An Introduction to Social Network Analysis

Nema Dean

School of Mathematics & Statistics,
University of Glasgow

Glasgow Social Statistics Group
18th October 2012

What is Social Network Analysis?

Simple defintion: Methodical analysis of social networks
The analytical study of (social) relations among a set of actors

What is a social network?

- Any set of data where you have information about the relationships between objects
- Objects can be people, postcodes, corporations, molecules anything really
- Can have information on the object level as well (attribute information)
- But emphasis is on relationships, associations, dependencies

What is Social Network Analysis?

Often think of Facebook and similar sites when social networks are mentioned but this is only one example of social network data

- The term “social network” used loosely to describe complex sets of relationships between individuals or groups
- The methodologies for analysing social networks fall under the term Social Network Analysis (SNA)
- Used in anthropology, biology, communication studies, economics, geography, information science, marketing, organizational studies, social psychology, and sociolinguistics

Terminology

A lot of methodology and models for SNA arose from graph theory so graph terminology is quite common

- Nodes or Vertices - objects in the network under study
- Edges - relationships between objects

You will also see the term *actor* used interchangeably with node and *ties* used interchangeably with edges.

Edges

- Edges are often binary,
  i.e. a tie exist between two actors or not {0,1}
- Alternatively can be weighted, counts, any form
- Also can have multiple graphs defined on the same set of nodes (where each graph represents measurement of a different type of relationship)
- Edges can be reciprocal, also known as undirected edges/ties, or not (known as directed)
- A pair of nodes is often known as a *dyad*

Examples

Social relations can be thought of as dyadic attributes (regular analysis works with monadic attributes)

- Examples of such relations include:
  - Kinship: brother of, father of
  - Social Roles: boss of, teacher of, friend of
  - Affective: likes, respects, hates
  - Cognitive: knows, views as similar
  - Actions: talks to, has lunch with, attacks
  - Flows: number of cars moving between
  - Distance: number of miles between
  - Co-occurrence: is in the same club as, has the same color hair as
  - Mathematical: is two links removed from

What is Social Network Analysis?

Two kinds of network data leading to different kinds of analysis

- Ego network analysis: random sampling,
  looks at the quality of a person's network
  - Convenient and allows classical statistical techniques to be applied
- Complete network analysis: all relationships in a set of respondents
  - Requires new types of analysis and metrics

Problems in Social Network Analysis

- Prior to computing boom - not possible to test statistical hypotheses on complete graphs, now possible
- In spite of growth in computing power, large datasets (order of 1,000's nodes) are still problematic to analyse
- Creates a problem:
  - Could artificially bound network but can distort the data
  - Without bounds network may get too large to be processed

