Sampling and incomplete network data

567 Statistical analysis of social networks

Peter Hoff

Statistics, University of Washington
Network sampling methods

It is sometimes difficult to obtain a complete network dataset:

- the population nodeset is too large;
- gathering all relational information is too costly;
- population nodes are hard to reach.

In such cases, we need to think carefully how to

- gather the data (i.e. design the survey);
- make inference (i.e. estimate and evaluate parameters).
Network sampling methods

It is sometimes difficult to obtain a complete network dataset:

- the population nodeset is too large;
- gathering all relational information is too costly;
- population nodes are hard to reach.

In such cases, we need to think carefully how to

- gather the data (i.e. design the survey);
- make inference (i.e. estimate and evaluate parameters).
Network sampling methods

It is sometimes difficult to obtain a complete network dataset:

- the population nodeset is too large;
- gathering all relational information is too costly;
- population nodes are hard to reach.

In such cases, we need to think carefully how to

- gather the data (i.e. design the survey);
- make inference (i.e. estimate and evaluate parameters).
Network sampling methods

It is sometimes difficult to obtain a complete network dataset:

- the population nodeset is too large;
- gathering all relational information is too costly;
- population nodes are hard to reach.

In such cases, we need to think carefully how to

- gather the data (i.e. design the survey);
- make inference (i.e. estimate and evaluate parameters).
Network sampling methods

It is sometimes difficult to obtain a complete network dataset:

- the population nodeset is too large;
- gathering all relational information is too costly;
- population nodes are hard to reach.

In such cases, we need to think carefully how to

- gather the data (i.e. design the survey);
- make inference (i.e. estimate and evaluate parameters).
Network sampling methods

It is sometimes difficult to obtain a complete network dataset:

- the population nodeset is too large;
- gathering all relational information is too costly;
- population nodes are hard to reach.

In such cases, we need to think carefully how to

- gather the data (i.e. design the survey);
- make inference (i.e. estimate and evaluate parameters).
Common sampling methods

1. node-induced subgraph sampling
2. edge-induced subgraph sampling
3. egocentric sampling
4. link tracing designs
5. censored nomination schemes
Node-induced subgraph sampling

Procedure:

1. Uniformly sample a set \( s = \{s_1, \ldots, s_{n_s}\} \) of nodes

\[ s \subset \{1, \ldots, n\}. \]

2. Observe relations \( y_s \) between sampled nodes

\[ Y_s = \{y_{i,j} : i \in s, j \in s\}. \]
Node-induced subgraph sampling

Procedure:

1. Uniformly sample a set \( s = \{s_1, \ldots, s_{n_s} \} \) of nodes

\[
s \subset \{1, \ldots, n\}.\]

2. Observe relations \( y_s \) between sampled nodes

\[
Y_s = \{y_{i,j} : i \in s, j \in s\}.
\]
Node-induced subgraph sampling

Procedure:
1. Uniformly sample a set $s = \{s_1, \ldots, s_{n_s}\}$ of nodes
   $$s \subset \{1, \ldots, n\}.$$  
2. Observe relations $y_s$ between sampled nodes
   $$Y_s = \{y_{i,j} : i \in s, j \in s\}.$$
Node-induced subgraph sampling

**Procedure:**

1. Uniformly sample a set $s = \{s_1, \ldots, s_{n_s}\}$ of nodes
   $$s \subset \{1, \ldots, n\}.$$ 

2. Observe relations $y_s$ between sampled nodes
   $$Y_s = \{y_{i,j} : i \in s, j \in s\}.$$
Node-induced subgraph sampling
Node-induced subgraph sampling
Node-induced subgraph sampling
Node-induced subgraph sampling
Node-induced subgraph sampling
Node-induced subgraph sampling
In what ways does $Y_s$ resemble $Y$?

For what functions $g()$ will $g(Y_s)$ estimate $g(Y)$?

Consider the following setup:

- $n \times n$ sociomatrix $Y$
- $n \times n$ dyadic covariate $X_d$
- $n \times 1$ nodal covariate $X_n$

Can we estimate the following from a sample?

\[ \bar{y} = \frac{1}{n(n-1)} \sum_{i \neq j} y_{i,j}, \quad \bar{x}_d = \frac{1}{n(n-1)} \sum_{i \neq j} x_{d,i,j}, \quad \bar{x}_n = \frac{1}{n} \sum x_{n,i} \]

\[ \bar{y}x_d = \frac{1}{n(n-1)} \sum_{i \neq j} y_{i,j}x_{d,i,j}, \quad \bar{y}x_n = \frac{1}{n(n-1)} \sum x_{n,i} \bar{y}_i. \]
In what ways does $Y_s$ resemble $Y$?

For what functions $g()$ will $g(Y_s)$ estimate $g(Y)$?

