ERGM for snowball sampled network data

Network sampling designs

- traditional random sampling is not sensible for a network study (except for ego-net studies).
  - Dependencies among individuals are lost by random sampling
  - an understanding of network connectivity is better obtained with some form of snowball sampling or link tracing design.

- Two broad motivations
  - to obtain information about network structures in larger communities
  - to obtain information about individuals in populations that are “hidden” or “hard to reach” (e.g. drug users).

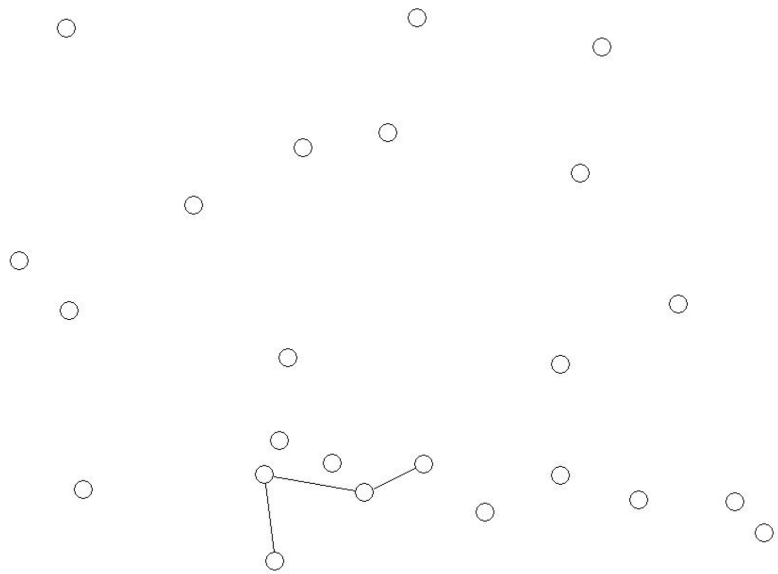
Snowball sampling

- the preferred approach for investigating network structure using sampled data
- A snowball sample is obtained by starting from an initial set of respondents (*the seed set*), determining their network partners (*wave 1*), determining the new network partners of wave 1 respondents (*wave 2*) and so on until stopping at an agreed number of waves.

Network sampling

○ Farmers

□ Agribusiness



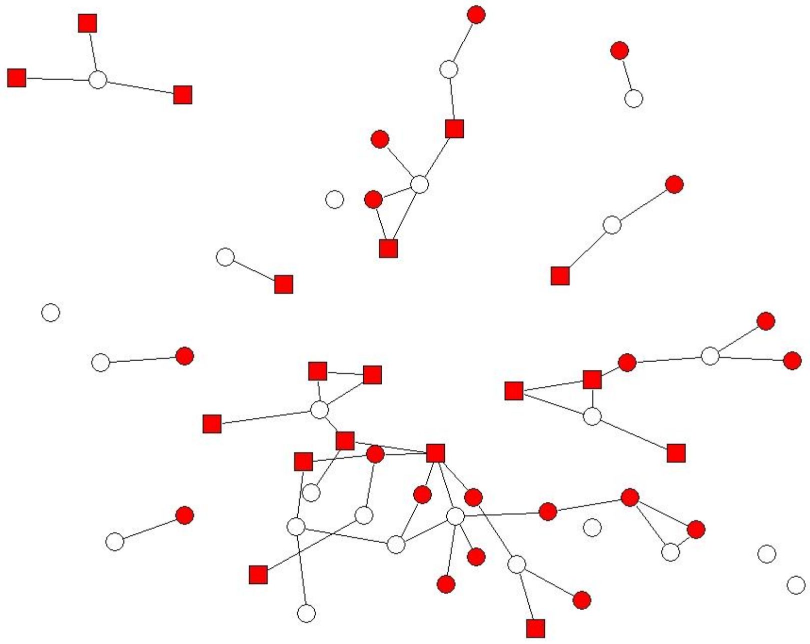
See set (wave 0)