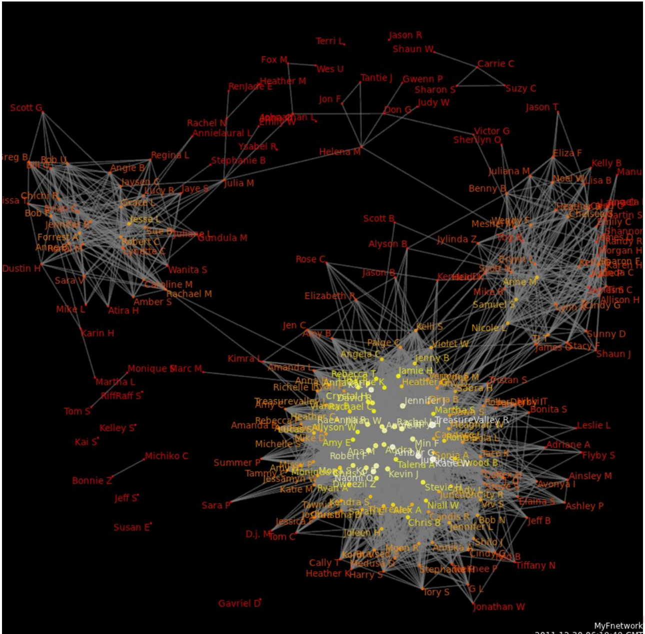
Storing Network Data

- As well as visualizing social networks as a graph (or sociogram) we also work with them in different forms
- One such form: Sociomatrix (matrix representation), also known as adjacency matrix
  - $N \times N$ matrix, N = number of nodes/actors
  - $\{i, j\}^{th}$ element gives information about the (directed) edge from the i^{th} to the j^{th} actor
  - 0 usually indicates absence of an edge but the values depend on the type of relationship being measured
- Undirected graphs will result in symmetric sociomatrices
- Other means of recording graph is a list of edges/dyads

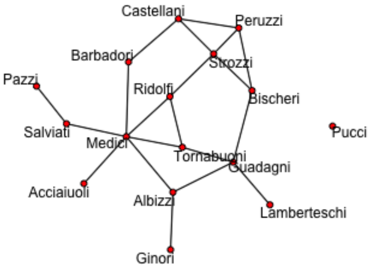
Interesting Examples

- HIV transmission: actors are people, edges could be needle-sharing or unprotected sex
- International Relations network: countries are nodes, edges are number of interactions over a certain period
- Facebook: nodes are accounts (not necessarily unique individuals) and edges exist between nodes if the accounts are friends

Facebook Example

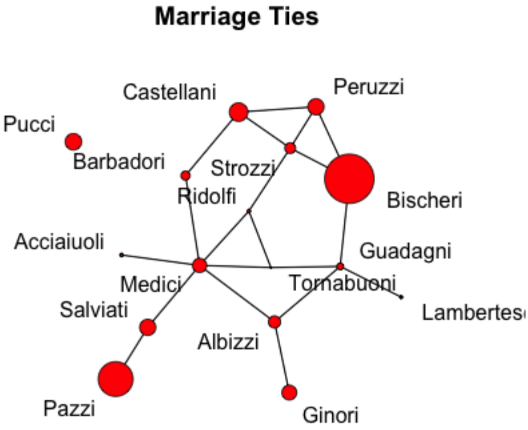


Florentine Marriage Data



Looks at the marriage links between 16 prominent Florentine families in the 15th century
(Actors = families, (undirected) ties exist if a marriage exists between two families)

Florentine Marriage Data



Node size proportional to family wealth

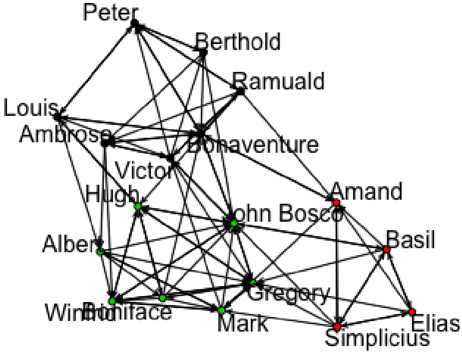
Sampson's Monk Data

- In 1969 Samuel Sampson recorded the social interactions among a group of monks while resident as an experimenter on vision, and collected numerous sociometric rankings.
- During his stay, a political "crisis in the cloister" resulted in the expulsion of four monks and the voluntary departure of several others.
- Data on positive affect relations ("liking"), in which each monk was asked if they had positive relations to each of the other monks
- Data from 3 different time periods, capturing change in group sentiment over time

Sampson's Monk Data

- Each member ranked only his top three choices on "liking." (Some subjects offered tied ranks for their top four choices). A (directed) tie from monk A to monk B exists if A nominated B as one of his three best friends at that that time point.
- Other information:
  - Groups of novices as classified by Sampson: "Loyal", "Outcasts", and "Turks"
  - An indicator if attendance the minor seminary of "Cloisterville" before coming to the monastery

Sampson's Monk Data



Black=loyal, red=outcasts, green=Turks

Why is Social Network Analysis different?

