



The Abdus Salam
International Centre for Theoretical Physics



SMR.1656 - 26

School and Workshop on Structure and Function of Complex Networks

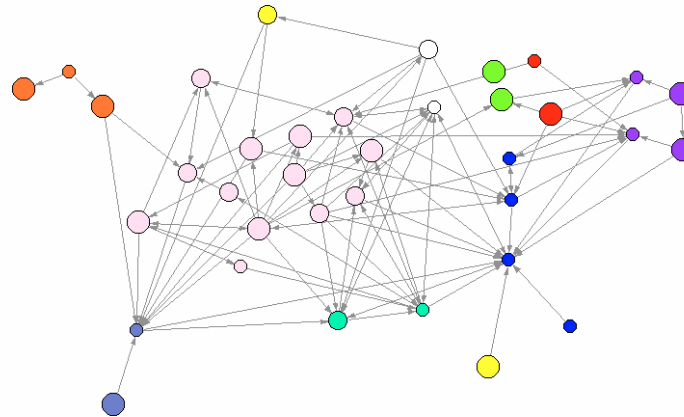
16 - 28 May 2005

Recent research in Statistical Models for Social Networks

**Stanley WASSERMAN
University of Indiana
Ballantine Hall 744
1020 E. Kirkwood Avenue
Bloomington, IN 47405-7103
U.S.A.**

These are preliminary lecture notes, intended only for distribution to participants

Recent Research in Statistical Models for Social Networks



Stanley Wasserman
Indiana University

ICTP, Trieste, May, 2005

(a subset of the powerpoint presentation given on Monday 23 May 2005)

Outline

- 1. Why statistical modeling?**
- 2. An exponential family of random graphs**
- 3. Model specification and homogeneous Markov random graphs**
- 4. Some new ideas**

1. Random graph models

Why is it important to *model* networks as completely as possible?

Modelling allows

precise inferences about the nature of regularities in networks and network-based processes from empirical observations

Quantitative estimates of these regularities (and their uncertainties) are important

small changes in these regularities can have substantial effects on global system properties

Modelling allows

an understanding of the relationship between (local) interactive network processes and aggregate (subsets of actors, group, community) outcomes

Modelling allows

formal assessment of goodness-of-fit (sorely lacking in much work today)

Approach to modelling networks

Guiding principles:

- 1. Network ties are the outcome of unobserved processes that tend to be local, interactive, and stochastic*
- 2. There are both regularities and irregularities in these local interactive processes. Goal is to model both. Usually irregularities dominate*
- 3. Focusing attention on just the degree (usually outdegree) distribution implicitly assumes a uniform random graph model, conditional on the degrees (which then implies that ties for a given individual actor are completely random (!!!!))*

Hence we aim for a stochastic model formulation in which:

local interactivity is permitted and assumptions about “locality” are explicit
regularities are represented by model parameters and estimated from data
consequences of local regularities for global network properties can be understood
(and can also provide an exacting approach to model evaluation)
contains more structure than usually assumed by standard power law models

Models for interactive systems of variables (Besag, 1974; Frank and Strauss, 1986; Wasserman and Pattison, 1996)

Two variables are *associated* if they are conditionally dependent given the observed values of all other variables

A *neighborhood* is a set of mutually dependent variables

A model for a system of variables has a form determined by its neighborhoods

Hammersley-Clifford theorem

This general approach leads to:

$\Pr(\mathbf{X} = \mathbf{x})$ *p^* -- an exponential family of random graph models*

Extension to directed dependence assumptions:

$\Pr(\mathbf{X} = \mathbf{x} | \mathbf{Y} = \mathbf{y})$ social selection models Robins et al 2001

$\Pr(\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{x})$ social influence models Robins et al 2002 (and many others)

$\Pr(\mathbf{X} = \mathbf{x} | \mathbf{S} = \mathbf{s})$ setting-dependent models Pattison & Robins 2002 (new!)

2. An exponential family of random graph (p^*) models

The Hammersley-Clifford Theorem uses defined neighborhoods
(or cliques in a dependence graph) and states:

$$\Pr(\mathbf{X} = \mathbf{x}) = (1/c) \exp\{\sum_Q \gamma_Q z_Q(\mathbf{x})\}$$

normalizing quantity

parameter

network statistic

the summation is over all neighborhoods Q

$z_Q(\mathbf{x}) = \prod_{X_{ij} \in Q} x_{ij}$ is an indicator reflecting whether
all ties in Q are observed in \mathbf{x}

$$c = \sum_{\mathbf{x}} \exp\{\sum_Q \gamma_Q z_Q(\mathbf{x})\}$$

3. Tools, Methods & Models

There are (at least) three general classes of statistical models for static networks:

1) Models of the network itself

The statistical question is how an observed network fits into the class of all possible random graphs with a given set of topological characteristics. The whole network is the substantive unit of analysis.

