

Estimating linear and generalized linear models using Respondent Driven Sampling (RDS)

Märt Möls¹,

Krista Fischer², Anneli Uusküla¹

¹Tartu University, Estonia; ²MRC Biostatistics Unit

Introduction

- Estonian Intravenous Drug Users Survey (Data collection 2005)

In co-operation with Imperial College London (Lucy Platt, Natalia Bobrova and Tim Rhodes)

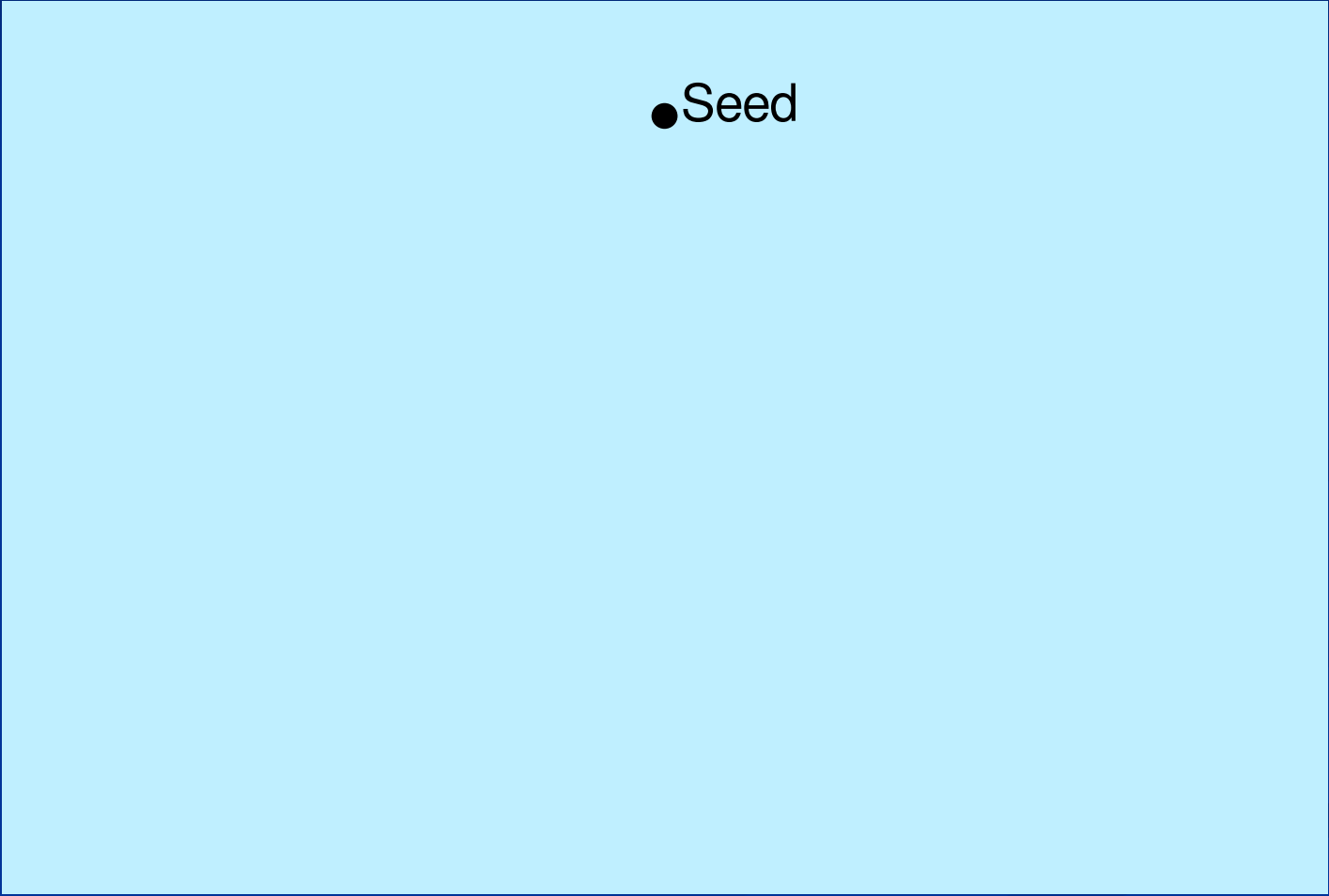
- Estonian Commercial Sex Workers Survey (Data collection 2006)

- Estonian Intravenous Drug Users (IDU) Survey 2 (Data collection 2007)

In co-operation with Don DesJarlais (Beth Israel Medical Center), partly funded by CRDF and NIH (NIDA).

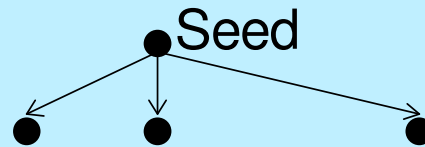
At least 125 studies worldwide using RDS methodology – drug users, men having sex with men, sex workers, high risk heterosexual men, homeless people, jazz musicians,

Respondent-Driven Sampling (RDS)

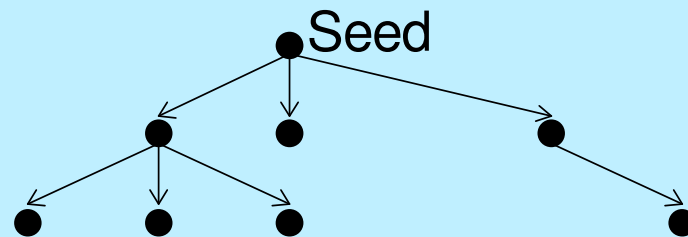


●Seed

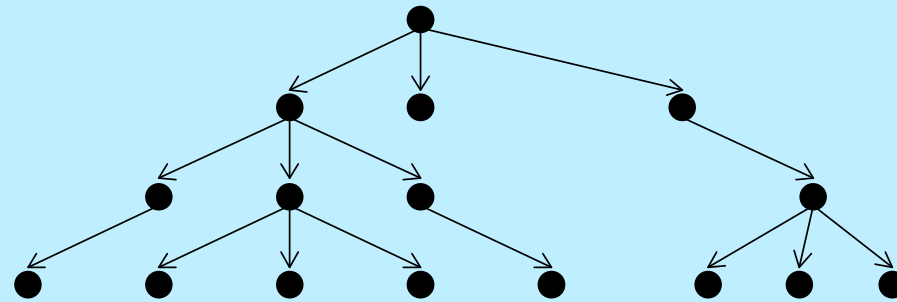
Respondent-Driven Sampling (RDS)



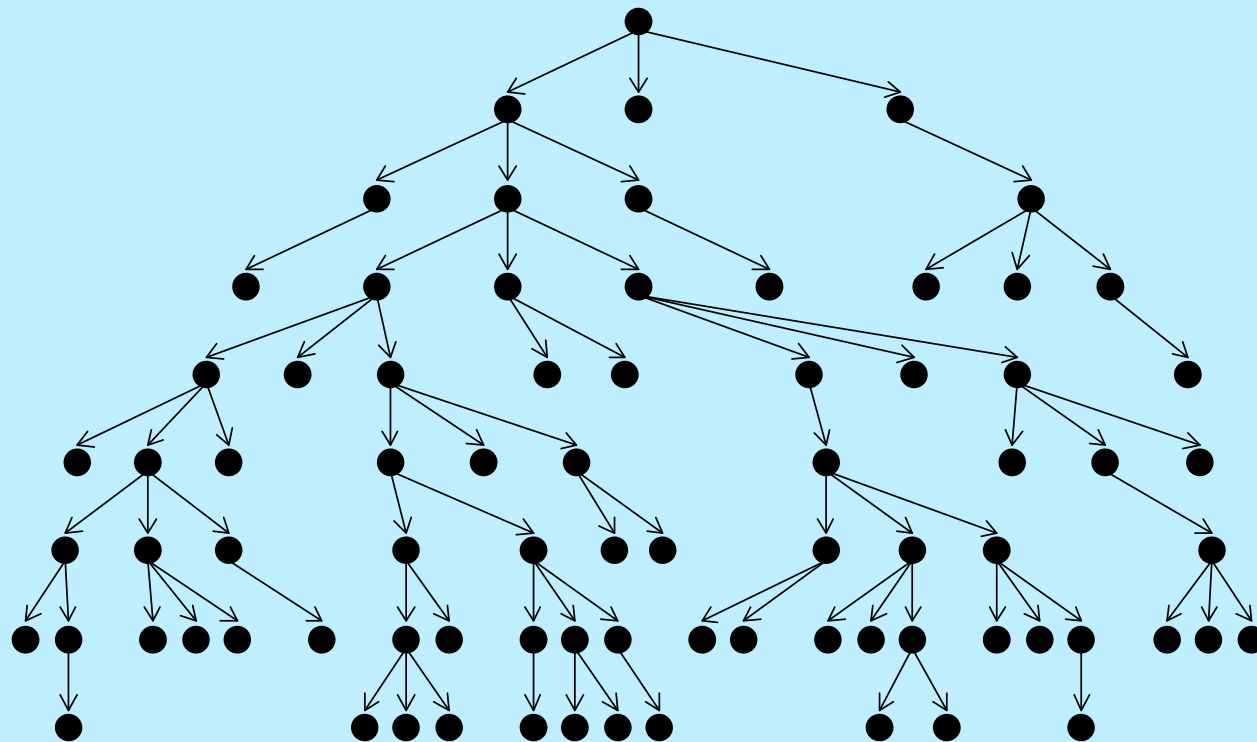
Respondent-Driven Sampling (RDS)