- Ordinarily dependencies between objects of study are a nuisance, make life difficult
- Try to make assumptions of conditional independence or ignore them to make them go away
- In SNA, dependencies are the focus of the analysis, so can't make simplifying assumptions/ignore them
- Naive approach to modelling binary edges $\Rightarrow$ logistic regression but this assumes edges are conditionally independent
- Naive approach leads to bias

Networks summaries/metrics of connectivity

- Homophily - tendency for similar (versus dissimilar) nodes to have edges
  - Examples of characteristics on which similarity can be measured:
    - Gender,
    - Race,
    - Sex,
    - Age,
    - Educational achievement, etc.
- Reciprocity (or Mutuality) - only defined really for graphs with directed edges, tendency edges that are reciprocal
- Transitivity - tendency to form triangles, e.g. friend of a friend is also a friend

Metrics of network distribution

- Bridge - individual with few ties but is the only link between two or more larger groups
- Centrality - measure of “importance” of a node,
  - Degree centrality - count of number of in or out (or both) edges a node has
  - Betweenness centrality - the number of shortest paths for all pairs of nodes that pass through that node
  - Alpha centrality, closeness centrality, eigenvector centrality, etc.
- Density - proportion of edges observed relative to the number possible for that number of nodes

Metrics of network distribution

- Distance - Minimum number of edges needed to connect 2 nodes - ‘six degrees of separation’
- Structural holes - absences of edges between 2 or more parts of a network
- Edge strength - number of different definitions depending on type of network

Segmentation metrics

- Clique - set of nodes with every node having an edge connecting it to every other node in the clique
- Clustering coefficient - measure of likelihood that 2 nodes connecting to one node are themselves connected by an edge
  Higher clustering coefficient implies greater cliqueishness
- Structural Cohesion - minimum number of actors that would have to be remove to disconnect the network

Components

- A component is a the maximum set of nodes where all pairs of nodes are connected by a path
- Connected graph: one component

SNA Goals

- The previous metrics useful for measuring properties of a network/node/subset of nodes
- Summarizing network patterns
  In general, social networks are self-organizing, emergent, and complex,
  local interaction of the elements $\Rightarrow$ globally coherent pattern
- Identifying important nodes (e.g. bridges or nodes with high centrality)
- Finding groups of highly connected nodes
- Characterizing the likelihood of edges in terms of node attributes and other edge patterns

SNA Goals

- We want MORE!
- We also want to model networks in a principled way
- Why?
  - Social behaviour - very complex $\Rightarrow$ stochastic models capture both the regularities in the process producing the network and also allows for variability
  - Can understand the uncertainty associated with the observed outcomes
  - Can estimate the parameters of the hypothesised model proposed for the data generation process
  - Can make inference about whether certain sub-structures are observed more commonly than expected by chance and can develop hypotheses about the social processes that might produce these
  - Example: clustering in a network may come about due to homophily or from endogenous structural effects, to decide between the two possible causes, need a model incorporating both effects which can be then assessed

Exponential Random Graph Models

- Exponential Random Graph Models (ERGMs): One of the most powerful, commonly used set of statistical models for networks (Frank & Strauss 1986)
- Also know as the p^* model
- Generalization beyond restrictive dyadic independence assumption
- Based on the idea of using network metrics to capture the structure/dependence in the model
- Aim: Describe parsimoniously the local selection forces that shape the global structure of a network

ERGMs

- Simple example: binary graph Y on n nodes
- $Y_{ij} = 1$ if the node i is connected to node j and 0 otherwise, $i, j = 1, \ldots, n$
- Observed graph y, set of metrics based on observed network and nodal attributes $\mathbf{s}(y)$
  ERGM for y is given by

$$P(Y = y|\theta) = \frac{\exp(\theta^T \mathbf{s}(y))}{c(\theta)}$$

  where θ is a vector of model parameters associated with $\mathbf{s}(y)$ and $c(\theta)$ is a normalising constant

