Transitivity and Triads

Tom A.B. Snijders

University of Oxford

May 14, 2012

Outline

Local Structure – Transitivity

Markov Graphs

Local Structure in Social Networks

From the standpoint of structural individualism,
one of the basic questions in modeling social networks is,
how the global properties of networks can be understood
from local properties.

A major example of this is the theory of clusterability
of balanced signed graphs.

Harary's theorem says
that a complete signed graph is balanced
if and only if the nodes can be partitioned into two sets
so that all ties within sets are positive,
and all ties between sets are negative.

This was generalized by Davis and Leinhardt
to conditions for clusterability of signed graphs
and structures of ranked clusters;
see Chapter 6 in Wasserman and Faust (1994).

These theories are about the question,
how triadic properties of signed graphs,
i.e., aggregate properties of all subgraphs of 3 nodes,
can determine global properties of signed graphs.

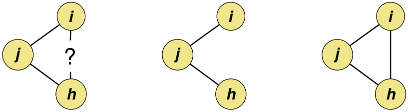
This presentation is about such questions
for graphs without signs.

Transitivity

Transitivity of a relation means that
when there is a tie from i to j, and also from j to h,
then there is also a tie from i to h:

friends of my friends are my friends.

Transitivity depends on *triads*, subgraphs formed by 3 nodes.



Potentially transitive | Intransitive | Transitive

Transitive graphs

One example of a (completely) transitive graph is evident:
the complete graph K_n, which has n nodes and density 1.
(The K is in honor of Kuratowski, a pioneer in graph theory.)

Is the empty graph transitive?

Try to find out for yourself,
what other graphs exist that are completely transitive!

Measure for transitivity

A measure for transitivity is the (global) transitivity index, defined as the ratio

$$\text{Transitivity Index} \;=\; \frac{\sharp\text{Transitive triads}}{\sharp\text{ Potentially transitive triads}}\ .$$

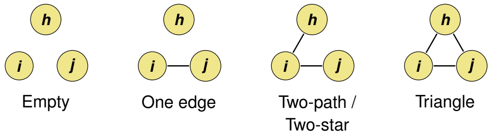
(Note that "$\sharp A$" means the number of elements in the set A.)
This also is sometimes called a *clustering* index.

This is between 0 and 1; it is 1 for a transitive graph.
For random graphs, the expected value of the transitivity index is close to the density of the graph (why?);
for actual social networks,
values between 0.3 and 0.6 are quite usual.

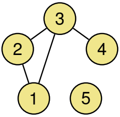
Local structure and triad counts

The studies about transitivity in social networks led Holland and Leinhardt (1975) to propose that the *local structure* in social networks can be expressed by the *triad census* or *triad count*, the numbers of triads of any kinds.

For (nondirected) graphs, there are four triad types:



A simple example graph with 5 nodes.



i	j	h	triad type
1	2	3	triangle
1	2	4	one edge
1	2	5	one edge
1	3	4	two-star
1	3	5	one edge
1	4	5	empty
2	3	4	two-star
2	3	5	one edge
3	4	5	one edge

In this graph, the triad census is $(1, 5, 2, 1)$
(ordered as: empty – one edge – two-star – triangle).

It is more convenient to work with *triplets* instead of triads: triplets are like triads, but they refer only to the presence of the edges, and do not require the absence of edges.

E.g., the number of two-star triplets is the number of potentially transitive triads.

The triplet count for a non-directed graph is defined by the number of edges, the total number of two-stars (irrespective of whether they are embedded in a triangle), and the number of triangles.

In the 5-node example graph, the triplet-based summary is:

$L = 4$ edges: $(1-2);\ (2-3);\ (1-3);\ (3-4)$.
$S_2 = 5$ two-stars:

$\Big(1-(2,3)\Big);\ \Big(2-(1,3)\Big);\ \Big(3-(1,2)\Big);\ \Big(3-(1,4)\Big);\ \Big(3-(2,4)\Big)$.

$T = 1$ triangle: $(1,2,3)$.

(The fourth degree of freedom:
for $n = 5$ nodes there are $\binom{5}{3} = 10$ triads.)

Formulae

Triplet counts can be defined
by more simple formulae than triad counts.
If the edge indicator (or tie variable) from i to j is denoted Y_{ij}
(1 if there is an edge, 0 otherwise)
then the formulae are:

$$
\begin{aligned}
L &= \tfrac{1}{2}\sum_{i,j} Y_{ij} && \text{edges} \\
S_2 &= \tfrac{1}{2}\sum_{i,j,k} Y_{ij}\, Y_{ik} && \text{two-stars} \\
T &= \tfrac{1}{6}\sum_{i,j,k} Y_{ij}\, Y_{ik}\, Y_{jk} && \text{triangles}
\end{aligned}
$$

Some algebraic manipulations can be used to show that
the *degree variance*, i.e.,
the variance of the degrees Y_{i+}, can be expressed as

$$\text{var}(Y_{i+}) = \frac{2}{n}S_2 + \frac{1}{n}L - \frac{1}{n^2}L^2 .$$

This shows that for non-directed graphs,
the triad census gives information equivalent to:
density, degree variance, and transitivity index.

