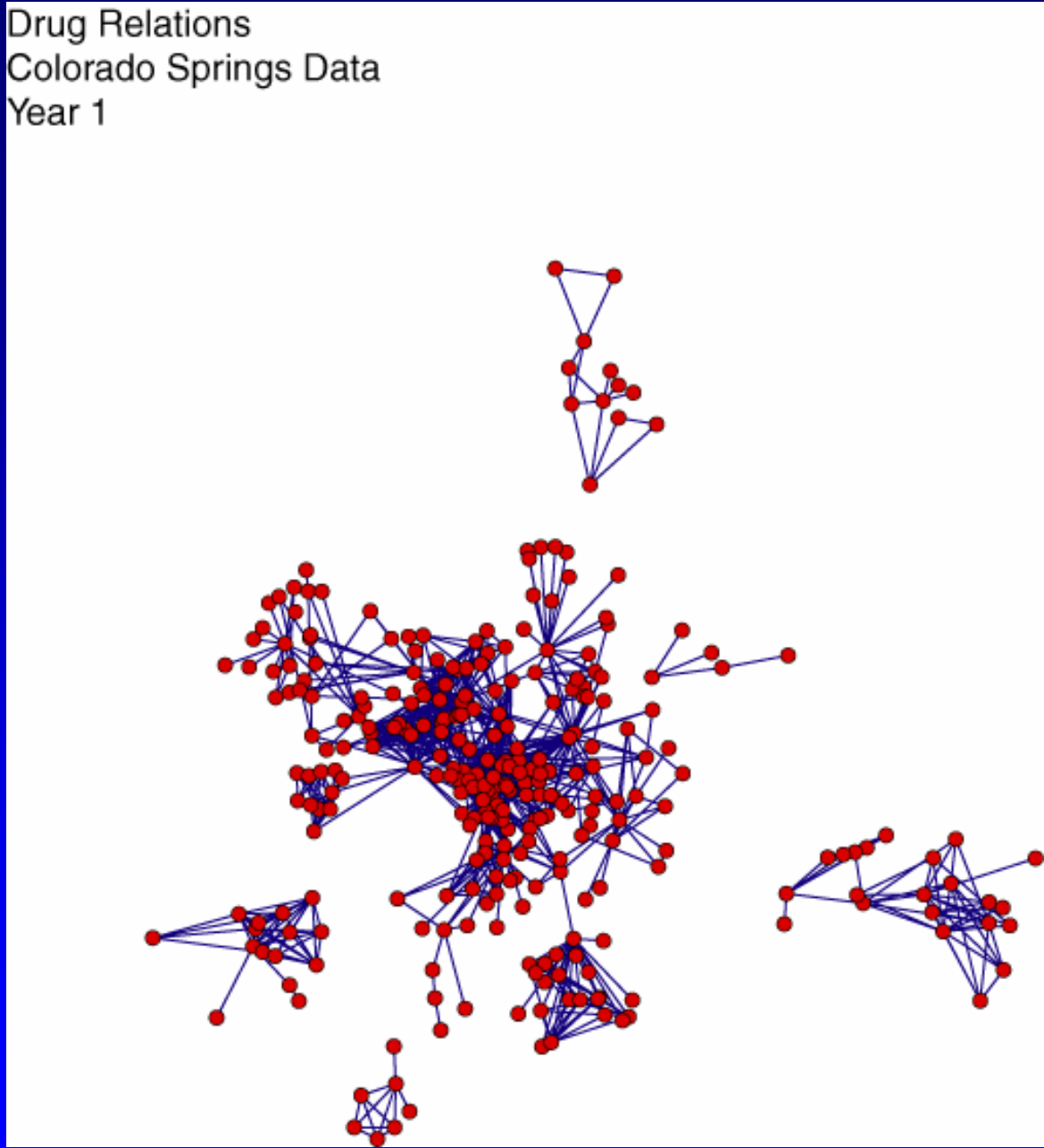
- Study a population of injecting drug users in Colorado Springs, CO from 1988-1992.
- Relationship is defined as needle sharing in the last six months
- A population of about 300 in the city.
- Data from a social survey
 - participants were asked to nominate sex and drug partners, to complete a face-to-face questionnaire, to provide a blood sample for HIV



Colorado Springs IVDU with race and sex

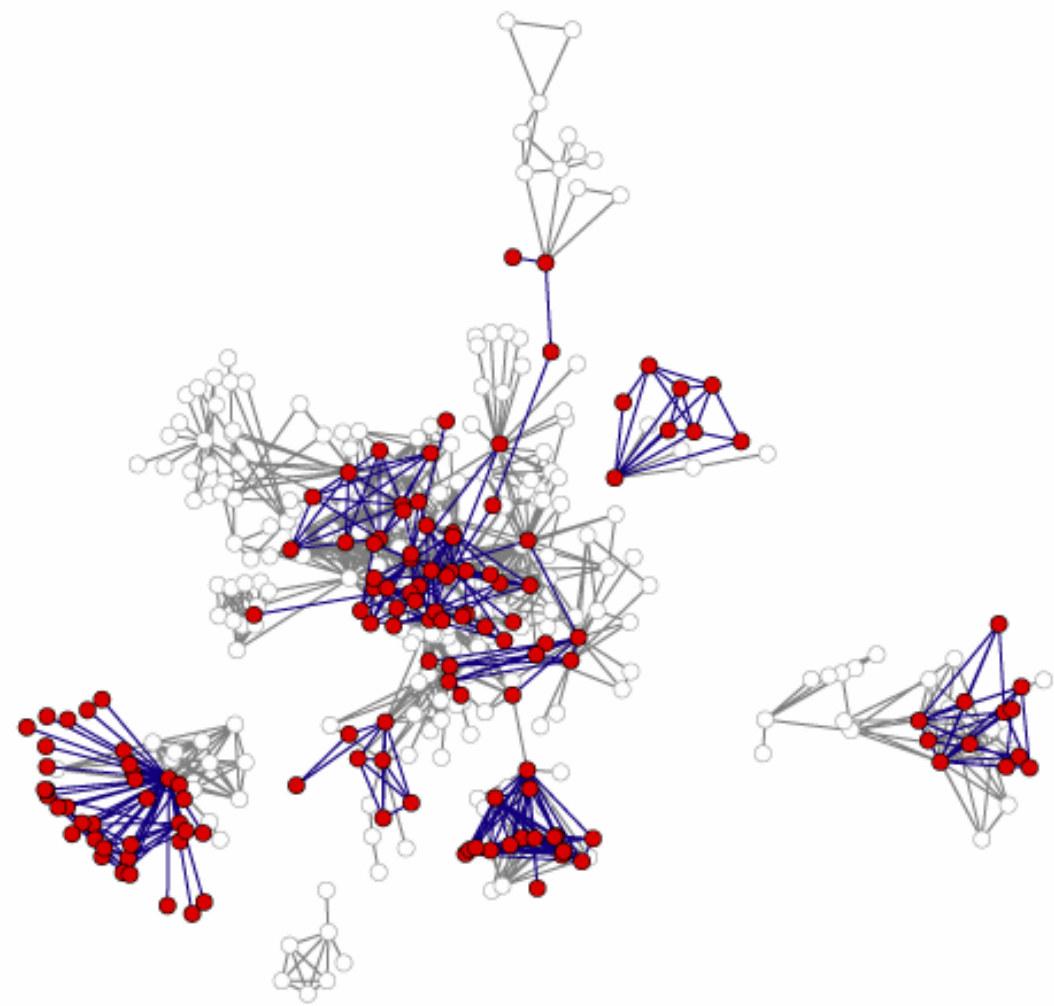
Drug Relations
Colorado Springs Data
Year 1

Data on drug users in
Colorado Springs, over
5 years



Data on drug users in Colorado Springs, over 5 years

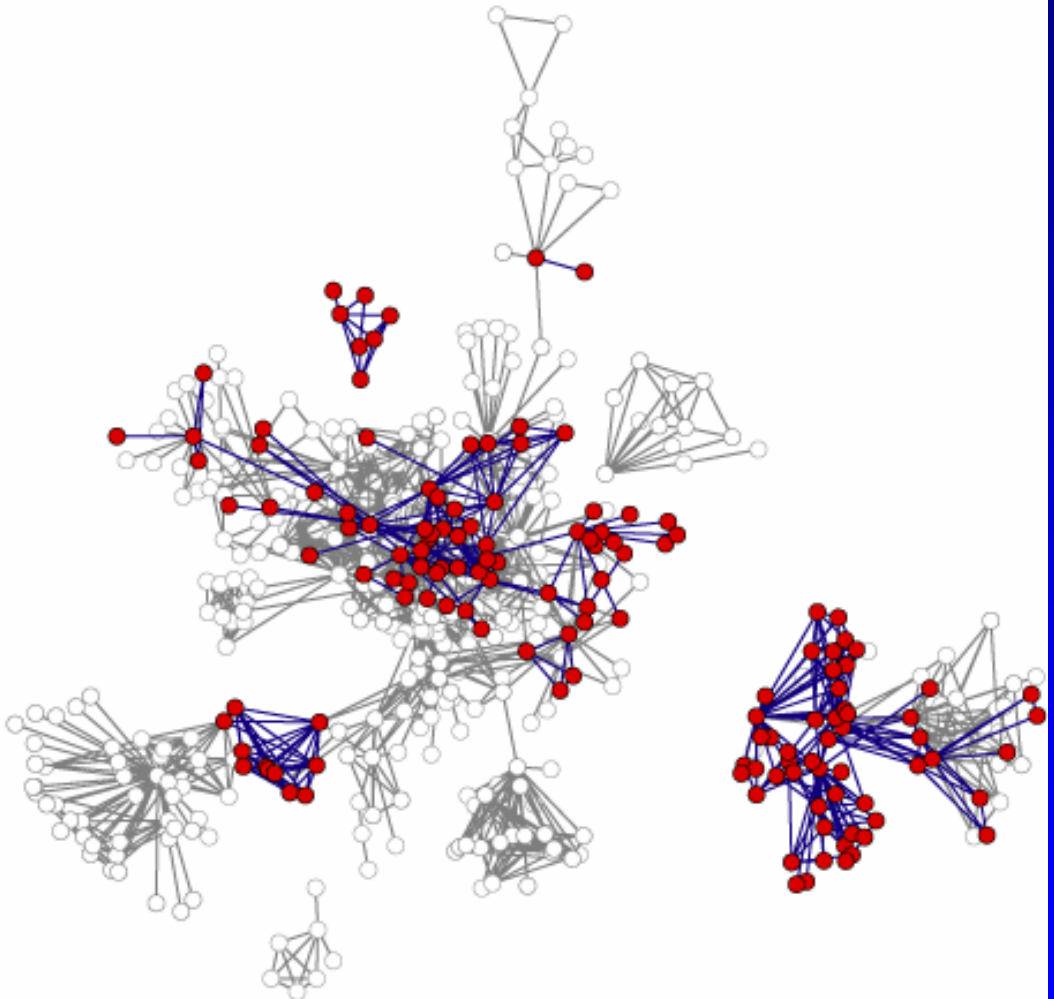
Drug Relations
Colorado Springs Data
Year 2