ERGMs

More general ERGM for y with covariate information $\mathbf{X}$ is given by

$$P(Y = y|\theta) = \frac{\exp(\theta^T \mathbf{s}(y, X))}{c(\theta)}$$

- ERGMs are a models giving a probability distribution on each possible network of n nodes
- Number of potential graphs for directed binary graph with n nodes is $2^{n(n-1)}$
- So $c(\theta)$ usually not available
- So MCMC sampling is used or Pseudo-Likelihood Estimation (PLE)
- In `R` the package `ergm` has an `ergm` command that automatically fits a specified example from a wide range of ERGMs

Interpreting ERGMs

- Define change statistic $\delta_s(y)$ by

$$\delta_s(y)_{ij} = \mathbf{s}(y_{ij}^{+}) - \mathbf{s}(y_{ij}^{-})$$

  where y_{ij}^{+} and y_{ij}^{-} represent networks with y_{ij} fixed to be 1 or 0, respectively
- The ERGM model then implies the following for the Bernoulli variable y_{ij} conditional on the rest of the network:

$$\text{logit}[P(Y_{ij} = 1 | Y_{ij}^{c} = y_{ij}^{c})] = \theta^{T} \delta_s(y)_{ij}$$

  where $\text{logit}(p) = \log(p/(1-p))$ and Y_{ij}^{c} is the rest of the network except for Y_{ij}
- Note: RHS depends only on the change statistic $\delta_s(y)_{ij}$ not on the $\mathbf{s}(y_{ij}^{+})$ and $\mathbf{s}(y_{ij}^{-})$ themselves

Interpreting ERGMs

- Each component in the vector θ may be interpreted as the increase in conditional log-odds of the network per unit increase in the corresponding component of $\mathbf{s}(y)$, resulting from switching a particular Y_{ij} from 0 to 1, leaving the rest of the network fixed
- Note the dimension of θ is at most $2^{n(n-1)}$, usually much smaller

Special cases of ERGMs

- Special case ERGM: Bernoulli and Erdös-Rényi network
  All dyads are independent and have a common probability of a tie
- Well understood but unrealistic
- Fit with

  ```
  ergm(y ~ edges)
  ```

Special cases of ERGMs

- Other special case of ERGM is p_1 model where each dyad has its own probability distribution with arbitrary nodal indegree and outdegree marginal distributions and strength of reciprocity within dyads

$$log[P(Y = y)] = \sum\sum_{i<j} \rho_{ij} y_{ij} y_{ji} + \sum\sum_{i \neq j} \phi_{ij} y_{ij} - \log(c(\rho, \phi))$$

- In general restricted to $\rho_{ij} = \rho \; \forall \, i, j$ and $\phi_{ij} = \theta + \alpha_i + \beta_j$
- This is an ERGM with an edge, sender, receiver and reciprocity effect,
- Fit with

```
ergm(y ~ edges + sender + receiver + mutual)
```

More general ERGMs

- Want to include covariates
- Avoid making strong independence assumptions

Simple example

- Suppose we are modelling friendship as a directed tie
- We would like to see if there is a greater amount of reciprocity in the observed network than would be observed by chance
- $s(y) = \sum_{i<j} y_{ij} y_{ji}$ and θ will be a reciprocity parameter (0 indicating reciprocation in the graph is random, positive indicating more reciprocity than expected)
- Fit model and examine the estimate $\hat{\theta}$

Application: Florentine Marriage Network

- First model fit where propensity to form ties between families depends on the absolute difference in wealth:
  $s_1(y)$ = number of edges in the network
  $s_2(y)$ = sum of the absolute differences in wealth between all pairs of nodes connected by an edge
- Using the `R` language and `ergm` package

```
flomarriage <- network(flo,directed=FALSE)
flomarriage %v% "wealth" <- c(10,36,27,146,55,44,20,8,
42,103,48,49,10,48,32,3)
library(ergm)
gest <- ergm(flomarriage ~ edges + absdiff("wealth"))
summary(gest)
```