Examples: p^* models (Wasserman, Pattison, Robins, Snijders, Handcock)

2) Models of individual behavior that incorporate network characteristics

The statistical question is whether or not network properties affect individual behaviors.... Social influence models

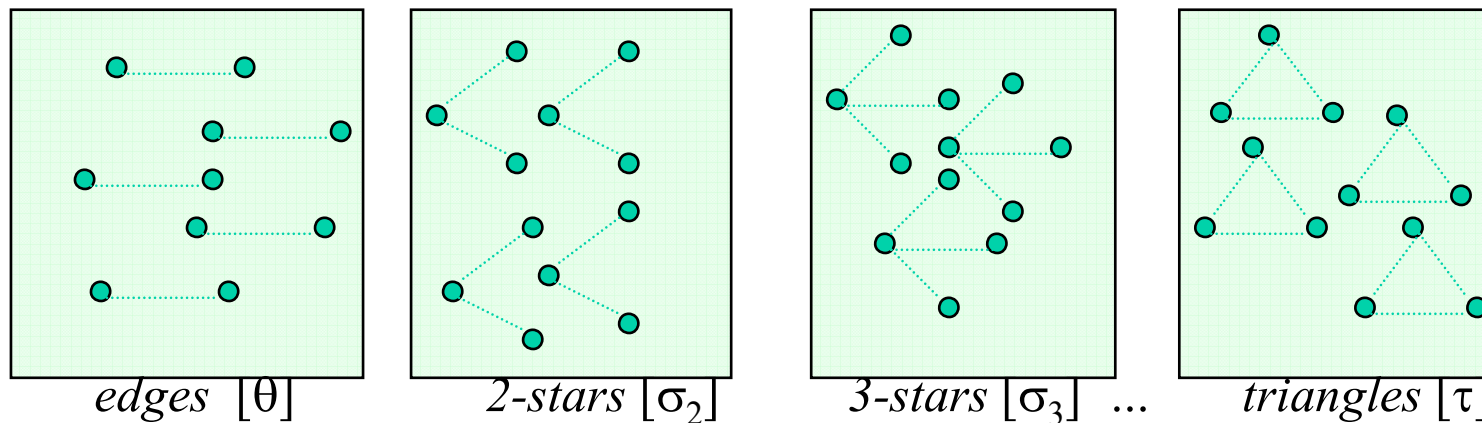
Examples: Network regressive-autoregressive models (Doriean), Peer influence models (Friedkin)

3) Models of degree distributions

3. Homogeneous network models

$$\Pr(\mathbf{X} = \mathbf{x}) = (1/c) \exp\{\sum_{Q^*} \gamma_{Q^*} z_{Q^*}(\mathbf{x})\}$$

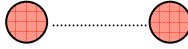
If we assume that parameters for *isomorphic* configurations are the same:

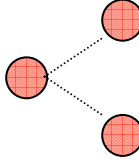


then there is one parameter γ_{Q^*} for each *class* Q^* of isomorphic configurations and the corresponding statistic $z_{Q^*}(\mathbf{x})$ is a *count* of such observed configurations in \mathbf{x}

Homogeneous Markov random graphs (Frank & Strauss, 1986)

$$\Pr(\mathbf{X} = \mathbf{x}) = (1/c) \exp\{\theta L(\mathbf{x}) + \sigma_2 S_2(\mathbf{x}) + \dots + \sigma_k S_k(\mathbf{x}) + \dots + \tau T(\mathbf{x})\}$$

where: $L(\mathbf{x})$ no of *edges* in \mathbf{x} 

$S_2(\mathbf{x})$ no of *2-stars* in \mathbf{x} 

...

$S_k(\mathbf{x})$ no. of *k-stars* in \mathbf{x} 

...

$T(\mathbf{x})$ no of *triangles* in \mathbf{x} 

References ...

- Chapters in Carrington, Scott, and Wasserman (2005). *Models and Methods in Social Network Analysis*. New York: Cambridge University Press.....

take a look at those chapters in this volume authored by Robins, Wasserman, Pattison, Koehly, Snijders, Huisman, and van Duijn

4. New ideas and new thoughts

a. Homogeneous Markov models

Handcock (2002) defines homogeneous Markov random graph models to be *degenerate* if most of the probability mass is concentrated in small parts of the state space

Regions of parameter space that are not degenerate may be quite small (Handcock, 2002)

Star parameters are particularly problematic

Simulation-based parameter estimation methods often wander into degenerate regions of parameter space (unless steering is excellent). Pseudo-likelihood estimation can be disastrous

Robins (2003) showed empirically that parameters estimated from data (using SIENA [Snijders, 2002]) can be quite close to degenerate regions

There are theoretical reasons to doubt the adequacy of a homogeneous Markov assumption

b. Better parameter estimation

“Logistic” Maximum Pseudo-Likelihood estimation:

<http://www.sfu.ca/~richards/Pages/pspar.html>

Markov chain Monte Carlo maximum likelihood estimation:

Ongoing work by Mark Handcock ([statnet](#)), Tom Snijders ([SIENA](#))

c. New parameter specifications

1. *alternating k -star statistics*
2. *other degree functions*
3. *more complicated “neighborhood/settings” parameters*