Year 2 points in red, previous points in gray

Data on drug users in Colorado Springs, over 5 years

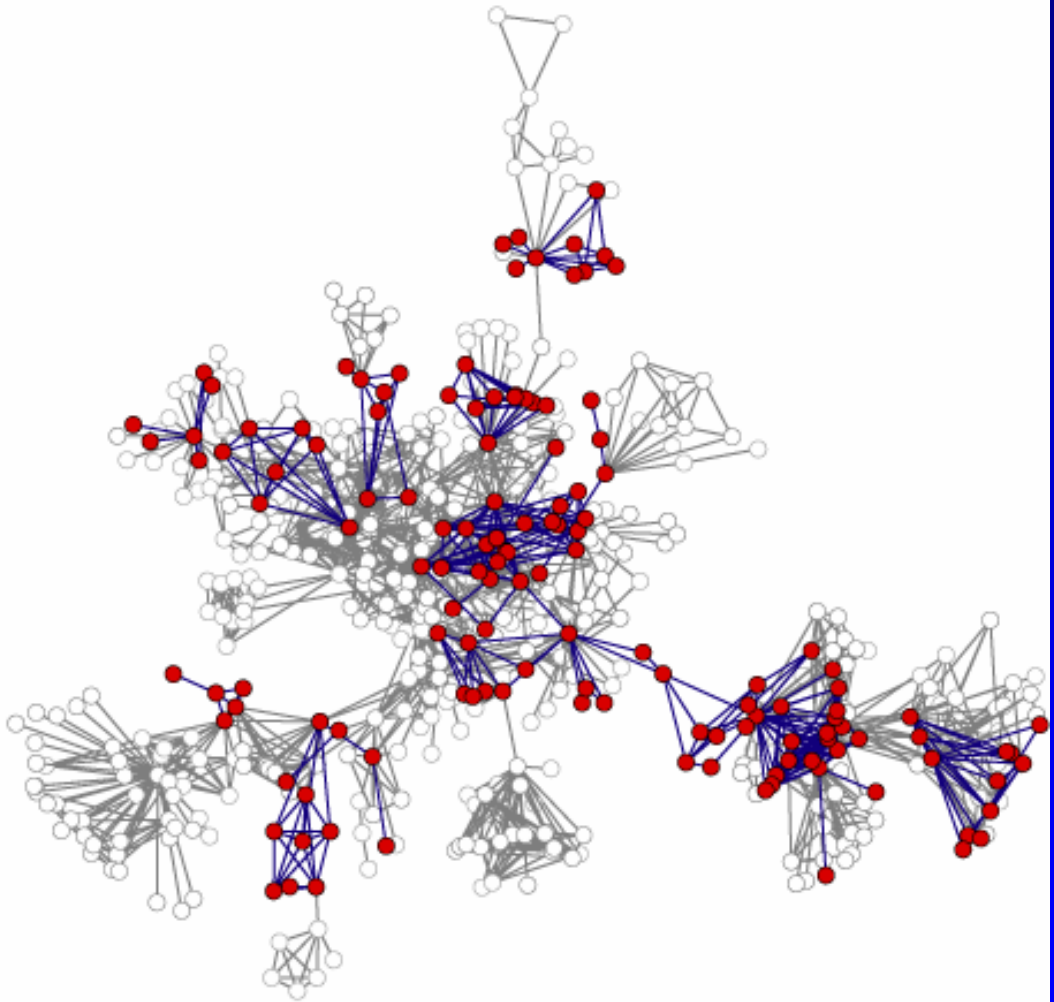
Drug Relations
Colorado Springs Data
Year 3



Year 3 points in red, previous points in gray

Data on drug users in
Colorado Springs, over
5 years

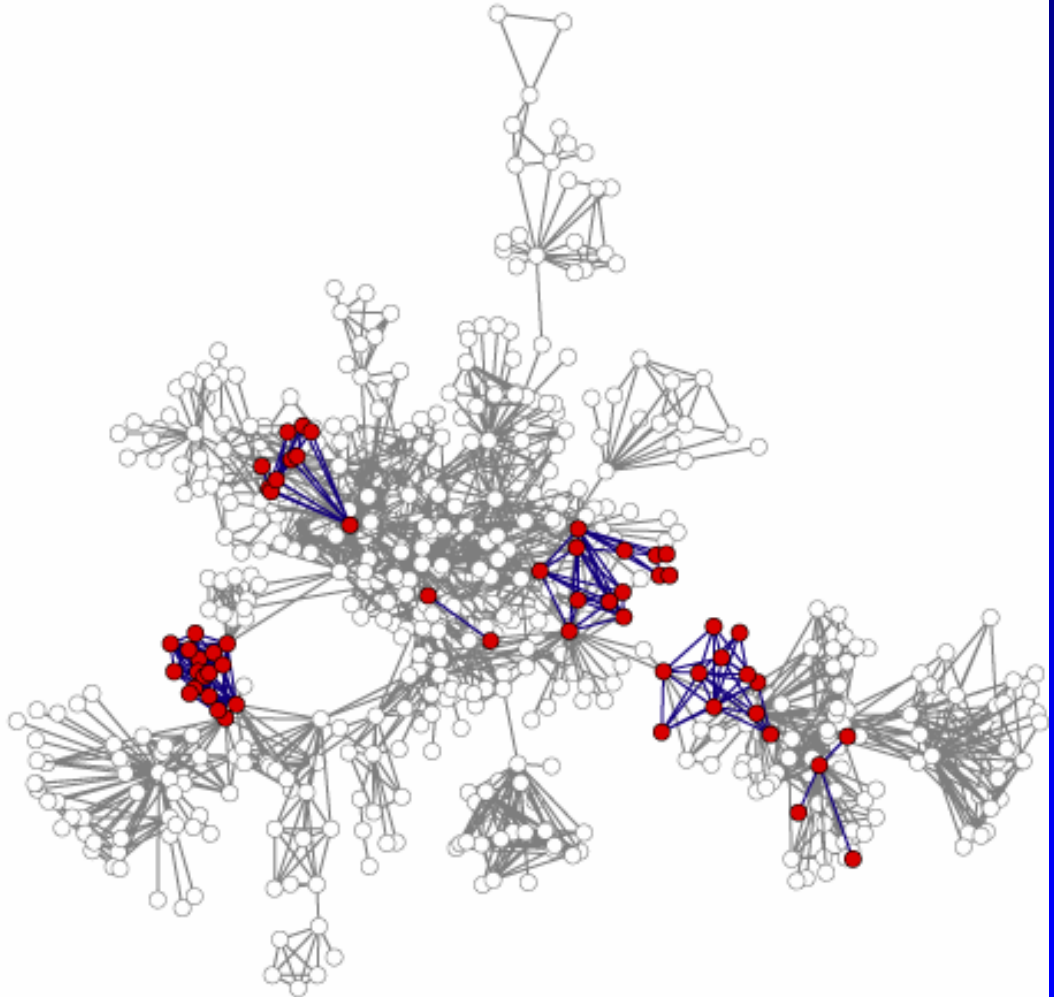
Drug Relations
Colorado Springs Data
Year 4



Year 4 points in red, previous points in gray

Data on drug users in
Colorado Springs, over
5 years

Drug Relations
Colorado Springs Data
Year 5



Year 5 points in red, previous points in gray

Some Networks types receiving attention

- Social networks
 - interpersonal relations (kinship, roles, cognitive, affiliations)
 - transactions
- Information networks
 - WWW, citations of scientific articles, recommender systems
- Technology networks
 - typically designed for distribution
 - electric power grid, roads, Internet
- Biological networks
 - used to represent biological systems
 - metabolic pathways, genetic regulatory networks, food web

Perspectives to keep in mind

- Network-specific verses Population-process
 - *Network-specific*: interest focuses only on the actual network under study
 - *Population-process*: the network is part of a population of networks and the latter is the focus of interest
 - the network is conceptualized as a realization of a social process
- Social relations verses Nodal Attribute index
 - Interdependence/endogeneity verses atomistic essentialism
 - structure verses composition

