

Models for cross-sectional network data

p2 and ergm in StOCNET

Marijtje van Duijn

Dept. of Sociology / ICS
University of Groningen

Groningen
Jan 14, 2010

Cross-sectional network data

One single observation of the network

- No longitudinal data available (yet)
- The first observation of the network starting point of the longitudinal analysis

Statistical analysis

- Simple model (Bernoulli graph) based on (single) ties
- Loglinear (p_1) model recognizing dyads (Y_{ij}, Y_{ji})
- Random effects (p_2) model recognizing dependence between dyads with same nodes (actors)
- Exponential random graph model (p^*) recognizing larger dependence structures
- Note that other statistical models are available

Simple distributions for directed graphs

Uniform distribution of ties

- A directed graph of size n has a finite number of outcomes: $2^{n(n-1)}$
- Equivalently: $P(Y_{ij} = 1) = P(Y_{ij} = 0) = 0.5$
- Generalization to Bernoulli graph where $P(Y_{ij} = 1) = P_{ij}$

Under the uniform distribution it is possible

- to derive distribution of network statistics (e.g. dyad census)
- to test the uniform distribution
- to estimate the (constant) probability of a tie

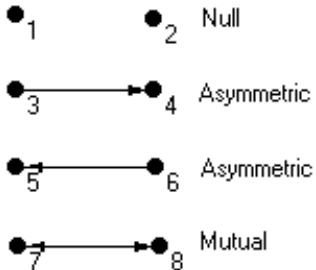
Conditional distributions for directed graphs

Limiting the number of possible outcomes

Conditioning on

- number of arcs ($g(g-1)$)
- in- and/or outdegrees
- number of mutual, asymmetric and null dyads (U—MAN)

Dyad has four possible outcomes



Plausible types of dyad dependence

Dyad is defined as

the pair of ties between two actors (Y_{ij} , Y_{ji}).

Both actors are ego and alter, or sender and receiver

- *reciprocity* dependence between Y_{ij} and Y_{ji} .
- dependence within each sender:
outgoing relations of the same actor
- dependence within each receiver:
Incoming relations of the same actor

Notation

Matrix Y is *adjacency matrix* of digraph.

$$Y_{ij} = \begin{cases} 1 & \text{if relation is present} \\ 0 & \text{if relation is absent.} \end{cases}$$

(Diagonal values Y_{ii} usually meaningless.)

ρ_1 model (Holland and Leinhardt, 1981)

Important first step

- models the four possible dyadic outcomes
- with *fixed* effects of
 - ρ reciprocity
 - α sending
 - β receiving

Important first step

- models the four possible dyadic outcomes
- with *fixed* effects of
 - ρ reciprocity
 - α sending
 - β receiving

$$P\{Y_{ij} = y_1, Y_{ji} = y_2\} = \exp\{y_1(\mu + \alpha_i + \beta_j)\} * \exp\{y_2(\mu + \alpha_j + \beta_i)\} * \exp\{y_1 y_2 \rho\} / k_{ij}$$

$$y_1, y_2 = 0, 1$$

μ : density

k_{ij} : normalizing constant

p_2 model (Van Duijn, Snijders & Zijlstra, 2004; Zijlstra, 2008)

Random effects version of p_1

- Fixed approach not so attractive because of $2n$ (actor) parameters
- Replaced by latent random variables
- Allows inclusion of explanatory actor and dyad variables

ρ_2 model (Van Duijn, Snijders & Zijlstra, 2004; Zijlstra, 2008)

Random effects version of ρ_1

- Fixed approach not so attractive because of $2n$ (actor) parameters
- Replaced by latent random variables
- Allows inclusion of explanatory actor and dyad variables

$$\alpha_i = X_i \gamma_1 + A_i, \quad i = 1 \dots n$$

$$\beta_j = X_j \gamma_2 + B_j, \quad j = 1 \dots n$$

$$\mu_{ij} = \mu + Z_{ij1} \delta_1, \quad i \neq j$$

$$\rho_{ij} = \rho + Z_{ij2} \delta_2, \quad Z_{ij2} = Z_{ji2}, \quad i \neq j$$

$$\begin{aligned} \text{var}(A_i) = \sigma_A, \text{var}(B_i) = \sigma_B, \text{cov}(A_i, B_i) = \sigma_{AB} \text{ for all } i \\ \text{cov}(A_i, A_j) = \text{cov}(B_i, B_j) = \text{cov}(A_i, B_j) = 0 \text{ for } i \neq j. \end{aligned}$$

ρ_2 model parameters

Random effects

A_i and B_j are the *unexplained* parts of the sender effect of actor i and the receiver effect of actor j , given the actor-dependent explanatory variables X_i for the sender; X_j for the receiver.

A_i and B_j are assumed to be independent for different i and normally distributed;

A_i and B_i , referring to the same actor, are assumed to be correlated.

Parameters to be estimated

The statistical parameters of the ρ_2 model are the parameters μ and ρ , the regression coefficients γ , δ and the variances σ_A^2 and σ_B^2 and the correlation σ_{AB} of the actor effects.

