

SIMULATION-BASED  
STATISTICAL INFERENCE  
FOR EVOLUTION  
OF SOCIAL NETWORKS

*Tom A.B. Snijders*

*ICS*

*University of Groningen, The Netherlands*

Presentation at *IMA Workshop 3:*

*Networks and the Population Dynamics of Disease Transmission*, November 20, 2003

## Modeling social networks

Statistical paradigm for modeling:

aim is to use rather generally applicable models

(inherently stochastic)

including *a priori unknown parameters*

to be estimated from data;

explicit statements about parameter uncertainty;

models not more complex than empirically warranted;

important issues are *robustness*

and *goodness of fit / specification*.

Models for single (one-moment) observations of networks important for describing and capturing structure.

‘Mechanisms’ of network formation can be studied only by modeling *network dynamics* on the basis of *longitudinal* network data.

This presentation is about

1. models for network dynamics  
(Snijders *Soc. Meth.* 2001 and further work)
2. models for joint evolution of networks and behavior  
(ongoing work).

Substantive background: sociology.

Relevance to disease transmission:

⇒ social network is substrate for transmission process  
(for some diseases)

⇒ methods could be used for, or adapted to, disease transmission.

“Network effects” to be included in the dynamic model are typically endogenous feedback effects:  
e.g., reciprocity, transitivity, popularity, subgroup formation.  
These can be well expressed by microsimulation models.

Here: *microsimulation model used as statistical model,*

~ relatively simple for microsimulation models

~ relatively complicated for statistical models;

aiming to capture the essentials in a simple and robust model.

Currently: modeling applies only to relatively small groups;  
for more than a few hundred nodes, more thought must be given  
to specification of effects of settings/opportunity structures.

Group composition is fixed, but model can be extended  
to changing composition (exogenous / endogenous).

Methodological requirement:

make microsimulation model amenable for statistical analysis;  
use statistical procedures that can handle  
models expressed as simulation models.

Purpose of statistical inference:

investigate network & behavior evolution as function simultaneously  
of

1. structural effects
2. explanatory actor variables
3. explanatory dyadic variables

and test such effects, controlling for the other effects.

Longitudinal data collection and modeling of social networks has an *important advantage* over use of one-moment observations:

for modeling a single observation of a network,  
“everything depends on everything else”,  
which leads to big problems in modeling, statistical inference,  
and interpretation.

For longitudinal modeling of social networks,  
the first observation may be taken as given rather than modeled,  
and then the remaining dependence is unidirectional in time  
and less difficult to model.

*But: complete network at first time point required.*

*Statistical Methodology for the simultaneous evolution of networks  $X(t)$  and behavior  $Z(t)$ .*

Actors are influenced in their behavior and attitudes by other actors to whom they are tied

(e.g., N. Friedkin, *A Structural Theory of Social Influence*, C.U.P., 1998).

At the same time, actors choose relation partners on the basis of their behavior and other characteristics (similarity, opportunities for future rewards, etc.).

Constant or exogenous actor or dyad characteristics also possible, not mentioned explicitly.

### Examples:

- ⇒ Risky social behaviors (like smoking, taking alcohol or drugs) are 'contagious' among friends but also operative in friendship formation.
- ⇒ How hard pupils and employees work often is subject to social control.
- ⇒ Firms choose partners for collaboration based on complementary expertise, reputation, trust, etc.

This leads to a feedback relation in the dynamics of relational networks and actor behavior / performance.

*Digression:* how might this be adapted to infectious diseases?

Interpretation of the two components:

1. network  $X(t)$  of acquaintance / friendship / sexual relations defines opportunity structure for disease transmission
2.  $Z(t)$  is state of infection  
e.g.  $S - I - R$  where  $R$  actors are Removed  
(perhaps new actors might come in)  
(there could also be additional behavior variables, e.g. risk taking).

Model should express how dynamics of  $X(t)$  and  $Z(t)$  are interrelated.

The investigation of such social feedback processes is difficult:

- \* Both the *network*  $\Rightarrow$  *behavior* and the *behavior*  $\Rightarrow$  *network* effects may lead to an association current behavior — network: “friends of smokers are smokers”, “high-reputation firms don’t collaborate with low-reputation firms”.

It is hard to ascertain the strengths of the causal relations in the two directions.