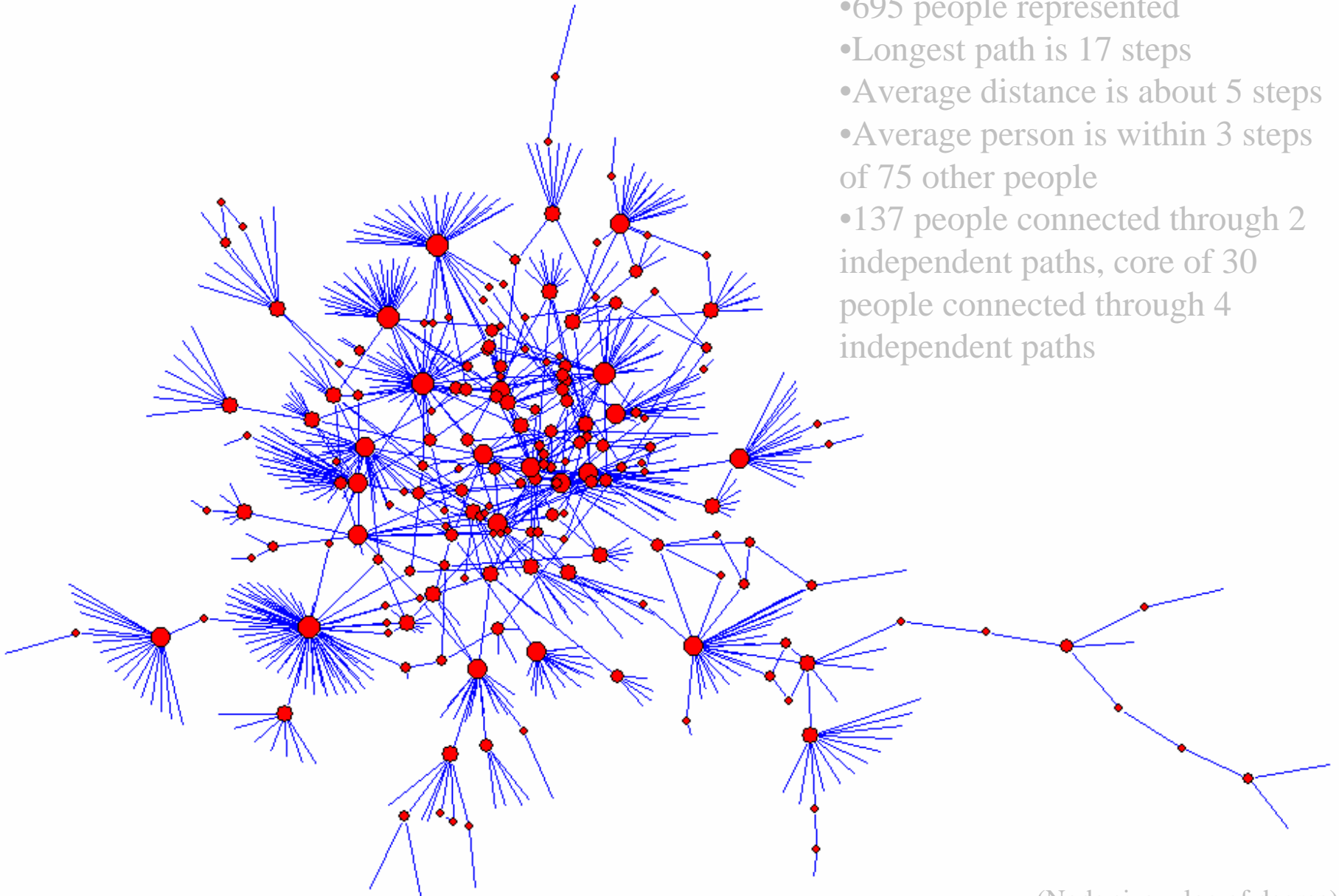
What do social networkers study?

- **Flows within Networks:** diffusion, transmittal
 - *Topology:* structure or shape of the network
 - * *Connectivity:* between pairs of actors
 - *reachability:* Is there a path between given actors?
 - *geodesic distance:* If reachable, how long is the path?
 - *Number of paths:* How many paths of each length?
 - * *Centrality:* concepts of a node's location within the network
 - * *Activity:* the distribution of the number of edges of each node
 - *Timing:* Network topology changes over time
 - * timing of edge formation and dissolution
 - * flow very sensitive to timing

Reachability in Colorado Springs

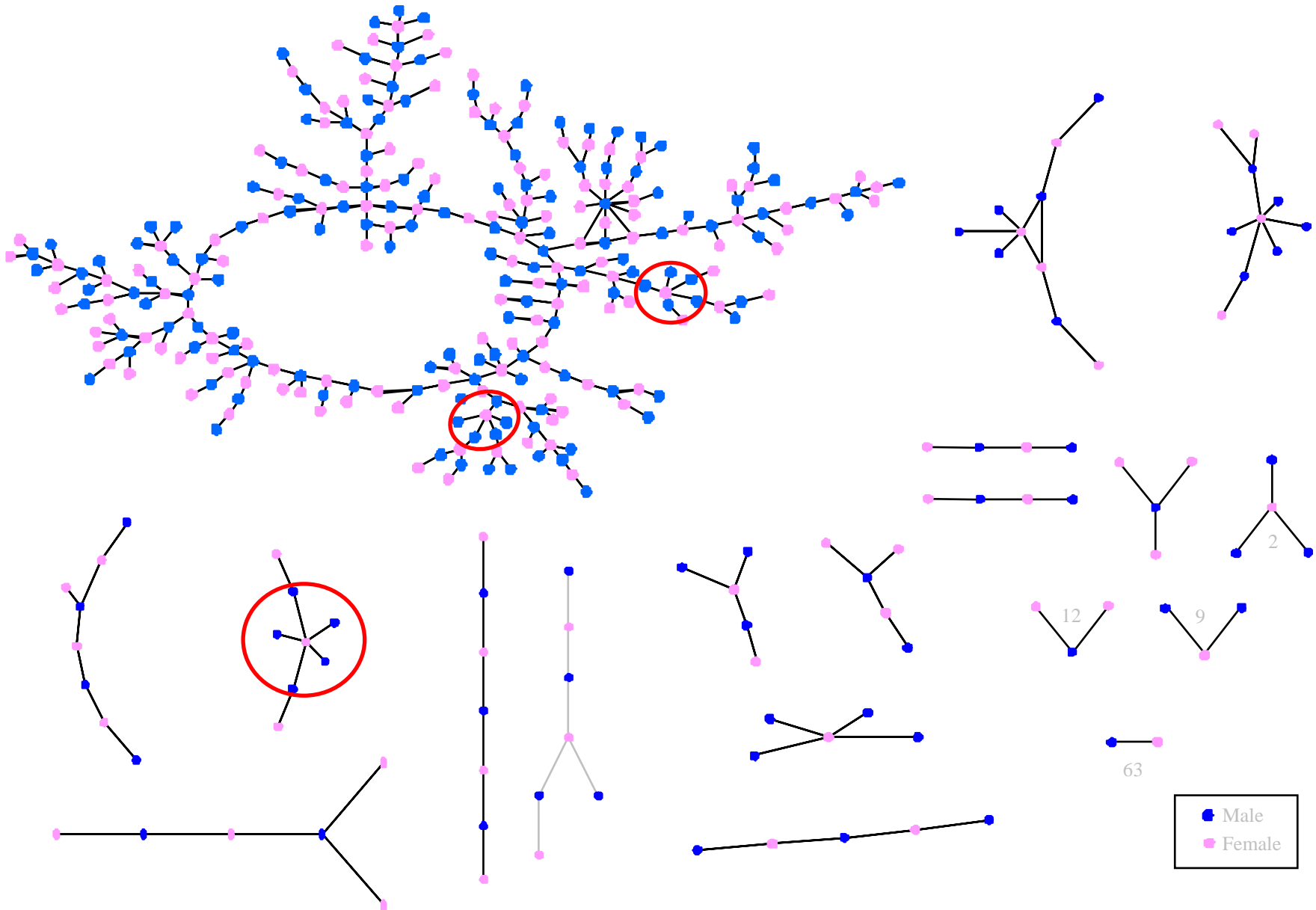
(Sexual contact only)

- High-risk actors over 4 years
- 695 people represented
- Longest path is 17 steps
- Average distance is about 5 steps
- Average person is within 3 steps of 75 other people
- 137 people connected through 2 independent paths, core of 30 people connected through 4 independent paths



(Node size = log of degree)

Reachability example: All romantic contacts reported ongoing in the last 6 months in a moderate sized high school (AddHealth)