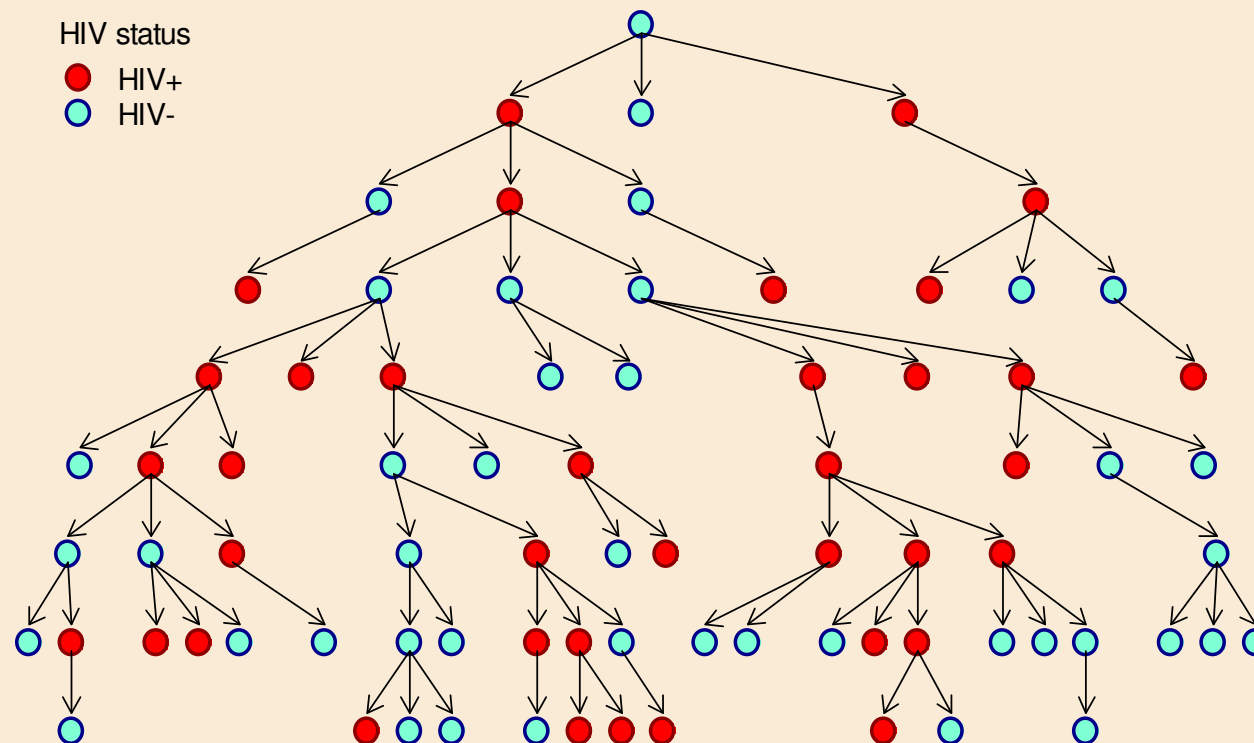
Respondent-Driven Sampling (RDS)



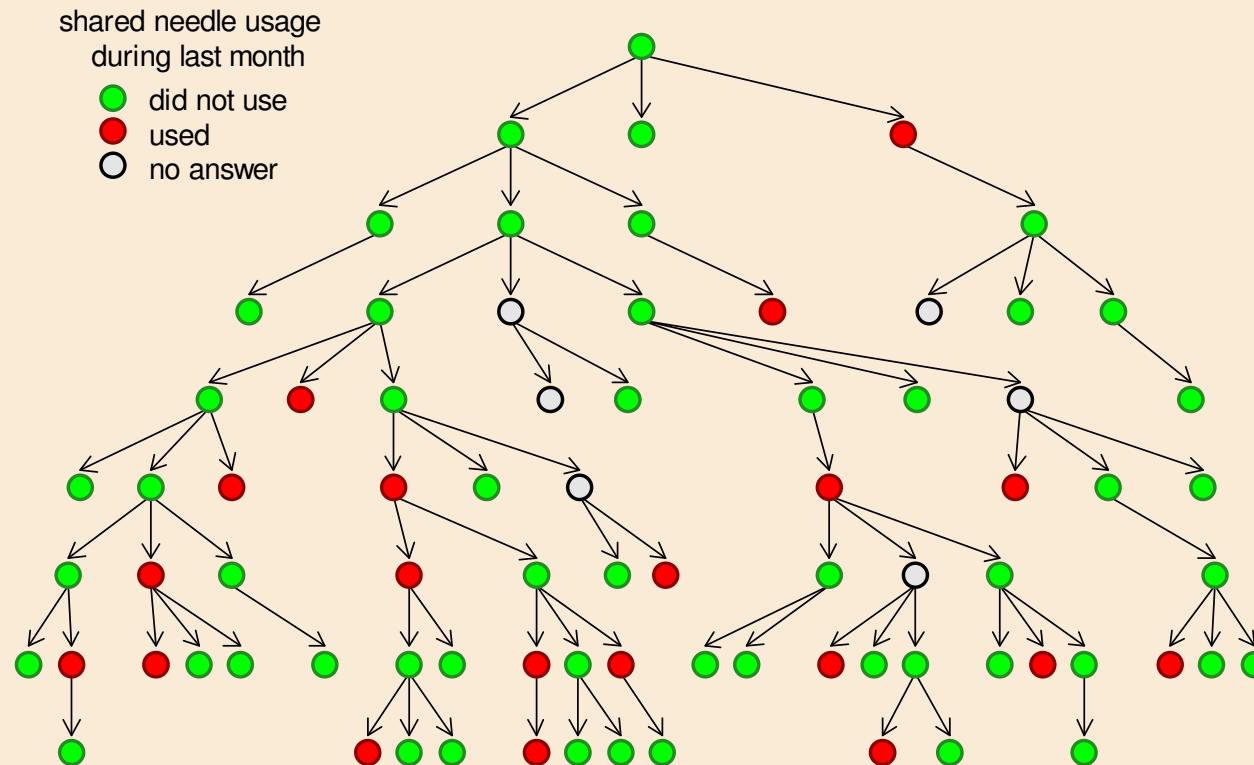
Respondent-Driven Sampling (RDS)



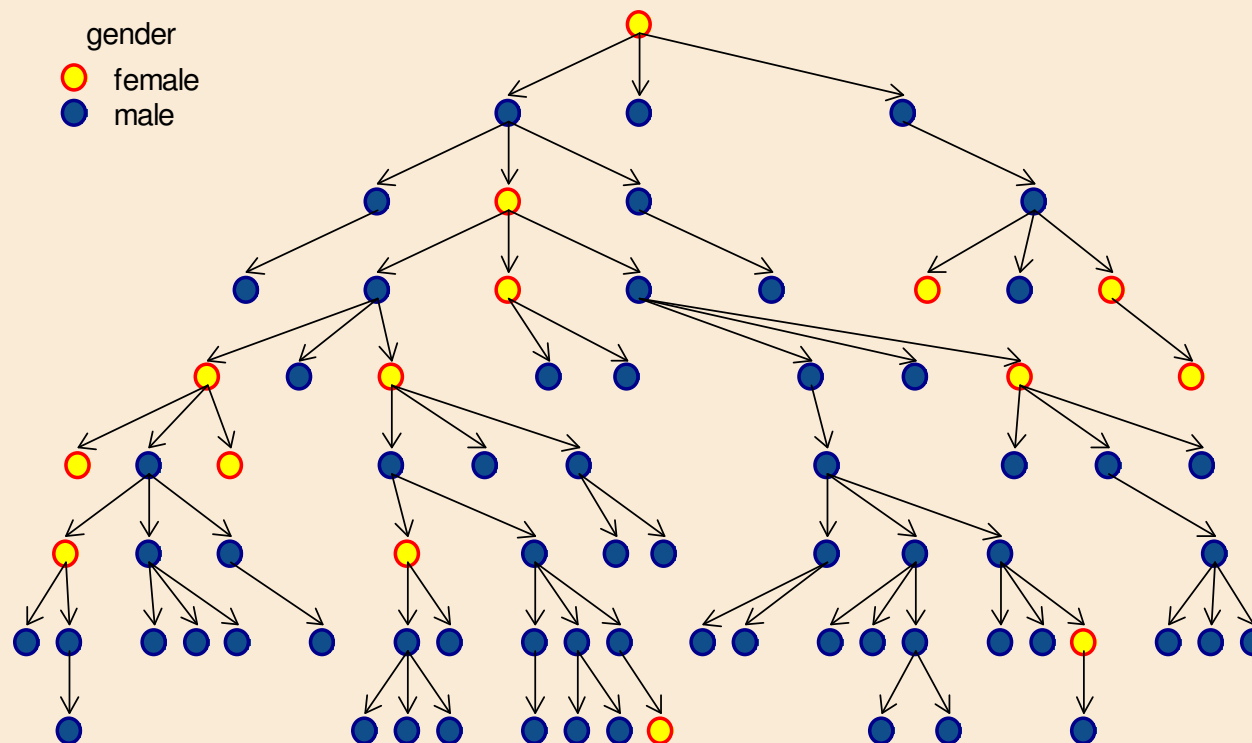
Introduction



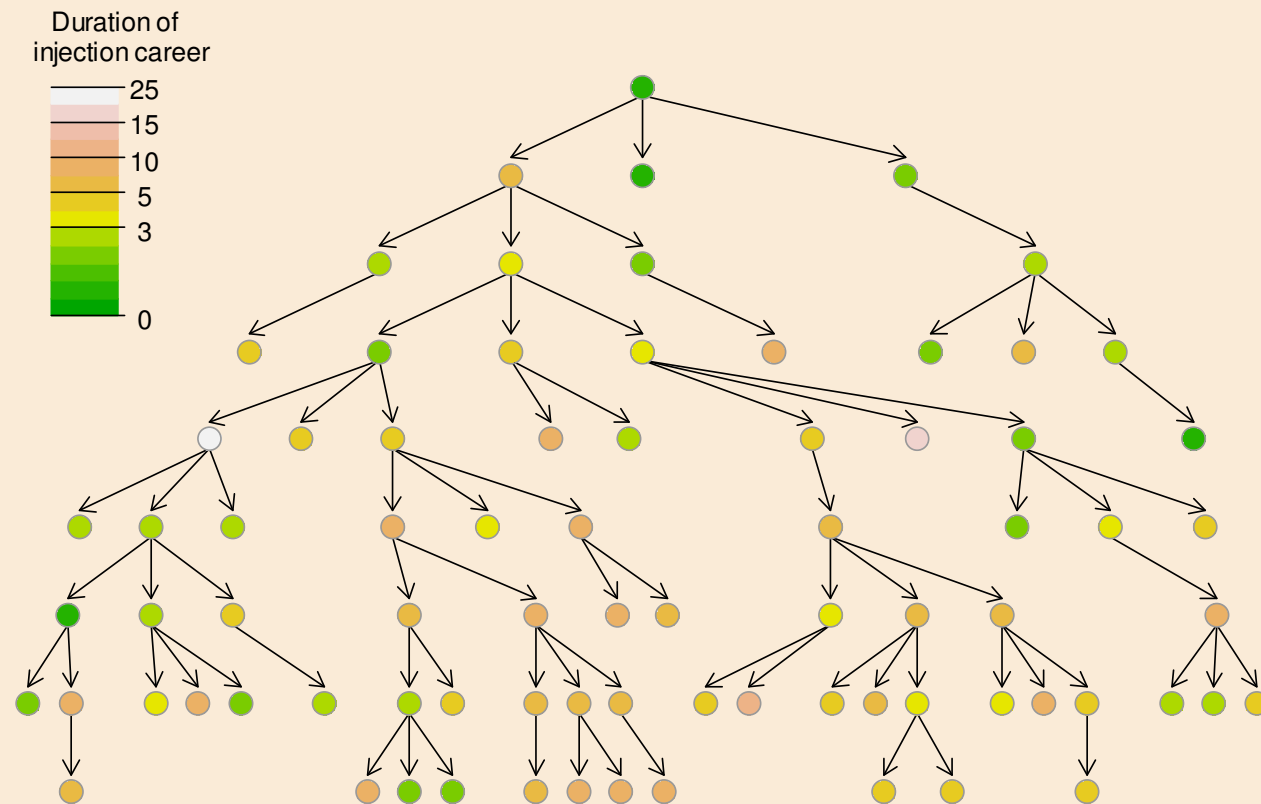
Introduction



Introduction



Introduction



Basic description of the Studies

	Intervenous Drug Users	Commercial Sex Workers	IDU 2
Sample size	450 Tallinn - 350; Kohtla-Järve-100	227	700
Number of seeds	9	43	11
Median network size	50	6	50
Number of “waves”	9	9	17
HIV+	62%	7,5%	63%

Wishlist from epidemiologists

- Prevalence of HIV+ (probability to be HIV+), mean duration of injection career, ... (with Confidence Intervals)
- Test whether sharing needles/type of drug/ ... is a risk factor for HIV+; how much the odds to get HIV increase if someone adopts riskier behaviour (tests, Odds Ratios – OR – with CI)
- Models for predicting the probability of HIV;
-

How to analyse a RDS study?