- \* For many phenomena quasi-continuous longitudinal observation is infeasible. Instead, it may be possible to observe networks and behaviors at a few discrete time points. Such an observation design is the point of departure here.

The repeated measurements must show enough change  
(period from first to last observation not too short)  
but not too much  
(time lapse between subsequent observations not too long)  
in order to give information about rules of network dynamics.

*Principle:* regard these observations as discrete observations of a process developing in *continuous time*, where actors can make unobserved changes between the observation moments, being each others' changing environment:  
*actor-oriented model.*

This seems natural for many processes on directed graphs.

*For non-directed relations:* two actors must agree on the tie.

*Alternative:* tie-oriented models.

An advantage of using continuous-time models, even if observations are made at a few discrete time points, is that a more natural and simple representation may be found, especially in view of the endogenous dynamics.

(Cf. Coleman, 1964).

No problem with irregularly spaced data.

For *discrete data*:

Kalbfleisch & Lawless, JASA, 1985;

for *continuous data*:

mixed state space modelling well-known in engineering,  
in economics e.g. Bergstrom (1976, 1988),  
in social science Tuma & Hannan (1984),  
work by H. Singer in the 1990s.

## Basic ideas for probability model:

- \* Embed the observations in a latent continuous-time stochastic process. This represents the continuous dynamics in networks and behavior, independent of being observed.
- \* Represent the (unobserved) changes in this continuous-time process as the result of the optimization of actor 'utilities' plus a random component (cf. random utility modeling for discrete choices).
- \* Integrate the *influence/contagion* (network  $\Rightarrow$  characteristics) and *selection* (characteristics  $\Rightarrow$  network) processes.

**Notation:**

set of  $n$  actors, with a dichotomous ('on/off') relation, represented as a *directed graph (digraph)*.

Tie variable from  $i$  to  $j$  indicated by  $X_{ij}(t)$  :

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

(Diagonal values  $X_{ii}(t)$  meaningless.)

$X_{ij}(t)$  is *arc* or *directed line* from  $i$  to  $j$ .

Matrix  $X(t)$  is *adjacency matrix* of digraph.

In addition, associated to each actor  $i$  there is a vector  $Z_i(t)$  of actor characteristics indexed by  $h = 1, \dots, H$ .

For the moment: ordered discrete

(simplest case: one dichotomous variable).

## Actor-oriented models :

each actor “controls” his outgoing ties,  
collected in the row vector  $(X_{i1}(t), \dots, X_{in}(t))$ ,  
and his characteristic  $Z_i(t) = (Z_{i1}(t), \dots, Z_{iH}(t))$   
( $H$  is the number of dependent actor variables).

At stochastic times

(*rate functions*  $\lambda^X$  for changes in network,

$\lambda_h^Z$  for changes in characteristic  $h$ ),

the actors may change a tie or a characteristic: *mini-step*.

The actors try to attain a rewarding configuration of the network  
expressed in *objective functions*  $f^X$  and  $f^Z$  .

Actions propelled also by a *random component*.

### Mini-step for change in network:

At random moments occurring at a rate  $\lambda^X$ , a random actor is designated to make a change in one tie variable: the *mini-step* (on  $\Rightarrow$  off, or off  $\Rightarrow$  on.)

### Mini-step for change in characteristic:

At random moments occurring at a rate  $\lambda_h^Z$ , a random actor is designated to make a change in characteristic  $h$  (one component of  $Z_i$ , assumed to be ordinal): the *mini-step* is a change to an adjacent category.

Many mini-steps can *accumulate* to big differences between consecutive observations.

When actor  $i$  'may' change an outgoing tie to some other actor  $j$ , he/she chooses the 'best'  $j$  by maximizing the objective function of the situation obtained after the coming network change plus a random component representing unexplained influences;

and when this actor 'may' change behavior  $h$ , he/she chooses the "best" change (up, down, nothing) by maximizing the objective function of the situation obtained after the coming behavior change plus a random component representing unexplained influences.

This leads to multinomial logit models for the results of the ministeps.

*Digression* for infectious diseases:

A 'utility maximizing' model seems less applicable to humans than to viruses.

Rather: if  $Z_{ih} = 0/1$  denotes susceptible / infected, then at a rate  $\lambda_h^Z$ , if  $z_{ih} = 1$ ,  $z_{jh} = 0$ ,  $x_{ij} = 1$ ,  $i$  will infect  $j$ .