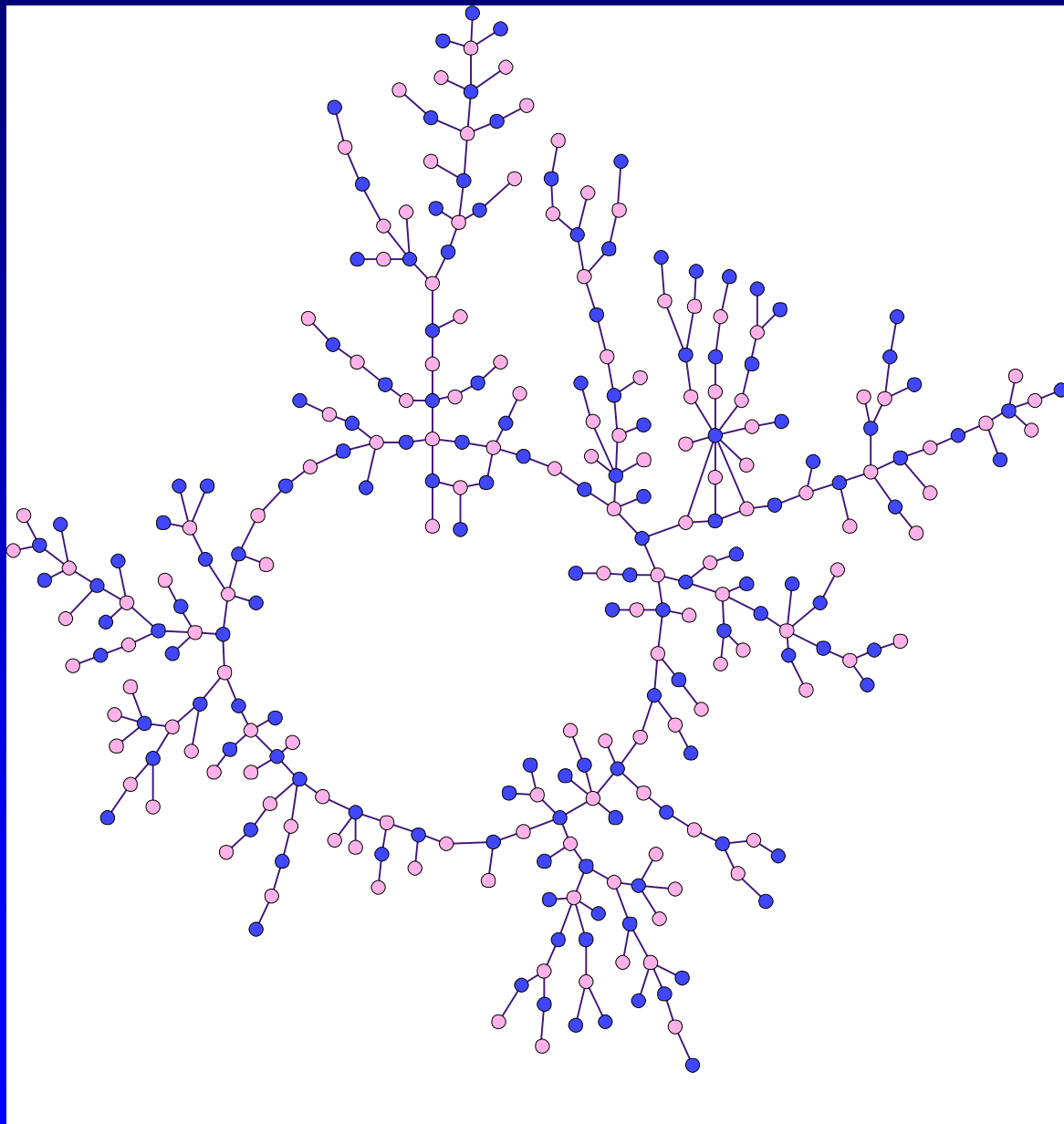


(From Bearman, Moody and Stovel, n.d.)

Mark S. Handcock

Slide courtesy of Jim Moody

Social Networks Overview

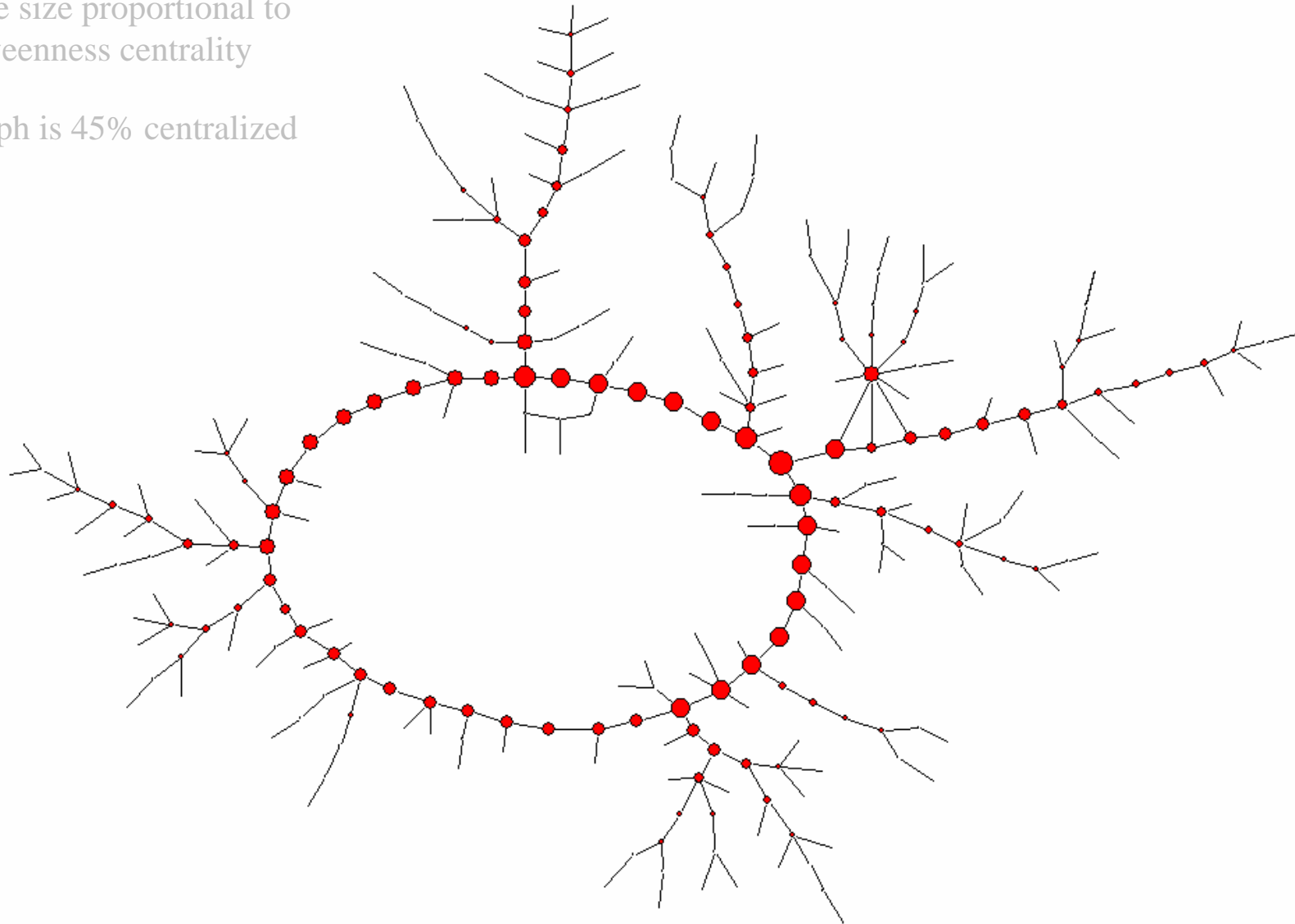


- 288 People in largest component
- 42 steps maximum distance
- Mean distance between non-connected pairs is 16 steps
- Mean number within 3 steps is: 9.7
- 45 people are biconnected (in the center ring).