- Naive Analysis - Treat the data as from usual random sample. Convenient, often used, but... ;
- Heckathorn's method and software – use Markov Chains to derive asymptotically unbiased (under plausible assumptions) estimates of proportions;
- Use Linear Mixed Models/Generalized Linear Mixed Models with suitable covariance structure (our proposal).

Naive Analysis - problems

1. Undersampling of respondents with few friends (small network size);
2. Bias due to non-random selection of seeds;
3. Friends are “more similar” to each other than two randomly chosen persons from the target population

Association between recruiter and recruited (from IDU study):

HIV status	p-value=0,0009
age	p-value=0,03
type of drug used	p-value=0,00002

Heckathorn's method

- "Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations." By Douglas D. Heckathorn. Social Problems, 1997.
- "Respondent-Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations." By Douglas D. Heckathorn. Social Problems, 2002.
- www.respondentdrivensampling.org

Heckathorn's method

Nearly-unbiased estimation of a distribution with few possible categories (for example: infected/not-infected), under understandable set of assumptions.

Bootstrapping methods to calculate approximate standard errors and confidence intervals for prevalence estimates

Heckathorn's method: limitations

Limited possibilities for more complex models;

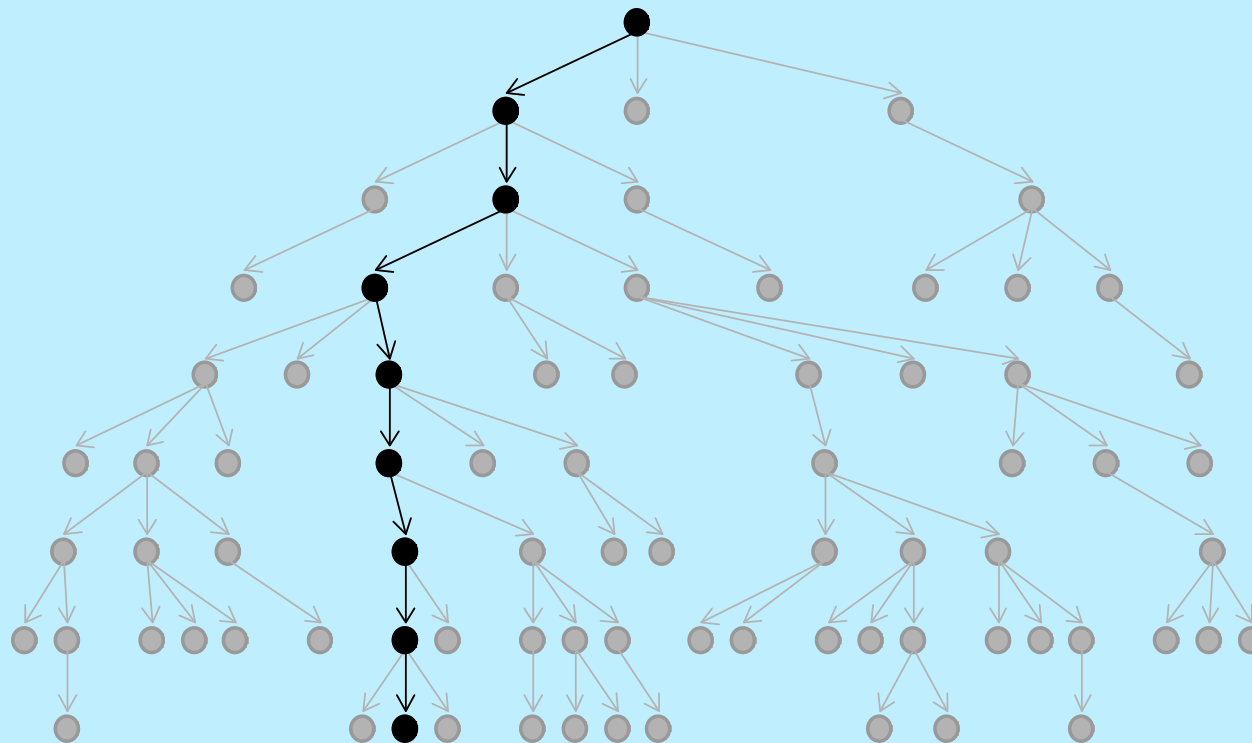
Limited possibilities to handle continuous variables;

Now we will continue with our proposals (which are actually just some old classical methods)

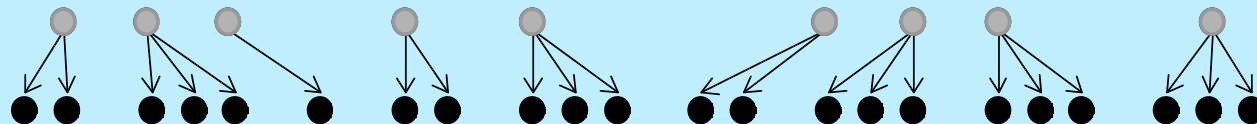
There is nothing new in analysing
correlated observations.

For example, one can use

Time Series Analysis



Repeated Measures / Multilevel Analysis



Analysing Correlated Data

- Estimate the correlation/covariance matrix V by using ML or REML;
- Estimate parameter vector via Estimated BLUE (GLS):

$$\hat{\beta} = (X^T \hat{V}^{-1} X)^{-} X^T \hat{V}^{-1} Y$$

- Test hypothesis etc:

$$L\hat{\beta} \pm t_{\hat{v}, \alpha/2} \sqrt{L\hat{C}L^T}$$
$$F = \frac{\hat{\beta}^T L^T (L\hat{C}L^T)^{-1} L\hat{\beta}}{\text{rank}(L)}$$
$$\hat{C} = (X^T \hat{V}^{-1} X)^{-}$$

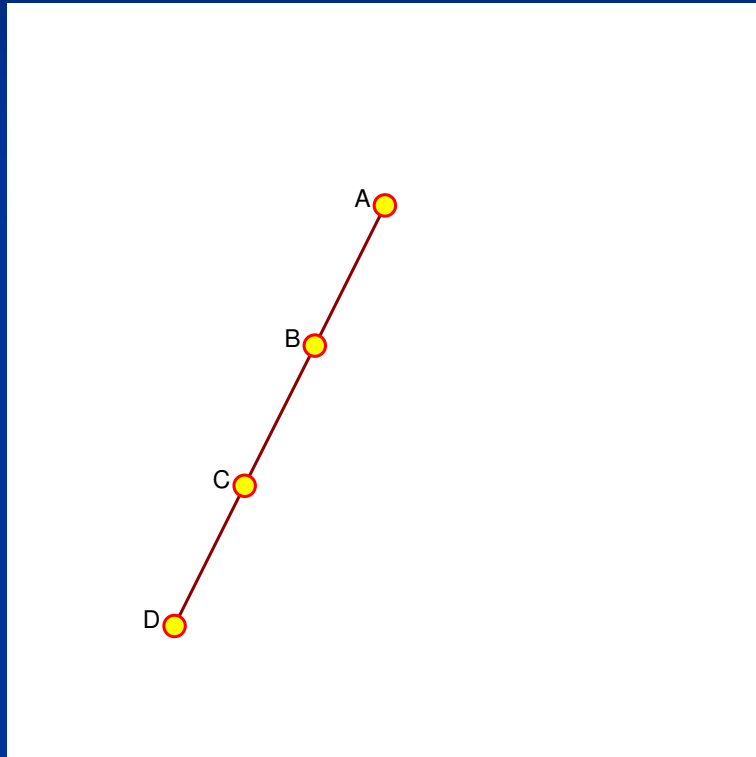
See:

SAS Online Documentation, PROC MIXED, PROC GLIMMIX;

Searle et al – Variance Components; Generalized, Linear, and Mixed Models

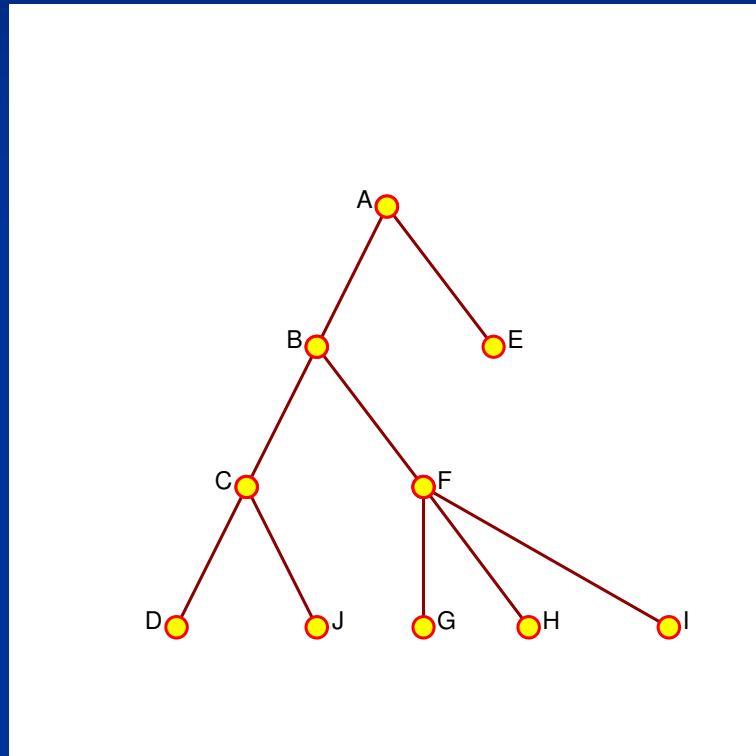
Other textbooks about Mixed models/linear models