Interpretation of ρ_2 parameters

Actor-related effects

- pos/neg sender effect in/decreases outgoing tie probability
- pos/neg receiver effect in/decreases ingoing tie probability

Dyad-related effects

- pos/neg density effect in/decreases any dyadic tie probability
- pos/neg reciprocity effect in/decreases mutual tie probability
- in addition to density effect - like an interaction effect

MCMC estimation of the p_2 model (Zijlstra,

Van Duijn & Snijders, 2009)

No closed form of the likelihood function

- Prior distributions for model parameters
- Fixed effects normal distribution
- Covariance matrix Wishart distribution

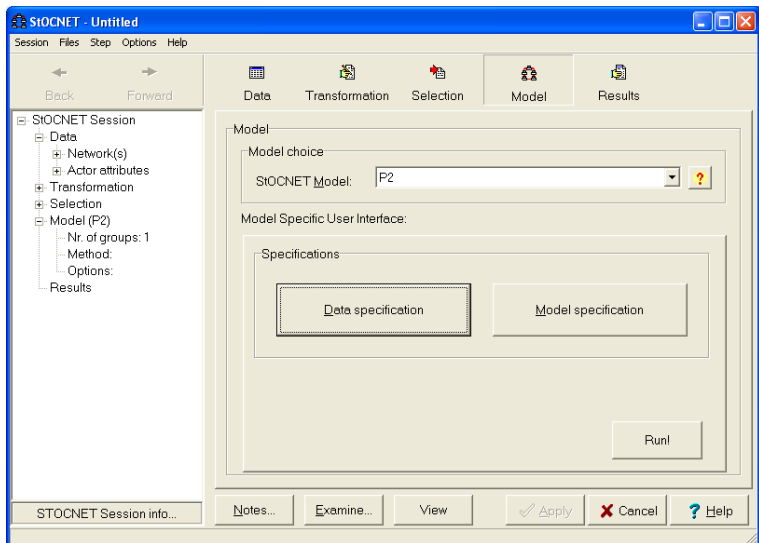
Posterior distributions

- Fixed (and random) effects not tractable
- Covariance matrix Wishart

MCMC algorithms

- Random Walk with Metropolis steps
- Independence Chain with Metropolis-Hastings steps

p_1 and p_2 in StOCNET



Application (Lazega & Van Duijn, 1997; Van Duijn, Zijlstra, & Snijders, 2004)

Research interest

Structure of professional relationships and effect of formal hierarchy.

Data

Advice network of 71 lawyers in US law firm
36 partners, 35 associates
3 offices; 2 specialties (litigator, corporate)

Results

	Parameter	Estimate	(S.E.)
Sender	Variance σ_A^2	0.75	(0.11)
	Senior Partner	0.76	(0.10)
Receiver	Variance σ_B^2	0.49	(0.08)
	Seniority Associate	-0.50	(0.06)
Sender-Receiver	Covariance σ_{AB}	-0.05	(0.06)
Density	μ	-3.98	(0.22)
	Similarity Office	1.79	(0.11)
	Similarity Specialty	1.60	(0.12)
	Superiority Seniority	-0.29	(0.11)
Reciprocity	ρ	1.42	(0.13)
	Similarity Specialty	-0.81	(0.28)

Other statistical models

Extensions to

- Multiple networks (schoolclasses) Zijlstra et al., 2006
- Multiple relations Zijlstra et al., 2008

Other statistical models

- p_2 is a bivariate logistic regression model
- p_2 is a generalized linear mixed model
- p_2 is a cross-nested multilevel model
- p_2 is the dichotomous counterpart of the Social Relations Model (Snijders & Kenny, 1996)
- p_2 belongs to the exponential family but is not an ERGM

Dyadic (in)dependence sufficient?

Exponential Random Graph Models

Main differences between p_1/p_2 and *ERGM*

- p_1/p_2 : models (four possible) dyadic outcomes (connected between actors through random effects)
- *ERGM* models complete network where the dependence structure has yet to be determined (no random effects)

Exponential random graph models (p^*)

Family of probability functions

$$P_{\theta}\{Y = y\} = \exp(\theta' u(y) - \psi(\theta))$$

- $u = u(x)$ vector of sufficient statistics
- $\psi(\theta)$ norming constant

The choice of (sufficient) statistics reflects the dependence structure in the network.

Simplest dependence structure

- Independent ties
- Bernoulli graph where every tie has the same probability of occurrence (cf. binomial distribution)
- Sufficient statistic: number of observed ties in the network y_{++} .