Centrality example: Add Health

Node size proportional to
betweenness centrality

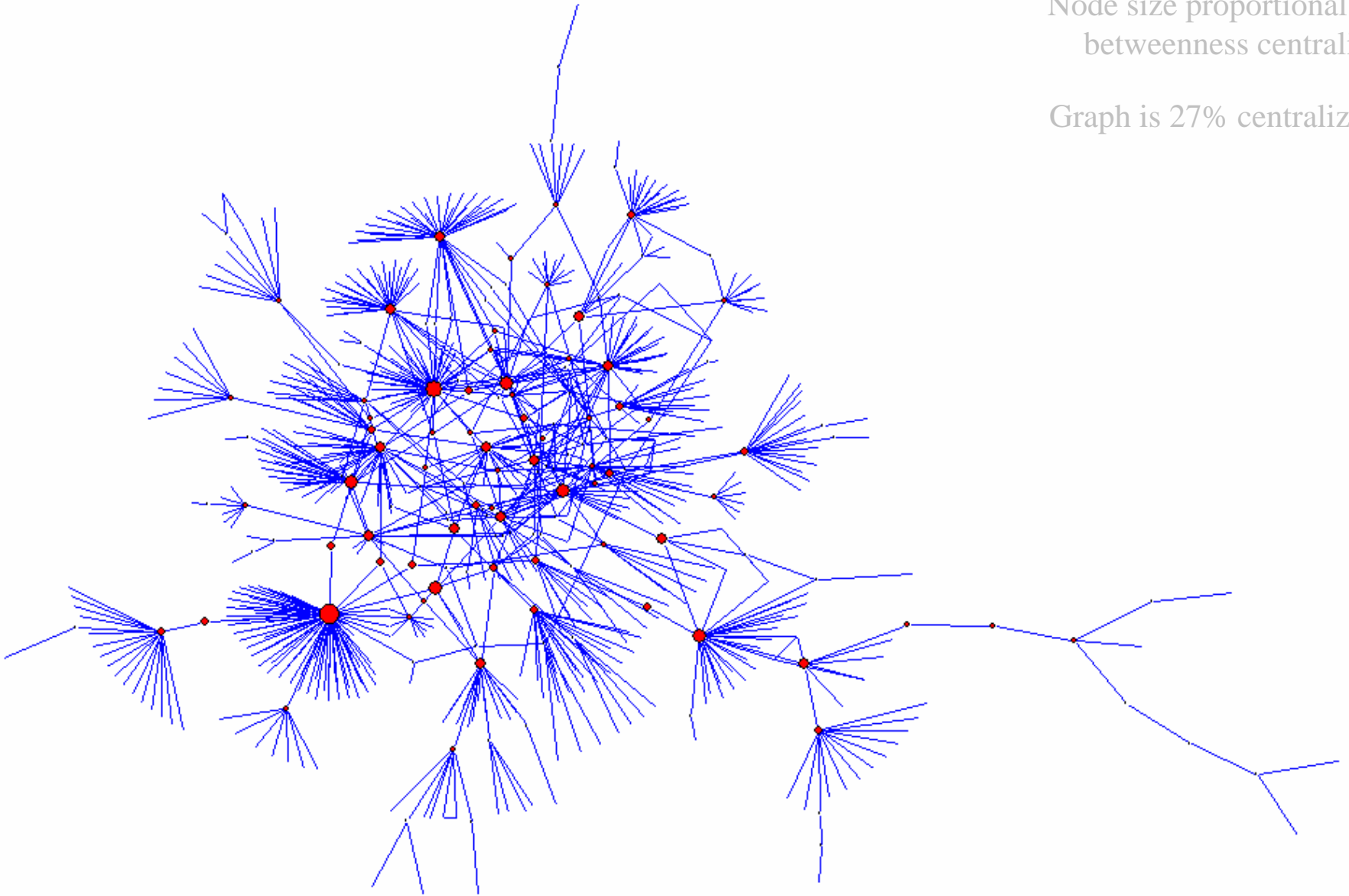
Graph is 45% centralized



Centrality example: Colorado Springs

Node size proportional to
betweenness centrality

Graph is 27% centralized



- **Structure of Social Space**

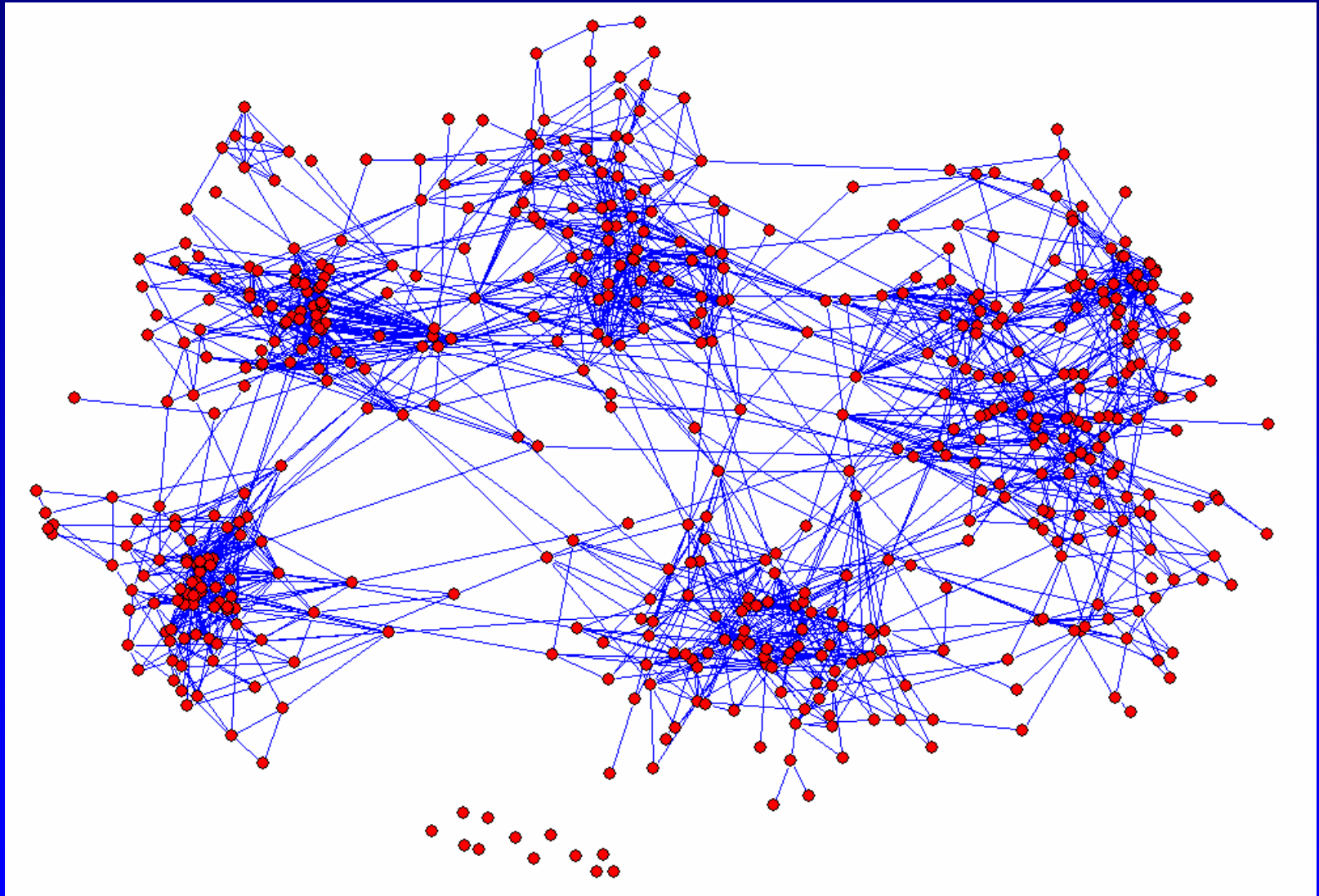
- *Cohesive groups*: “clustering” of nodes based on relatively strong internal relations
 - * insiders, outsiders, bridges
- *Attribute based mixing*: ties between exogenous types of nodes
- *Hierarchy of nodes*: positions within a network indicated via patterns of directional ties

- **Roles**: endogenous latent types that partition the social system

- Roles indicated by patterns of relationships
- Two actors are *structurally equivalent* if they have the same types of ties to the same (types of) people.

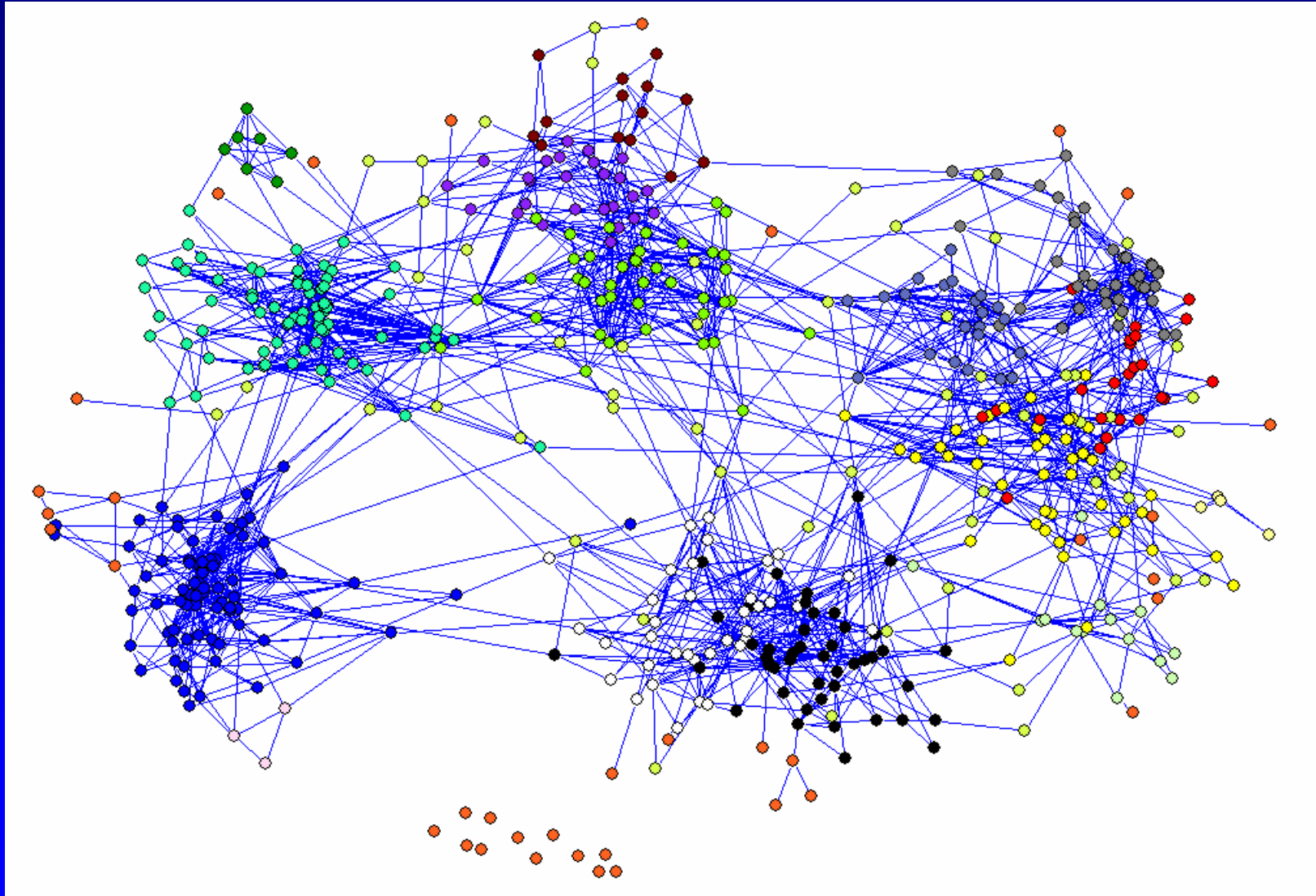
Cohesive Group Structure

“Immaculate Preparatory High School”