A possible correlation structure for a RDS Design



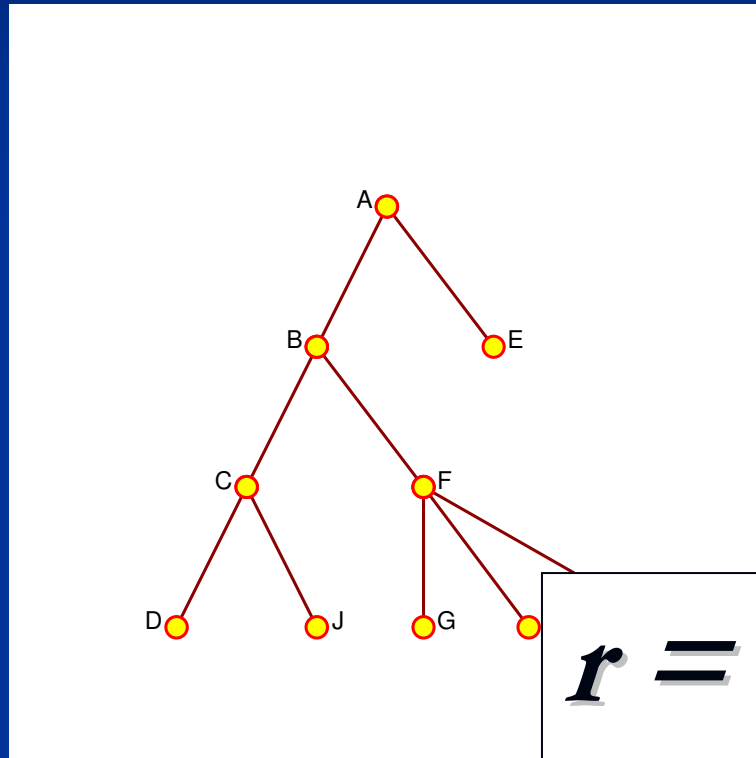
	A	B	C	D
A	1	r	r^2	r^3
B	r	1	r	r^2
C	r^2	r	1	r
D	r^3	r^2	r	1

Possible correlation structure for a RDS Design



	A	B	C	D	E	F	G	H	I
A	1	r	r^2	r^3	r	r^2	r^3	r^3	r^3
B	r	1	r	r^2	s	r	r^2	r^2	r^2
C	r^2	r	1	r	sr	s	sr	sr	sr
D	r^3	r^2	r	1	sr^2	sr	sr^2	sr^2	sr^2
E	r	s	sr	sr^2	1	sr	sr^2	sr^2	sr^2
F	r^2	r	s	sr	sr	1	r	r	r
G	r^3	r^2	sr	sr^2	sr^2	r	1	s	s
H	r^3	r^2	sr	sr^2	sr^2	r	s	1	s
I	r^3	r^2	sr	sr^2	sr^2	r	s	s	1

Possible correlation structure for a RDS Design



	A	B	C	D	E	F	G	H	I
A	1	r	r^2	r^3	r	r^2	r^3	r^3	r^3
B	r	1	r	r^2	s	r	r^2	r^2	r^2
C	r^2	r	1	r	sr	s	sr	sr	sr
D	r^3	r^2	r	1	sr^2	sr	sr^2	sr^2	sr^2
E	r	s	sr	sr^2	1	sr	sr^2	sr^2	sr^2
F	r^2	r	s	sr	sr	1	r	r	r
G	r^3	r^2	r	sr^2	sr^2	r	1	s	s
H	r^3	r^2	r	sr^2	sr^2	r	s	1	s
I	r^3	r^2	r	sr^2	sr^2	r	s	s	1

$$r = 0,01$$

$$s = 0,12$$

Logistic regression / Generalized Linear Mixed Model

SAS PROC GLIMMIX documentation:

$$g(EY)=X\beta; \quad EY=g^{-1}(X\beta)$$

$$\text{Var}(Y)=A^{1/2}VA^{1/2}$$

The matrix A is a diagonal matrix and contains the variance functions of the model. The variance function expresses the variance of a response as a function of the mean.

Undersampling

1. Undersampling can be corrected using weighted averages or related techniques (Horwitz-Thompson estimator). Corrections are regularly applied if stratified random sampling has been used, for example.

$$P(\text{Sampled} \mid \text{Network Size} = i) \propto i$$

Correction for undersampling

1. Include network size to the model of interest, eg.:

$$\text{logit}(P(\text{HIV+} | NS)) = c + f(NS)$$

2. Integrate NS out from the final result based on the estimated proportions of network sizes, eg:

$$P_{\text{est}}(\text{HIV+}) = \sum_i P_{\text{est}}(\text{HIV+} | NS=i) P_{\text{est}}(NS=i)$$

3. One can use for example delta method to calculate standard error for the estimate (is some better method available?)

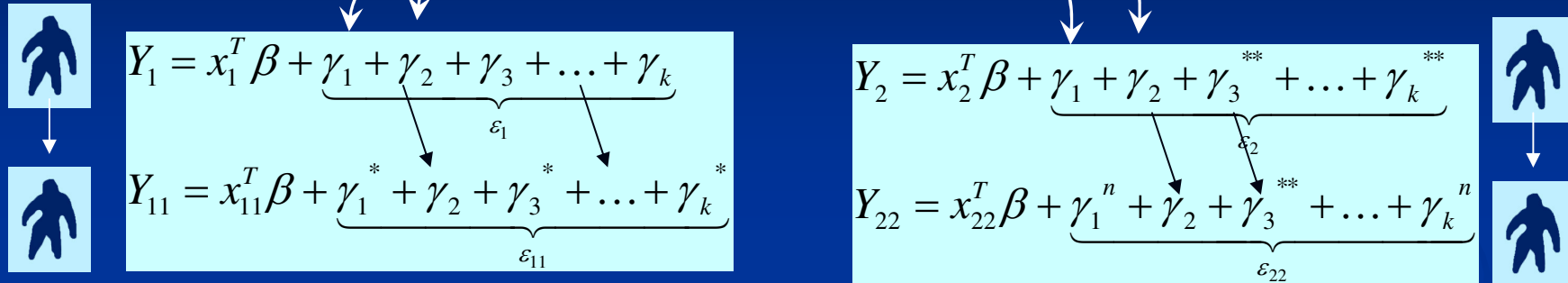
Correction for undersampling

A note:

Most applications in epidemiology (probably) do not require the complicated procedure to calculate the standard error. If one is interested in estimating the effect of a risk factor and there is no interaction effect between the risk factor (RF) and network size, one can just fit a model

$\text{logit}(P(\text{HIV+} | NS)) = c + f(NS) + \text{RF} + \text{confounders}$
to the data (+correct covariance structure) and the delivered inference for RF is still valid.

Seed selection bias I



	Y_1	Y_{11}	Y_2	Y_{22}
Y_1	σ^2	r	b	br
Y_{11}	r	σ^2	br	br^2
Y_2	b	br	σ^2	r
Y_{22}	br	br^2	r	σ^2

Seeds are correlated, friends are correlated, but correlation goes to zero as the distance within recruitment chain increases.

Seed selection II



$$Y_1 = x_1^T \beta + \lambda_1 + \varepsilon_1$$

$$Y_{11} = x_{11}^T \beta + \lambda_1 + \varepsilon_{11}$$

$$Y_2 = x_2^T \beta + \lambda_2 + \varepsilon_2$$

$$Y_{22} = x_{22}^T \beta + \lambda_2 + \varepsilon_{22}$$



	Y_1	Y_{11}	Y_2	Y_{22}
Y_1	$\sigma^2 + f$	$r + f$	0	0
Y_{11}	$r + f$	$\sigma^2 + f$	0	0
Y_2	0	0	$\sigma^2 + f$	$r + f$
Y_{22}	0	0	$r + f$	$\sigma^2 + f$

Different non-communicating populations, one recruitment chain will never jump to a different population (respondents within one recruitment chain will always remain correlated), but seeds can be considered as a random sample from target population

Seed selection bias III – Problem!



$$Y_1 = x_1^T \beta + \underbrace{\gamma_1 + \gamma_2 + \gamma_3 + \dots + \gamma_k}_{\varepsilon_1}$$

$$Y_{11} = x_{11}^T \beta + \underbrace{\gamma_1 + \gamma_2 + \gamma_3^* + \dots + \gamma_k^*}_{\varepsilon_{11}}$$

$$Y_2 = x_2^T \beta + \underbrace{\gamma_1 + \gamma_2 + \gamma_3^{**} + \dots + \gamma_k^{**}}_{\varepsilon_2}$$

$$Y_{22} = x_{22}^T \beta + \underbrace{\gamma_1 + \gamma_2 + \gamma_3^{**} + \dots + \gamma_k^n}_{\varepsilon_{22}}$$



	Y_1	Y_{11}	Y_2	Y_{22}
Y_1	$\sigma^2 + b$	$r + b$	b	b
Y_{11}	$r + b$	$\sigma^2 + b$	b	b
Y_2	b	b	$\sigma^2 + b$	$r + b$
Y_{22}	b	b	$r + b$	$\sigma^2 + b$

Different non-communicating populations, one recruitment chain will never jump to a different population, seeds are also correlated – not truly a random sample from target population

Does it work?

- Comparison with other methods
- Simulations
- Does the proposed model fit to the real data?

Some results I

CSW Study:

	Naive	Heckathorn	GLMM
HIV+ (%)	7,5%	4,7%	4,7%
se	0,01747	-	0,0169
95%-CI	4%..12%	3%..9%	1%..8%

Some results II

IDU Study:

	Naive	Heckathorn	GLMM
Tallinn, HIV+ (%)	54%	47%	47%
Kohtla-Järve, HIV+ (%)	89%	91%	90%
Average Age	24,2	NA	23,7

Simulation results

(True population prevalence about 0,45)

Method	MSE	95%-CI	
		coverage	bias
Naive	0,313	76,0%	0,046
Heckatorn	0,141	96,5%	0,003
GLMM	0,143	96,0%	-0,006

Does the model fit?