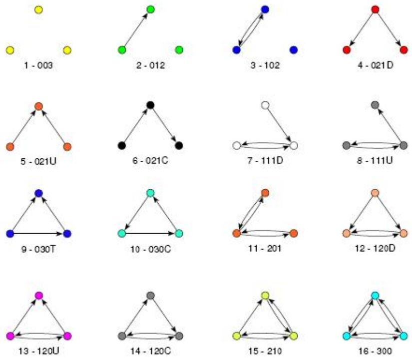
This can be regarded as a basic set of descriptive statistics
for a non-directed network.

Holland and Leinhardt's (1975) proposition was,
that many important theories about social relations
can be tested by means of hypotheses about the triad census.

They focused on directed rather than non-directed graphs.

The following picture gives the 16 different triads
for directed graphs.

The coding refers to
the numbers of *mutual*, *asymmetric*, and *null dyads*,
with a further identifying letter: Up, Down, Cyclical, Transitive.
E.g., 120D has 1 mutual, 2 asymmetric, 0 null dyads,
and the Down orientation.



Probability models for networks

The statistical approach proposed by Holland and Leinhardt now is obsolete.

Since 1986, statistical methods have been proposed
for probability distributions of graphs
depending primarily on the triad or triplet counts,
complemented with star counts and nodal variables.

It has been established recently
that, in addition, inclusion of higher-order configurations
(subgraphs with more nodes)
is essential for adequate modeling of empirical network data.

In the statistical approach to network analysis,
the use of probability models is *model based*
instead of *sampling based*.

If we are analyzing one network,
then the statistical inference is about this network only,
and it is supposed that the network
observed between these actors could have been different:

the ties are regarded as
the realization of a probabilistic social process
where 'probability' comes in as a result of influences
not represented by nodal or dyadic variables ('covariates')
and of measurement errors.

Markov graphs

In probability models for graphs, usually the set of nodes is fixed
and the set of edges (or arcs) is random.

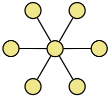
Frank and Strauss (1986) defined that
a probabilistic graph is a *Markov graph* if
for each set of 4 *distinct* actors i, j, h, k,
the tie indicators Y_{ij} and Y_{hk} are *independent*,
conditionally on all the other ties.

This generalizes the concept of Markov dependence for
time series, where random variables are ordered by time,
to graphs where the random edge indicators are ordered
by pairs of nodes.

Frank and Strauss (1986) proved that
a probability distribution for graphs,
under the assumption that the distribution
does not depend on the labeling of nodes,
is Markov if and only if it can be expressed as

$$P\{Y = y\} = \frac{\exp\left(\theta L(y) + \sum_{k=2}^{n-1} \sigma_k S_k(y) + \tau T(y)\right)}{\kappa(\theta, \sigma, \tau)}$$

where L is the edge count,
T is the triangle count,
S_k is the k-star count, and
$\kappa(\theta, \sigma, \tau)$ is a normalization constant
to let the probabilities sum to 1.



a 6-star

It is in practice not necessary to use all k-star parameters,
but only parameters for lower-order stars,
like 2-stars and 3-stars.

Varying the parameters leads to quite different distributions.
E.g., when using k-stars up to order 3, we have:

- higher θ gives more edges $\Rightarrow$ higher density;
- higher σ_2 gives more 2-stars $\Rightarrow$ more degree dispersion;
- higher σ_3 gives more 3-stars $\Rightarrow$ more degree skewness;
- higher τ gives more triangles $\Rightarrow$ more transitivity.

But note that having more triangles and more k-stars
also implies a higher density!

Small and other worlds

Robins, Woolcock and Pattison (2005)
studied these distributions in detail and investigated their potential to generate *small world networks* (Watts, 1999) defined as networks with many nodes, limited average degrees, low geodesic distances and high transitivity.

(Note that high transitivity in itself will lead to long geodesics.)

They varied in the first place the parameters τ, σ_k
and then adjusted θ to give a reasonable average degree.
All graphs have 100 nodes.