Consider the following setup:
- $n \times n$ sociomatrix $Y$
- $n \times n$ dyadic covariate $X_d$
- $n \times 1$ nodal covariate $X_n$

Can we estimate the following from a sample?

$$
\bar{y} = \frac{1}{n(n-1)} \sum_{i \neq j} y_{i,j}, \quad \bar{x}_d = \frac{1}{n(n-1)} \sum_{i \neq j} x_{d,i,j}, \quad \bar{x}_n = \frac{1}{n} \sum x_{n,i}
$$

$$
\bar{y}x_d = \frac{1}{n(n-1)} \sum_{i \neq j} y_{i,j}x_{d,i,j}, \quad \bar{y}x_n = \frac{1}{n(n-1)} \sum x_{n,i}\bar{y}_i.
$$
Estimation from sampled data

In what ways does $Y_s$ resemble $Y$?

For what functions $g()$ will $g(Y_s)$ estimate $g(Y)$?

Consider the following setup:

- $n \times n$ sociomatrix $Y$
- $n \times n$ dyadic covariate $X_d$
- $n \times 1$ nodal covariate $X_n$

Can we estimate the following from a sample?

$$
\bar{y} = \frac{1}{n(n-1)} \sum_{i \neq j} y_{i,j}, \quad \bar{x}_d = \frac{1}{n(n-1)} \sum_{i \neq j} x_{d,i,j}, \quad \bar{x}_n = \frac{1}{n} \sum x_{n,i}
$$

$$
\bar{y}x_d = \frac{1}{n(n-1)} \sum_{i \neq j} y_{i,j} x_{d,i,j}, \quad \bar{y}x_n = \frac{1}{n(n-1)} \sum_i x_{n,i} \bar{y}_i.
$$
Estimation from sampled data

In what ways does $Y_s$ resemble $Y$?

For what functions $g()$ will $g(Y_s)$ estimate $g(Y)$?

Consider the following setup:

- $n \times n$ sociomatrix $Y$
- $n \times n$ dyadic covariate $X_d$
- $n \times 1$ nodal covariate $X_n$

Can we estimate the following from a sample?

$$
\bar{y} = \frac{1}{n(n-1)} \sum_{i \neq j} y_{i,j}, \quad \bar{x}_d = \frac{1}{n(n-1)} \sum_{i \neq j} x_{d,i,j}, \quad \bar{x}_n = \frac{1}{n} \sum_i x_{n,i}
$$

$$
\bar{y}x_d = \frac{1}{n(n-1)} \sum_{i \neq j} y_{i,j} x_{d,i,j}, \quad \bar{y}x_n = \frac{1}{n(n-1)} \sum_i x_{n,i} \bar{y}_i.
$$
Estimation from sampled data

In what ways does \( Y_s \) resemble \( Y \)?

For what functions \( g() \) will \( g(Y_s) \) estimate \( g(Y) \)?

Consider the following setup:

- \( n \times n \) sociomatrix \( Y \)
- \( n \times n \) dyadic covariate \( X_d \)
- \( n \times 1 \) nodal covariate \( X_n \)

Can we estimate the following from a sample?

\[
\bar{y} = \frac{1}{n(n-1)} \sum_{i \neq j} y_{i,j}, \quad \bar{x}_d = \frac{1}{n(n-1)} \sum_{i \neq j} x_{d,i,j}, \quad \bar{x}_n = \frac{1}{n} \sum x_{n,i}
\]

\[
\bar{y}x_d = \frac{1}{n(n-1)} \sum_{i \neq j} y_{i,j}x_{d,i,j}, \quad \bar{y}x_n = \frac{1}{n(n-1)} \sum x_{n,i} \bar{y}_i.
\]
Node-induced subgraph sampling
Node-induced subgraph sampling

For some functions $g$, the sample value $g(Y_s)$ is an unbiased estimator of the population value $g(Y)$:

$$g(Y) = \text{an average of subgraphs of size } k, \text{ for } k \leq n_s$$

$$g(Y) = \frac{1}{\binom{n}{2}} \sum_{i<j} h(y_{i,j}, y_{j,i})$$

$$g(Y) = \frac{1}{\binom{n}{3}} \sum_{i<j<k} h(y_{i,j}, y_{j,i}, y_{i,k}, y_{k,i}, y_{j,k}, y_{k,j}) \text{ if } n_s \geq 3$$

Why does it work?:
Each subgraph of size $k$ appears in the sample with equal probability (although the subgraphs that appear are dependent).

Some functions of interest are not of this type:
- in and outdegree distributions;
- geodesics, distances, number of paths, etc.
Node-induced subgraph sampling

For some functions $g$, the sample value $g(Y_s)$ is an unbiased estimator of the population value $g(Y)$:

$$g(Y) = \text{an average of subgraphs of size } k, \text{ for } k \leq n_s$$

$$g(Y) = \frac{1}{\binom{n}{2}} \sum_{i<j} h(y_{i,j}, y_{j,i})$$

$$g(Y) = \frac{1}{\binom{n}{3}} \sum_{i<j<k} h(y_{i,j}, y_{j,i}, y_{i,k}, y_{k,i}, y_{j,k}, y_{k,j}) \text{ if } n_s \geq 3$$

Why does it work?:
Each subgraph of size $k$ appears in the sample with equal probability (although the subgraphs that appear are dependent).