Network sampling

○ Farmers

□ Agribusiness

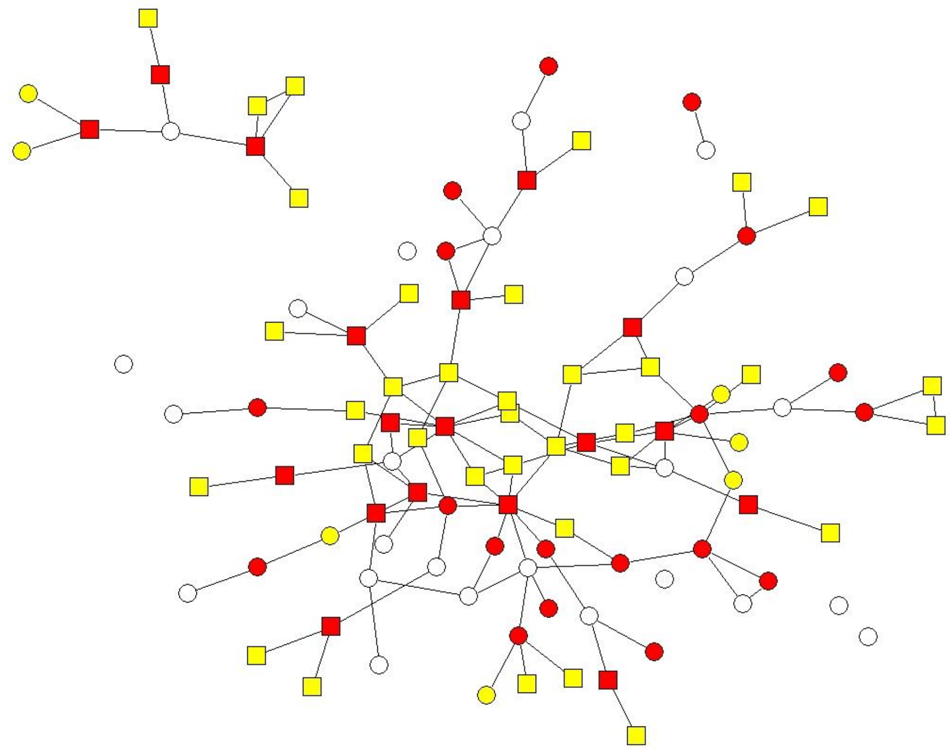
Seed set + 1



Network sampling

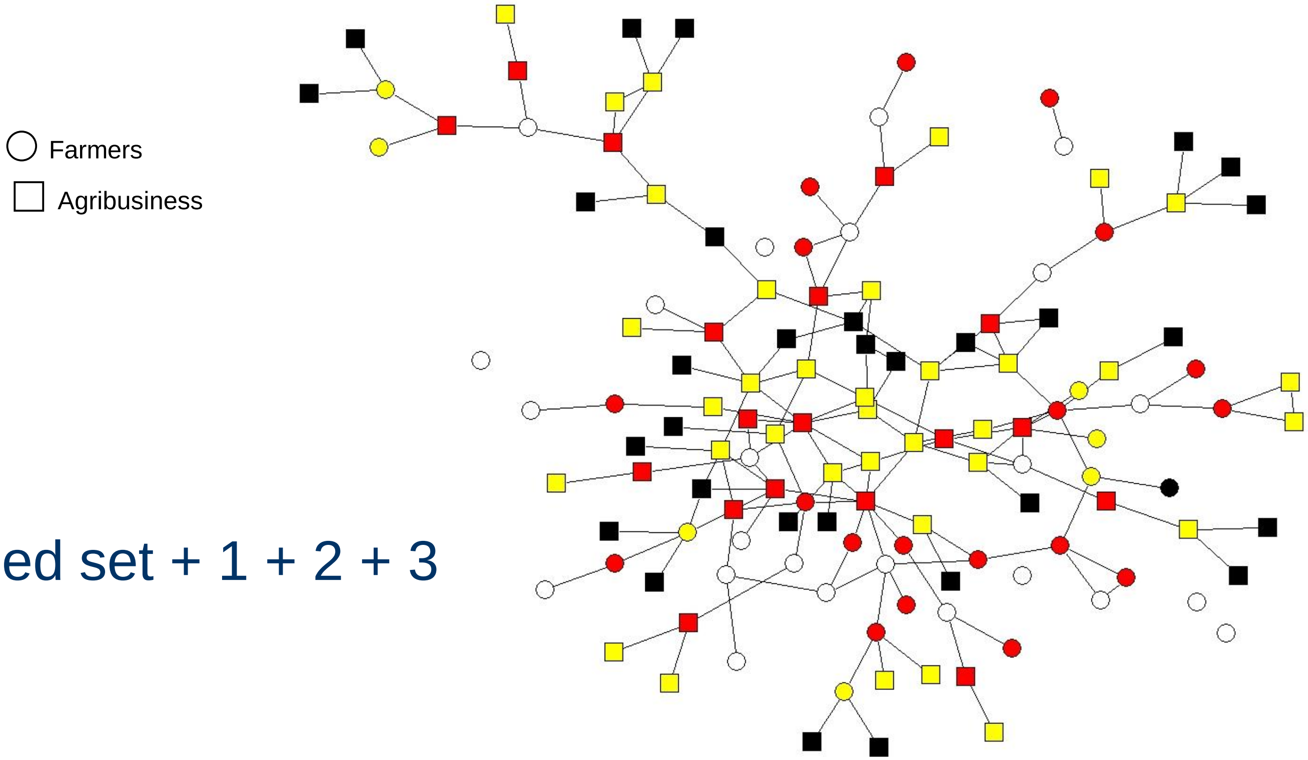
Farmers

Agribusiness



Seed set + 1 + 2

Network sampling



Seed set + 1 + 2 + 3

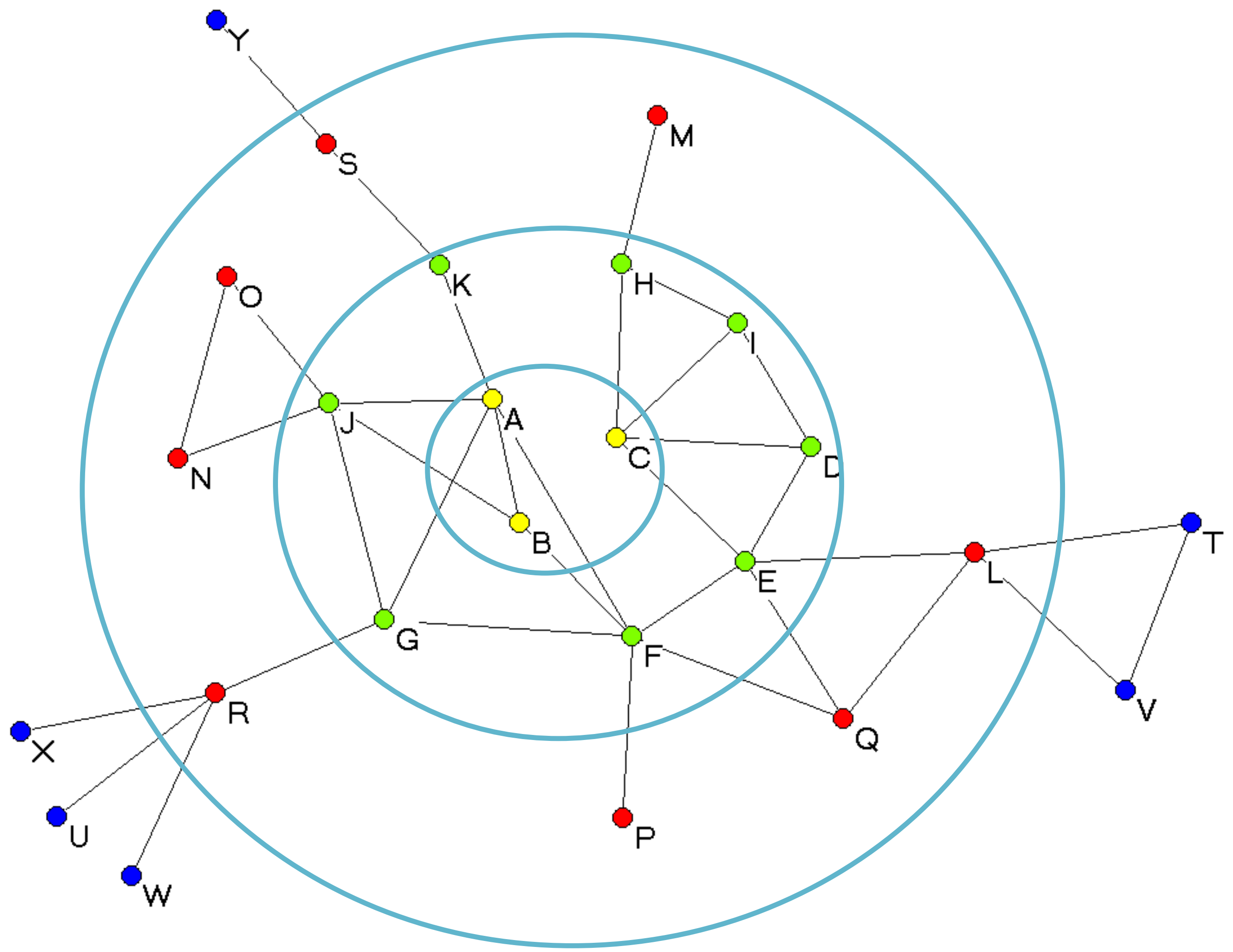
Estimation from snowball samples

Can we estimate exponential random graph models (ERGMs) when:

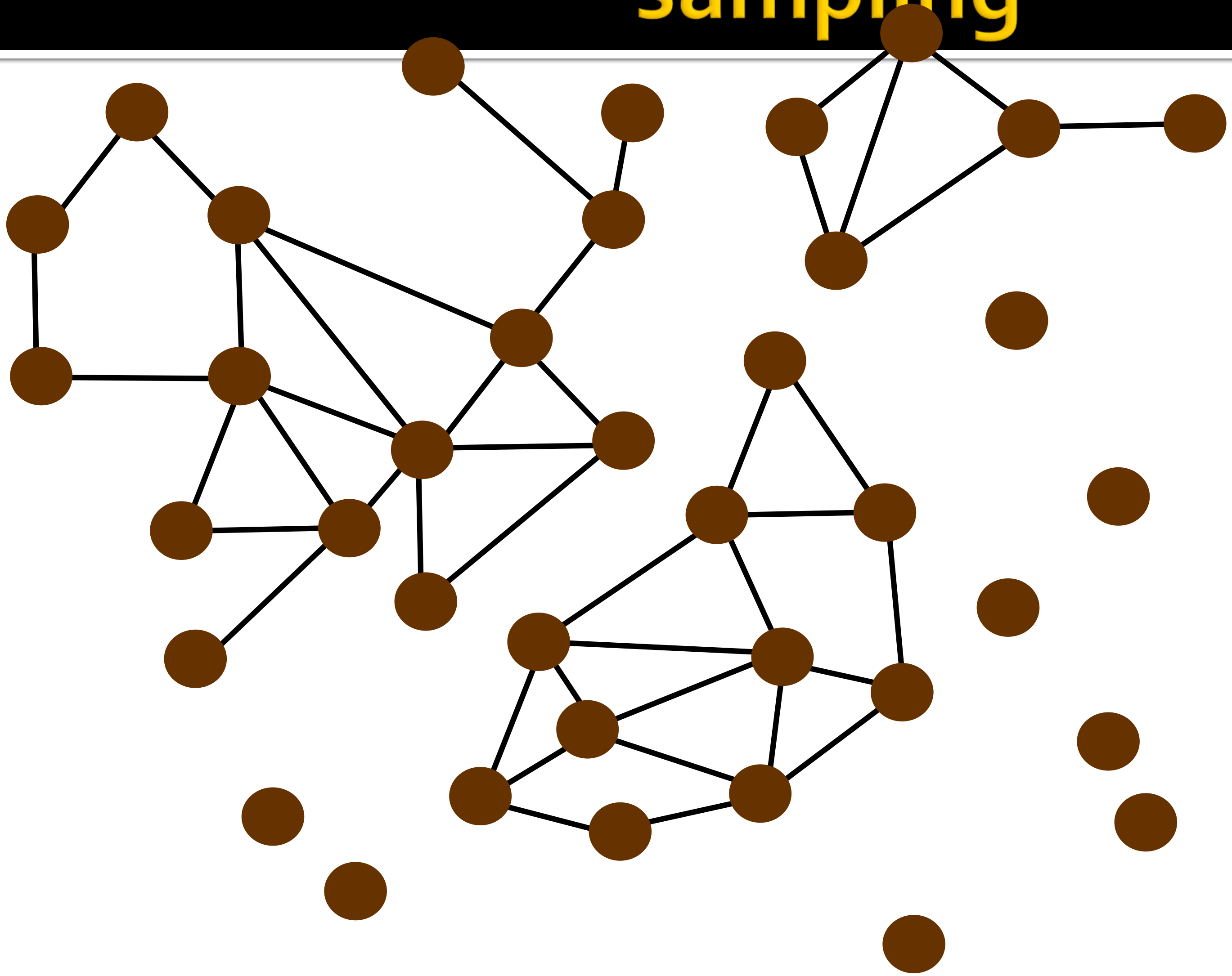
- **we have a sample rather than a census of network ties (specifically, a snowball sample)?**
- **the network is large?**
- **the network size is unknown?**
- **we have various model specifications in mind?**

This work complements Handcock and Gile (2010), who develop general likelihood inference for partial network data with known network size

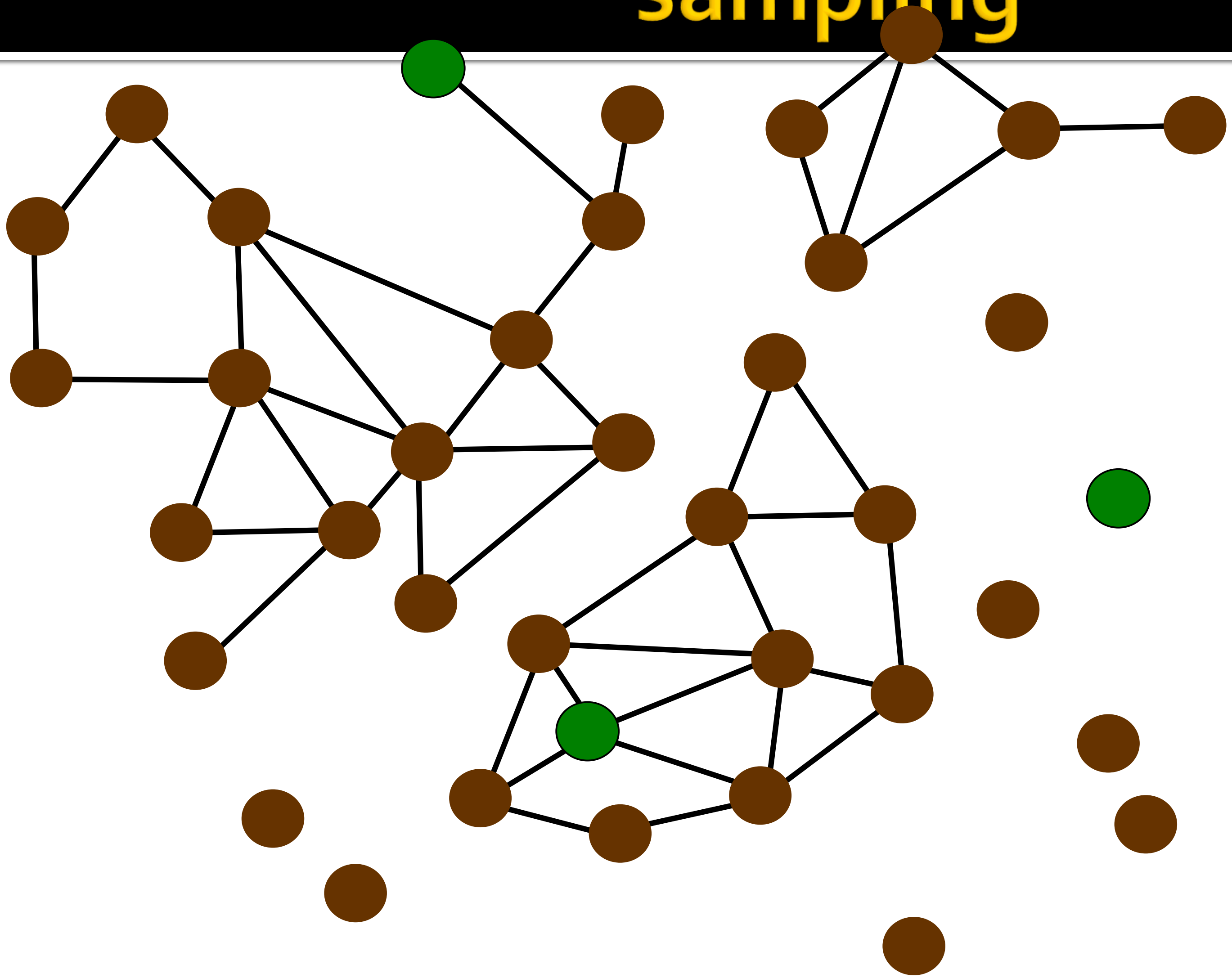
Multi-wave snowball designs (seed set = Z_0, k-wave snowball sample gives rise to the neighbourhood N_{k+1} of the seed set).