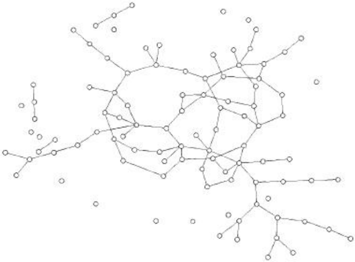


FIG. 9.—A Bernoulli graph

Bernoulli graph: random

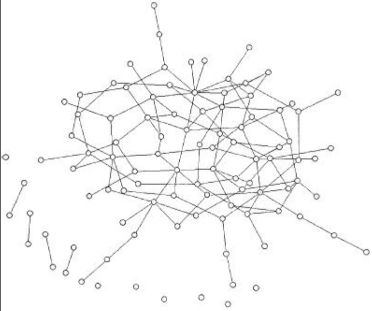


FIG. 5.—A small world graph

$(\theta, \sigma_2, \sigma_3, \tau) =$

$(-4, 0.1, -0.05, 1.0)$

small-world graph:
high transitivity,
short geodesics

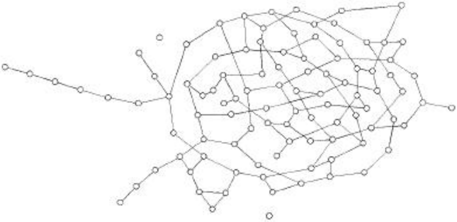


FIG. 7.—A graph with long median paths

$(\theta, \sigma_2, \sigma_3, \tau) = (-1.2, 0.05, -1.0, 1.0)$

long paths; few high-order stars

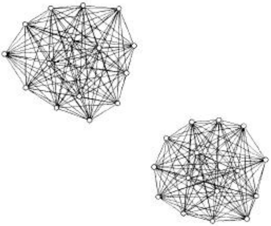
FIG. 8.—A long path graph with low clustering

$(\theta, \sigma_2, \sigma_3, \tau) =$

$(-2.0, 0.05,$

$-2.0, 1.0)$

long paths
low transitivity



$(\theta, \sigma_2, \sigma_3, \tau) =$

$(-3.2, 1.0, -0.3, 3.0)$

caveman world

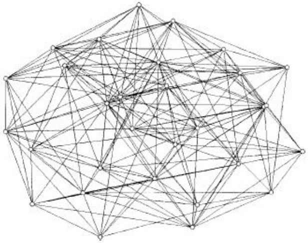


FIG. 11.—Effects of parameter scaling for two temperatures

$(\theta, \sigma_2, \sigma_3, \tau) =$

$(-0.533, 0.167,$

$-0.05, 0.5)$

heated
caveman world
(all parameters
divided by 6)

Thus we see that by varying the parameters,
many different graphs can be obtained.

This suggests that the Markov graphs will provide
a good statistical model for modeling observed social networks.

For some time, so-called *pseudo-likelihood methods* were used
for parameter estimation;
but these were shown to be inadequate.

Snijders (2002) and Handcock (2003) elaborated
maximum likelihood estimation procedures
using the Markov chain Monte Carlo (MCMC) approach.
These are now implemented in the software packages
SIENA, statnet, and pnet.

More general specifications

Markov graph models, however,
turn out to be not flexible enough to represent
the degree of transitivity observed in social networks.

It is usually necessary
for a good representation of empirical data
to generalize the Markov model and include in the exponent
also higher-order subgraph counts.

This means that the Markov dependence assumption
of Frank and Strauss is too strong, and less strict
conditional independence assumptions must be made.

The new models still remain in the framework of so-called
exponential random graph models (ERGMs),

$$\mathrm{P}_\theta\{Y = y\} \;=\; \frac{\exp\big(\sum_k \theta_k s_k(y)\big)}{\kappa(\theta)}$$

also called p^* models,
see Frank (1991), Wasserman and Pattison (1996),
Snijders, Pattison, Robins, and Handcock (2006).
Here the $s_k(y)$ are *arbitrary* statistics of the network,
including covariates, counts of edges, k-stars, and triangles,
but also counts of higher-order configurations.

Tutorials: both papers Robins et al. (2007).

Literature

- Frank, Ove, and David Strauss. 1986.
  "Markov Graphs." *Journal of the American Statistical Association*, 81: 832 – 842.
- Holland, P.W., and Leinhardt, S. 1975.
  "Local structure in social networks." In D. Heise (ed.), *Sociological Methodology*. San Francisco: Jossey-Bass.
- Snijders, Tom A.B. 2002.
  "Markov Chain Monte Carlo Estimation of Exponential Random Graph Models." *Journal of Social Structure*, 3.2.
- Snijders, T.A.B., Pattison, P., Robins, G.L., and Handock, M. 2006.
  "New specifications for exponential random graph models." *Sociological Methodology*, 99–153.

- Robins, G., Pattison, P., Kalish, Y., and Lusher, D. 2007.
  "An introduction to exponential random graph (p^*) models for social networks." *Social Networks*, 29, 173–191.
- Robins, G., Snijders, T., Wang, P., Handcock, M., and Pattison, P. 2007.
  "Recent developments in Exponential Random Graph (p^*) Models for Social Networks." *Social Networks*, 29, 192–215.
- Robins, G.L., Woolcock, J., and Pattison, P. 2005. "Small and other worlds: Global network structures from local processes." *American Journal of Sociology*, 110, 894–936.
- Wasserman, Stanley, and Katherine Faust. 1994.
  Social Network Analysis: Methods and Applications. New York and Cambridge: Cambridge University Press.
- Wasserman, Stanley, and Philippa E. Pattison. 1996.
  "Logit Models and Logistic Regression for Social Networks: I. An Introduction to Markov Graphs and p^*." *Psychometrika*, 61: 401 – 425.