Some functions of interest are not of this type:
- in and outdegree distributions;
- geodesics, distances, number of paths, etc.
Node-induced subgraph sampling

For some functions $g$, the sample value $g(Y_s)$ is an unbiased estimator of the population value $g(Y)$:

$$g(Y) = \text{an average of subgraphs of size } k, \text{ for } k \leq n_s$$

$$g(Y) = \frac{1}{(n \choose 2)} \sum_{i<j} h(y_{i,j}, y_{j,i})$$

$$g(Y) = \frac{1}{(n \choose 3)} \sum_{i<j<k} h(y_{i,j}, y_{j,i}, y_{i,k}, y_{k,i}, y_{j,k}, y_{k,j}) \text{ if } n_s \geq 3$$

**Why does it work?:**
Each subgraph of size $k$ appears in the sample with equal probability (although the subgraphs that appear are dependent).

Some functions of interest are not of this type:
- in and outdegree distributions;
- geodesics, distances, number of paths, etc.
Node-induced subgraph sampling

For some functions $g$, the sample value $g(Y_s)$ is an unbiased estimator of the population value $g(Y)$:

$$g(Y) = \text{an average of subgraphs of size } k, \text{ for } k \leq n_s$$

$$g(Y) = \frac{1}{\binom{n}{2}} \sum_{i<j} h(y_{i,j}, y_{j,i})$$

$$g(Y) = \frac{1}{\binom{n}{3}} \sum_{i<j<k} h(y_{i,j}, y_{j,i}, y_{i,k}, y_{k,i}, y_{j,k}, y_{k,j}) \text{ if } n_s \geq 3$$

Why does it work?:

Each subgraph of size $k$ appears in the sample with equal probability (although the subgraphs that appear are dependent).

Some functions of interest are not of this type:

- in and outdegree distributions;
- geodesics, distances, number of paths, etc.
Edge-induced subgraph sampling

Procedure:

1. Uniformly sample a set $e = \{e_1, \ldots, e_{n_e}\}$ of edges

   $$e \subset \{(i,j) : y_{i,j} = 1\}$$

2. Let $Y_s$ be the edge-generated subgraph of $e$. 
**Edge-induced subgraph sampling**

**Procedure:**

1. Uniformly sample a set $\mathbf{e} = \{e_1, \ldots, e_{n_e}\}$ of edges

   $$\mathbf{e} \subset \{(i, j): y_{i,j} = 1\}$$

2. Let $\mathbf{Y}_s$ be the edge-generated subgraph of $\mathbf{e}$. 
Edge-induced subgraph sampling
Edge-induced subgraph sampling
Edge-induced subgraph sampling
Edge-induced subgraph sampling
Edge-induced subgraph sampling
Edge-induced subgraph sampling
Edge-induced subgraph sampling

How well do these subgraphs represent $Y$?

Can you infer anything about $Y$ from these data?
Edge-induced subgraph sampling

How well do these subgraphs represent $Y$?
Can you infer anything about $Y$ from these data?
Edge-induced subgraph sampling
Egocentric sampling

**Procedure:**

1. Uniformly sample a set $s_1 = \{s_{1,1}, \ldots, s_{1,n_s}\}$ of nodes

   $$s_1 \subseteq \{1, \ldots, n\}.$$ 

2. Observe the relations for each $i \in s_1$, i.e. observe $\{y_{i,1}, \ldots, y_{i,n}\}$.

3. Let $s_2$ be the set of nodes having a link from anyone in $s_1$. Observe the relations of anyone in $s_2$ to anyone in $s_1 \cup s_2$.

   $$Y_s = \{y_{i,j} : i, j \in s_1 \cup s_2\}$$

For large graphs, these data can be obtained (with high probability) by asking each $i \in s_1$ the following:

1. Who are your friends?
2. Among your friends, which are friends with each other?
Egocentric sampling

**Procedure:**

1. Uniformly sample a set $s_1 = \{s_{1,1}, \ldots, s_{1,n_s}\}$ of nodes

$$s_1 \subset \{1, \ldots, n\}.$$

2. Observe the relations for each $i \in s_1$, i.e. observe \{\(y_{i,1}, \ldots, y_{i,n}\}\}.

3. Let $s_2$ be the set of nodes having a link from anyone in $s_1$. Observe the relations of anyone in $s_2$ to anyone in $s_1 \cup s_2$.

$$Y_s = \{y_{i,j} : i, j \in s_1 \cup s_2\}$$

For large graphs, these data can be obtained (with high probability) by asking each $i \in s_1$ the following:

1. Who are your friends?
2. Among your friends, which are friends with each other?
Egocentric sampling

Procedure:

1. Uniformly sample a set \( s_1 = \{s_{1,1}, \ldots, s_{1,n_s}\} \) of nodes

\[
s_1 \subset \{1, \ldots, n\}.
\]

2. Observe the relations for each \( i \in s_1 \), i.e. observe \( \{y_{i,1}, \ldots, y_{i,n}\} \).

3. Let \( s_2 \) be the set of nodes having a link from anyone in \( s_1 \). Observe the relations of anyone in \( s_2 \) to anyone in \( s_1 \cup s_2 \).

\[
Y_s = \{y_{i,j} : i, j \in s_1 \cup s_2\}
\]

For large graphs, these data can be obtained (with high probability) by asking each \( i \in s_1 \) the following:

1. Who are your friends?
2. Among your friends, which are friends with each other?
Egocentric sampling

Procedure:

1. Uniformly sample a set $s_1 = \{s_{1,1}, \ldots, s_{1,n_s}\}$ of nodes

   $$s_1 \subset \{1, \ldots, n\}.$$ 

2. Observe the relations for each $i \in s_1$, i.e. observe $\{y_{i,1}, \ldots, y_{i,n}\}$.

3. Let $s_2$ be the set of nodes having a link from anyone in $s_1$. Observe the relations of anyone in $s_2$ to anyone in $s_1 \cup s_2$.

   $$\mathbf{Y}_s = \{y_{i,j} : i, j \in s_1 \cup s_2\}$$

For large graphs, these data can be obtained (with high probability) by asking each $i \in s_1$ the following:

1. Who are your friends?

2. Among your friends, which are friends with each other?
Egocentric sampling

Procedure:
1. Uniformly sample a set \( s_1 = \{s_{1,1}, \ldots, s_{1,n_s}\} \) of nodes
   \[ s_1 \subset \{1, \ldots, n\}. \]
2. Observe the relations for each \( i \in s_1 \), i.e. observe \( \{y_{i,1}, \ldots, y_{i,n}\} \).
3. Let \( s_2 \) be the set of nodes having a link from anyone in \( s_1 \). Observe the relations of anyone in \( s_2 \) to anyone in \( s_1 \cup s_2 \).
   \[ Y_s = \{y_{i,j} : i,j \in s_1 \cup s_2\} \]

For large graphs, these data can be obtained (with high probability) by asking each \( i \in s_1 \) the following:
1. Who are your friends?
2. Among your friends, which are friends with each other?
**Snowball sampling**: Iteratively repeat the egocentric sampler, obtaining the stage-$k$ nodes $s_k$ from the links of $s_{k-1}$.

This is a type of link-tracing design. The links of the current nodes determine who is next to be included in the sample.

How will such subgraphs $\mathbf{Y}_s$ be similar to $\mathbf{Y}$? How will they differ?
**Snowball sampling**: Iteratively repeat the egocentric sampler, obtaining the stage-$k$ nodes $s_k$ from the links of $s_{k-1}$.

This is a type of link-tracing design. The links of the current nodes determine who is next to be included in the sample.

How will such subgraphs $Y_s$ be similar to $Y$?
How will they differ?
Link-tracing designs

**Snowball sampling**: Iteratively repeat the egocentric sampler, obtaining the stage-$k$ nodes $s_k$ from the links of $s_{k-1}$.

This is a type of link-tracing design. The links of the current nodes determine who is next to be included in the sample.

How will such subgraphs $Y_s$ be similar to $Y$?
How will they differ?
Egocentric sampling
Egocentric sampling
Egocentric sampling
Egocentric sampling
Egocentric sampling
Egocentric sampling
Egocentric sampling
Egocentric sampling
Egocentric sampling
Egocentric sampling
Egocentric sampling
• \( Y_s \) is not generally representative of \( Y \).
  • For some statistics, weighted averages based on \( Y_s \) can be unbiased (Horwitz-Thompson estimator).
  • For many statistics, part of \( Y_s \) can be used to obtain good estimates:
    • degree distributions can be estimated from degrees of egos;
    • covariate distributions can be estimated from those of the egos;

However, use of data from \( s_2 \) generally requires a reweighting scheme.
Inference with egocentric samples

- \( Y_s \) is not generally representative of \( Y \).
- For some statistics, weighted averages based on \( Y_s \) can be unbiased (Horwitz-Thompson estimator).
- For many statistics, part of \( Y_s \) can be used to obtain good estimates:
  - degree distributions can be estimated from degrees of egos;
  - covariate distributions can be estimated from those of the egos;

However, use of data from \( s_2 \) generally requires a reweighting scheme.
Inference with egocentric samples

- $Y_s$ is not generally representative of $Y$.
- For some statistics, weighted averages based on $Y_s$ can be unbiased (Horwitz-Thompson estimator).
- For many statistics, part of $Y_s$ can be used to obtain good estimates:
  - degree distributions can be estimated from degrees of egos;
  - covariate distributions can be estimated from those of the egos;

However, use of data from $s_2$ generally requires a reweighting scheme.
Inference with egocentric samples

• $Y_s$ is not generally representative of $Y$.
• For some statistics, weighted averages based on $Y_s$ can be unbiased (Horwitz-Thompson estimator).
• For many statistics, part of $Y_s$ can be used to obtain good estimates:
  • degree distributions can be estimated from degrees of egos;
  • covariate distributions can be estimated from those of the egos;

However, use of data from $s_2$ generally requires a reweighting scheme.
Inference with egocentric samples

- $Y_s$ is not generally representative of $Y$.
- For some statistics, weighted averages based on $Y_s$ can be unbiased (Horwitz-Thompson estimator).
- For many statistics, part of $Y_s$ can be used to obtain good estimates:
  - degree distributions can be estimated from degrees of egos;
  - covariate distributions can be estimated from those of the egos;