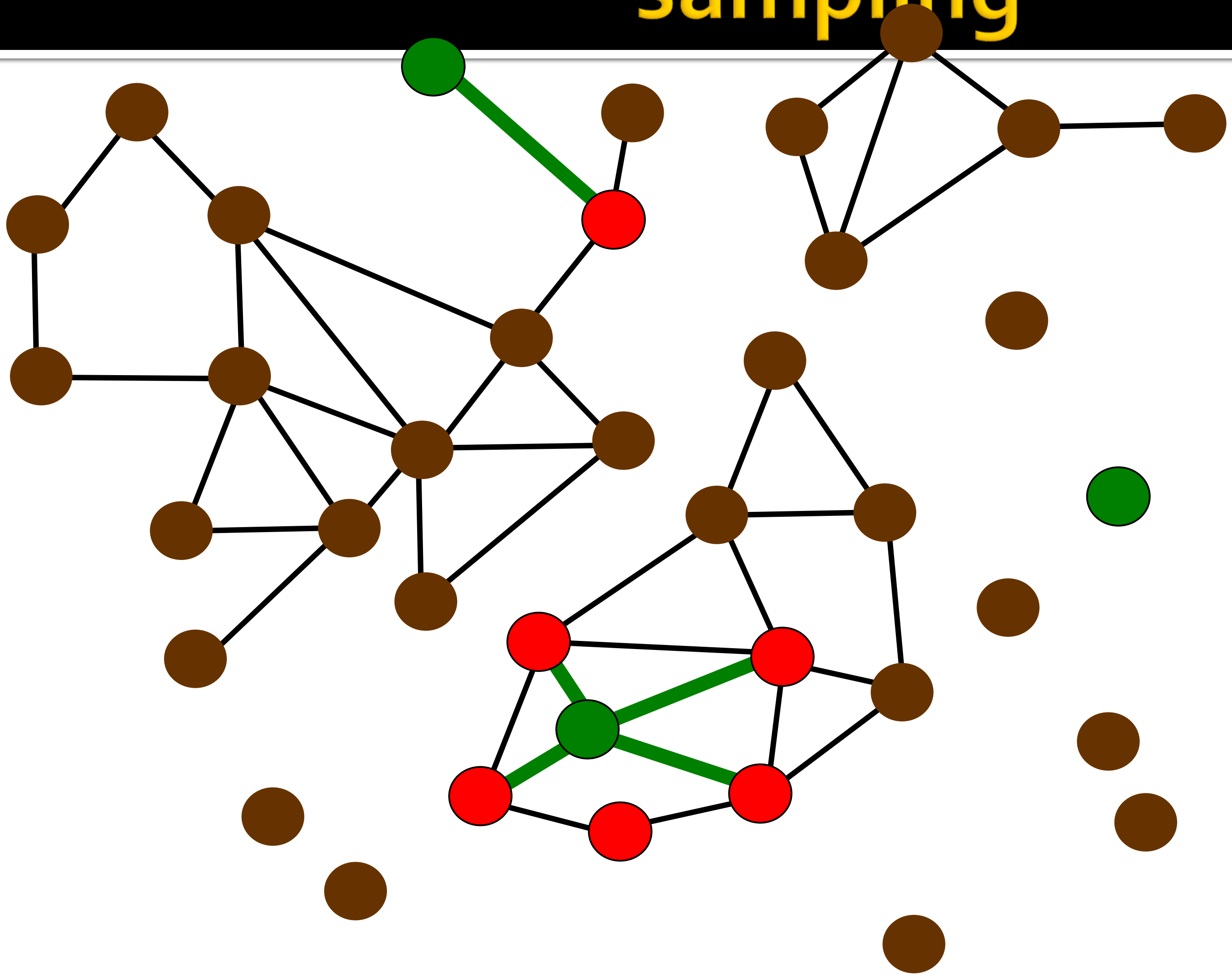
unobserved data: snowball sampling



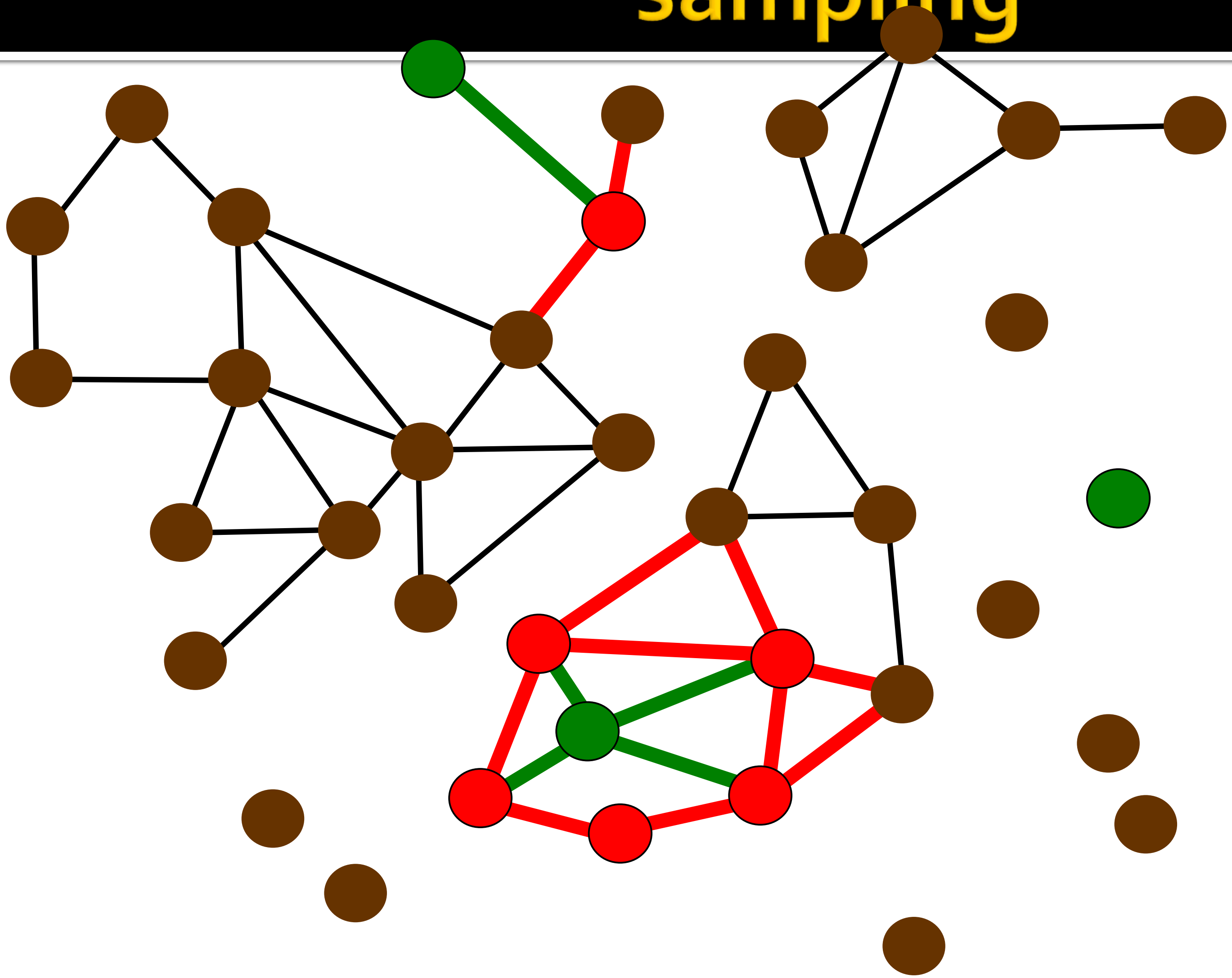
unobserved data: snowball sampling



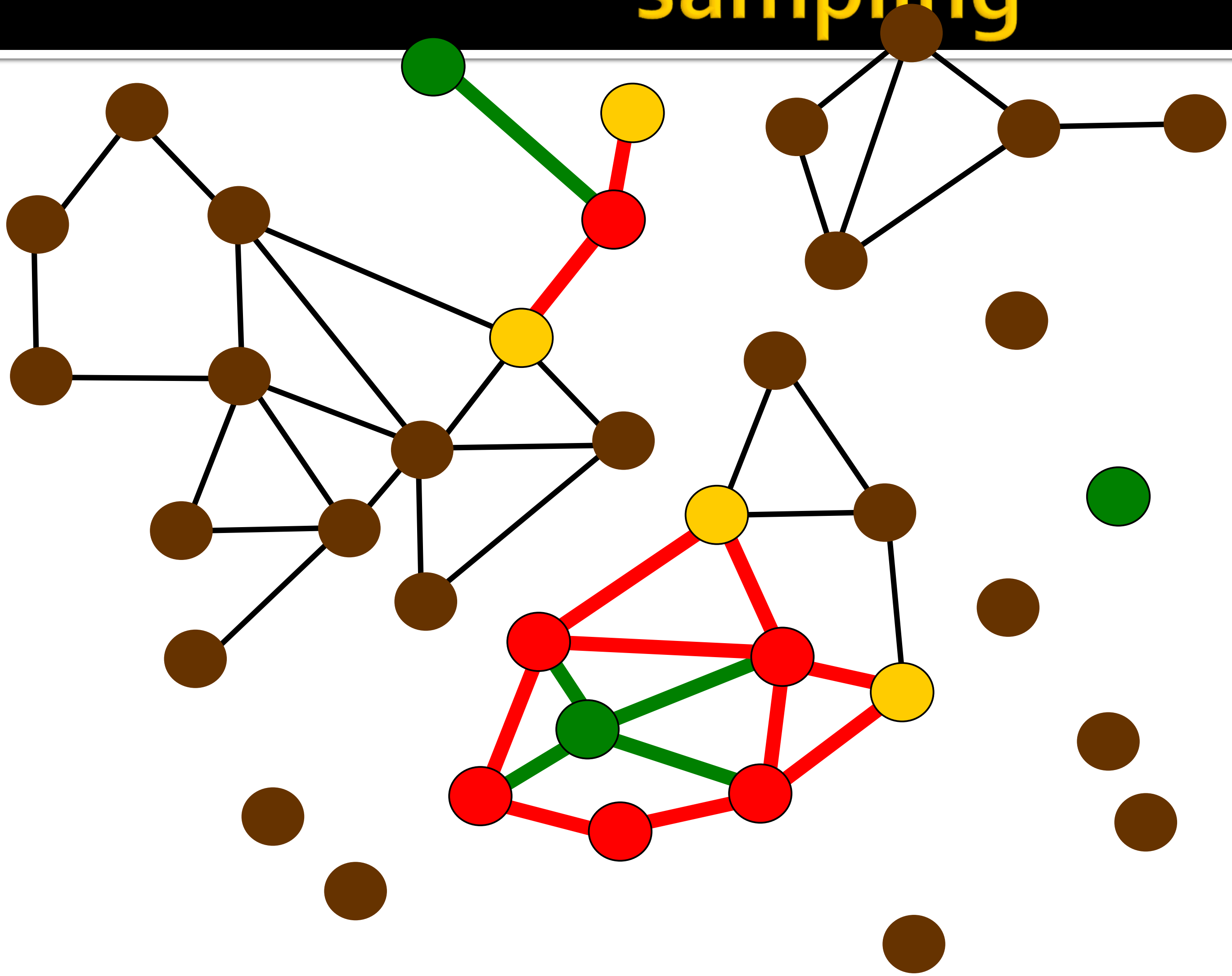
unobserved data: snowball sampling



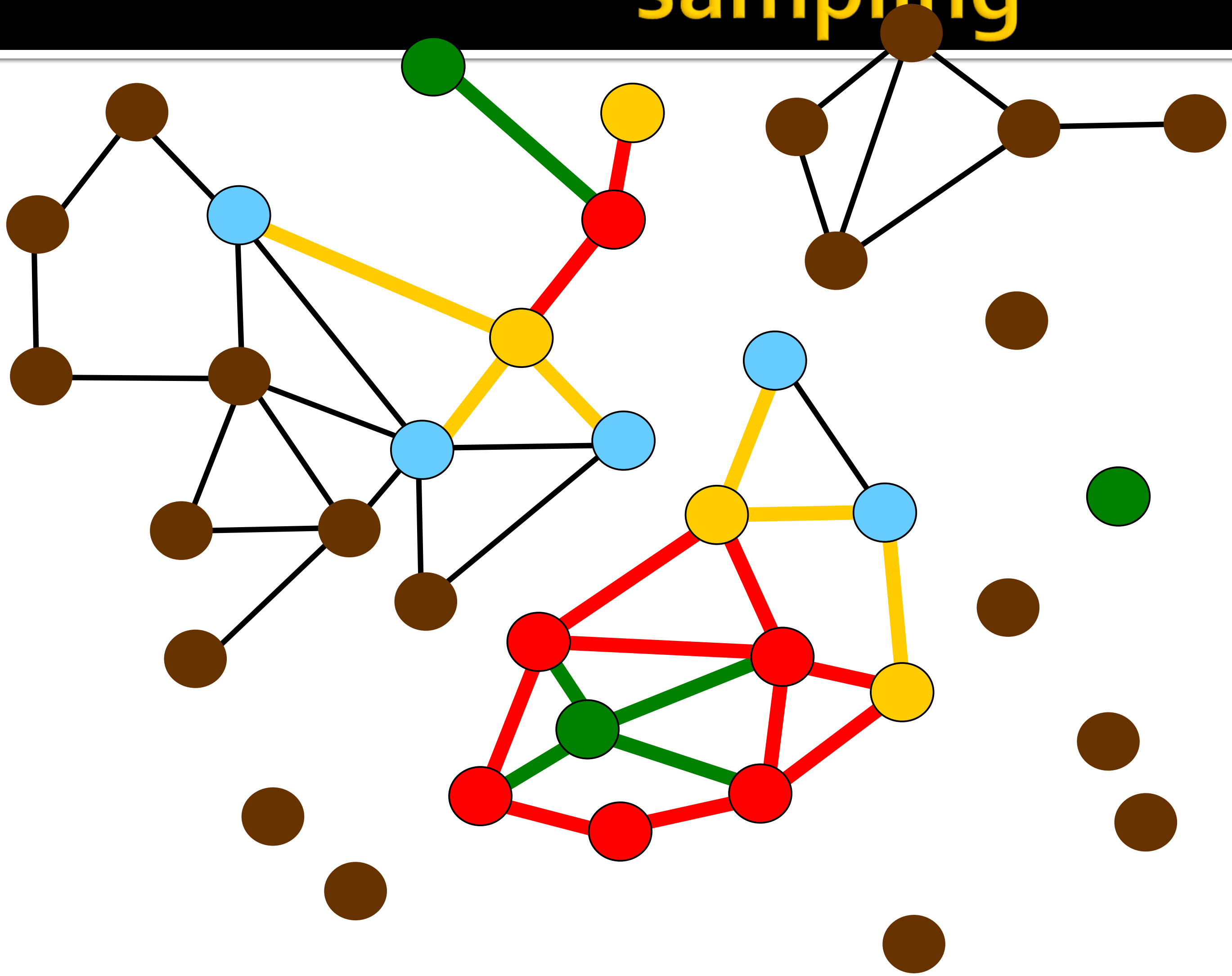
unobserved data: snowball sampling



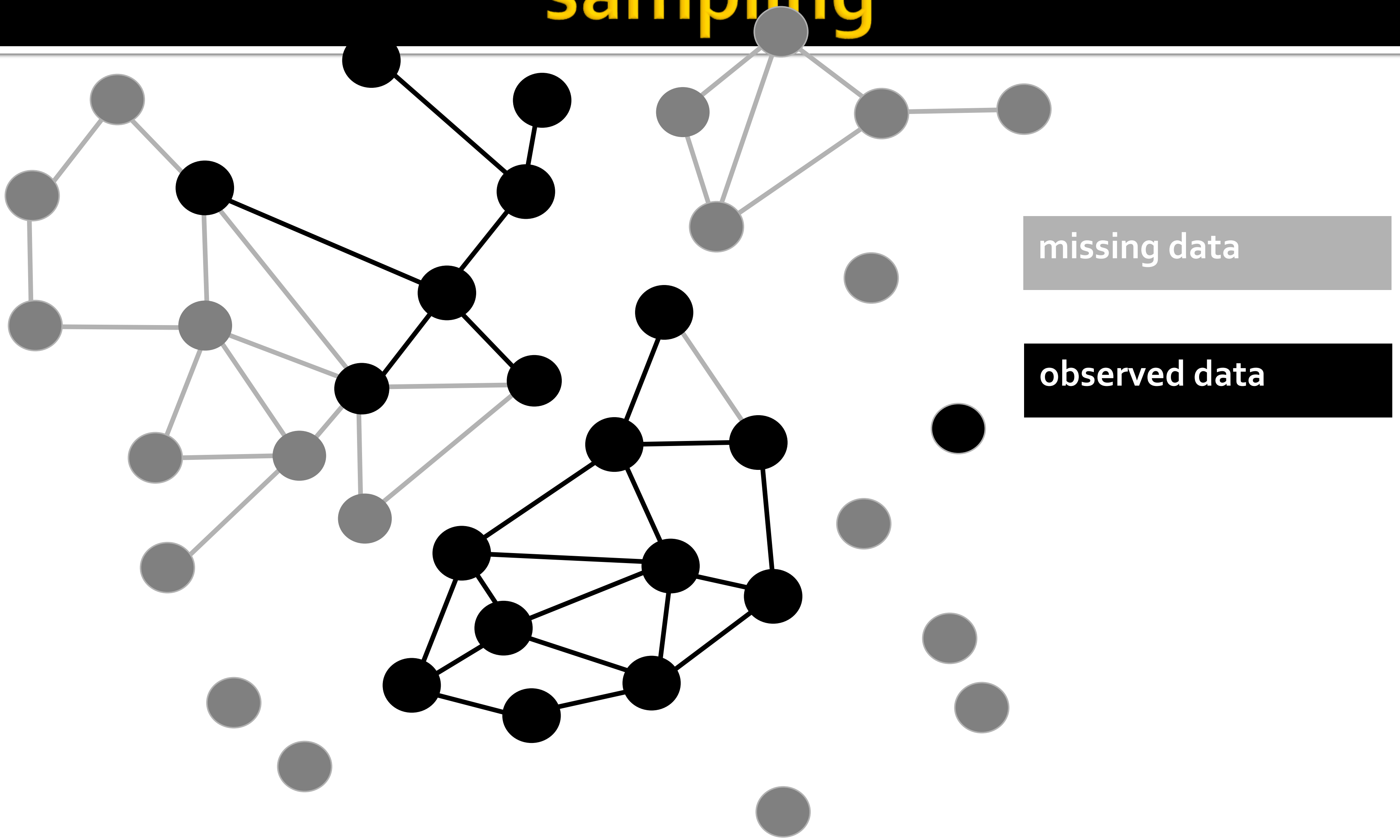
unobserved data: snowball sampling



unobserved data: snowball sampling



unobserved data: snowball sampling



unobserved data: snowball sampling

$$x = \left[\begin{array}{cc|c} - & 0 & 1 \\ & & \\ \hline 1 & & \\ & & \\ & & \end{array}\right]$$

unobserved data: snowball sampling

$$x = \left[\begin{array}{cc|ccc}
- & 0 & 1 & & \\
0 & - & 0 & 1 & 1 \\
\hline
1 & 0 & & & \\
 & 1 & & & \\
 & 1 & & & \\
 & & & &
\end{array}\right]$$

unobserved data: snowball sampling

$$x = \left[\begin{array}{cc:ccc}
- & 0 & 1 & 0 & 0 \\
0 & - & 0 & 1 & 1 \\
\hline
1 & 0 & & & \\
0 & 1 & & & \\
0 & 1 & & & \\
 & & & &
\end{array}\right]$$

