

MODELS FOR SINGLE OBSERVATIONS OF SOCIAL NETWORKS

Tom A.B. Snijders

ICS

University of Groningen, The Netherlands

Modeling social networks:

the p_1 and p_2 models

and the p^* = ERGM = Exponential Random Graph Model

for single observations of social networks

These presentations are about modeling a social network (binary; one observation, i.e., not longitudinal) as a *dependent variable*.

¿How to explain an observed network. ?

Notation:

set of n actors, with a dichotomous = binary relation, represented as a *directed graph (digraph)*.

Tie variable from i to j indicated by Y_{ij} :

$$Y_{ij} = \begin{cases} 1 & \text{if there is a tie} \\ 0 & \text{if there is no tie.} \end{cases}$$

(Diagonal values Y_{ii} meaningless, formally defined as $Y_{ii} = 0$.)

Y_{ij} is the indicator of the *arc* or *directed line* from i to j .

Matrix Y is *adjacency matrix* of digraph.

Set of all digraphs, or all Y , denoted by \mathcal{Y} .

Same approach is possible for non-directed graphs.

Existence of ties can be explained on the basis of

1. Explanatory variables = covariates;
can be function of individual actors (actor-based covariate)
or of directed or undirected pairs of actors
(dyad-based covariates).
2. Patterns of further ties in the network;
this reflexivity / endogenous feedback is the big difficulty
in modeling social networks.

Complicated dependence structure between ties.

Some interesting kinds of dependence between ties:

★ *Reciprocity* :

dependence between Y_{ij} and Y_{ji} .

The pair (Y_{ij}, Y_{ji}) is called a *dyad*.

★ The dependence within each *row*:

outgoing relations of the same actor.

★ The dependence within each *column*:

incoming relations of the same actor.

★ *Transitivity:*

"a friend of my friend is also my friend",

⇒ dependence between triples of actors.

★ *Popularity:*

choices lead to more choices.

★ *Balance:*

preference for others who make the same choices
as the actor him/herself

(looks a bit like transitivity, not the same).

The difficulty with stochastic models for social networks is that they have to represent *dependence*, and cannot be built – like most stochastic models – on broad *independence* assumptions.

How could one model a social network, while forgetting about this independence?

Y_{ij} is a binary (0–1) dependent variable,
so the most straightforward model would be *logistic regression*:

$$\text{logit} \left(P \{ Y_{ij} = 1 \} \right) = \gamma_0 + \gamma_1 w_{1ij} + \dots + \gamma_p w_{pij}$$

where W_1, \dots, Z_p are explanatory variables, and where

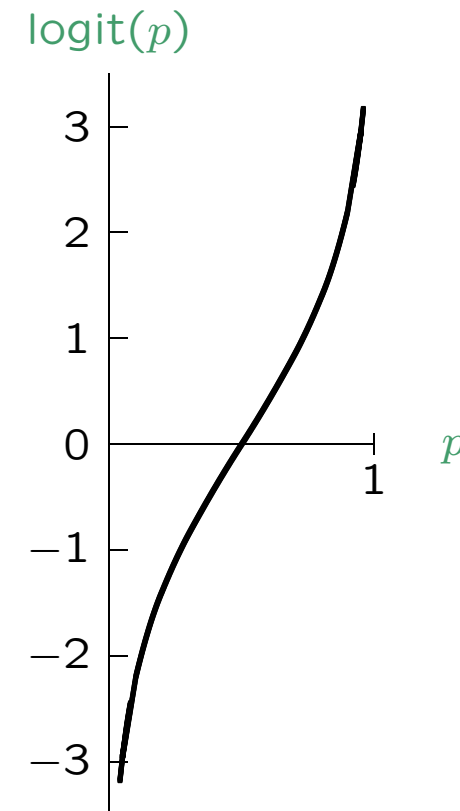
$$\text{logit}(p) = \log \left(\frac{p}{1-p} \right) .$$

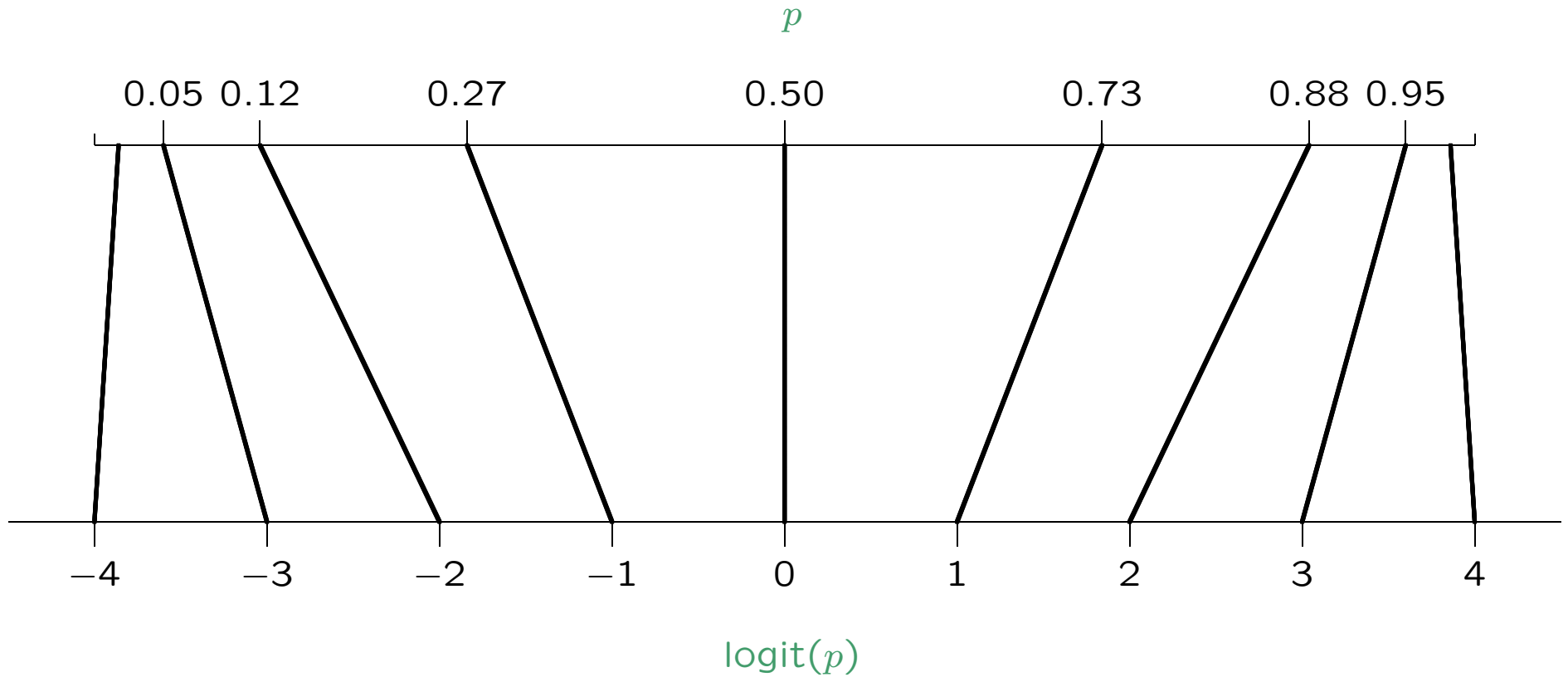
The *logarithm* transforms a multiplicative to an additive scale and transforms the set of positive real numbers to the whole real line. Indeed, one of the most widely used transformations of probabilities is the *log odds*, defined by