Example: linear model (for age) from the IDU study

Correlation parameter estimates:

$$r = 0,084$$

$$s = 0,152$$

$$AIC = 2700 \quad (\text{RDS correlation structure})$$

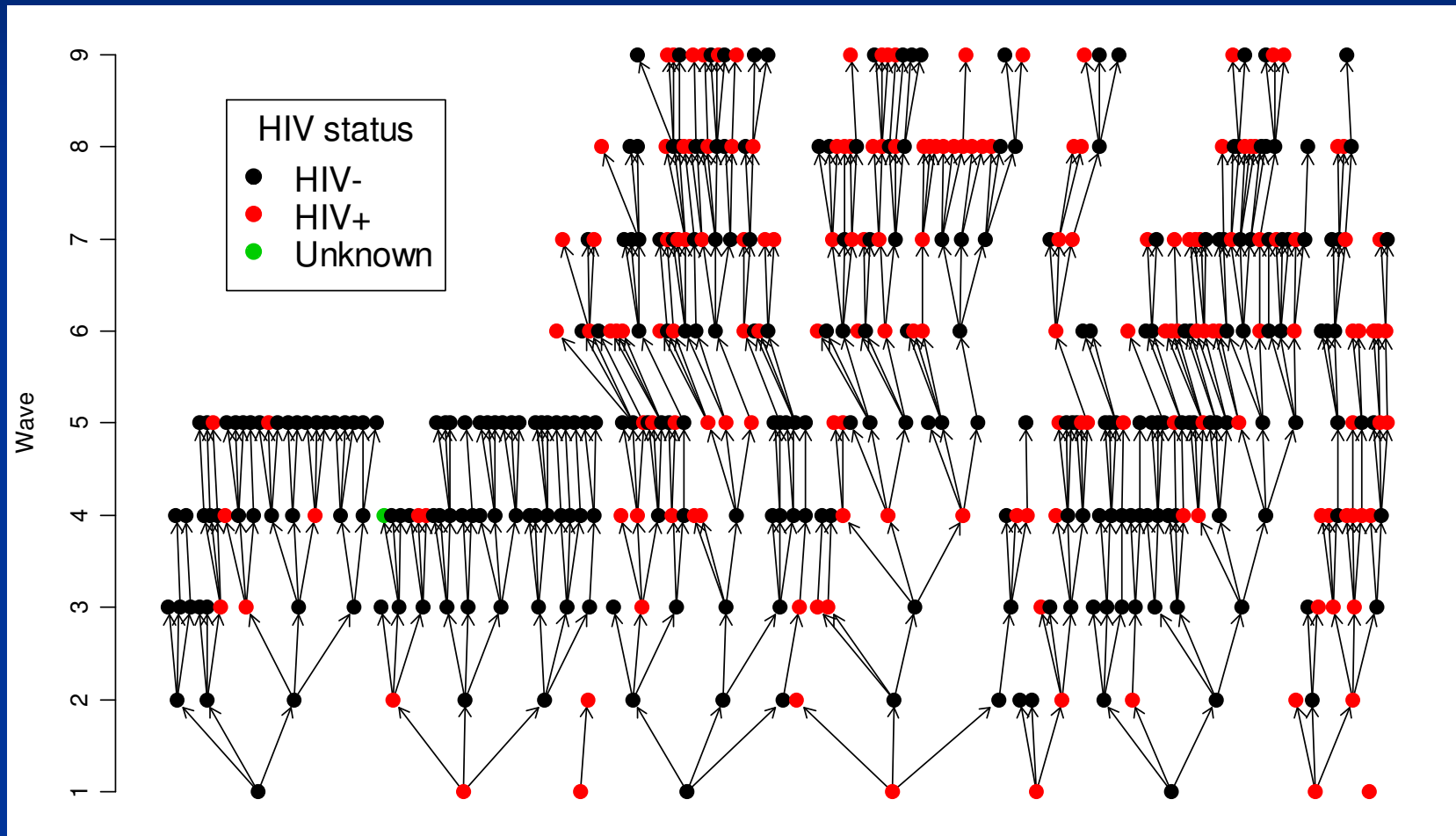
$$AIC = 2704 \quad (\text{Independence})$$

Conclusions

- Regression and logistic regression models can be fit and valid inference can be made (for RDS samples)
- One can correctly analyse an RDS study by using standard software (for example R)
- However: untestable(?) assumptions; using the software can be tricky (lack of documentation) etc.

Few additional slides, just in case...

Estonian IDU Study



HIV risk factors - CSW

	Naive Analysis OR (95% CI) (unadjusted)	GLMM OR (95% CI) (unadjusted)
Category of SW		
Brothel	4,0 (1,2...15,3)	3,6 (1,2...10,4)
Street	12,6 (3,0...56,8)	3,1 (0,7...14,1)
Other	1	1
Drug use: No	1	1
Yes	8,3 (2,3...27,8)	2,3 (0,6...8,5)
Years of CSW	0,84 (0,70...0,95)	0,76 (0,67..0,87)

HIV risk factors - IDU

	Naive Analysis OR (95% CI) (unadjusted)	GLMM OR (95% CI) (unadjusted)
Gender: male	1	1
female	1.2 (0.7...2.1)	1.3 (0.8...2.3)
Years of drug use:		
2 years or less	1	1
3..5 years	2.8 (1.4...5.5)	2.4 (1.2...4.9)
6..9 years	4.8 (2.4...9.3)	3.8 (1.9...7.4)
10 years or more	3.5 (1.7...7.0)	2.5 (1.2...5.2)

IDU – final model for HIV status

	log (OR)	OR	95%-CI	p-value
Network size (Ref: 100+)				
less than 100	-0.75	0.47	0.30...0.75	0.002
Duration of injection career (Ref: 0-2 years)				
3-5 years	0.94	2.55	1.16...5.62	0.020
6-9 years	1.61	5.00	2.19..11.38	0.001
10 years or more	1.31	3.69	1.42...9.61	0.008
Place of residence (Ref: Tallinn)				
Kohtla-Järve	1.79	5.99	2.53..14.19	0.005
Number of sexual partners during last year (Ref: 0)				
one	-1.37	0.25	0.09...0.69	0.007
more than one	-1.07	0.34	0.13...0.89	0.028
Age group (Ref: <20)				
20-24	-0.43	0.65	0.33...1.29	0.218
25-29	-0.90	0.41	0.18...0.89	0.026
30 or more	-1.26	0.28	0.11...0.75	0.012

Software

- Most major statistical software packages (SAS, STATA, S-plus, R,...) can correctly analyse data with correlated errors – one just needs to specify correct correlation structure.
- However, no package offers a ready-made correlation structure suitable for analysing RDS-designs. Some packages allow user to supply their own user-specified correlation structure (R, S-Plus)

Software example: R

R is a free statistical package (www.r-project.org)

To create a new correlation structure (corNew) one has to write five new functions:

1. corNew - constructor function creates a new object
2. Initialize.corNew - initializes the new object
3. **corMatrix.corNew – returns the correlation matrix
(based on current parameter values)**
4. Functions “coef.corNew” and “coef<-.corNew” to extract and change the correlation parameters.

Created correlation structure can then be used in other functions (gls, lme, glmmPQL,...)

Heckathorn's method

1. Estimate the matrix of transition probabilities:

		Recruited	
		HIV-	HIV+
Recruiter	HIV-	$p_{--} = 0,946$	$p_{-+} = 0,054$
	HIV+	$p_{+-} = 0,769$	$p_{++} = 0,231$

Heckathorn's method

2. Estimate the average network size for all groups

	HIV-	HIV+
Average network size (ANS)	3,40	5,12

Heckathorn's method

3. Assumption:

The total number of ties from HIV- group to HIV+ group is the same as the total number of ties from HIV+ group to HIV- group (If A knows B then B also knows A).

$$N_+ \text{ANS}_+ p_{+-} = N_- \text{ANS}_- p_{-+}$$

Heckathorn's method

From where:

$$N_+ / (N_+ + N_-) = \frac{ANS_- p_{-+}}{ANS_- p_{-+} + ANS_+ p_{+-}}$$

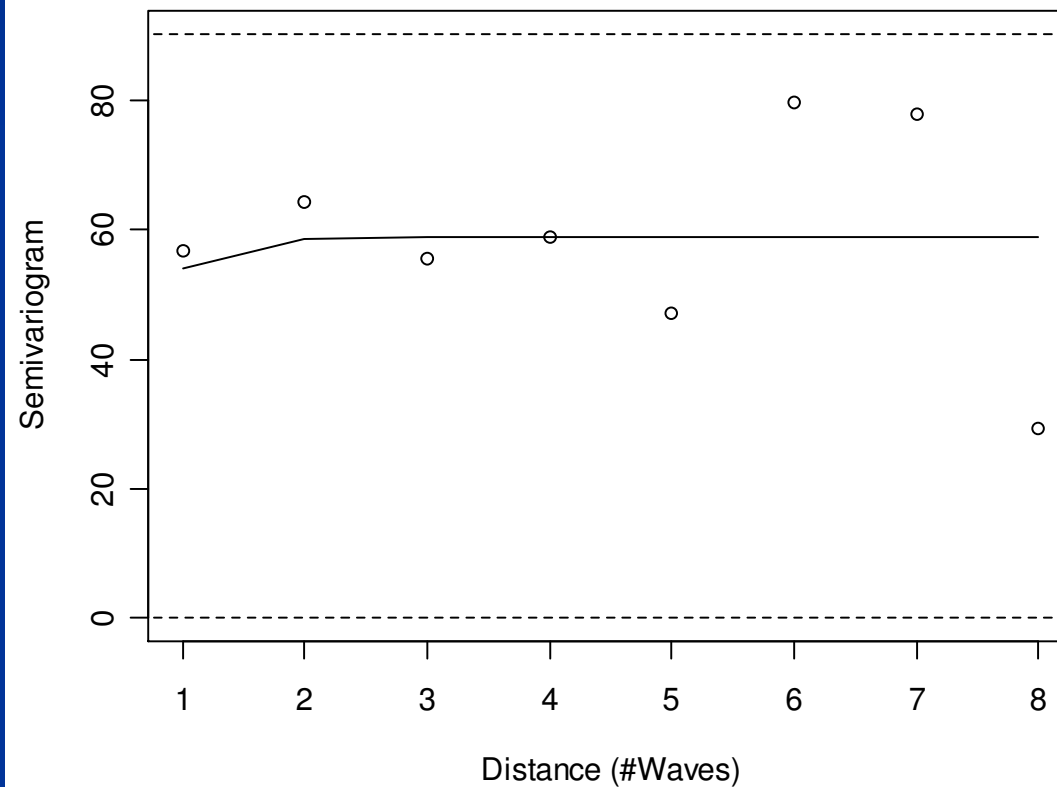
$$3,40 \cdot 0,054 / (5,12 \cdot 0,769 + 3,40 \cdot 0,054) = 0,045$$