However, use of data from $s_2$ generally requires a reweighting scheme.
Inference with egocentric samples

- $Y_s$ is not generally representative of $Y$.
- For some statistics, weighted averages based on $Y_s$ can be unbiased (Horwitz-Thompson estimator).
- For many statistics, part of $Y_s$ can be used to obtain good estimates:
  - degree distributions can be estimated from degrees of egos;
  - covariate distributions can be estimated from those of the egos;

However, use of data from $s_2$ generally requires a reweighting scheme.
Inference with egocentric samples

- $Y_s$ is not generally representative of $Y$.
- For some statistics, weighted averages based on $Y_s$ can be unbiased (Horwitz-Thompson estimator).
- For many statistics, part of $Y_s$ can be used to obtain good estimates:
  - degree distributions can be estimated from degrees of egos;
  - covariate distributions can be estimated from those of the egos;
However, use of data from $s_2$ generally requires a reweighting scheme.
Inference with egocentric samples

- $Y_s$ is not generally representative of $Y$.
- For some statistics, weighted averages based on $Y_s$ can be unbiased (Horwitz-Thompson estimator).
- For many statistics, part of $Y_s$ can be used to obtain good estimates:
  - degree distributions can be estimated from degrees of egos;
  - covariate distributions can be estimated from those of the egos;

   However, use of data from $s_2$ generally requires a reweighting scheme.
References

- Snijders (1992), “Estimation on the basis of snowball samples: How to Weight?”
Parameter estimation with incomplete sampled data

**Model:** \( \Pr(Y = y | \theta), \theta \in \Theta. \)

**Complete data:** \( Y \)

**Observed data:** \( Y[O] \), where \( O \) is a set of pairs of indices:

\[
O = \begin{pmatrix}
i_1 & j_1 \\
i_2 & j_2 \\
i_3 & j_3 \\
\vdots & \vdots \\
i_s & j_s
\end{pmatrix}
\]

How can we make inference for \( \theta \) based on \( Y[O] \)?
Parameter estimation with incomplete sampled data

**Model:** \( \Pr(Y = y|\theta), \theta \in \Theta. \)

**Complete data:** \( Y \)

**Observed data:** \( Y[O], \) where \( O \) is a set of pairs of indices

\[
O = \begin{pmatrix}
  i_1 & j_1 \\
  i_2 & j_2 \\
  i_3 & j_3 \\
  \vdots & \vdots \\
  i_s & j_s 
\end{pmatrix}
\]

How can we make inference for \( \theta \) based on \( Y[O] \)?
Parameter estimation with incomplete sampled data

Model: \( \Pr(Y = y|\theta), \theta \in \Theta. \)

Complete data: \( Y \)

Observed data: \( Y[O] \), where \( O \) is a set of pairs of indices

\[
O = \begin{pmatrix}
  i_1 & j_1 \\
  i_2 & j_2 \\
  i_3 & j_3 \\
  \vdots & \vdots \\
  i_s & j_s 
\end{pmatrix}
\]

How can we make inference for \( \theta \) based on \( Y[O] \)?
Parameter estimation with incomplete sampled data

Model: \(\Pr(Y = y | \theta), \theta \in \Theta.\)

Complete data: \(Y\)

Observed data: \(Y[O]\), where \(O\) is a set of pairs of indices

\[
O = \begin{pmatrix}
i_1 & j_1 \\
i_2 & j_2 \\
i_3 & j_3 \\
\vdots & \vdots \\
i_s & j_s
\end{pmatrix}
\]

How can we make inference for \(\theta\) based on \(Y[O]\)?
### Node-induced subgraph sampling

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NA</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>NA</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>NA</td>
</tr>
</tbody>
</table>
### Node-induced subgraph sampling

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NA</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>NA</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>NA</td>
</tr>
</tbody>
</table>
## Node-induced subgraph sampling

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NA</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>NA</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>NA</td>
</tr>
</tbody>
</table>
**Study design and missing data**

**Node-induced subgraph sampling:** Observed data

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>2</td>
<td>NA</td>
<td>NA</td>
<td>1</td>
<td>NA</td>
<td>0</td>
<td>NA</td>
</tr>
<tr>
<td>3</td>
<td>NA</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>0</td>
<td>NA</td>
</tr>
<tr>
<td>4</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>5</td>
<td>NA</td>
<td>0</td>
<td>1</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>
### Study design and missing data

#### Edge-induced subgraph sampling

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NA</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>NA</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>NA</td>
</tr>
</tbody>
</table>
## Edge-induced subgraph sampling

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NA</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>NA</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>NA</td>
</tr>
</tbody>
</table>
**Edge-induced subgraph sampling:** Observed data

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td><strong>1</strong></td>
</tr>
<tr>
<td>2</td>
<td>NA</td>
<td>NA</td>
<td><strong>1</strong></td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>3</td>
<td><strong>1</strong></td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>4</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>5</td>
<td>NA</td>
<td>NA</td>
<td><strong>1</strong></td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6</td>
<td>NA</td>
<td>NA</td>
<td><strong>1</strong></td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>
Egocentric sampling

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NA</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>NA</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>NA</td>
</tr>
</tbody>
</table>
### Study design and missing data