$$\text{logit}(p) = \ln \left(\frac{p}{1-p} \right),$$

where $\ln(x)$ denotes the natural logarithm of the number x . The logit function, of which graphs are shown here, is an increasing function defined for numbers between 0 and 1, and its range is from minus infinity to plus infinity.

For example, $p = 0.269$ is transformed to $\text{logit}(p) = -1$ and $p = 0.982$ to $\text{logit}(p) = 4$. The logit of $p = 0.5$ is exactly 0.





Correspondence between p and $\text{logit}(p)$.

The logistic regression model is a model where $\text{logit}(p)$ is a linear function of the explanatory variables. In spite of the attractive properties of the logit function, it is by no means the only suitable function for transforming probabilities to arbitrary real values.

The general term for such a transformation function is the *link function*, as it links the probabilities (or more generally, the expected values of the dependent variable) to the explanatory variables. The probit function (which is the inverse cumulative distribution function of the standard normal distribution) also is often used as a link function for dichotomous variables. A generalized linear model for a dichotomous outcome with the probit link function is called a probit regression model.

For still other link functions see, e.g., the textbooks Long (1997) or McCullagh and Nelder (1989).

Statistical theory tells us the following about such a procedure, which neglects the dependence structure in the model for the data:

1. The parameter estimates are *consistent*, i.e., for large n they are reasonable;
2. however, they are not *efficient*, i.e., their precision is not optimal;
3. the standard errors are totally unreliable.

Example: Friendship between Lazega's lawyers

Example based on Lazega's research on a New England law firm (Lazega, 2001; Lazega and Pattison, 1999).

Friendship relation between the 36 partners.

Actor covariates:

- ⇒ seniority (rank number of entry in the firm)
- ⇒ gender
- ⇒ office (three different cities)
- ⇒ years in the firm
- ⇒ age

⇒ practice (litigation or corporate law)

⇒ law school attended (Ivy League, non-I.L. in the region, other).

For categorical covariates (office and law school),
similarity represented by within-dyad identity:

$$I\{x_i = x_j\} = \begin{cases} 1 & \text{if } x_i = x_j, \\ 0 & \text{otherwise.} \end{cases}$$

Density = 0.21, average degree = 7.4.

In-degrees vary from 2 to 16, out-degrees from 0 to 21.

For the actor covariates, we distinguish between the following effects on Y_{ij} :

1. the activity or out-degree effect, leading to correlation between the covariate W_i and the out-degrees Y_{i+}
2. the popularity or in-degree effect, leading to correlation between the covariate W_i and the in-degrees Y_{+i}
3. similarity or dissimilarity effects, leading to correlation between the absolute difference $|W_i - W_j|$ and the tie indicators Y_{ij} .

Parameter	est.	s.e.
Constant term	-1.750	0.097
Same office	2.129	0.183
Seniority popularity	0.003	0.008
Seniority activity	0.023	0.008
Seniority dissimilarity	-0.077	0.011
Practice (corp. law) popularity	0.184	0.156
Practice (corp. law) activity	0.565	0.157
Same practice	0.396	0.155

Logistic regression estimates of covariate effects
for Lazega's friendship data (partners).

A next step could be to include row ('*sender*') and column ('*receiver*') effects in the model:

$$\text{logit} \left(P \{ Y_{ij} = 1 \} \right) = \alpha_i + \beta_j + \gamma_1 w_{1ij} + \dots \gamma_p w_{pij} .$$

Then the variables W_1, \dots, W_p all must be dyadic covariates, because actor-dependent covariates, depending only on i or j , would be completely collinear with the sender and receiver effects.

The sender effects α_i represent the *activity / outgoingness* of actor i , while the receiver effects β_i represent the *popularity / attractiveness* of actor j .

Holland and Leinhardt (1981) proposed the p_1 model, which does not include covariate effects, but which does include *sender and receiver effects* as well as *reciprocity*.