Distribution of network sizes

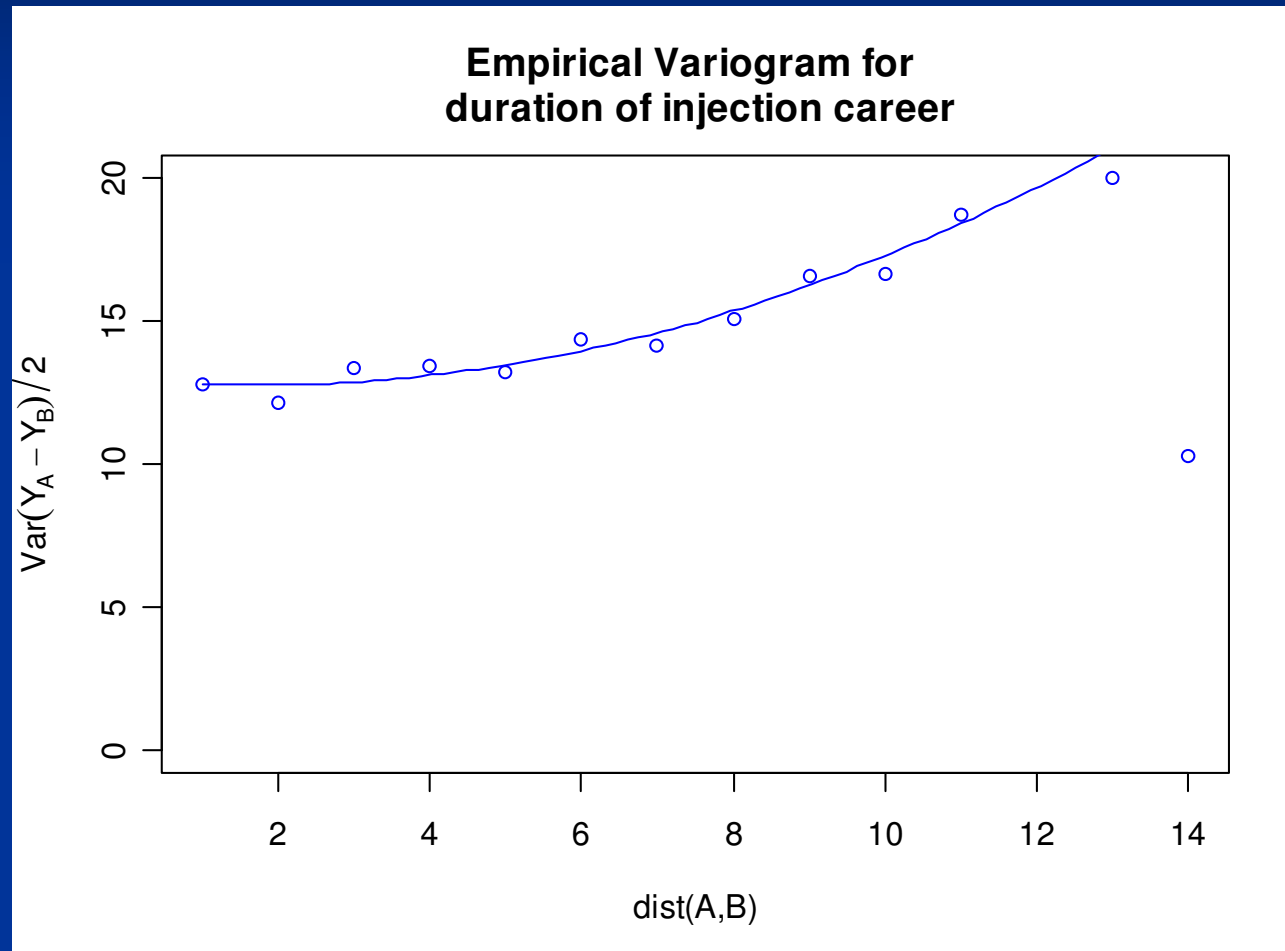
$$P(S = 1 \mid NS = i) = \frac{P(NS = i \mid S = 1)P(S = 1)}{P(NS = i)}$$