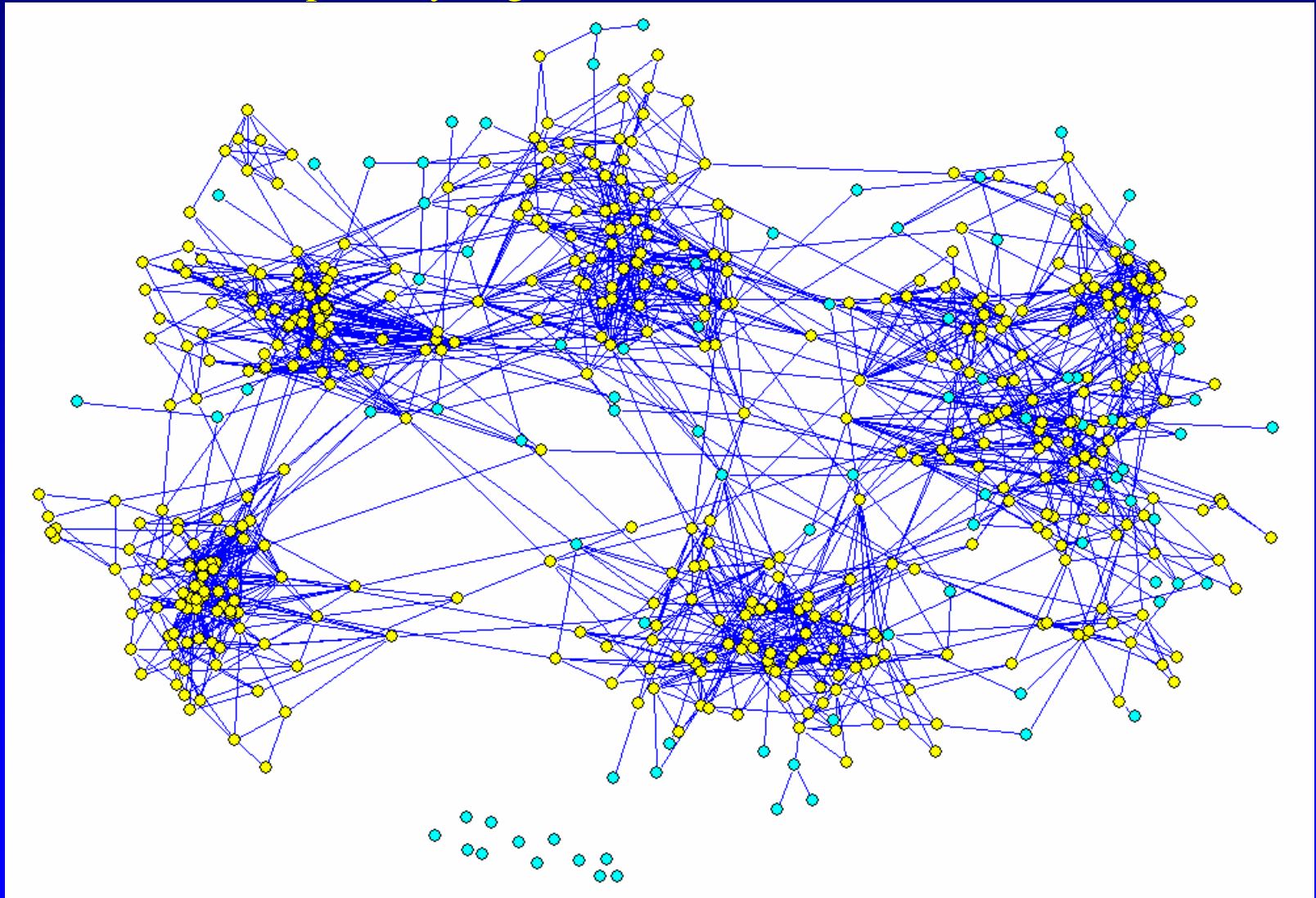
Cohesive Group Structure: 3 types of positions

“Immaculate Preparatory High School”



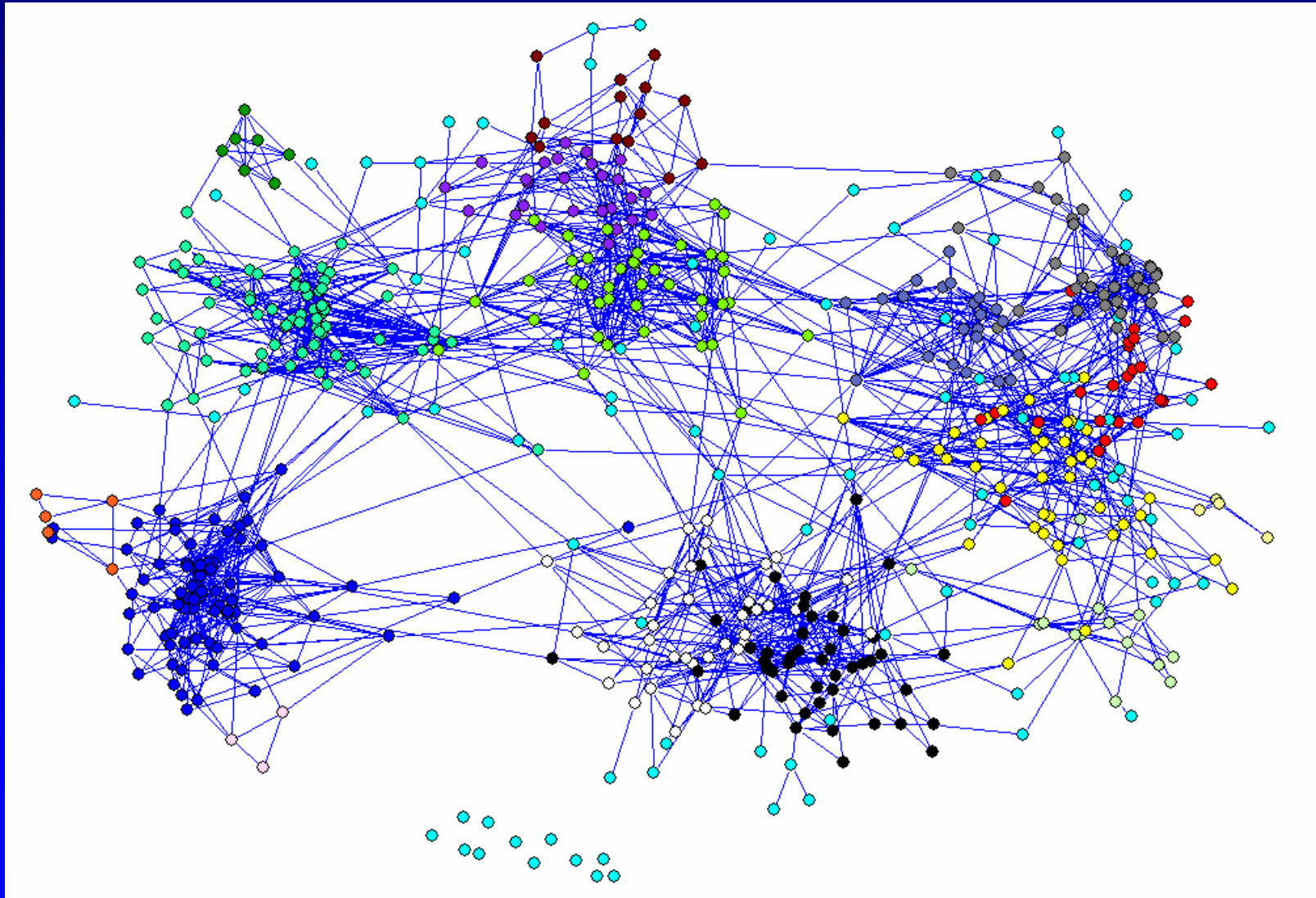
Cohesive Group Structure: Group member

“Immaculate Preparatory High School”