The model includes the following parameters:

- μ , parameter for the density (μ higher \Rightarrow more ties)
- α_i , activity parameter for actor i
(α_i higher $\Rightarrow i$ has more outgoing ties)
restriction $\sum_i \alpha_i = 0$
- β_i , popularity parameter for actor i
(β_i higher $\Rightarrow i$ has more incoming ties)
restriction $\sum_i \beta_i = 0$
- ρ , reciprocity parameter
(ρ higher $\Rightarrow Y_{ij}$ and Y_{ji} tend to be more alike)

The probability distribution for each dyad (Y_{ij}, Y_{ji}) is defined by

$$P\{(Y_{ij}, Y_{ji}) = (a, b)\} = k_{ij} \exp \left(a(\mu + \alpha_i + \beta_j) + b(\mu + \alpha_j + \beta_i) + ab\rho \right) ;$$

this formula holds if a and b are 0 or 1
(possible outcomes of Y_{ij} and Y_{ji});

k_{ij} is a number (not depending on (a, b))

ensuring that the four probabilities

(for outcomes $(0,0)$, $(0,1)$, $(1,0)$, $(1,1)$) sum to 1.

Different dyads are assumed to be statistically independent.

Note 1:

This form of the probability distribution, consisting of an exponential function of a linear function of the parameters, is well-known in mathematical statistics: a so-called *exponential family of distributions*.

Exercise:

Using the properties of the exponential function and the definition of statistical independence, prove that the relations Y_{ij} and Y_{ji} are independent if $\rho = 0$.

Note 2:

The p_1 model can be regarded as a bivariate logistic regression model for the dyads, where the different dyads have different parameters depending on the actors i and j involved in the dyad.

To see the correspondence with the usual logistic regression model, write out the logit of the probability; recall that $\text{logit}(p) = \ln(p/(1 - p))$; the result is

$$\begin{aligned}\text{logit}(P(Y_{ij} = 1 \mid Y_{ji} = b)) &= \ln \left(\frac{P\{Y_{ij} = 1, Y_{ji} = b\}}{P\{Y_{ij} = 0, Y_{ji} = b\}} \right) \\ &= \mu + \alpha_i + \beta_j + b\rho.\end{aligned}$$

This makes clear that

α_i is a main effect of the "sender" (i) of the tie,

β_j is a main effect of the "receiver" (j) of the tie,

and ρ is the effect of the other relation (Y_{ji}) within the dyad.

For a single dyad, this would be a rather foolish model:
four outcomes and six parameters...;
but we have $n(n - 1)/2$ dyads to estimate the parameters from.

Note, however,
that there are two parameters associated with every actor.
The total number of parameters is $2 + 2n$;
since there are two equality constraints ($\sum_i \alpha_i = \sum_i \beta_i = 0$),
the number of independent parameters is $2n$.

The p_1 model can be estimated using methods for estimating loglinear models.

Estimation techniques and elaborations of the p_1 model were developed by Wasserman and co-workers. (See Wasserman and Faust, 1994.)

In this model, *dyads* (Y_{ij}, Y_{ji}) are independent. This is a severe limitation.

However, this model was an important first step for the statistical modeling of social networks.

Actor	Alpha	Beta
1	-0.365	-0.515
2	-1.202	1.022
3	-	0.291
4	1.182	0.496
5	-1.000	0.058
6	-	-0.575
7	-0.916	-1.109
8	-2.420	0.426
9	-1.125	1.856
10	1.802	-1.773
11	-1.099	1.346
12	2.774	-1.240
13	0.178	0.923
14	0.428	-0.650
15	-0.375	-1.346
16	-0.198	0.589
17	1.843	0.561
18	0.311	-0.879

Actor	Alpha	Beta	Rho	T
19	-1.155	-0.544	3.574	-2.
20	1.094	-0.555		
21	-0.610	1.319		
22	0.537	-0.456		
23	-2.739	0.965		
24	1.707	0.024		
25	0.696	0.279		
26	-0.800	2.049		
27	0.152	1.264		
28	0.640	-0.285		
29	0.495	0.012		
30	-0.001	-0.703		
31	2.116	-1.373		
32	-1.000	0.058		
33	1.489	-2.439		
34	-1.344	0.665		
35	-0.358	0.033		
36	-0.735	0.207		

It is not very attractive that differences between actors are represented in the statistical model by *parameters*, two parameters for each actor.

E.g., this precludes the possibility of testing effects of actor-dependent explanatory variables, because the differences between actors are already represented completely by the parameters α_i and β_i .

This is similar to the choice between representing group differences by fixed or by random effects in multilevel analysis.

To overcome this problem,
van Duijn (1995) proposed the p_2 model,
(also see van Duijn and Lazega, 1997;
van Duijn, Snijders, and Zijlstra, 2004)
which can be regarded as a *random effects* version,
or also as a *multilevel* version, of the p_1 model.

In this model, the sender and receiver effects α_i and β_i
are regarded not as statistical parameters
but as *latent (i.e., unobserved) random variables* ;
and these latent variables
can be explained by explanatory actor-dependent variables.
Dyad-dependent explanatory variables can also be included.

Since it is usual to denote random variables by capital instead of greek letters, we now write

U_i and V_i , respectively, as the *unexplained* parts of the sender and receiver effects of actor i

– unexplained, that is, given the explanatory variables $W^{(1)}, \dots, W^{(p)}$, where $W^{(h)}$ has values $W_{ij}^{(h)}$, which could depend on i (the sender) only or on j (the receiver) only, but also on i and j simultaneously.

The vectors (U_1, \dots, U_n) and (V_1, \dots, V_n) are denoted U and V , respectively.

The probability distribution for each dyad (Y_{ij}, Y_{ji}) is defined under the p_2 model by

$$\begin{aligned} P\{(Y_{ij}, Y_{ji}) = (a, b) \mid U, V\} &= k_{ij} \exp \left(a \left(\sum_h \gamma_h W_{ij}^{(h)} + \mu + U_i + V_j \right) \right. \\ &\quad \left. + b \left(\sum_h \gamma_h W_{ji}^{(h)} + \mu + U_j + V_i \right) + ab\rho \right) \end{aligned}$$

where a and b again are 0 or 1 (possible outcomes of Y_{ij} and Y_{ji}).

The numbers γ_h are statistical parameters which are completely analogous to coefficients in logistic regression.

For the random variables U_i and V_i it is assumed that they are independent for different i and normally distributed; however, U_i and V_i , which variables refer to the same actor, can be correlated.