$$P(NS = i \mid S = 1) \propto P(NS = i) i$$

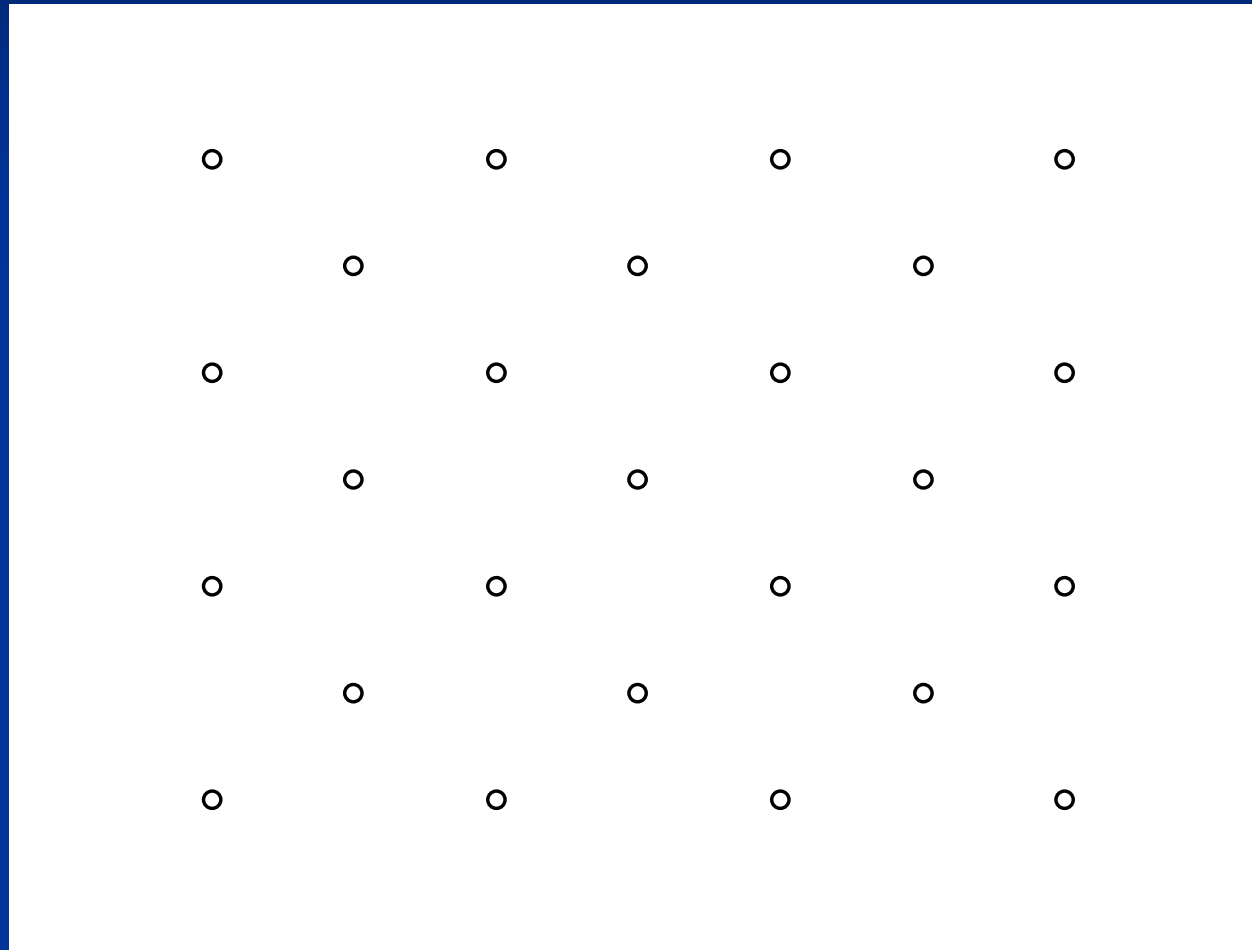
Problems with the CSW-study



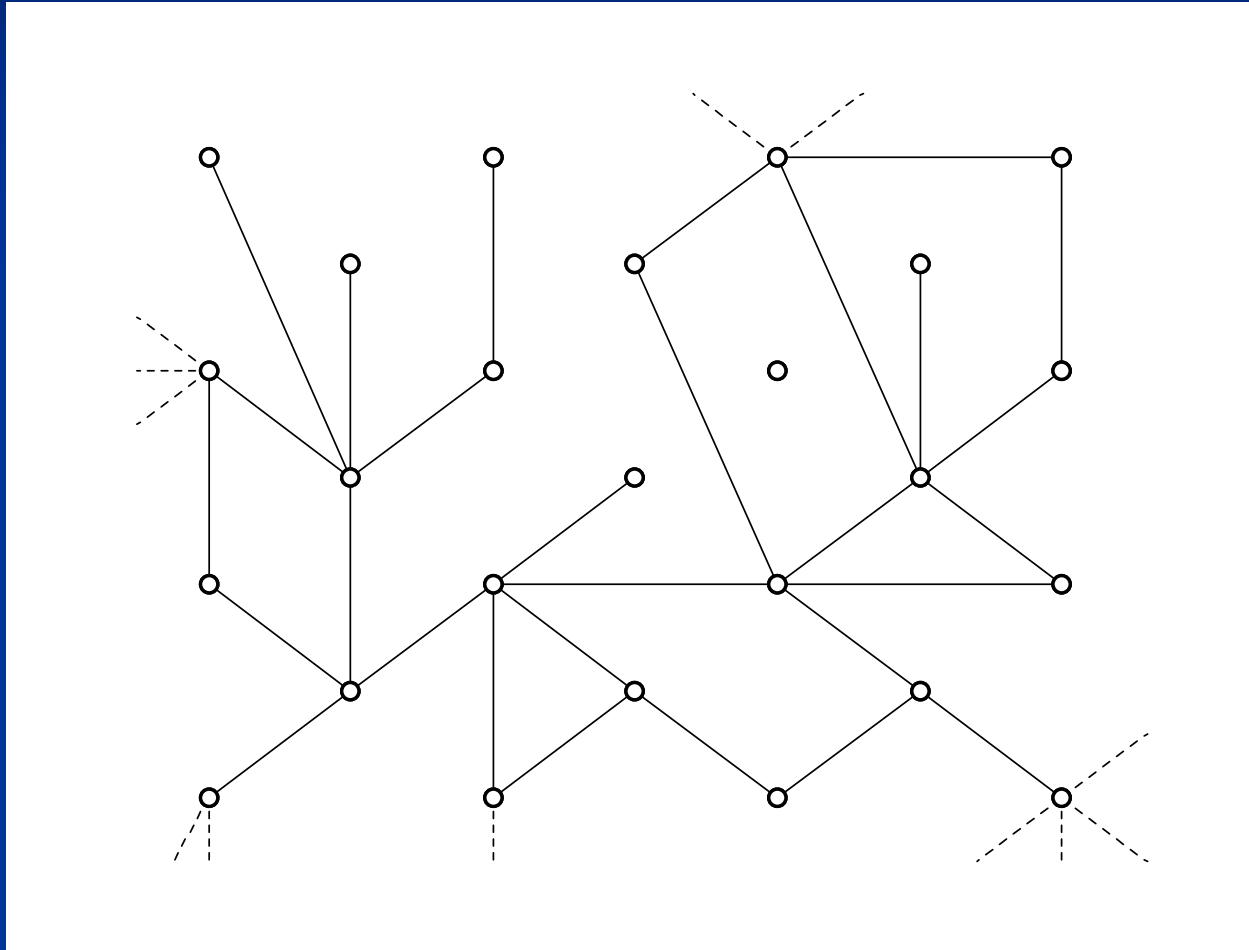
Problems



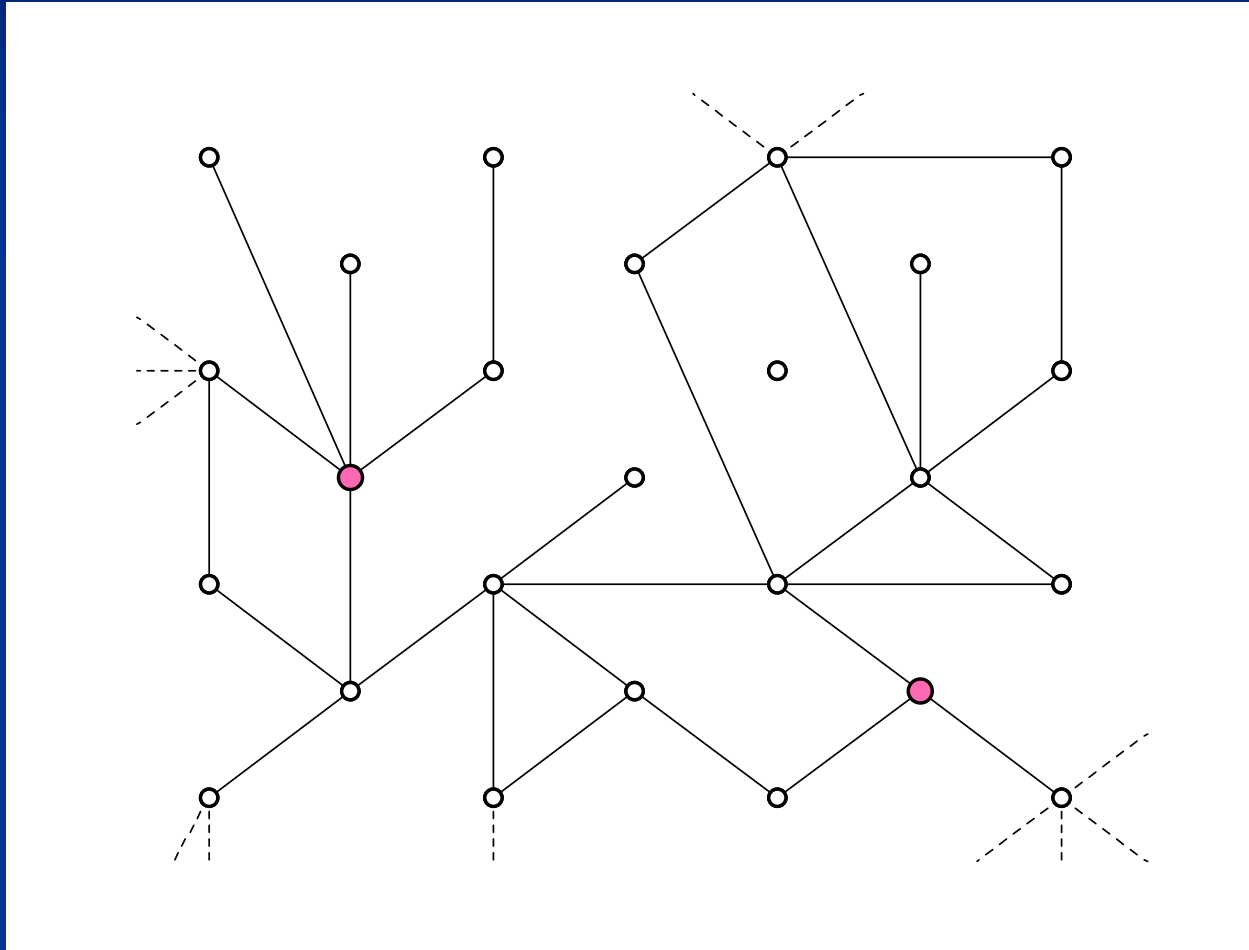
Simulation – creating a population



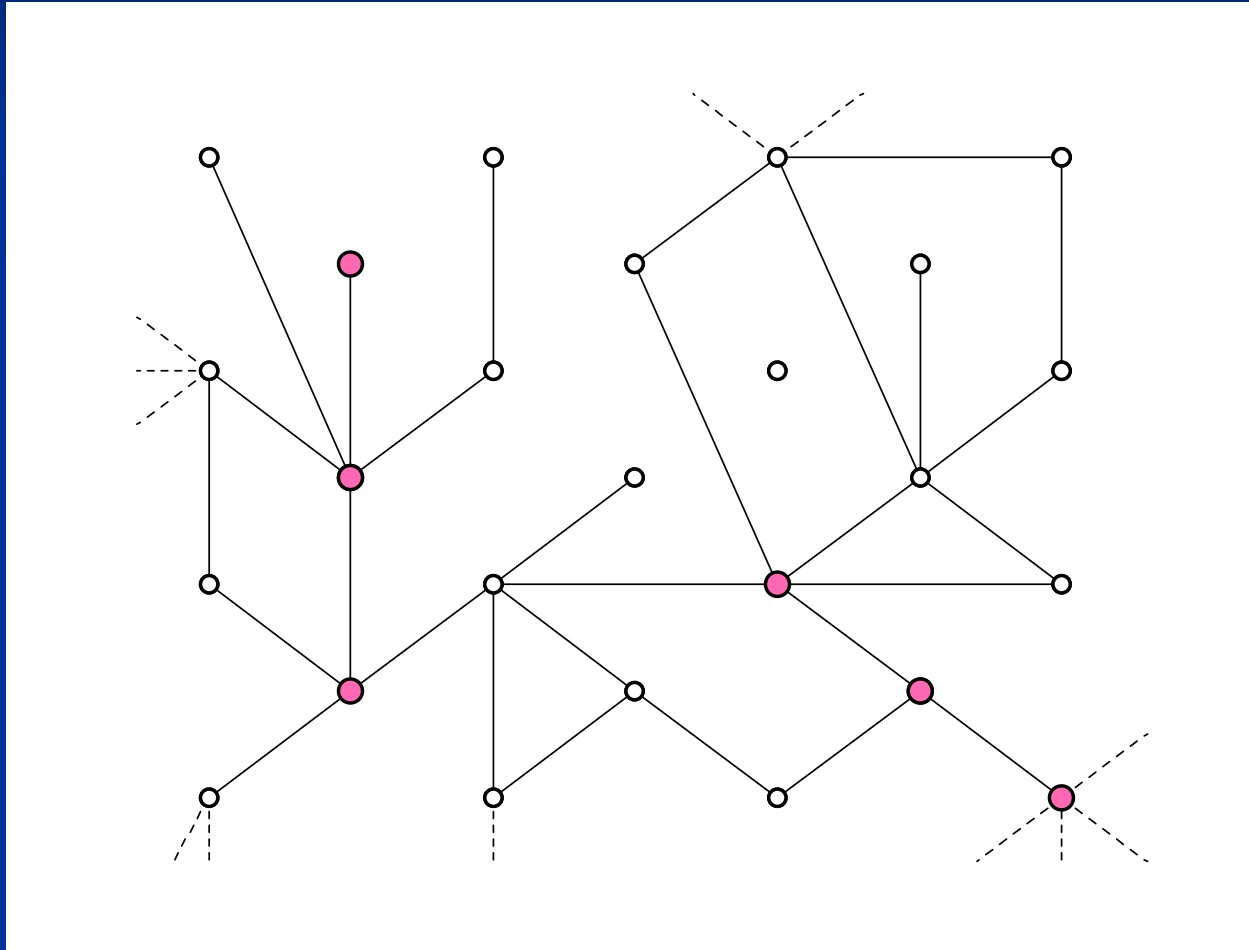
Simulation – creating a population



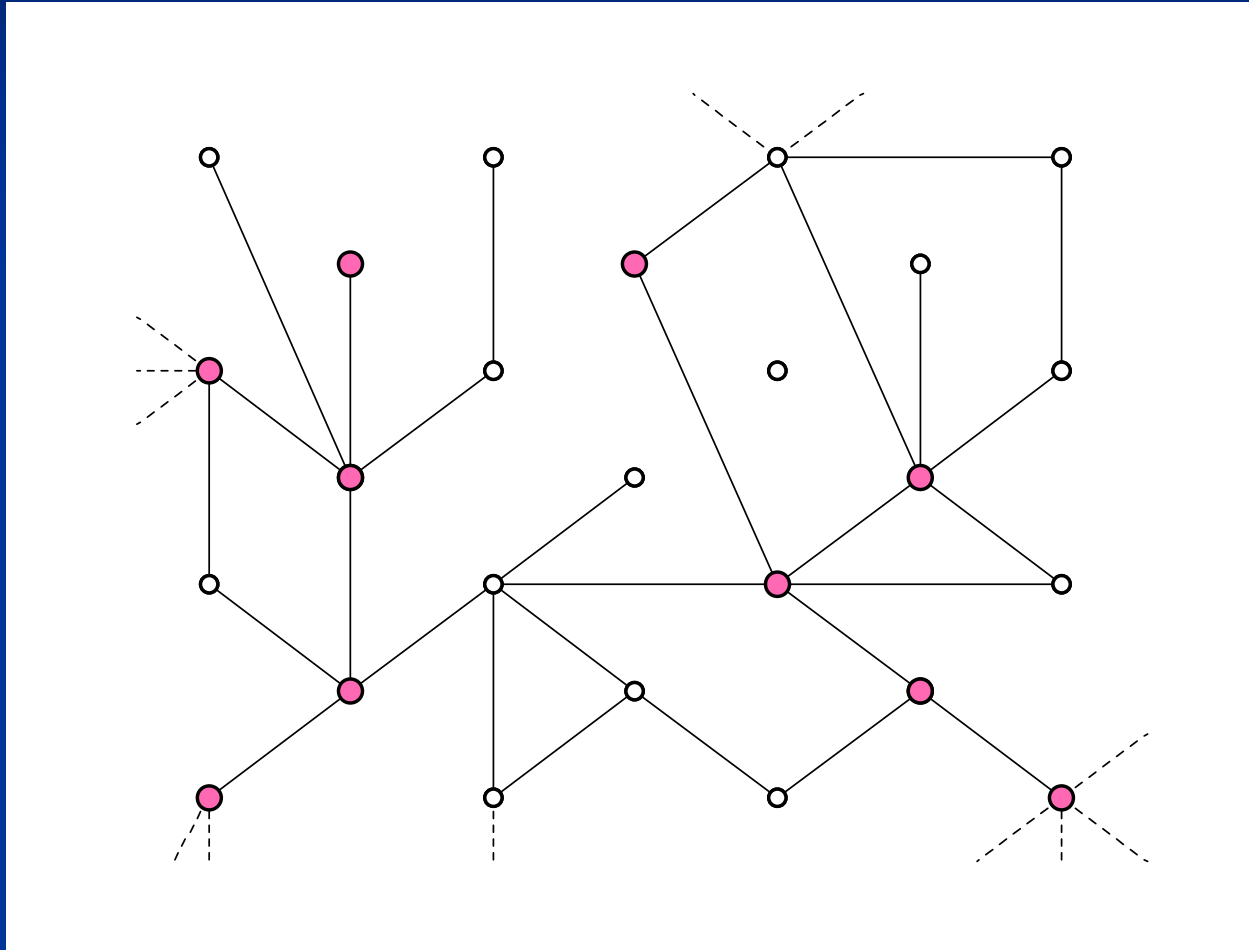