(*Alternative:*

or, if  $z_{ih} = 1$ ,  $i$  infects one of his partners, randomly chosen.)

The rates  $\lambda^X$ ,  $\lambda_h^Z$  may also depend on the characteristics  $z_i$ , or the network position of  $i$ .

With this specification, the model  $(X(t), Z(t))$  is a continuous-time Markov process with a discrete (but very large) state space.

It can be simulated on computer in a rather straightforward way. Change probabilities can be calculated as in multinomial logit modeling.

No assumption is made of stationarity of the distribution.

A variety of models should be considered to obtain a good fit.

## Simple model specification:

- \* *Constant rate functions:*

The actors all change their relationships at constant rate  $\lambda^X$ , and each of their characteristics  $h$  at constant rates  $\lambda_h^Z$ .

- \* When changing  $X$ , actor  $i$  tries to optimize an *objective function* with respect to the network configuration,

$$f_i^X(\theta, x, z),$$

and when changing  $Z$ , this actor tries to optimize an *objective function* with respect to the characteristics,

$$f_i^Z(\theta, x, z).$$

## Specification of the model

Objective functions  $f_i^X$  and  $f_i^Z$  :

both reflect effects of network and of individual characteristics.

Convenient definition of objective functions

$f_i^X$  and  $f_i^Z$  is a weighted sum, e.g.,

$$f_i^X(\beta, x) = \sum_{k=1}^L \beta_k^X s_{ik}(x, z),$$

where weights  $\beta_k^X$  are statistical parameters indicating strength of effect  $s_{ik}(x, z)$  on network evolution.

Some pure network effects for actor  $i$ :  
(others to whom actor  $i$  is tied are called here  $i$ 's 'friends')

First two basic effects:

1. *density effect*,

out-degree

$$s_{i1}(x) = x_{i+} = \sum_j x_{ij} \quad ;$$

2. *reciprocity effect*,

number of reciprocated relations

$$s_{i2}(x) = \sum_j x_{ij} x_{ji} \quad ;$$

A well-known effect is network closure / clustering / transitivity:

“friends of my friends are my friends”

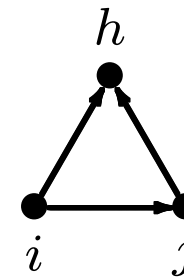
⇒ ties tend to be concentrated in loose subgroups.

Network closure can be modeled by different effects, e.g.

3. *transitivity effect*, number of transitive triplets

$$(i \rightarrow j, j \rightarrow h, i \rightarrow h)$$

$$s_{i3}(x) = \sum_{j,h} x_{ij} x_{jh} x_{ih}$$



transitive triplet

4. *indirect relations effect*,

number of actors to whom  $i$  is indirectly related  
(through one intermediary),

= number of geodesic distances equal to 2,

$$s_{i4}(x) = \#\{j \mid x_{ij} = 0, \max_h (x_{ih} x_{hj}) > 0\}$$

5. *balance* or structural equivalence,  
similarity between out-relations of  $i$   
with out-relations of his friends,

$$s_{i5}(x) = \sum_{j=1}^n x_{ij} \sum_{\substack{h=1 \\ h \neq i,j}}^n \left(1 - |x_{ih} - x_{jh}|\right),$$

(Also compare the alternating  $k$ -triangles discussed yesterday  
by P. Pattison and G. Robins – *embarras de choix ...* )

Differences between these three network closure effects:

⇒ transitive triplets effect:

$i$  more attracted to  $j$  if there are  
*more* indirect ties  $i \rightarrow h \rightarrow j$  ;

⇒ negative indirect connections effect:

$i$  more attracted to  $j$  if there is  
*at least one* such indirect connection ;

⇒ balance effect:

$i$  prefers others  $j$  who make same choices as  $i$ .

Non-formalized theories usually do not distinguish between these different closure effects.

It is possible to 'let the data speak for themselves' and see what is the best formal representation of closure effects.

Empirical experience has shown up to now,  
that negative indirect connections effect often is more important  
in driving the dynamics of friendship networks  
than the number of transitive triplets.