The statistical parameters of this model are

1. the parameters μ and ρ ,
2. the regression coefficients γ_h ,
3. and the variances σ_U^2 and σ_V^2 ,
4. and the correlation ρ_{UV} of the actor effects.

Note that ρ_{UV} has nothing to do with ρ , although the same letter is used.

Remark:

The complete specification of the p_2 model also includes explanatory variables associated with the reciprocity effect ρ ,

i.e., interactions between explanatory variables and reciprocity.

This is omitted here.

Parameter estimation methods for the p_2 model are rather complicated.

Better ('MCMC') estimation methods are now being developed.

The p_1 and p_2 models represent structural network effects only to a very limited extent.

Random effects:

	parameter estimate	standard error
sender variance:	1.1405	0.2356
receiver variance:	0.5615	0.1375
sender receiver covariance:	-0.3464	0.1375

Fixed effects:

Overall effects:

	parameter estimate	standard error
Density:	-3.5695	0.7524
Reciprocity:	1.9784	0.2687

Overall covariate effects:

Overall effects of covariates including diff and absdiff manipulations.

Covariate	Wald test		
	statistic	df	P
Seniority	30.7112	3	0.0000
practice	6.1933	3	0.1026
Same Office	73.8774	1	0.0000

Specific covariate effects:

Sender covariates:

	parameter	standard
Covariate	estimate	error
Seniority	0.0342	0.0196
practice	0.6004	0.4003

Receiver covariates:

	parameter	standard
Covariate	estimate	error
Seniority	-0.0091	0.0152
practice	-0.0062	0.3067

Density covariates:

	parameter	standard
Covariate	estimate	error
abs_diff_Seniority	-0.0523	0.0099
abs_diff_practice	-0.2573	0.1327
Same Office	1.5320	0.1782

Frank (1991) and Wasserman and Pattison (1996) proposed the p^* model for social networks, generalizing the Markov graph distribution of Frank and Strauss (1986), also called the *Exponential Random Graph Model, ERGM*.

The probability distribution for the *ERGM* can be defined by

$$P\{Y = y\} = \frac{\exp(\beta' z(y))}{\kappa(\beta)}$$

where $z(y)$ is a vector of statistics of the digraph, β is a vector of statistical parameters, and $\kappa(\beta)$ is a normalization factor, ensuring that the probabilities sum to 1.

Note that this is again an exponential family of distributions.

This formula is extremely general, because $z(y)$ could be anything. In this general form, the formula could be used to represent any probability distribution for a directed graph.

In practice, $z(y)$ will contain effects of covariate effects as well as structural graph effects representing the dependence of the existence of ties within the network.

Much work has been done since 1996 to elaborate this model (by Wasserman, Pattison, Robins, Handcock, Snijders, and others).

The earlier way of parameter estimation is a so-called pseudo-likelihood procedure, which reduces the estimation to a logistic regression procedure.

However, this procedure has poor statistical properties and is unreliable.

A better procedure was proposed by Snijders (2002).

This procedure uses *Markov chain Monte Carlo* (MCMC) methods to approximate the maximum likelihood estimator.

This procedure also does not always work well; Handcock (in preparation) and Snijders (2002) argued that this is because, depending on the specification, the ERGM can have a *degenerate nature*: e.g., for many parameter values, this distribution can be concentrated with a high probability on the full (all $y_{ij} = 1$) or empty (all $y_{ij} = 0$) graph.

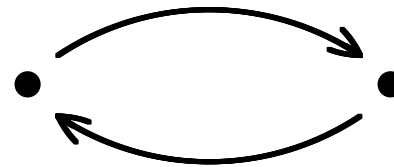
Therefore, the specification of the structural statistics, which express the dependencies of ties within the network, must be chosen in such a way that this degeneracy does not occur.

The mostly used statistics $z(y)$ depend on the *density* or (equivalently) the total number of ties,

$$Y_{++} = \sum_{i,j} Y_{ij} ,$$

the *number of mutual dyads*,

$$R = \sum_{i < j} Y_{ij} Y_{ji} ,$$



mutual = reciprocated dyad

various statistics expressing *transitivity*, and explanatory variables.

The structures representing transitivity are based on *triad counts*, which indicate the number of times that a given triad occurs as a subgraph in the entire graph.

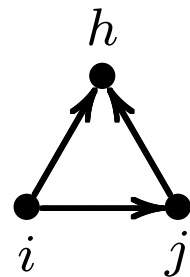
Triads are directed graphs with three nodes; e.g., the 3-cycle defined by $y_{12} = y_{23} = y_{31} = 1$, $y_{21} = y_{32} = y_{13} = 0$.

A triad count is the number of times that a given triad occurs among the $\binom{n}{3}$ triads generated by all subsets of three nodes within the total digraph y (where the order of the three nodes is not taken into account).

A *triplet* is an incomplete triad:
it defines the values of some, but not all,
of the six relations $y_{12}, y_{21}, y_{13}, y_{31}, y_{23}, y_{32}$.

The triplet count is defined in analogous fashion.

Examples are the transitive triplet defined by $y_{12} = y_{23} = y_{13} = 1$
and the intransitive triplet defined by $y_{12} = y_{23} = 1, y_{13} = 0$.



transitive triplet

Note that a triplet count can be expressed as a sum of triad counts,
because each triplet corresponds to a set of triads.

The definition of transitivity is:

if i is tied to j and j is tied to h , then i is tied to h .

Thus, the configuration $Y_{ij} = Y_{jh} = 1$, called a *two-path*, is a condition which must be fulfilled in order for transitivity to be a non-vacuous property.

If one wishes to test for transitivity, then it is relevant to control for this condition.