unobserved data: snowball sampling

$$
x = \left[\begin{array}{cc|ccc|cc}
- & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & - & 0 & 1 & 1 & 0 & 0 \\
\hline
1 & 0 & & & & & \\
0 & 1 & & & & & \\
0 & 1 & & & & & \\
\hline
0 & 0 & & & & & \\
0 & 0 & & & & &
\end{array}\right]
$$

unobserved data: snowball sampling

$$x = \begin{bmatrix} - & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & - & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & - & 0 & 1 & & \\ 0 & 1 & 0 & - & - & & \\ 0 & 1 & 1 & 0 & - & & \\ 0 & 0 & & & & & \\ 0 & 0 & & & & & \end{bmatrix}$$

unobserved data: snowball sampling

$$
=
$$

$$
x = \left[\begin{array}{cc|ccc|cc}
- & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & - & 0 & 1 & 1 & 0 & 0 \\
\hline
1 & 0 & - & 0 & 1 & ? & ? \\
0 & 1 & 0 & - & - & ? & ? \\
0 & 1 & 1 & 0 & - & ? & ? \\
\hline
0 & 0 & ? & ? & ? & - & ? \\
0 & 0 & ? & ? & ? & ? & -
\end{array}\right]
$$

A conditional estimation strategy for a social circuit model: the basic idea