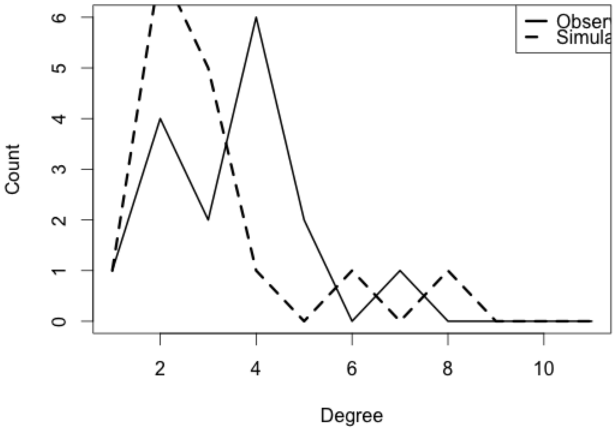
Application: Florentine Marriage Network

```
==========================
Summary of model fit
==========================

Formula:   flomarriage ~ edges + absdiff("wealth")

Iterations:  20

Monte Carlo MLE Results:
                Estimate Std. Error MCMC % p-value
edges          -1.457666   0.354532     NA  <1e-04 ***
absdiff.wealth -0.004176   0.007387     NA   0.573
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null  Deviance: 166.355  on 120  degrees of freedom
 Residual Deviance: 107.798  on 118  degrees of freedom
          Deviance:  58.557  on   2  degrees of freedom

AIC: 111.8    BIC: 117.37
```

Checking fit

- Suppose we wanted to see how well the model fit
- We could simulate another network based on the fitted ERGM and compare them

```
sim2 <- simulate(mod1, burnin = 1e+6, verbose = TRUE, seed = 9)
```



Application: Florentine Marriage Network

- Could count how many 2-stars there are in each:

```
c(flomarriage=summary(flomarriage~kstar(2)),
sim2 = summary(sim2~kstar(2)))
flomarriage.kstar2         sim2.kstar2
                47                  39
```

- Add terms to allow for the propensity to form 2-stars and triangles of families

```
gest <- ergm(flomarriage ~ kstar(1:2) + absdiff("wealth")
+ triangle)
#Note kstar(1) = edges
summary(gest)
```

Application: Florentine Marriage Network

```
==========================
Summary of model fit
==========================

Formula:   flomarriage ~ kstar(1:2) + absdiff("wealth") + triangle

Iterations:  20

Monte Carlo MLE Results:
                Estimate Std. Error MCMC % p-value
kstar1         -0.698208   0.715583     53   0.331
kstar2         -0.034389   0.308750     47   0.912
absdiff.wealth -0.004113   0.007873     13   0.602
triangle        0.224775   0.836199     12   0.789

    Null  Deviance: 166.355  on 120  degrees of freedom
 Residual Deviance: 107.716  on 116  degrees of freedom
          Deviance:  58.639  on   4  degrees of freedom

AIC: 115.72    BIC: 126.87
```

Application: Monk Data

Let's see if friendship links were more likely to be reciprocated in the Monk's data

```
==========================
Summary of model fit
==========================

Formula:   samplike ~ edges + mutual

Iterations:  20

Monte Carlo MLE Results:
       Estimate Std. Error MCMC % p-value
edges   -1.7583     0.2046      0  <1e-04 ***
mutual   2.3142     0.4080      0  <1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null  Deviance: 424.206  on 306  degrees of freedom
 Residual Deviance: 332.250  on 304  degrees of freedom
          Deviance:  91.956  on   2  degrees of freedom

AIC: 336.25    BIC: 343.7
```