The number of two-paths can be expressed as

$$\begin{aligned} \sum_{\substack{i,j,h=1 \\ i \neq h}}^n Y_{ij} Y_{jh} &= \sum_{j=1}^n \sum_{\substack{i,h=1 \\ i \neq h}}^n Y_{ij} Y_{jh} \\ &= \sum_{j=1}^n \left(Y_{+j} Y_{j+} - \sum_{i=1}^n Y_{ij} Y_{ji} \right) = \left(\sum_{j=1}^n Y_{+j} Y_{j+} \right) - 2R, \end{aligned}$$

where R is the number of mutual dyads.

This shows that the number of two-paths is a function of the degrees and the number of mutual dyads.

Thus, when testing for transitivity using the number of transitive triplets as a test statistic, it is meaningful either to control for the number of two-paths, or for the degrees and the number of mutual dyads.

If the elements of the vector $z(y)$ are the following:

1. total number of ties Y_{++}
2. number of reciprocated ties R
3. out-degrees Y_{i+}
4. in-degrees Y_{+i} ,

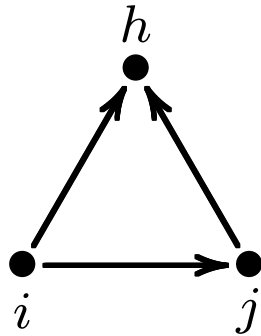
then the *ERGM* is just the p_1 model.

This requires some equality restriction among the parameters, because there are linear equalities among the statistics.

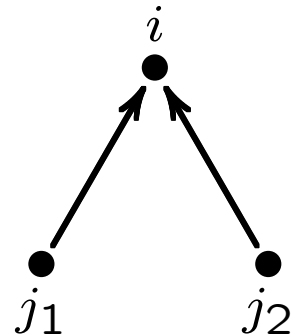
A more fruitful set of statistics is the following, which is a particular case of a *Markov graph* :

1. total number of ties Y_{++}
2. number of reciprocated ties R
3. number of out-2-stars $\sum_{i,j,h} Y_{ij} Y_{ih}$
4. number of in-2-stars $\sum_{i,j,h} Y_{ji} Y_{hi}$
5. number of 2-paths $\sum_{i,j,h} Y_{ij} Y_{jh}$
6. number of transitive triplets $\sum_{i,j,h} Y_{ij} Y_{jh} Y_{ih}$.

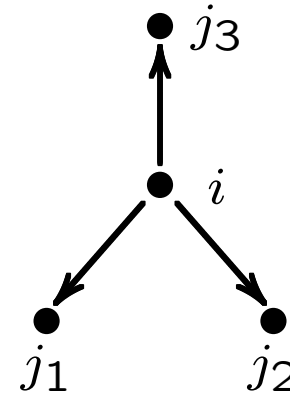
An explanation follows:



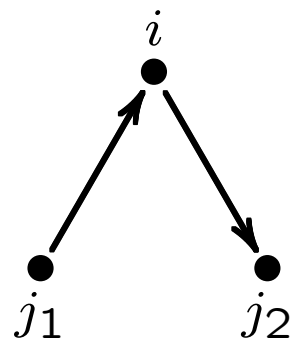
triangle
= transitive triad



in-two-star



out-three-star



two-path
= mixed two-star

The number of out-2-stars is

$$\begin{aligned} \sum_{\substack{i,j,h=1 \\ i \neq h}}^n Y_{ij} Y_{ih} &= \sum_{i=1}^n \sum_{\substack{j,h=1 \\ j \neq h}}^n Y_{ij} Y_{ih} \\ &= \sum_{j=1}^n \left(Y_{+j}^2 - \sum_{i=1}^n Y_{ij} \right) = n \operatorname{var}\{Y_{+j}\} + \frac{1}{n} Y_{++}^2 - Y_{++} . \end{aligned}$$

This implies that if the total number of ties, or equivalently the average degree ($= Y_{++}/n$) is fixed, then the number of out-2-stars is a function of the out-degree variance.

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

This implies that the numbers of out- and in-2-stars represent the variability among the out- and in-degrees.

The number of two-paths represent the covariance between out- and in-degrees.

These degree variances and covariance are to some extent functionally similar to the variance and covariance of U_i and V_i in the p_2 model.

The preceding specification of the *ERGM* represents popularity and activity differences between actors as well as the two primary structural effects: reciprocity and transitivity.

A graph is a *Markov graph* if for each set of 4 distinct actors i, j, h, k , the tie indicators Y_{ij} and Y_{hk} are *independent, conditionally* on all the other ties.

This is a very strong conditional independence assumption, and assumption easily leads to the problems of degeneracy of the random graph distribution.

Therefore, Snijders, Pattison, and Robins (2004) proposed a new specification for the ERGM model, defined by other choices of the statistics $z(y)$.

The new statistics are:

1. alternating in/out- k -star combinations
2. alternating independent 2-paths combinations
3. alternating k -transitive triplets combination.

The alternating in/out- k -star combinations represent the distribution of the in/out-degrees;

the alternating independent 2-paths represent the covariance between in-and out-degrees,

and the *conditions* for transitivity;

the alternating k -transitive triplets represent transitivity.

These are explained as follows.

Alternating in/out- k -star combinations

We saw above that the number of out-2-stars is a function of the out-degree variance.

Therefore, if we wish to represent graphs with a high dispersion of the out-degrees, then we wish to have many out-2-stars; but not too many, since that could bring us too close to a complete graph or another type of degenerate graph where some nodes have out-degree 0 and all others have the same out-degree K for some high value K .

This can be represented as follows:

we wish many 2-stars, but not too many 3-stars.

More generally, an out-degree equal to $k + h$ contributes $\binom{k+h}{k}$ k -stars, which is very large if h is large.

Therefore, if we wish not too have degenerate graphs, a high number of k -stars should be balanced by a not-too-high number of $(k + 1)$ -stars.

If we denote the number of out- k -stars by S_k , this leads to the idea of combining the k -star counts with *alternating signs*.

This is implemented by the following definition of *alternating out- k -star combinations* :

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

The number $\lambda \geq 1$ is a constant, chosen by the researcher, which ensures that the contribution of higher-order star counts is downweighted.