There are many different theoretical reasons why networks are important for behavior: e.g.,

1. *social capital* :

individuals may use resources of others;

2. *coordination* :

individuals can achieve some goals only by concerted behavior;

3. *imitation* :

individuals imitate others

(basic drive; uncertainty reduction);

4. *strategy*:

individuals try to reach relational goals  
by supposedly purposive behavior.

In this presentation, only imitation is considered, but other reasons can be of eminent importance.

There are many kinds of effect associated with actor characteristic  $z_{ih}$  ; the usually strongest effect, which can represent imitation, is the (dis)similarity effect,

6. *covariate-related dissimilarity* for binary or numerical  $z_{ih}$ ,  
sum of absolute covariate differences  
between  $i$  and his friends

$$s_{i6}(x, z) = \sum_j x_{ij} |z_{ih} - z_{jh}| .$$

## Statistical estimation

Observation moments  $t_1, t_2, \dots, t_M$   
with complete data on network and behavior.

Statistical parameters in this model:  
change rates  $\lambda$  ; weights  $\beta$  .

How to estimate  $\theta = (\lambda, \beta)$  ?

*Condition on  $X(t_1), Z(t_1)$  :*

the first observation is accepted as given,  
contains in itself no observation about  $\theta$  .

*No assumption of a stationary network distribution.*

Various estimation procedures possible.

- \* *Method of moments / estimating equations*
- \* *maximum likelihood* in progress (work with Johan Koskinen).

### *Method of moments :*

determine value of  $\theta = (\lambda, \beta)$  for which observed and expected values of suitable statistic are equal (statistic with as many components as there are unknown parameters).

To define expected values:

treat dynamics in each time interval from  $t_m$  to  $(t_{m+1} - \epsilon)$  as independent processes, starting at observed values  $x(t_m), z(t_m)$ .

## Questions:

- \* How to determine a suitable statistic?
- \* How to find this value of  $\theta$ ?  
By stochastic approximation (Robbins-Monro process) based on repeated simulations of the dynamic process, with parameter values getting closer and closer to the moment estimates.

Skip details.

## Suitable statistics for method of moments

E.g. for change rate in network in first observation period:  
relevant statistic is amount of change

$$\sum_{\substack{i, j=1 \\ i \neq j}}^n | X_{ij}(t_2) - x_{ij}(t_1) | .$$

Similar for the other change rates.

For weight parameters  $\beta_k^X$  in network objective function

$$f_i^X(\beta, x) = \sum_{k=1}^L \beta_k^X s_{ik}(x, z),$$

define the statistic

$$S_k^X = \sum_{m=2}^M \sum_{i=1}^n s_{ik}(X(t_m)).$$

Then a high  $\beta_k^X$  is expected to lead to high values of  $S_k^X$ .

This works well for effects

that do not depend on the endogenous characteristics  $Z$ .

Similar for effects in objective function for behavior,

that do not depend on network (e.g., trend).

To get some feeling for distinguishing effects network  $\Rightarrow$  behavior from effects behavior  $\Rightarrow$  network, focus on the *dissimilarity effect* :

sum of absolute covariate differences between  $i$  and his friends

$$s_{i6}(x, z) = \sum_j x_{ij} |z_{ih} - z_{jh}|.$$

A negative coefficient for this effect means that the actors prefer friends with similar  $Z_h$  values (*network autocorrelation*).

Actors can attempt to attain this by changing their own  $Z_h$  value to the average value of their friends (*network influence, contagion*),

or by becoming friends with those with similar  $Z_h$  values (*selection on similarity*).

The two different effects,  
network  $\Rightarrow$  behavior and behavior  $\Rightarrow$  network,  
both lead to  
a contemporaneous network autocorrelation of behavior;  
but they can be (in principle)  
distinguished empirically by the time order:

effects network  $\Rightarrow$  behavior :

$\sim$  association between ties at  $t_{m-1}$  and behavior at  $t_m$  ;

effects behavior  $\Rightarrow$  network:

$\sim$  association between behavior at  $t_{m-1}$  and ties at  $t_m$  .

For component

$$\beta_k^X s_{ik}(x, z),$$

in objective function for network, a relevant statistic is

$$S_k^X = \sum_{m=2}^M \sum_{i=1}^n s_{ik}(X(t_m), z(t_{m-1})) .$$

For component

$$\beta_k^Z s_{ik}(x, z),$$

in objective function for behavior, a relevant statistic is

$$S_k^Z = \sum_{m=2}^M \sum_{i=1}^n s_{ik}(x(t_{m-1}), Z(t_m)) .$$

Moment estimation will be based on the vector  $S$  of these statistics.