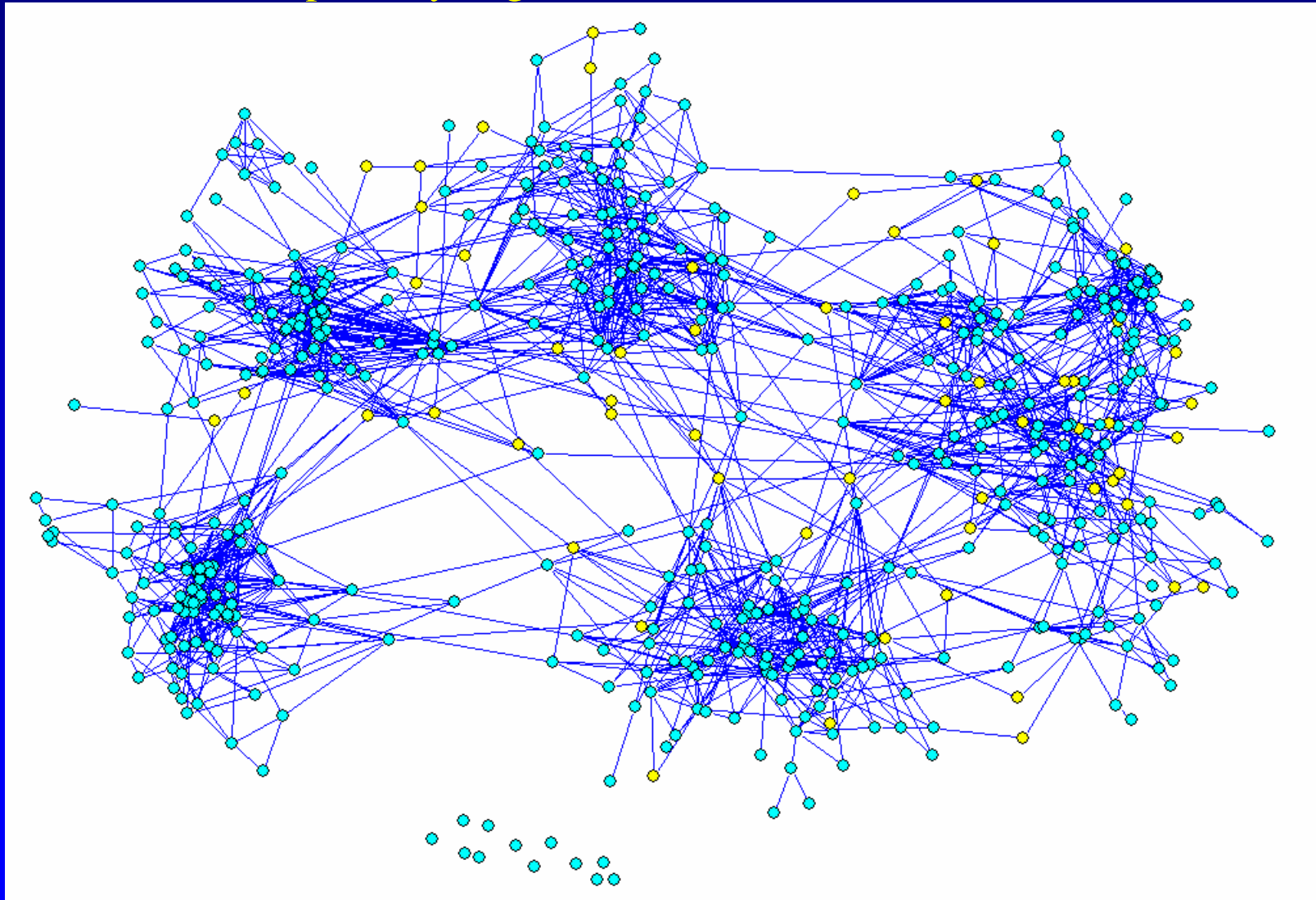
Cohesive Group Structure: Group Member

“Immaculate Preparatory High School”



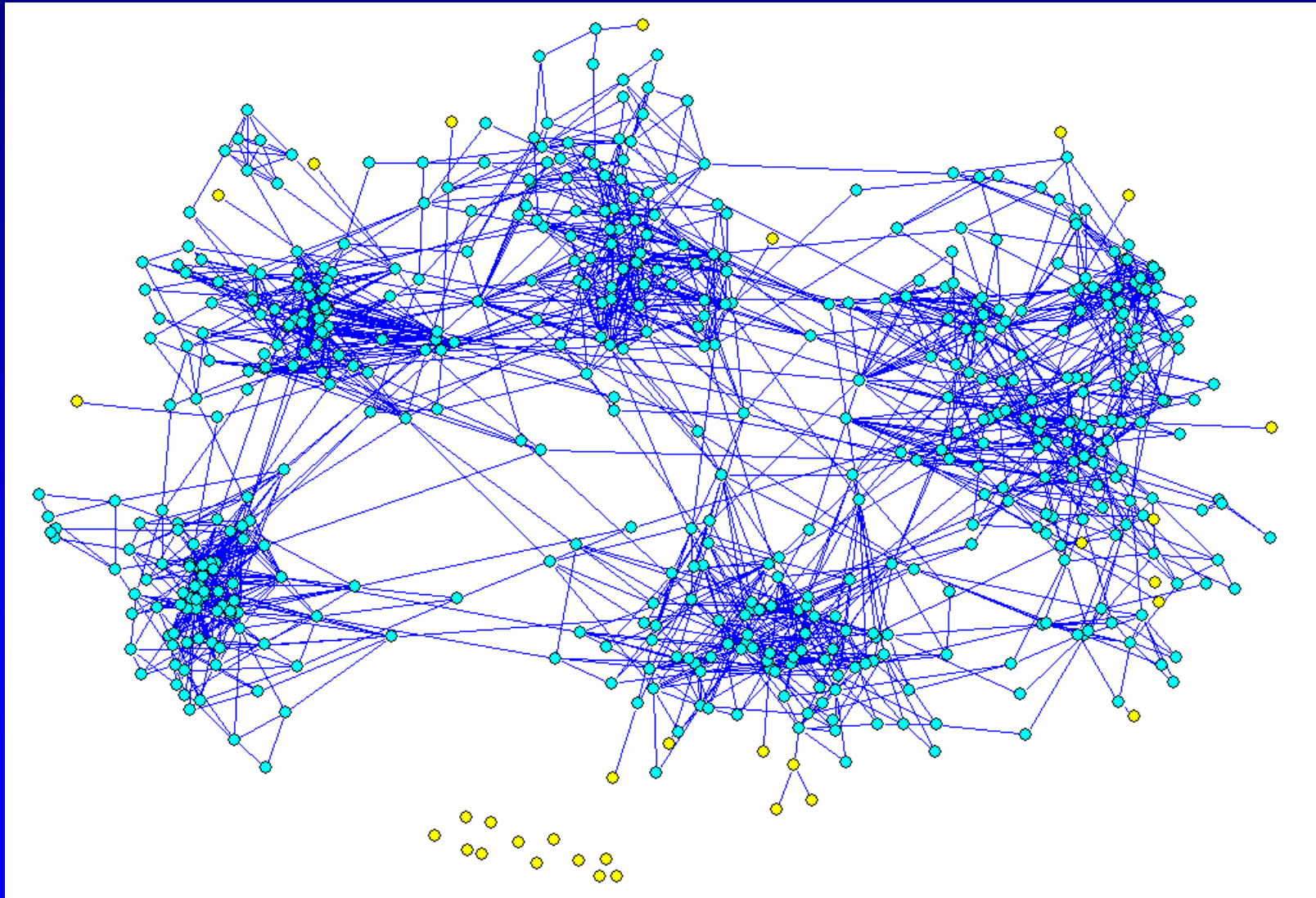
Cohesive Group Structure: Bridge between groups

“Immaculate Preparatory High School”



Cohesive Group Structure: Outsider

“Immaculate Preparatory High School”



Models for Social Networks

A *social network* is defined as a set of g social “actors” and a social relationship between each pair of actors.

- To fix ideas and easy of presentation:
 - label the actors $1, \dots, g$.
 - focus on a single binary relationship
 - ⇒ valued and metric relationships ok
 - the number of actors g is fixed and known
 - ⇒ can be generalized
 - The presence or absence of a relationship is observed for each pair of actors
 - ⇒ census: no sampling or missing data
 - cross-sectional (time aggregate/static) viewpoint

$$X_{ij} = \begin{cases} 1 & \text{relationship from actor } i \text{ to actor } j \\ 0 & \text{otherwise} \end{cases}$$

- call $X \equiv [X_{ij}]_{g \times g}$ a *sociomatrix*;
call the graphical representation of X a *sociogram*
 - a $N = g(g - 1)$ array of binary random variables
 - X represents a random network with nodes the actors and edges the relationship
- The basic problem of stochastic modeling is to specify a distribution for X i.e., $P(X = x)$

Random Graph Distributions

Let \mathcal{X} be the sample space of X e.g. $\{0, 1\}^N$

Any model for the multivariate distribution of X can be *parametrized* in the form:

$$P(X = x) = \frac{\exp\{\theta^T t(x)\}}{c(\theta)} \quad x \in \mathcal{X}$$

Besag (1974), Bahadur (1961), Frank and Strauss (1986)

- $\theta \in \Theta \subset R^q$ q -vector of parameters
- $t(x)$ q -vector of *network statistics*.
- For a “saturated” model $q = 2^N - 1$
- $c(\theta)$ distribution normalizing constant

$$c(\theta) = \sum_{x \in \mathcal{X}} \exp\{\theta^T t(x)\}$$

Simple models for social networks

- Bernoulli graph
 - X_{ij} are independent but have arbitrary distributions

$$P(X = x) = \frac{\exp \left\{ \sum_{i,j} \theta_{ij} x_{ij} \right\}}{c(\theta)} \quad x \in \mathcal{X}$$

$$t_{i,j}(x) = x_{ij}, \quad i, j = 1, \dots, g \quad q = N$$

$$\theta_{ij} = \text{logit}[P(X_{ij} = 1)]$$

$$c(\theta) = \prod_{i,j} [1 + \exp(\theta_{ij})]$$

- Homogeneous Bernoulli graph (Renyi-Erdos model)
 - X_{ij} are independent and equally likely with log-odds $\theta = \text{logit}[P(X_{ij} = 1)]$

$$P(X = x) = \frac{e^{\theta \sum_{i,j} x_{ij}}}{c(\theta)} \quad x \in \mathcal{X}$$

where $q = 1$, $t(x) = \sum_{i,j} x_{ij}$, $c(\theta) = [1 + \exp(\theta)]^g$

- homogeneity means it is unlikely to be proposed as a model for real phenomena

Some history of models for social networks

Holland and Leinhardt (1981) proposed a general dyad independence model

– Also an homogeneous version they refer to as the “ p^1 ” model

$$P(X = x) = \frac{\exp\{\rho \sum_{i < j} x_{ij} x_{ji} + \theta x_{++} + \sum_i \alpha_i x_{i+} + \sum_j \beta_j x_{+j}\}}{c(\rho, \alpha, \beta, \theta)}$$

where

- θ controls the expected number of edges
- ρ represent the expected tendency toward *reciprocation*
- α_i *productivity* of node i ; β_j *attractiveness* of node j
- Much related work and generalizations
 - Wasserman (1980), Fienberg, Meyer, and Wasserman (1985), Wasserman and Faust (1994), Wasserman and Pattison (1996)
 - Frank and Strauss (1986)