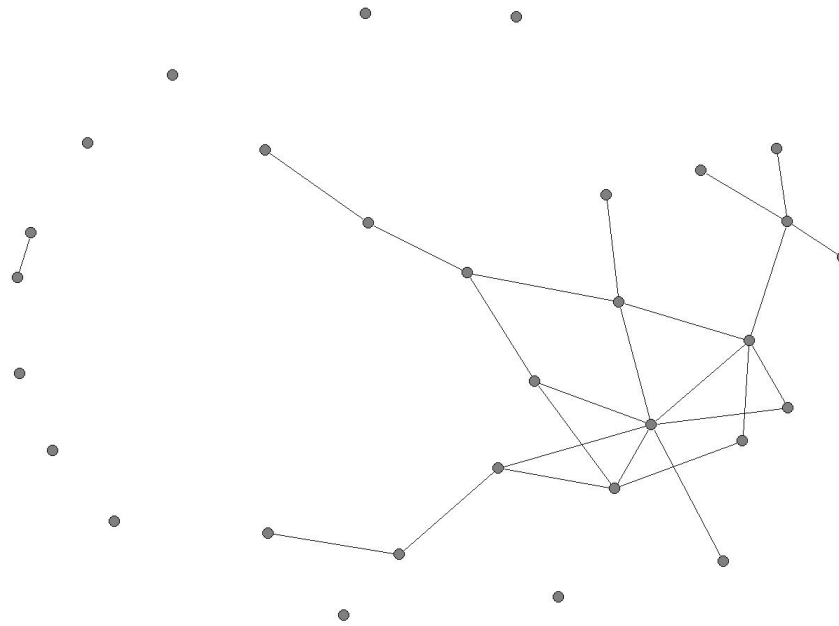
The formula above shows that this linear combination of k -star counts is a function of the out-degrees y_{i+} ; the function

$$\sum_{i=1}^n \left\{ \left(1 - \frac{1}{\lambda}\right)^{y_{i+}} + \frac{y_{i+}}{\lambda} - 1 \right\}$$

is an increasing function of the degrees y_{i+} , and high values of the degrees y_{i+} are downweighted more strongly when the parameter λ is closer to 1.

Therefore, including this statistic in the vector $z(\mathbf{y})$ of the ERGM can help to give a good fit to the distribution of the out-degrees. This can be in addition to, or in replacement of, the number of out-2-stars.

The following graph is the outcome of simulating
with an edge parameter of -4.3
and an alternating k -star parameter ($\lambda = 2$) of 0.65 .
This is a low density graph with 25 edges and a density of 0.06 .

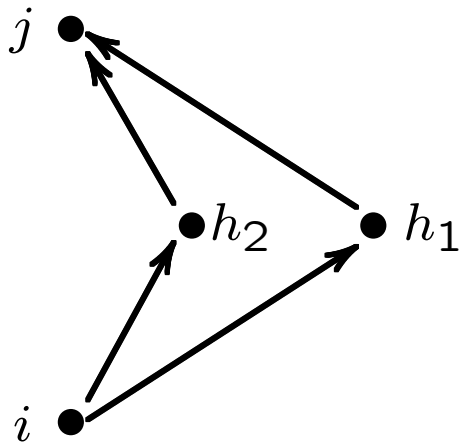


Alternating independent 2-paths

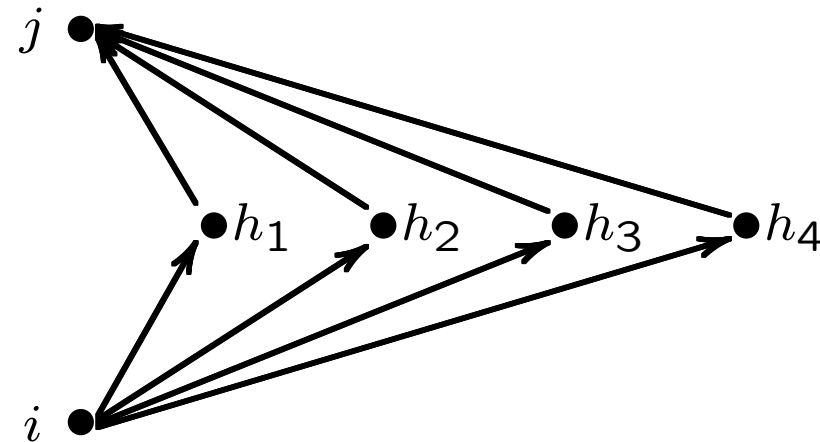
We saw earlier that two-paths are a precondition for a meaningful (non-vacuous) concept of transitivity; and the number of two-paths is also related to the covariance between in- and out-degrees.

Similar to the alternating k -star combinations, having many but not too many two-paths between two nodes i and j can be obtained by combining two-path counts of higher order with alternating signs.

This statistic is chosen in a way that is combined with the following statistic (alternating k -transitive triplets) and must be understood together with the latter statistic.



independent 2-two-paths



independent 4-two-paths

The graph-theoretical definition is that two-paths are independent if they are non-intersecting.

Define L_{2ij} as the number of two-paths connecting i and j ,

$$L_{2ij} = \sum_{h \neq i, j} y_{ih} y_{hj} .$$

Then the number of k -independent 2-paths can be defined by the formula

$$\begin{aligned} U_k &= \# \left\{ \left((i, j), \{h_1, h_2, \dots, h_k\} \right) \mid \right. \\ &\quad \left. i \neq j, y_{ih_\ell} = y_{h_\ell j} = 1 \text{ for } \ell = 1, \dots, k \right\} \\ &= \sum_{i, j} \binom{L_{2ij}}{k} . \end{aligned}$$

The corresponding statistic with alternating and geometrically decreasing weights is

$$\begin{aligned} u(y) &= U_1 - \frac{1}{\lambda} U_2 + \dots + \left(\frac{-1}{\lambda}\right)^{n-3} U_{n-2} \\ &= \lambda \sum_{i < j} \left\{ 1 - \left(1 - \frac{1}{\lambda}\right)^{L_{2ij}} \right\}. \end{aligned}$$

The term

$$\left\{ 1 - \left(1 - \frac{1}{\lambda}\right)^{L_{2ij}} \right\}$$

increases from 0 for $L_{2ij} = 0$ to almost 1 for $L_{2ij} = n - 2$.