Denote by  $s$  the observed value for  $S$ .

The moment estimate  $\hat{\theta}$  is defined as the parameter value for which the expected value of the statistic is equal to the observed value:

$$E_{\hat{\theta}} S = s .$$

## Robbins-Monro algorithm

The moment equation cannot be solved by analytical or the usual numerical procedures, because

$$E_{\theta} S$$

cannot be calculated explicitly.

However, the solution can be approximated by versions of the Robbins-Monro (1951) method for stochastic approximation.

*Basic iteration step:*

$$\hat{\theta}_{N+1} = \hat{\theta}_N - a_N D^{-1}(s_N - s) , \quad (1)$$

where  $D$  is a suitable matrix,

and  $s_N$  is a simulation of  $S$  with parameter  $\hat{\theta}_N$ ,

and  $a_N$  is a sequence  $a_N \rightarrow 0$  .

## Covariance matrix

The method of moments yields the covariance matrix

$$\text{cov}(\hat{\theta}) \approx D_{\theta}^{-1} \Sigma_{\theta} D_{\theta}'^{-1}$$

where

$$\begin{aligned} \Sigma_{\theta} &= \text{cov}\{S | X(t_1) = x(t_1)\} \\ D_{\theta} &= \frac{\partial}{\partial \theta} \text{E}\{S | X(t_1) = x(t_1)\}. \end{aligned}$$

(This is a bit simplified w.r.t. conditioning on the starting observations for all unobserved periods  $[t_m, t_{m+1})$  . )

After the presumed convergence of the algorithm for approximately solving the moment equation, extra simulations are carried out

- (a) to check that indeed  $E_{\hat{\theta}} S \approx s$ ,
- (b) to estimate  $\Sigma_{\theta}$ ,
- (c) and to estimate  $D_{\theta}$   
using difference quotients and common random numbers.

Other estimation procedure: *Maximum Likelihood*.

This is still in development (work with Johan Koskinen),  
based on  
using the sequence of all unobserved changes as 'missing data'  
that are stochastically reconstructed in a MCMC procedure.

Advantages:

- ⇒ greater statistical efficiency,  
welcome for these closely related effects;
- ⇒ better possibilities for eventual combination  
with different model elements;
- ⇒ model comparison.

## Example

Study of smoking initiation and friendship  
in a Scottish secondary school

(following up on earlier work by P. West, M. Pearson & others,  
see Pearson & Michell, *Drugs: educ., prev. and policy*, 2000.)

One school year group from a Scottish secondary school  
starting at age 12-13 years,  
was monitored over 3 years, 129 pupils present at all 3 observations,  
with sociometric & behavior questionnaires  
at three moments, at appr. 1 year intervals.

What does this data set tell us about the mutual effects  
between friendship and smoking?

First, results for a model  
for dynamics in networks and in smoking behavior  
under the assumption that both are unrelated.

Smoking measured in three categories:  
1 = no, 2 = occasionally, 3 = regularly.

There is more network change than behavior change;  
⇒ more power for discovering effects acting on network dynamics.

In this group, girls smoke more than boys.

Parameter estimates for network dynamics:  
assumed independent of smoking.

	Effect	Estimate	Standard error
<i>Rate function</i>			
$\lambda_0^X$	Rate parameter $t_1-t_2$	11.63	1.42
$\lambda_1^X$	Rate parameter $t_2-t_3$	9.27	0.96
<i>Objective function</i>			
$\beta_1^X$	Density / out-degree	-2.16	0.06
$\beta_2^X$	Reciprocity	2.03	0.08
$\beta_3^X$	Number of distances 2	-0.68	0.014
$\beta_4^X$	Transitive triplets	0.22	0.011
$\beta_5^X$	Gender ( $F$ ) popularity	-0.23	0.08
$\beta_6^X$	Gender ( $F$ ) activity	0.17	0.08
$\beta_7^X$	Gender similarity	0.82	0.11

Parameter estimates for smoking dynamics:  
assumed independent of friendship.