- Consider ties $X_{[k,k]}$ among nodes in N_k
- We use the fact that if $i, j \in N_k$, then $X(i,j)$ is conditionally independent of any tie involving a node outside N_{k+1}

	N_k	Z_{k+1}	$N_{k+1}{}^c$
N_k	$X_{[k,k]}$	$X_{k,k+1}$	0
Z_{k+1}	$X_{k+1,k}$	$X_{k+1,k+1}$	Z
$N_{k+1}{}^c$	0	Z	W

- More complex models can be accommodated: modelled ties must not be conditionally dependent on what has not been observed

Conditional estimation based on a *k*-wave sample

More generally:

Let $X_{[k,k]}$ denote tie variables on N_k. Then:

$$\log \Pr(X_{[k,k]}=x_{[k,k]} \mid Z_0, Z_1, \cdots Z_{k+1}, X_{[k,k]}^{C}=x_{[k,k]}^{C}) \\ = C + \sum_p \theta_p z_p(x_{[k+1,k+1]})$$

for a constant C that is independent of $x_{k,k]}$

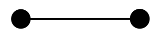
Conditionality on Z_0, Z_1, …, Z_{k+1} entails:

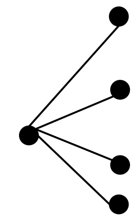
$X_{hm} = 0$ for $|h\text{-}m| \geq 2$ and all arrays of the form $X_{h,h+1}$ satisfy the condition that each node in Z_{h+1} can be reached from some node in Z_h