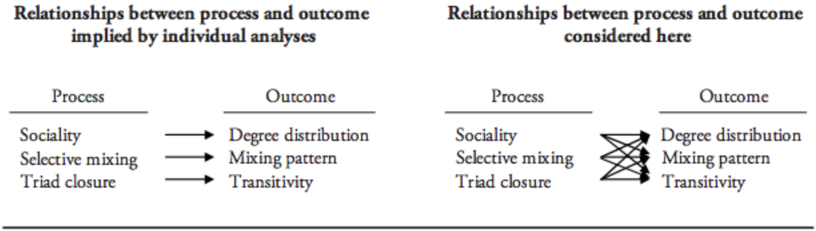
Application: Monk Data

Let's see if friendship links were more likely to be reciprocated in the Monk's data and if the likelihood of a friendship depended on group membership

```
==========================
Summary of model fit
==========================

Formula:   samplike ~ edges + mutual + nodematch("group", diff = T)

Iterations:  20

Monte Carlo MLE Results:
                          Estimate Std. Error MCMC %  p-value
edges                      -2.2489     0.2300      0  < 1e-04 ***
mutual                      1.3712     0.4932      0 0.005776 **
nodematch.group.Turks       2.2540     0.4027      0  < 1e-04 ***
nodematch.group.loyal       1.7084     0.3637      0  < 1e-04 ***
nodematch.group.outcasts    2.7998     0.7517      1 0.000233 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
    Null  Deviance: 424.21  on 306  degrees of freedom
 Residual Deviance: 264.76  on 301  degrees of freedom
          Deviance: 159.45  on   5  degrees of freedom

AIC: 274.76    BIC: 293.38
```

Relating process and outcome in Social Network Models

From “Birds of a feather, or friend of a friend? Using Exponential Random Graph Models to investigate adolescent social networks.” Steven M. Goodreau, James A. Kitts and Martina Morris



Notes: Specific forms of selective mixing include assortative mixing and disassortative mixing. Corresponding specific forms of mixing pattern include homophily and heterophily.

Problems with degneracy

- Some models specified have very counterintuitive implications
- Say we want to measure clustering/transitivity
- Natural model: network $\sim$ edge count + triangle count
- Distribution of networks from this model bimodal: all ties exist or none
- Excluding the parameters that give these degenerate models still leads to bimodal distribution: one mode with low density and high triad closure, the other with high density and low triad closure
- Observed network doesn't fit into these patterns?!
- Termed model degeneracy: if we specify a model unlikely to produce the observed network, either
  - the MLEs do not exist and the MCMC doesn't converge, or
  - the MLEs exist but do not provide a good fit for the data

Latent Space Network Models

- Presence/absence of a tie between two members of the network *i* and *j* is independent of the other ties in the network given the positions of *i* and *j* in a latent “social space”
- This means we take a conditional independence approach:

$$P(Y|X, Z, \theta) = \prod_{i \neq j} P(y_{ij}|z_i, z_j, x_{ij}, \theta)$$

- “Social Space refers to a space of unobserved latent characteristics that represent potential transitive tendencies in network relations” (Hoff et al, 2002)

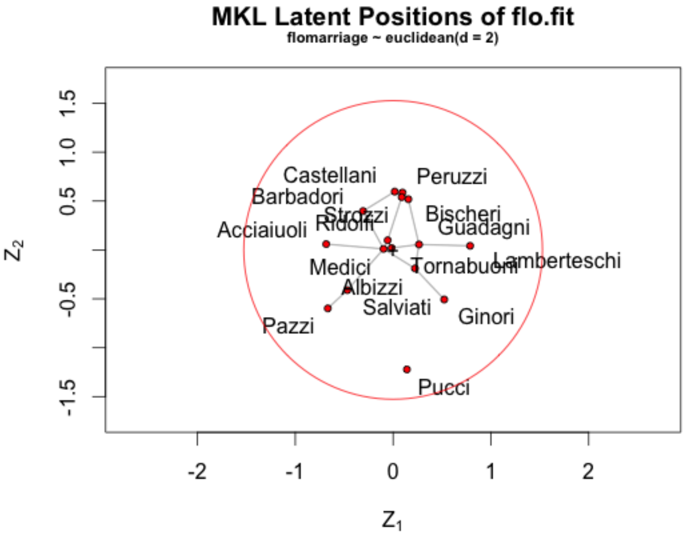
Latent Space Model

- log-odds$(y_{ij} = 1 | z_i, z_j, \beta) = \alpha + \beta^T x_{ij} - |z_i - z_j|$ where α is an intercept and $|z_i - z_j|$ is the Euclidean distance between the latent positions of nodes i and j
- This model has a simple interpretation: for two actors j and k equidistant from i, the log odds ratio of a tie between i and j versus i and k is $\beta^T(x_{ij} - x_{ik})$
- Problem: z_i's not identifiable: $|z_i - z_j|$ could be replaced by an arbitrary set of distances $\{d_{ij}\}$, satisfying the triangle inequality $d_{ij} \leq d_{ik} + d_{kj}$
- Prefer to have d_{ij}'s as distances in low-dimensional Euclidean space