	Effect	Estimate	Standard error
<i>Rate function</i>			
$\lambda_0^Z$	Rate parameter $t_1-t_2$	0.68	0.22
$\lambda_1^Z$	Rate parameter $t_2-t_3$	0.72	0.24
<i>Objective function</i>			
$\beta_1^Z$	Tendency	-0.26	0.24
$\beta_2^Z$	Gender ( $F$ )	2.05	1.16
$\beta_3^Z$	Parents' smoking	0.70	1.20

Now a similar analysis,  
but with a model in which there is a mutual effect  
between smoking and friendship formation.

Parameter estimates for network dynamics:  
dependent of smoking.

	Effect	Estimate	Standard error
<i>Rate function</i>			
$\lambda_0^X$	Rate parameter $t_1-t_2$	11.74	1.25
$\lambda_1^X$	Rate parameter $t_2-t_3$	9.53	1.07
<i>Objective function</i>			
$\beta_1^X$	Density / out-degree	-2.17	0.05
$\beta_2^X$	Reciprocity	2.06	0.08
$\beta_3^X$	Number of distances 2	-0.80	0.013
$\beta_4^X$	Transitive triplets	0.17	0.009
$\beta_5^X$	Gender ( $F$ ) popularity	-0.20	0.08
$\beta_6^X$	Gender ( $F$ ) activity	0.18	0.08
$\beta_7^X$	Gender similarity	0.80	0.09
$\beta_8^X$	Smoking similarity	0.17	0.05

Parameter estimates for smoking dynamics:  
dependent of friendship.

	Effect	Estimate	Standard error
<i>Rate function</i>			
$\lambda_0^Z$	Rate parameter $t_1-t_2$	0.84	0.30
$\lambda_1^Z$	Rate parameter $t_2-t_3$	0.84	0.27
<i>Objective function</i>			
$\beta_1^Z$	Tendency	0.13	0.35
$\beta_2^Z$	Gender ( $F$ )	1.26	1.14
$\beta_3^Z$	Parents' smoking	1.00	1.28
$\beta_4^Z$	Friendship	0.33	0.37

### Conclusions :

Evidence for effect of smoking on friendship development;  
no evidence for effect of friendship on smoking initiation;  
note that

taking the mutual effect of smoking and friendship into account  
(even though the effect friendship  $\Rightarrow$  smoking was not significant)  
reduces strongly the estimated effect of gender:

in the first analysis

there seemed some evidence for a gender effect on smoking,  
but this can equally be explained as an effect of friendship  
(friendship is rather gender-homogeneous).

## Discussion

This talk tried to explicate statistical modeling approaches to the analysis of empirical data on the simultaneous evolution of relational networks and behavior.

Ongoing work; this is just the beginning.

Note:

- ⇒ continuous-time model more natural
- ⇒ actor-oriented model
- ⇒ stochastic micro-simulation model used as statistical model (gives access to the statistical paradigm for model assessment and improvement).

Some issues:

- \* different specifications;
- \* the *fit* of these models  
and the *robustness* of the conclusions;
- \* to what extent is this *causal* modeling?  
very modest claims: “as if” approach,  
we are describing data & testing substantive theories  
using models that express causality;  
effects of explanatory variables are ‘maximally’ controlled  
for structural network effects;
- \* richer modeling of network effects is important:  
e.g., of network positions of individual actors;  
work in progress (Christian Steglich, Mike Pearson);

- \* application of this type of model to larger groups requires separate modeling of settings / opportunity structure : parameters will be different for ties within and ties between settings;
- \* what is more important for disease dynamics: small-scale within-setting processes, or between-setting transmission?
- \* what is effects of infected status on behavior – e.g., within-setting relational behavior & between-settings movement;
- \* parameters governing creation of new ties may be different from those governing canceling existing ties (objective function not sufficient to express this).

Some mathematical questions:

- \* asymptotic distributions of this process ( $t \rightarrow \infty$ )  
(could be degenerate!)
- \* asymptotic distributions ( $n \rightarrow \infty$ ) of parameter estimates, log-likelihood-ratios
- \* efficiency, robustness of MoM and ML estimates  
(seems difficult)
- \* analytic properties and explicit estimators for simple cases  
to give intuition in what might happen in general cases
- \* scaling of parameters for increase in  $n$
- \* missing data in initial observation / adaptive data collection designs.