For $\lambda = 1$ the statistic reduces to

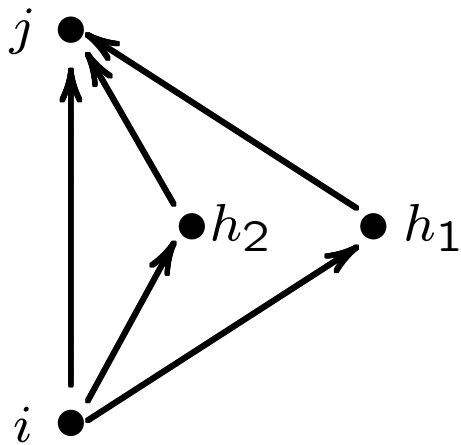
$$u(y) = \sum_{i,j} I\{L_{2ij} \geq 1\} ,$$

the number of ordered pairs (i, j)

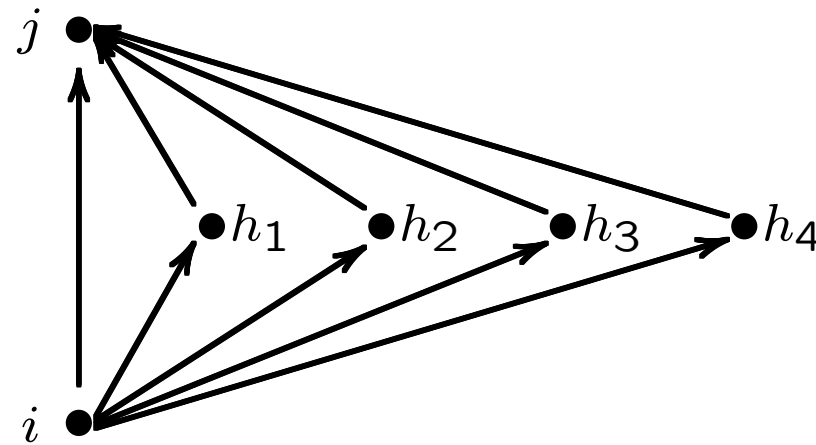
indirectly connected by at least one two-path.

Alternating k -transitive triplets

k -transitive triplets are defined as multiple transitive relations on the same “base tie” $i \rightarrow j$:



2-transitive triplet



4-transitive triplet

The formula for the number of k -transitive triplets is

$$\begin{aligned} T_k &= \# \left\{ \left((i, j), \{h_1, h_2, \dots, h_k\} \right) \mid \right. \\ &\quad \left. y_{ij} = 1 \text{ and } y_{ih_\ell} = y_{h_\ell j} = 1 \text{ for } \ell = 1, \dots, k \right\} \\ &= \sum_{i < j} y_{ij} \binom{L_{2ij}}{k} . \end{aligned}$$

The linear combination of k -transitive triplets with alternating signs is

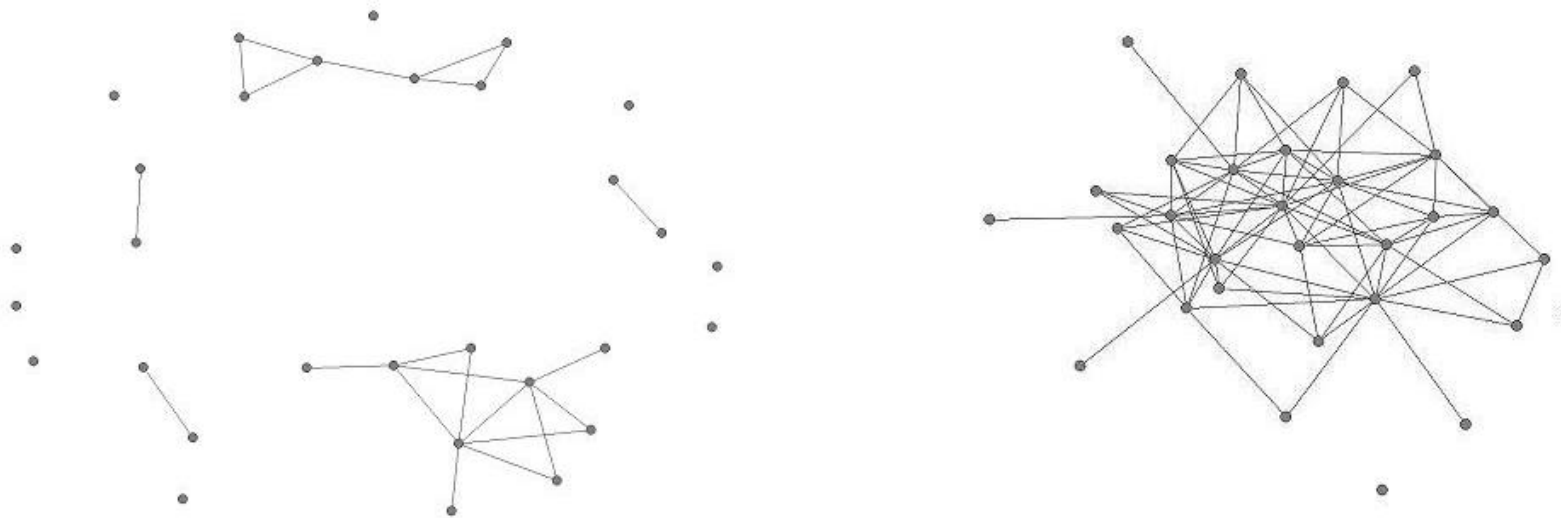
$$\begin{aligned} u(y) &= T_1 - \frac{T_2}{\lambda} + \frac{T_3}{\lambda^2} - \dots + (-1)^{n-3} \frac{T_{n-2}}{\lambda^{n-3}} \\ &= \lambda \sum_{i < j} y_{ij} \left\{ 1 - \left(1 - \frac{1}{\lambda} \right)^{L_{2ij}} \right\}. \end{aligned}$$

This counts the number of transitive triplets but downweights the large numbers of transitive triplets formed by the single tie $i \rightarrow j$ when there are many actors linked to i as well as j — i.e., when L_{2ij} is large.