**Egocentric sampling**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NA</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>NA</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>NA</td>
</tr>
</tbody>
</table>
**Egocentric sampling**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NA</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>NA</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>NA</td>
</tr>
</tbody>
</table>
Egocentric sampling

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NA</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>NA</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>NA</td>
</tr>
</tbody>
</table>
### Egocentric sampling

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>NA</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>NA</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>5</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6</td>
<td>NA</td>
<td>0</td>
<td>1</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>
Parameter estimation with missing data

If the data are missing at random, i.e. the value of $o$, what you get to observe,

- doesn’t depend on $\theta$
- doesn’t depend on values of $Y$,

then valid likelihood and Bayesian inference can be obtained from the observed-data likelihood:

$$l_{MAR}(\theta : y[o]) = Pr(Y[o] = y[o] : \theta) = \sum_{y[o^c]} Pr(Y = y : \theta)$$

Inference based on $l(\theta : y[o])$ is provided in amen:

- put NA’s in place of any non-observed relations.
Parameter estimation with missing data

If the data are missing at random, i.e. the value of \( o \), what you get to observe,

- doesn’t depend on \( \theta \)
- doesn’t depend on values of \( Y \),

then valid likelihood and Bayesian inference can be obtained from the observed-data likelihood:

\[
I_{MAR}(\theta : y[o]) = \Pr(Y[o] = y[o] : \theta) = \sum_{y[o^c]} \Pr(Y = y : \theta)
\]

Inference based on \( I(\theta : y[o]) \) is provided in \texttt{amen}:

- put NA’s in place of any non-observed relations.
If the data are missing at random, i.e. the value of $\mathbf{o}$, what you get to observe,

- doesn’t depend on $\theta$
- doesn’t depend on values of $\mathbf{Y}$,

then valid likelihood and Bayesian inference can be obtained from the observed-data likelihood:

$$l_{MAR}(\theta : y[\mathbf{o}]) = \Pr(\mathbf{Y}[\mathbf{o}] = y[\mathbf{o}] : \theta)$$

$$= \sum_{y[\mathbf{o}^c]} \Pr(\mathbf{Y} = y : \theta)$$

Inference based on $l(\theta : y[\mathbf{o}])$ is provided in amen:

- put NA’s in place of any non-observed relations.
Parameter estimation with missing data

If the data are missing at random, i.e. the value of \( \mathbf{o} \), what you get to observe,

- doesn’t depend on \( \theta \)
- doesn’t depend on values of \( \mathbf{Y} \),

then valid likelihood and Bayesian inference can be obtained from the observed-data likelihood:

\[
l_{MAR}(\theta : \mathbf{y}[o]) = \Pr(\mathbf{Y}[o] = \mathbf{y}[o] : \theta)
= \sum_{\mathbf{y}[o^c]} \Pr(\mathbf{Y} = \mathbf{y} : \theta)
\]

Inference based on \( l(\theta : \mathbf{y}[o]) \) is provided in \texttt{amen}:

- put NA’s in place of any non-observed relations.
Parameter estimation with missing data

If the data are missing at random, i.e. the value of $o$, what you get to observe,

- doesn’t depend on $\theta$
- doesn’t depend on values of $Y$,

then valid likelihood and Bayesian inference can be obtained from the observed-data likelihood:

$$l_{MAR}(\theta : \mathbf{y}[o]) = \Pr(\mathbf{Y}[o] = \mathbf{y}[o] : \theta) = \sum_{\mathbf{y}[o^c]} \Pr(\mathbf{Y} = \mathbf{y} : \theta)$$

Inference based on $l(\theta : \mathbf{y}[o])$ is provided in $\text{amen}$:

- put NA’s in place of any non-observed relations.
Missing at random designs

Which designs we've discussed correspond to MAR relations?

- Node-induced subgraph sampling?
- Edge-induced subgraph sampling?
- Egocentric sampling?
Which designs we've discussed correspond to MAR relations?

- Node-induced subgraph sampling?
- Edge-induced subgraph sampling?
- Egocentric sampling?
Missing at random designs

Which designs we've discussed correspond to MAR relations?

- Node-induced subgraph sampling?
- Edge-induced subgraph sampling?
- Egocentric sampling?
Missing at random designs

Which designs we've discussed correspond to MAR relations?

- Node-induced subgraph sampling?
- Edge-induced subgraph sampling?
- Egocentric sampling?
Ignorable designs

While egocentric and other link-tracing designs are not MAR, they still can be analyzed as if they were. The argument is as follows:

The “data” include

- $O = o$, the determination of which relations you get to see;
- $Y'O = y[o]$, the relationship values for the observable relations.

The likelihood is then

$$l_\theta(o, y[o]) = \Pr(Y[o] = y[o], O = o | \theta)$$

$$= \Pr(Y[o] = y[o] | \theta) \times \Pr(O = o | \theta, Y[o] = y[o])$$

$$= l_{MAR}(\theta : y[o]) \times \Pr(O = o | \theta, Y[o] = y[o])$$

If the design part doesn’t depend on $\theta$, then the observed likelihood is proportional to the MAR likelihood, and the design can be ignored.
Ignorable designs

While egocentric and other link-tracing designs are not MAR, they still can be analyzed as if they were. The argument is as follows:

The “data” include

- $O = o$, the determination of which relations you get to see;
- $Y[O] = y[o]$, the relationship values for the observable relations.