Simulation – creating a population



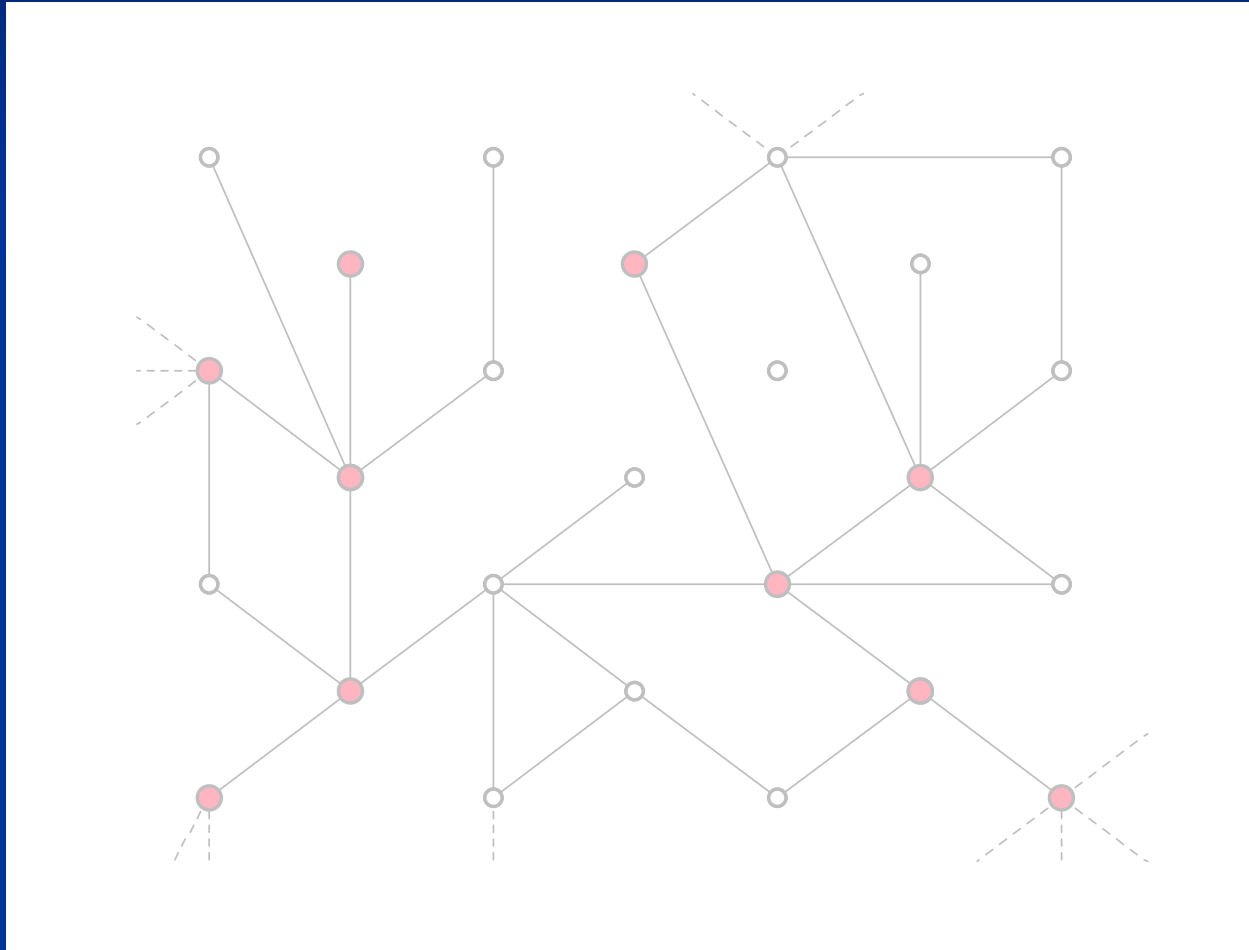
Simulation – creating a population



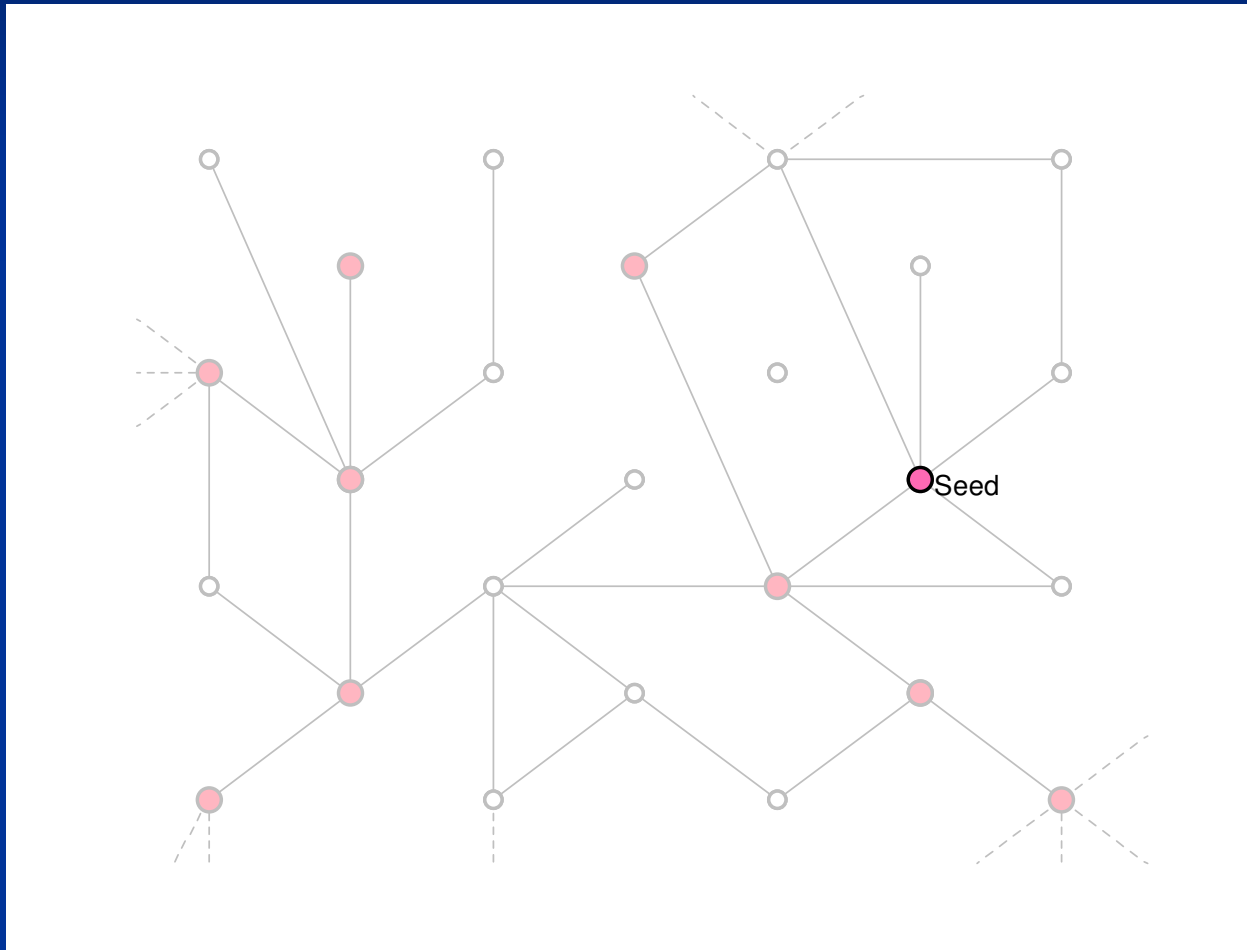
Simulation – creating a population



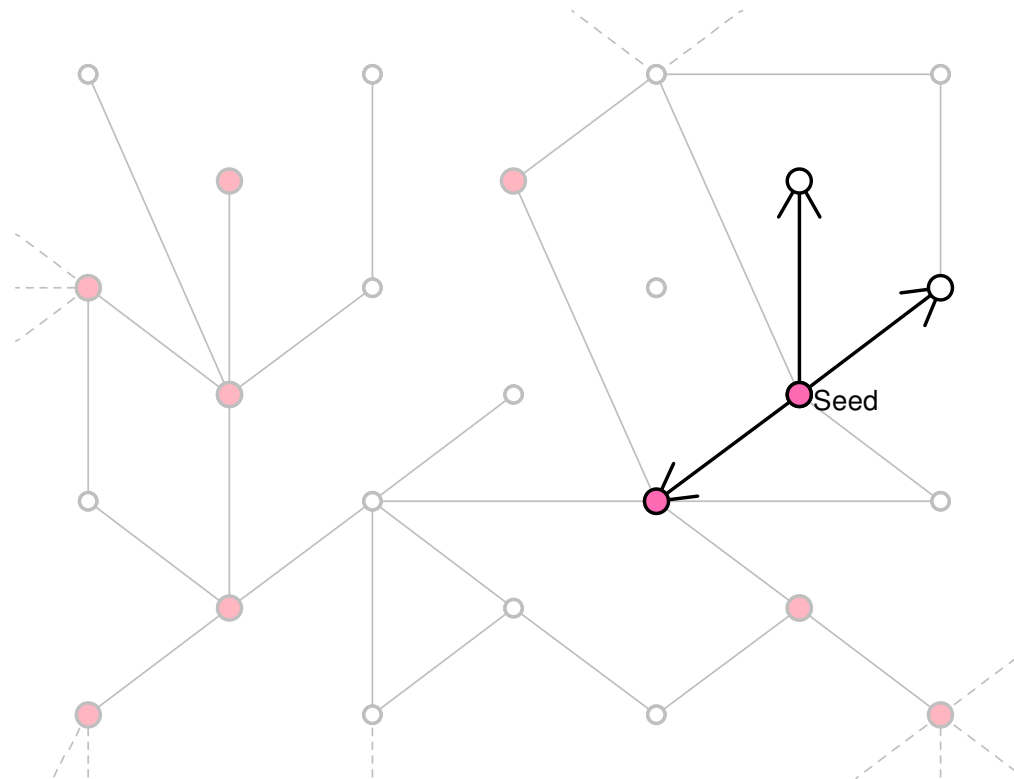
Simulation – sampling



Simulation – sampling



Simulation – sampling



Simulation – sampling