For $\lambda = 1$ the statistic is equal to

$$u(y) = \sum_{i < j} y_{ij} I\{L_{2ij} \geq 1\} ,$$

the number of ordered pairs (i, j) that are directly linked ($y_{ij} = 1$) but also indirectly linked ($y_{ih} = y_{hj} = 1$ for at least one other node h).



A low density and a higher density k -transitive triplets graphs

Edge parameter = -3.7 for both;

alternating k -transitive triplets parameter = 1.0 and 1.2 ,

respectively ($\lambda = 2$).

The proposed way of specifying the ERGM now is:

choose a value of λ — e.g., 1, 2, or 3;

and include the following statistics:

1. The total number of edges $S_1(y)$,
to reflect the density of the graph;
2. the alternating out- k -star combinations
and in- k -star combinations,
to reflect the heterogeneity of the degrees;
3. this can be supplemented, or replaced,
by the number of out-two-stars, in-two-stars, and two-paths,
provided that this leads to a good model fitting procedure;

4. the alternating k -transitive triplet combinations and the alternating combinations of k -independent two-paths, to reflect transitivity;
5. this can be supplemented with the triad count $T(y) = T_1(y)$, if a satisfactory estimate can be obtained for the corresponding parameter, and if this yields a better fit as shown from the t -statistic for this parameter.

The choice of a suitable value of λ depends on the data set.

Of course, actor and dyadic covariate effects can also be added. Three different effects are possible for each actor covariate v_i .

These are represented by the following statistics (potential elements of $z(y)$)

1. *covariate-related activity*,

sum of sender-covariate over all ties, $\sum_{i,j} y_{ij} v_i$;

2. *covariate-related popularity*,

sum of receiver-covariate over all ties $\sum_{i,j} y_{ij} v_j$;

3. *covariate-related dissimilarity*,

sum of absolute covariate differences over all ties

$$\sum_{i,j} y_{ij} |v_i - v_j|.$$

Graphs according to the ERGM can be simulated, and the parameters of the ERGM can be estimated by a Markov Chain Monte Carlo ('MCMC') procedure described in Snijders (2002) and implemented in the *SIENA* program. A requirement is that the data are complete (no missings!).

The stability of the estimation process is much improved if the total number of ties is kept fixed ('conditioned upon').

The following table presents MCMC parameter estimates for the friendship relation between Lazega's lawyers, with $\lambda = 2$.

Parameter	Model 1		Model 2	
	est.	s.e.	est.	s.e.
Mutual dyads	1.653	0.257	2.405	0.291
Out-two-stars	—	—	0.131	0.015
In-two-stars	—	—	0.149	0.024
two-paths	—	—	−0.090	0.019
Alternating out- k -stars	0.235	0.295	−0.333	0.341
Alternating in- k -stars	−1.143	0.551	−2.096	0.927
Alternating k -trans. triplets	0.661	0.142	0.704	0.146
Alternating indep. two-paths	−0.140	0.031	−0.065	0.034
Same office	0.582	0.125	0.859	0.169
Seniority popularity	−0.002	0.007	0.001	0.007
Seniority activity	0.013	0.007	0.010	0.006
Seniority dissimilarity	−0.039	0.008	−0.041	0.008
Practice (corp. law) popularity	−0.052	0.160	0.067	0.114
Practice (corp. law) activity	0.328	0.137	0.200	0.097
Same practice	0.279	0.125	0.292	0.129

Other effects, such as the number of transitive triplets and the numbers of three-stars and four-stars are also represented adequately, each with a difference between observed and estimated expected value of less than 1.5 standard deviation.

Software for parameter estimation
of these models is available in the
StOCNET package

at <http://stat.gamma.rug.nl/stocnet/> .

Estimation for p_1 is in the Examine option;
the p_2 model has a module all for itself;
and the MCMC estimation of the p^* model
is in the SIENA module.

References

Frank, O. 1991. Statistical analysis of change in networks. *Statistica Neerlandica*, 45, 283 – 293.

Frank, O., and Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, 81, 832 – 842.

Holland, P.W., and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs (with discussion). *Journal of the American Statistical Association*, 76, 33 – 65.

Lazega, E., and van Duijn, M.A.J. (1997). Position in formal structure, personal characteristics and choices of advisors in a law firm: a logistic regression model for dyadic network data. *Social Networks* 19, 375 – 397.

Robins, G.L., Woolcock, J., and Pattison, P. (2004). Small and other worlds: Global network structures from local processes. *American Journal of Sociology*, in press.

Snijders, T.A.B. 2002. Markov Chain Monte Carlo Estimation of Exponential Random Graph Models.

Journal of Social Structure, Vol. 3 (2002), No. 2. Available from <http://www2.heinz.cmu.edu/project/INSNA/joss/index1.html>.

Snijders, T.A.B., Pattison, P.E., and Robins, G. (2004). New specifications for exponential random graph models. *Manuscript in preparation*.

van Duijn, M.A.J. (1995). Estimation of a random effect model for directed graphs. *Symposium Statistische Software 1995*.

van Duijn, M.A.J., Snijders, T.A.B., & Zijlstra, B.H. (2004). p_2 : a random effects model with covariates for directed graphs. *Statistica Neerlandica*, in press.

Wasserman, S., and K. Faust (1994). *Social Network Analysis: Methods and Applications*. New York and Cambridge: Cambridge University Press.

Wasserman, S., and P. Pattison (1996). Logit models and logistic regression for social networks: I. An introduction to Markov graphs and p^* . *Psychometrika* 61, 401 – 425.