The likelihood is then

$$l(θ : o, y[o]) = Pr(Y[o] = y[o], O = o|θ)$$

$$= Pr(Y[o] = y[o]|θ) × Pr(O = o|θ, Y[o] = y[o])$$

$$= l_{MAR}(θ : y[o]) × Pr(O = o|θ, Y[o] = y[o])$$

If the design part doesn’t depend on $θ$, then the observed likelihood is proportional to the MAR likelihood, and the design can be ignored.
Ignorable designs

While egocentric and other link-tracing designs are not MAR, they still can be analyzed as if they were. The argument is as follows:

The "data" include

- \( O = o \), the determination of which relations you get to see;
- \( Y[O] = y[o] \), the relationship values for the observable relations.

The likelihood is then

\[
l(\theta : o, y[o]) = \Pr(Y[o] = y[o], O = o|\theta) \\
= \Pr(Y[o] = y[o]|\theta) \times \Pr(O = o|\theta, Y[o] = y[o]) \\
= l_{MAR}(\theta : y[o]) \times \Pr(O = o|\theta, Y[o] = y[o])
\]

If the design part doesn’t depend on \( \theta \), then the observed likelihood is proportional to the MAR likelihood, and the design can be ignored.
Ignorable designs

While egocentric and other link-tracing designs are not MAR, they still can be analyzed as if they were. The argument is as follows:

The “data” include

- \( O = o \), the determination of which relations you get to see;
- \( Y[O] = y[O] \), the relationship values for the observable relations.

The likelihood is then

\[
I(\theta : o, y[o]) = \Pr(Y[o] = y[o], O = o|\theta) \\
= \Pr(Y[o] = y[o]|\theta) \times \Pr(O = o|\theta, Y[o] = y[o]) \\
= I_{MAR}(\theta : y[o]) \times \Pr(O = o|\theta, Y[o] = y[o])
\]

If the design part doesn’t depend on \( \theta \), then the observed likelihood is proportional to the MAR likelihood, and the design can be ignored.
Ignorable designs

While egocentric and other link-tracing designs are not MAR, they still can be analyzed as if they were. The argument is as follows:

The “data” include

- $O = o$, the determination of which relations you get to see;
- $Y[O] = y[o]$, the relationship values for the observable relations.

The likelihood is then

$$l(\theta : o, y[o]) = \Pr(Y[o] = y[o], O = o|\theta)$$

$$= \Pr(Y[o] = y[o]|\theta) \times \Pr(O = o|\theta, Y[o] = y[o])$$

$$= l_{MAR}(\theta : y[o]) \times \Pr(O = o|\theta, Y[o] = y[o])$$

If the design part doesn’t depend on $\theta$, then the observed likelihood is proportional to the MAR likelihood, and the design can be ignored.
Ignorable designs

While egocentric and other link-tracing designs are not MAR, they still can be analyzed as if they were. The argument is as follows:

The “data” include

- \( O = o \), the determination of which relations you get to see;
- \( Y[O] = y[o] \), the relationship values for the observable relations.

The likelihood is then

\[
I(\theta : o, y[o]) = \Pr(Y[o] = y[o], O = o|\theta) = \Pr(Y[o] = y[o]|\theta) \times \Pr(O = o|\theta, Y[o] = y[o]) = I_{MAR}(\theta : y[o]) \times \Pr(O = o|\theta, Y[o] = y[o])
\]

If the design part doesn’t depend on \( \theta \), then the observed likelihood is proportional to the MAR likelihood, and the design can be ignored.
**Ignorable designs**

\[ l(\theta : o, y[o]) = l_{MAR}(\theta : y[o]) \times \Pr(O = o|\theta, Y[o] = y[o]) \]

When is the design ignorable?

**(MAR)** If the probability that \( O \) equals \( o \) doesn’t depend on \( \theta \) or \( Y \) (e.g., node-induced subgraph sampling), the design is ignorable.

**ID** If the probability that \( O \) equals \( o \)
* doesn’t depend on \( \theta \)
* only depends on \( Y \) through \( Y[o] \).

then the design is ignorable.

The latter conditions are often met for link tracing designs, like egocentric and snowball sampling.
Ignoring designs

\[ l(\theta : o, y[o]) = l_{MAR}(\theta : y[o]) \times \Pr(O = o|\theta, Y[o] = y[o]) \]

When is the design ignorable?

**MAR** If the probability that \( O \) equals \( o \) doesn’t depend on \( \theta \) or \( Y \) (e.g., node-induced subgraph sampling), the design is ignorable.

**ID** If the probability that \( O \) equals \( o \)
- doesn’t depend on \( \theta \)
- only depends on \( Y \) through \( Y[o] \).

then the design is ignorable.