Latent Space Model

- LSM inherently reciprocal and transitive
- Analogous with Multidimensional Scaling
- The command `ergmm` in the R package `latentnet` allows for fitting of latent space and ergm models

Application: Florentine Marriage Data

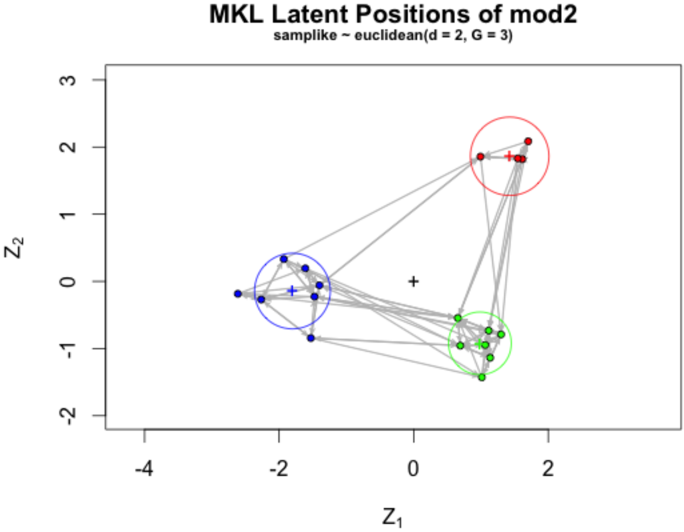


Latent Position Cluster Model

- Extension of the LCM
- Instead of all the z_i's coming from one distribution, they are clustered
- Each cluster has its own distribution $\rightarrow$ model-based clustering/gaussian mixture model clustering/latent class clustering

$$z_i \sim \sum_{g=1}^{G} \lambda_g MVN_d(\mu_g, \sigma_g^2 I_d)$$

Application: Sampson's Monk Data



Not covered

- Latent space models with random effects
- p_2 models: dyadic independence conditional on node-level attribute effects
- Stochastic block models: extension of p_1 model, which h includes parameters describing differential rates of between-group and within-group ties
- …

Useful Software for SNA

- Packages in R
  - `statnet` is a set of libraries with modelling/graphing tools (including library `ergm`)
  - `latentnet` is a library for fitting latent space models to networks (also produces graphical representations of the latent spaces)
- igraph: software for graphing/modelling networks (in C, so quite fast)
- Pajek: software for analysis/visualization of large scale networks
- SIENA, StOCNET, UCINET all free software with various graphical/analytical capabilities
- Look up "Social network analysis software" on Wikipedia for rafts more

Useful References for Introduction to Social Network Analysis

- Steve Borgatti's Instructional Social Network Analysis website: http://www.analytictech.com/networks/
- Classic text: Social Network Analysis: Methods and Applications (1994), Stanley Wasserman and Katherine Faust
- Online textbook: Introduction to social network methods (2005), Robert A. Hanneman and Mark Riddle http://faculty.ucr.edu/ hanneman/nettext/
- Good intro to ERGMs: An Introduction to Exponential Random Graph (p*) Models for Social Networks (2006), Garry Robins, Pip Pattison, Yuval Kalish and Dean Lusher
- Hoff, P., Raftery, A. E. and Handcock, M. S. (2002), "Latent space approaches to social network analysis", Journal of the American Statistical Assoication, 97, 1090 –1098
- Special edition of Journal of Statistical Software, Volume 24, Number 5, on `statnet`