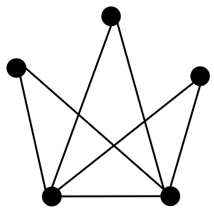
Estimation: in the MCMC during estimation, we propose only random changes to the entries in $X_{[k,k]}$ that respect this conditioning

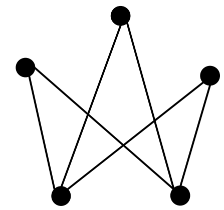
A simulation study

For the same fixed model:

edge -4.0 

alt-star 0.2 

alt-triangle 1.0 

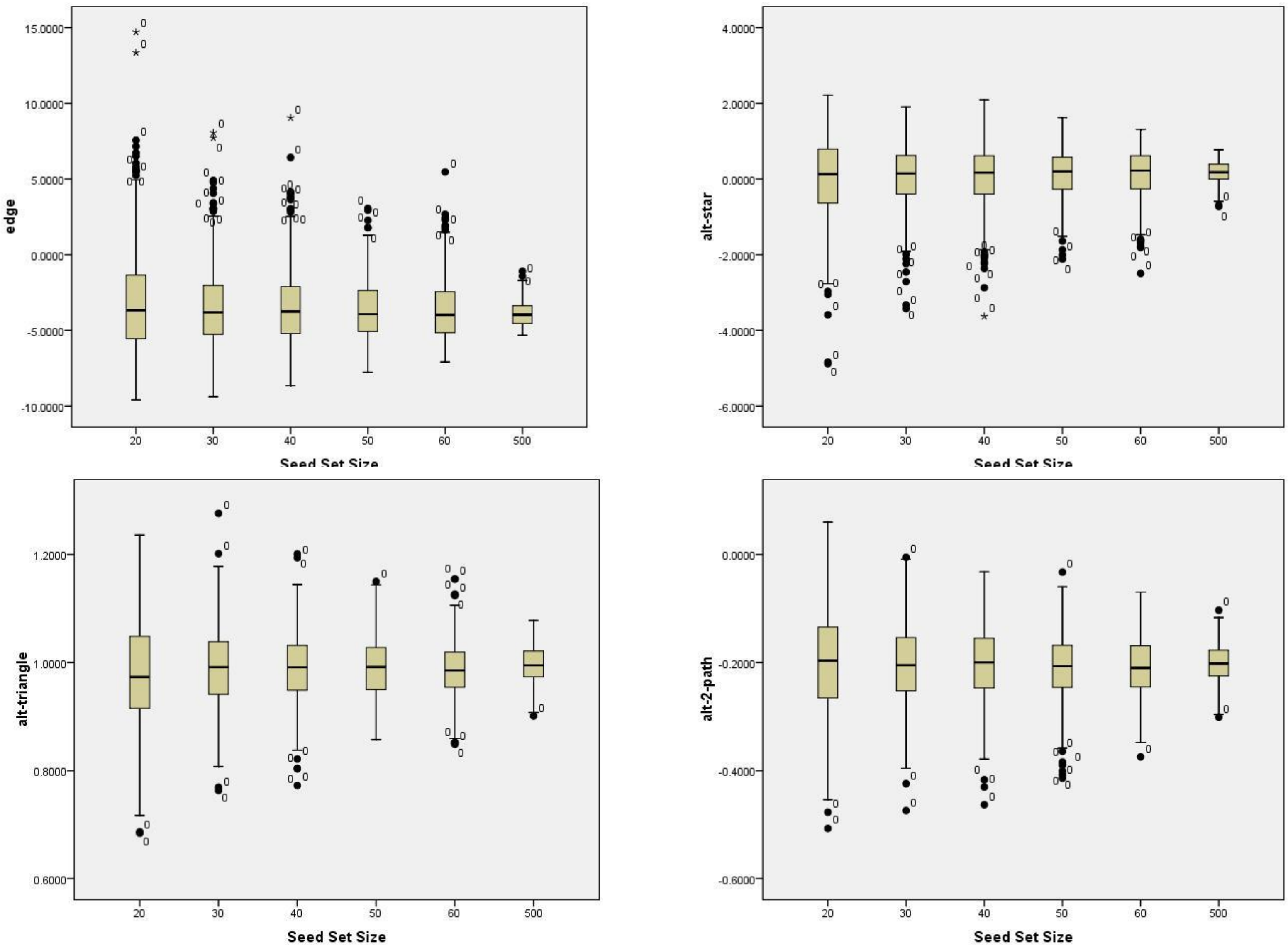
alt-2-path -0.2 

Size of network: 150, 500, 1000
Size of random seed sets: 20, 30, 40, 50, 60, **150** or **500**)

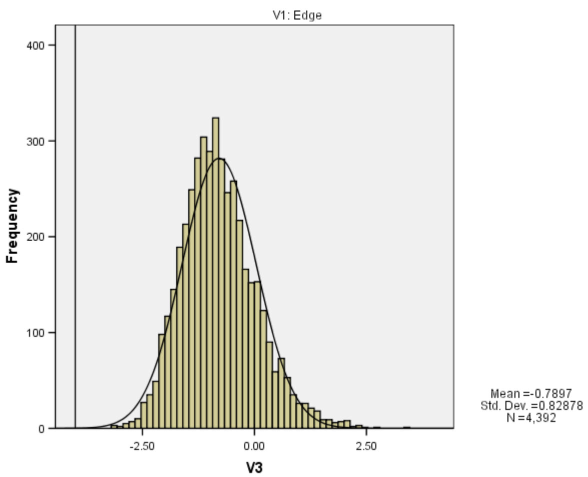
Complete networks

500 graphs sampled from the ERGM distribution
One snowball sample per graph

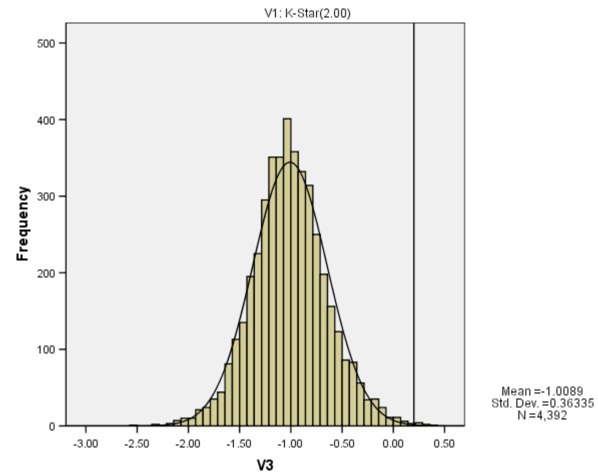
Distributions of estimates (*n* = 500)



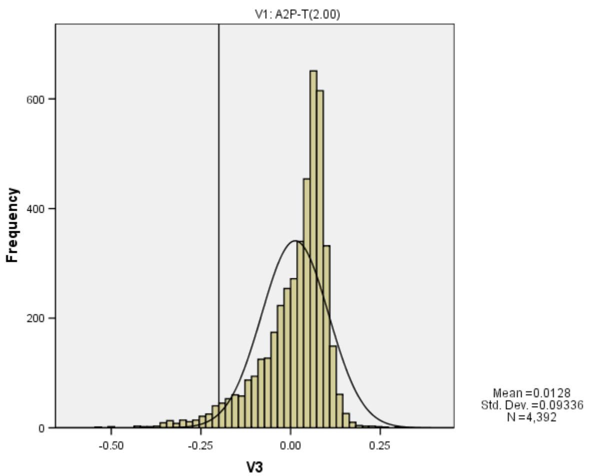
What if we ignore the sampling design?
MCMCMLEs from network on $Z_{[2]} = Z_0 \cup Z_1 \cup Z_2$