The latter conditions are often met for link tracing designs, like egocentric and snowball sampling.
**Ignorable designs**

\[ l(\theta : o, y[o]) = l_{MAR}(\theta : y[o]) \times \Pr(O = o|\theta, Y[o] = y[o]) \]

**When is the design ignorable?**

**MAR** If the probability that \(O\) equals \(o\) doesn’t depend on \(\theta\) or \(Y\) (e.g., node-induced subgraph sampling), the design is ignorable.

**ID** If the probability that \(O\) equals \(o\)

- doesn’t depend on \(\theta\)
- only depends on \(Y\) through \(Y[o]\).

then the design is ignorable.

The latter conditions are often met for link tracing designs, like egocentric and snowball sampling.
When is the design ignorable?

**(MAR)** If the probability that $O$ equals $o$ doesn’t depend on $\theta$ or $Y$ (e.g., node-induced subgraph sampling), the design is ignorable.

**ID** If the probability that $O$ equals $o$

- doesn’t depend on $\theta$
- only depends on $Y$ through $Y[o].$

then the design is ignorable.

The latter conditions are often met for link tracing designs, like egocentric and snowball sampling.
Ignorable designs

\[ l(\theta : o, y[o]) = l_{MAR}(\theta : y[o]) \times \Pr(O = o | \theta, Y[o] = y[o]) \]

When is the design ignorable?

**MAR** If the probability that \( O \) equals \( o \) doesn’t depend on \( \theta \) or \( Y \) (e.g., node-induced subgraph sampling), the design is ignorable.

**ID** If the probability that \( O \) equals \( o \)

- doesn’t depend on \( \theta \)
- only depends on \( Y \) through \( Y[o] \).

then the design is ignorable.

The latter conditions are often met for link tracing designs, like egocentric and snowball sampling.
Ignoreable designs

\[ l(\omega : \omega[\omega]) = l_{\text{MAR}}(\omega : \omega[\omega]) \times \Pr(O = \omega | \omega, Y[\omega] = \omega[\omega]) \]

When is the design ignorable?

**MAR** If the probability that \( O \) equals \( \omega \) doesn’t depend on \( \omega \) or \( Y \) (e.g., node-induced subgraph sampling), the design is ignorable.

**ID** If the probability that \( O \) equals \( \omega \)

- doesn’t depend on \( \omega \)
- only depends on \( Y \) through \( Y[\omega] \).

then the design is ignorable.

The latter conditions are often met for link tracing designs, like egocentric and snowball sampling.
Ignorable designs

\[ l(\theta : o, y[o]) = l_{MAR}(\theta : y[o]) \times Pr(O = o|\theta, Y[o] = y[o]) \]

When is the design ignorable?

**MAR** If the probability that \( O \) equals \( o \) doesn’t depend on \( \theta \) or \( Y \) (e.g., node-induced subgraph sampling), the design is ignorable.

**ID** If the probability that \( O \) equals \( o \)

- doesn’t depend on \( \theta \)
- only depends on \( Y \) through \( Y[o] \).

then the design is ignorable.

The latter conditions are often met for link tracing designs, like egocentric and snowball sampling.
Ignorable designs

\[ l(\theta : o, y[o]) = l_{MAR}(\theta : y[o]) \times \Pr(O = o|\theta, Y[o] = y[o]) \]

When is the design ignorable?

(MAR) If the probability that \(O\) equals \(o\) doesn’t depend on \(\theta\) or \(Y\) (e.g., node-induced subgraph sampling), the design is ignorable.

ID If the probability that \(O\) equals \(o\)

- doesn’t depend on \(\theta\)
- only depends on \(Y\) through \(Y[o]\).

then the design is ignorable.

The latter conditions are often met for link tracing designs, like egocentric and snowball sampling.
References

- Thompson and Frank (2000) “Model-based estimation with link-tracing sampling designs”
Simulation study - ID likelihoods

\[ y_{i,j} = \beta_0 + \beta_r x_{n,i} + \beta_c x_{n,j} + \beta_d i.j + a_i + b_j + \epsilon_{i.j} \]

\textbf{fit.pop} = fitted model based on complete network data

\textbf{fit.samp} = fitted model based on sampled network data

How do the parameter estimates of \textbf{fit.samp} compare to those of \textbf{fit.pop}?
Simulation study - ID likelihoods

\[ y_{i,j} = \beta_0 + \beta_{r} x_{n,i} + \beta_{c} x_{n,j} + \beta_{d,i,j} + a_i + b_j + \epsilon_{i,j} \]

\textit{fit.pop} = fitted model based on complete network data

\textit{fit samp} = fitted model based on sampled network data

How do the parameter estimates of \textit{fit samp} compare to those of \textit{fit pop}?
Simulation study - ID likelihoods

\[ y_{i,j} = \beta_0 + \beta_r x_{n,i} + \beta_c x_{n,j} + \beta_{d,i,j} + a_i + b_j + \epsilon_{i,j} \]

\textit{fit.pop} = fitted model based on complete network data

\textit{fit.samp} = fitted model based on sampled network data

How do the parameter estimates of \textit{fit.samp} compare to those of \textit{fit.pop}?
Node-induced subgraph sample

\( n_p = 32, n_s = 10 \)
Egocentric sample

\( n_p = 32, \ n_{s1} = 4 \)