Models based on the degree distribution only

$$P(X = x) = \frac{\exp\{\sum_{k=1}^{g-1} \alpha_k d_k(x)\}}{c(\alpha)} \quad x \in \mathcal{X}$$

where

- $d_k(x)$ = the proportion of actors with exactly k relationships
- α g -vector of degree parameters
 - Long-history in social network community
 - Wasserman (1977), Wasserman and Faust (1994), Snijders (1991)
 - Recent focus
 - Barabási & Albert (1999), Newman, Strogatz and Watts (2001)
 - Further direct parametrization of the degree distribution
 - Forms with power-law behavior: Albert and Barabási (1999)

A more realistic model for social networks

$$P(X = x) = \frac{\exp\{x^T Z\beta + \sum_{k=1}^{g-1} \alpha_k d_k(x) + \theta^T t(x)\}}{c(\alpha, \beta, \theta)} \quad x \in \mathcal{X}$$

where

- x is the N -vector of the unique elements of X
- $Z = \{z_{ij}\}_{N \times p}$ matrix of (exogenous) covariates on the ij^{th} dyad
- $d_k(x) =$ the proportion of actors with exactly k relationships
- $t(x)$ q -vector of additional *network statistics*
- β p -vector of regression parameters
- α g -vector of degree parameters
- θ q -vector of network structure parameters

Usually the additional statistics kept “simple”

– e.g., *clustering* or *transitivity*

If we add the *proportion of triangles amongst triads*

$$t(x) = \frac{1}{\binom{g}{3}} \sum_{\{i,j,k\} \in \binom{g}{3}} x_{ij}x_{ik}x_{jk}$$

these are the homogeneous nodal Markov graphs of Frank and Strauss (1986):

– edges in X that do not share an actor are conditionally independent given the rest of the network

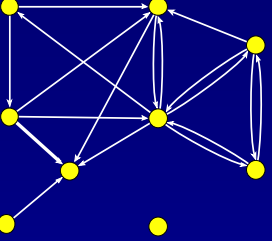
⇒ analogous to nearest neighbor ideas in spatial statistics

A smattering of other models

- Models of individual behavior that incorporate network characteristics
 - Network regressive-autoregressive models (Doreian)
 - Peer influence models (Friedkin)
- Spatial Models of Large-Scale Social Networks
 - Butts (2002)
- Measurement error models for informant reports
 - Killworth and Bernard (1976), Bernard et al (1984)
 - Romney and Faust (1982), Krackhardt and Kilduff (1999), Butts (2003)
- Random effects models: positing latent social structure
 - Latent class models: Nowicki and Snijders (2001)
 - Latent space models: Hoff, Raftery and Handcock (2001)

Models for dynamic social networks

- Continuous-time Markov models
 - Wasserman (1977), Holland and Leinhardt (1977)
 - Leenders (1995)
- Actor Oriented: fusion of rational choice and Continuous-time Markov models
 - Snijders (1996) (1977), Leenders (1995)
- Models of Network Growth
 - Motivated by Simon (1955)
 - Citation networks: Price (1965; 1976)
 - WWW: Albert and Barabási (1999)



Software for Social Network Analysis

INSNA links to network analysis software packages:

http://www.heinz.cmu.edu/project/INSNA/soft_inf.html

1) UCI-NET

- General Network analysis program, runs in Windows
- Good for computing measures of network topography for single nets
- Input-Output of data is a little clunky, but workable.
- Not optimal for large networks
- Available from:

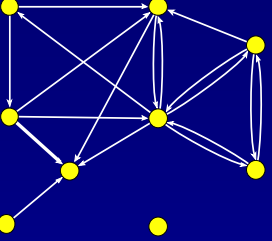
Analytic Technologies

Borgatti@mediaone.net

2) STRUCTURE

- “A General Purpose Network Analysis Program providing Sociometric Indices, Cliques, Structural and Role Equivalence, Density Tables, Contagion, Autonomy, Power and Equilibria In Multiple Network Systems.”
- DOS Interface w. somewhat awkward syntax
- Great for role and structural equivalence models
- Manual is a very nice, substantive, introduction to network methods
- Available from a link at the INSNA web site:

http://www.heinz.cmu.edu/project/INSNA/soft_inf.html



Software for Social Network Analysis

- 3) StOCNET <http://stat.gamma.rug.nl/stocnet/>
 - Actor-oriented, block and most of Tom Snijder's models

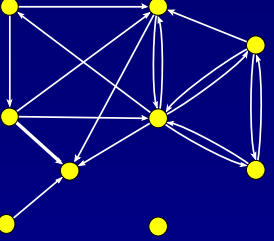
- 4) PREPSTAR <http://kentucky.psych.uiuc.edu/pstar/index.html>
 - Maximum Pseudo-likelihood estimates (MPLE) of exponentially parameterized random graph models

- 5) SNA: R package for Social Network Analysis
 - Numerical measures, etc
 - CRAN – Offered by Carter Butts

- 6) ERGM: R package
 - Fits exponentially parameterized random graph models
 - Calculates MPLE, MLE, and MAP
 - Latent space models
 - GRAPH: R package specifying R class for graph objects

Visualization of dynamic social networks

- Benefits:
 - Intuitive way to display networks (Moreno 1932; 1934)
 - Helps people see a map of social space
 - A concise presentation of a great deal of data.
- Costs:
 - Lack of standards for how to display can create misleading images
 - Displays of large networks tend to reveal only the roughest properties of the network
- Linton Freeman has contributed greatly to this
 - For history and development of the field, see <http://eclectic.ss.uci.edu/~lin/gallery.html>

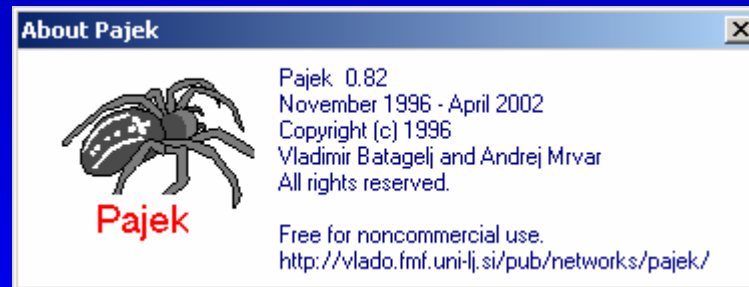


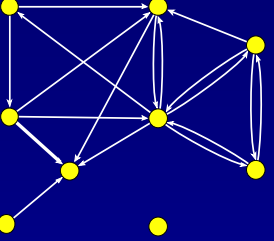
Tools, Methods & Models

Graphical Display: Software

PAJEK

- Program for analyzing and plotting very large networks
- Intuitive windows interface
- Used for most of the real data plots in this presentation
- Mainly a graphics program, but is expanding the analytic capabilities
- Free
- Available from:



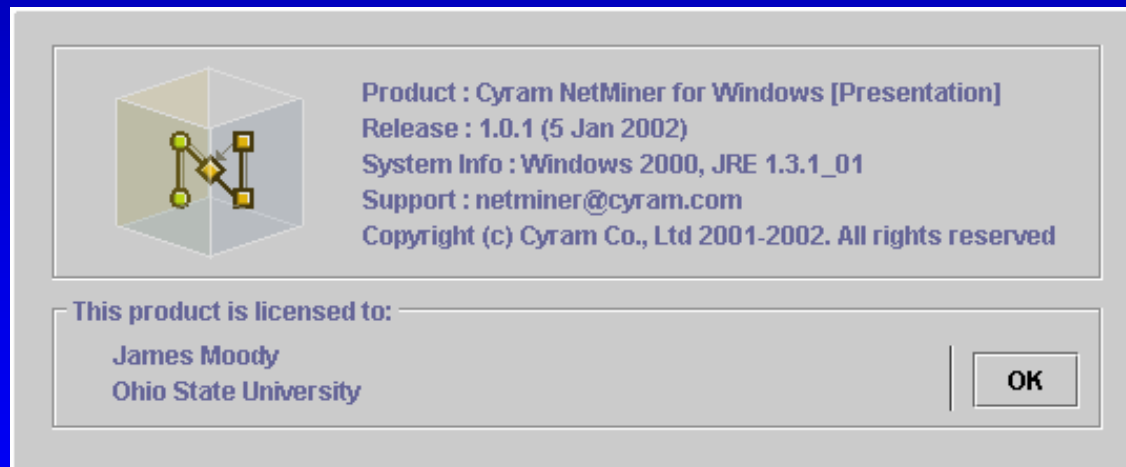


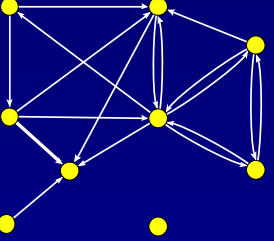
Tools, Methods & Models

Graphical Display: Software

Cyram Netminer for Windows

- Very new: largely untested
- Price range depends on application
- Limited to smaller networks $O(100)$





Tools, Methods & Models

Graphical Display: Software

NetDraw

- Also very new, but by one of the best known names in network analysis software.
- Free
- Limited to smaller networks $O(100)$



Some Key references:

- Leinhardt (1977) Book: “Social Networks: A Developing Paradigm”
- Holland and Leinhardt (1981) JASA
- Frank and Strauss (1986) JASA
- Wasserman and Faust (1994) Book
- Doreian and Stokman (1997) Book: “Evolution of Social Networks”
- Besag (1974-2000) JRSSB, etc
- Newman (2003) SIAM Review