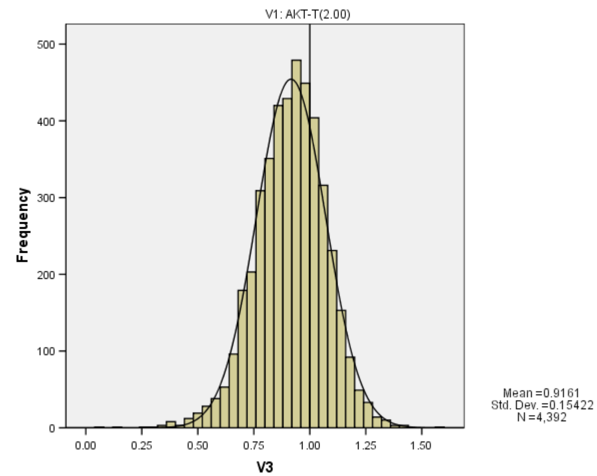
edge



alt-star

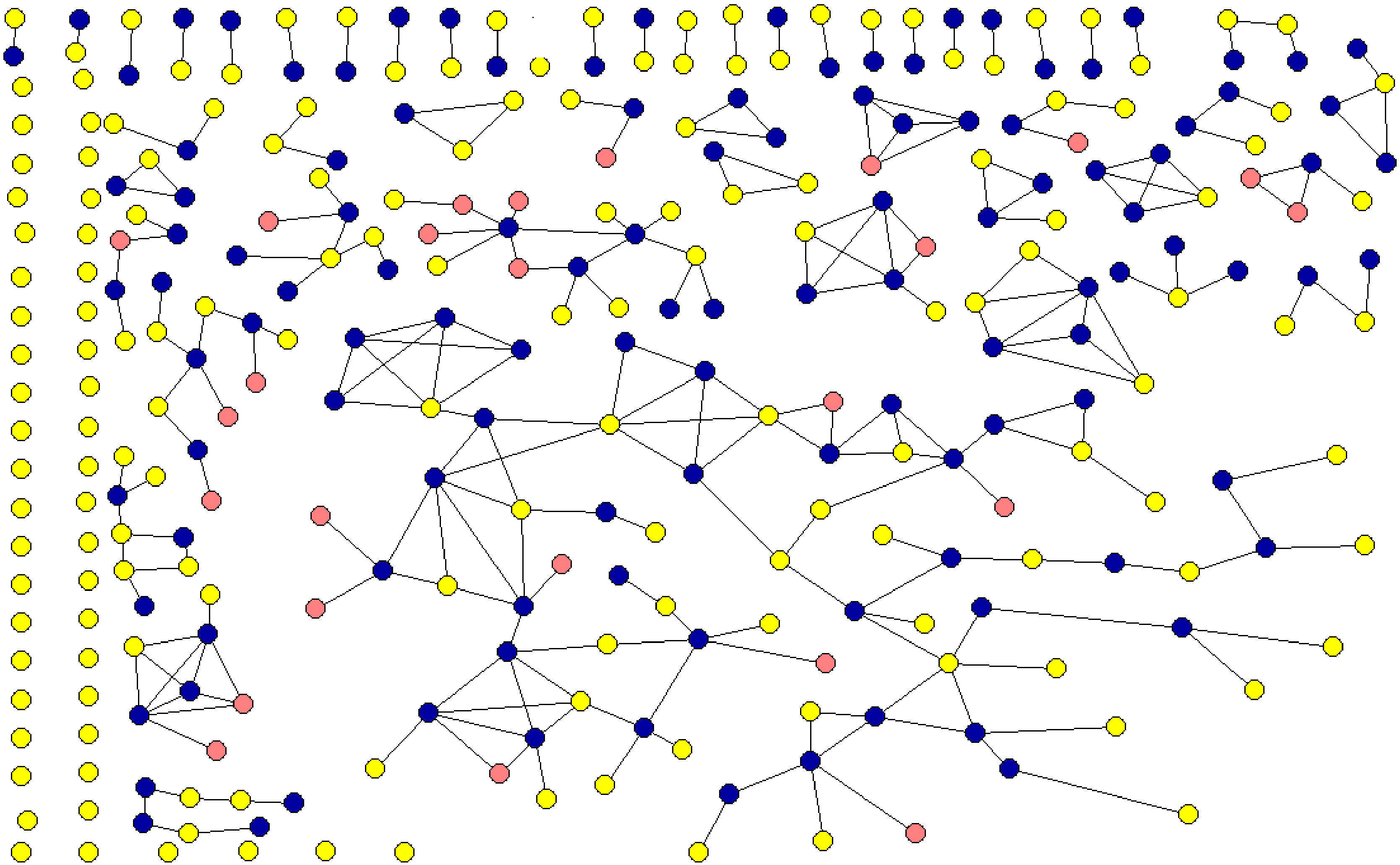


alt-2-path



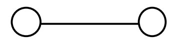
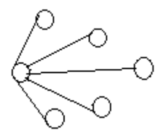
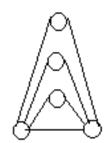
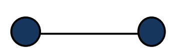
alt-triangle

Hellard, Aitken et al's *Networks II* study: injecting network for intravenous drug users in Melbourne

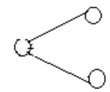
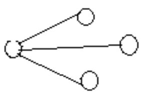
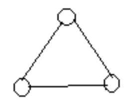
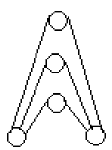


Zone 0
Zone 1
Zone 2

Conditional parameter estimates

Parameter	Configuration	Estimates (Standard error)
Edge		**-6.73*** (1.14)
Alt-star		-0.13 (0.33)
Alt-triangle		**1.42*** (0.27)
Gender		-0.85 (0.51)
Frequent user		-0.04 (0.39)
Different age		-**0.19*** (0.09)
# of non-identified partners		0.21 (0.27)
Same location		**1.72*** (0.59)
Same ethnicity		-0.06 (0.6)

Heuristic (conditional) goodness of fit

		
3-stars		0.0136
Triangles		0.2268
Isolates		0.0346
A-2-paths		-0.1923
Standard deviation of degree distribution		-0.0035
Skew degree distribution		0.0614
Global Clustering		0.2767
Mean local clustering		0.2357
Variance local clustering		0.6444

Potential applications

From the model estimates, we have obtained quantitative estimates (and estimated uncertainty) of:

- Density (low)
- Degree heterogeneity (not high)
- Form and level of clustering in the network (high)
- Homophily effects
  - Age, location (strong), Gender, no. of non-identified partners, frequency of use, ethnicity (weak)

Application:

- The model can be used in turn to build agent-based models of transmission of diseases such as HCV among IDUs ***at the population level***
- Such a model can be to used to assess potential impact (and uncertainty) of possible interventions ***(work in progress)***

Finally, other things you can do with ERGMs that we haven't talked about:

- Longitudinal models
- Social influence models
  - Autologistic actor attribute models
  - Including for multilevel networks
- Estimating network size
- New dependence assumptions
  - Brokerage: Edge-triangle configurations (and more)
- Snowball sampling for bipartite networks