Example of an exponential random graph model with dyadic independence

Sufficient statistics are

- number of ties y_{++}
- number of mutuals $\sum_{i < j} y_{ij} y_{ji}$
- in-degrees y_{i+}
- out-degrees y_{+j}

with some restrictions on the parameters

Markov graphs

Assumes dependence for relations from/to same node

- Note that this implies independence for ties not sharing an actor!
- Sufficient statistics
 - number of mutuals
 - triad counts (digraph of three nodes, 6 types)
 - transitive triads: $\sum_{i \neq j \neq k} y_{ij} y_{jk} y_{ik}$
 - cyclic triads: $\sum_{i \neq j \neq k} y_{ij} y_{jk} y_{ki}$
 - k -stars ($\sum_{i \neq j_1 \neq j_2 \neq \dots \neq j_k} y_{ij_1} y_{ij_2} \dots y_{ij_k}$)
- Extension with covariates possible

Maximum Likelihood estimation

Via general theory of exponential families

- $\mu(\theta) = E(u(Y))$ is gradient of $\psi(\theta)$
- $\Sigma(\theta) = \text{cov}(u(Y))$ is the matrix of derivatives of $\mu(\theta)$
- Maximum Likelihood by Newton-Raphson algorithm

PROBLEM

$\mu(\theta)$ and $\Sigma(\theta)$ not computable

Pseudo-likelihood estimation

(Frank & Strauss, 1986; Wasserman & Pattison, 1996)

Maximize

$$l(\theta) = \sum_{i,j} \ln(\mathbf{P}\theta\{Y_{ij} = y_{ij} | Y_{hk} = y_{hk} \text{ for all } (h,k) \neq (i,j)\})$$

- advantage: easily implemented with logistic regression (with change in sufficient statistics as covariates)
- disadvantage: not a function of $u(Y)$
 $\Rightarrow \hat{\theta}$ not admissible

Simulation-based estimation

Possible solution to the intractability of the likelihood function:
Markov Chain Monte Carlo simulation using (e.g.)
Robbins-Monro (1951) stochastic moment estimation
(Snijders, 2002)
solving $E(Z_\theta) = 0$ with $Z_\theta = u(Y) - u(y)$.
Solution is ML estimate

Necessary

Monte Carlo simulation of Y from exponential random graph distribution (for any θ)

Problems with simulation

e.g. with Gibbs sampler updating Y_{ij}

- bimodal distribution is obtained for many choices of $u(y)$
- two regimes lead to (too) slow convergence to target distribution
- large gradients cause instability for parameter estimation

especially estimation of transitivity effect turns out to be difficult

One of the explanations is the strong Markov assumption:

Y_{ij} and Y_{hk} are independent conditional on all other ties.

Additional sufficient statistics

The out-degree variance is a function of the number of out-2-stars and the total number of ties.

A similar relation holds for the in-2-stars.

Higher-order k -stars and k -triangles help.

This can be seen as a (better) to represent in/outdegree variability and transitivity.

Not too much, therefore *alternating* with λ chosen.

Alternating sufficient statistics

$$\begin{aligned}u(y) &= S_2 - \frac{S_3}{\lambda} + \frac{S_4}{\lambda^2} - \dots + (-1)^{n-2} \frac{S_{n-1}}{\lambda^{n-3}} \\ &= \sum_{k=2}^{n-1} (-1)^k \frac{S_k}{\lambda^{k-2}} \\ &= \lambda^2 \sum_{i=1}^n \left\{ \left(1 - \frac{1}{\lambda}\right)^{y_{i+}} + \frac{y_{i+}}{\lambda} - 1 \right\} .\end{aligned}$$

Results example Lazega analysis of friendship network

Note: usually not easy to get converged models

Note: no density parameter (conditioned on)

Parameter	Estimate	(S.E.)
Reciprocity	2.65	(0.35)
Alternating k-triangles	0.80	(0.11)
Similarity Office	0.45	(0.17)

Parameter	Estimate	(S.E.)
Reciprocity	3.20	(0.31)
Alternating out-2-stars	0.99	(0.28)
Alternating in-2-stars	0.66	(0.31)
Similarity Office	0.64	(0.23)

p_2 results

	Parameter	Estimate	(S.E.)
Sender	Variance σ_A^2	0.81	(0.34)
Receiver	Variance σ_B^2	0.26	(0.29)
Sender-Receiver	Covariance σ_{AB}	-0.31	(0.21)
Density	μ	-6.36	(0.32)
	Similarity Office	4.30	(0.34)
Reciprocity	ρ	-0.06	(1.23)

p_2 or ERGM ?

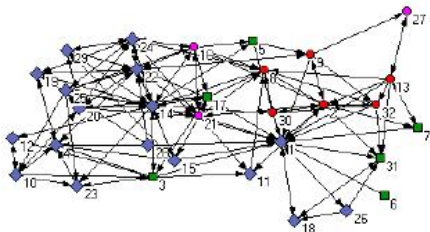
- dyadic dependence vs. complete network
- structural effects up to reciprocity vs. transitivity & more

Is the p_2 model (too) simple (enough) to model networks?

Example

Freeman's EIES network

- 32 actors: social network researchers at a conference
- tie: acquaintanceship (not knowing - friend)
- actor covariates: discipline, number of citations



StOCNET data

- eies.1 friendship time 1 (0-4)
- eies.2 friendship time 2 (0-2)
- eies.att attributes (discipline, number of citations)